Learning Discrete Distributions from Metastable Data via Pseudo-Likelihood

Anonymous Author(s)
Affiliation
Address
email

Abstract

Physically motivated stochastic dynamics are standard tools for sampling high-dimensional distributions but often mix slowly due to metastability. We show that for multivariate discrete distributions, *the true stationary model can be learned from i.i.d. samples drawn from a metastable distribution*. The key observation is that for strongly metastable states of a reversible chain with stationary distribution μ, the single-variable conditionals of the metastable law are, on average, close to those of μ, even when the two distributions are far in global metrics. This enables accurate parameter recovery with conditional-likelihood estimators such as pseudo-likelihood (PL). We formalize these guarantees and illustrate them numerically on the Curie-Weiss model, where PL succeeds while maximum likelihood fails.

Motivation and Overview

Markov chains used for sampling may become trapped for long periods in regions of state space, a hallmark of metastability and a principal cause of poor mixing. From a learning standpoint this is troubling: data collected from a natural dynamics or a sampling algorithm may reflect a metastable law ν that can be *globally* far from the true equilibrium μ in total variation or KL divergence. Global-metric-based estimators such as maximum likelihood are therefore ill-suited: they attempt to match sufficient statistics of ν, not μ. A more fruitful perspective is local: many efficient estimators for undirected graphical models learn by *matching single-variable conditionals* (e.g., pseudo-likelihood/logistic regressions) [2, 4]. Our central finding is that if ν is "strongly metastable" for a reversible chain with equilibrium μ, then the average discrepancy between the single-variable conditionals of ν and those of μ is small. Consequently, PL trained on metastable samples is nearly optimal for μ, and parameters of pairwise Ising models can be recovered (up to a small bias controlled by the metastability level). The effect is especially pronounced in slow-mixing regimes, where the bias decays with system size and is dominated by statistical error.

Setup and Main Definitions

We consider n discrete variables $\sigma = (\sigma_1, \dots, \sigma_n) \in \mathcal{Q}^n$ and a reversible Markov chain P with stationary distribution μ. We adopt two notions of metastability.

Definition 1 (Metastability) *A distribution ν is η-metastable (w.r.t. P) if $\|\nu - \nu P\|_{\mathrm{TV}} \leq \eta$.*

Definition 2 (Strong metastability) *A distribution ν is η-strongly metastable (w.r.t. P) if*

$$\frac{1}{2} \sum_{i,j} \left| P(i|j)\nu(j) - P(j|i)\nu(i) \right| \leq \eta.$$

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Strong metastability implies metastability and captures small violations of detailed balance. It arises naturally from bottlenecks in probability flow. For a set A with conductance $\Gamma(A)$, the truncated distribtuion $\mu_A(\sigma) \propto \mu(\sigma)\mathbf{1}\{\sigma \in A\}$ is η-strongly metastable with $\eta = \Gamma(A)$, hence slow-mixing chains admit exponentially small η on suitable A. In these cases, we should think of η as scaling as the inverse of the mixing time of the chain and hence as a quantity that decays fast as function of n.

We assume the usual single-site-update setting and a bounded spin-flip probability:

$$\frac{P(\sigma_u \to q \,|\, \sigma)}{\mu(q \,|\, \sigma_{\setminus u})} \;\geq\; \omega_P > 0, \qquad \text{e.g., } \omega_P = \tfrac{1}{n} \text{ for Glauber dynamics.}$$

Closeness of conditionals. The central technical statement of our work is the following:

Theorem 1 (Conditionals of strongly metastable states) *Let ν be η-strongly metastable for a reversible P with equilibrium μ. If the bounded spin-flip condition holds with constant ω_P, then*

$$\sum_{u=1}^{n} \mathbb{E}_{\sigma\sim\nu} \left\|\nu(\,\cdot\,|\,\sigma_{\setminus u}) - \mu(\,\cdot\,|\,\sigma_{\setminus u})\right\|_{\mathrm{TV}} \;\leq\; \frac{\eta}{\omega_P}. \tag{1}$$

Thus, even if ν and μ are far in global TV, their *local* conditionals are close on average. Via reverse Pinsker, the same holds in average conditional KL, which is precisely the PL loss.

Learning via Pseudo-Likelihood

We model $\mu(\sigma) \propto \exp(E^*(\sigma))$ and estimate within a parametric class $\{E(\sigma;\theta) : \theta \in \Theta\}$ using PL. Now we assume a finite interaction strength for the models in the parametric class. Now if the data comes from a metastable state of a Markov chain attempting to sample from $\mu(\sigma)$, then we can show that the true parameters of the model (θ^*) are almost optimal for the pseudo-likelihood loss function,

Near-optimality of PL on metastable data. With M' i.i.d. samples from ν, the PL estimator $\hat{\theta}$ satisfies

$$\frac{1}{M'}\sum_{t,u}\Big(L_u(\theta^*;\sigma^{(t)}) - L_u(\hat{\theta};\sigma^{(t)})\Big) \;\lesssim\; \underbrace{O(\frac{\eta}{\omega_P})}_{\text{bias from metastability}} + O(\underbrace{M'^{-\frac{1}{2}}}_{\text{statistics}}),$$

so in slow-mixing regimes (small η), the bias is dominated by sampling noise. (Constants depend on the size of the discrete alphabet and the interaction strength)

Ising models: couplings and structure. For the specific case of Ising models, we can use this result to show that PL learning from metastable data correctly recovers the parameters of the true distribution. In this case, $E^*(\sigma) = \sum_{i<j}\theta^*_{ij}\sigma_i\sigma_j + \sum_i \theta^*_i\sigma_i$, ℓ_1-constrained PL yields high-dimensional guarantees analogous to the i.i.d. case. With $M = \tilde{O}(e^{8\gamma}\gamma^4\varepsilon^{-4}\log(n))$ samples from ν,

$$\max_{v\neq u}|\widehat{\theta}_{uv} - \theta^*_{uv}| \;\leq\; \varepsilon \;+\; 4e^{2\gamma}\sqrt{\tfrac{(1+\gamma)\,\eta}{\omega_P}},$$

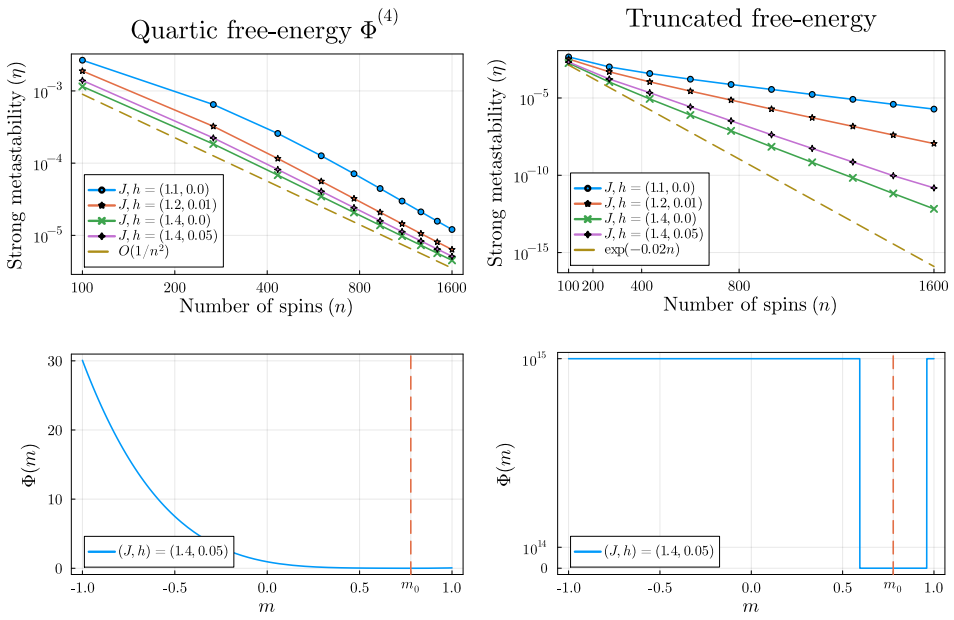
and edges are recovered by simple thresholding provided nonzero couplings exceed a (metastability-dependent) level. In particular, for Glauber dynamics $\omega_P = \frac{1}{n}$, any $\eta = o(n^{-1})$ yields vanishing bias with growing n. The proof here uses a perturbative generalization of the techniques in [3].

Illustration: Curie–Weiss Numerics

We illustrate the theory on the Curie–Weiss ferromagnet with spins $\sigma_i \in \{\pm 1\}$ as it is the canonical example of a system that exhibits metastability [1].

$$E(\sigma) = \frac{J}{n}\sum_{i<j}\sigma_i\sigma_j - h\sum_i \sigma_i, \qquad m(\sigma) = \frac{1}{n}\sum_i \sigma_i.$$

The equilibrium distribution over magnetization concentrates near minima of the free energy $\Psi(m) = -\frac{J}{2}m^2 + hm - S(m)$, where S is the entropy. In the low-temperature, weak-field regime ($J > 1$, small h), there are two well-separated minima, and Glauber dynamics started at all-plus becomes trapped near the positive-magnetization well for exponentially long times.



(a) (b)

Figure 1: Strongly metastable states in the CW model. Here we plot the violation in detailed balance condition (η in the definition of strong metastability) computed by projecting to the magnetization space. (a) Fourth-order approximation to the free energy at m_0 (b) truncated free energy around m_0.

Metastable states near free-energy minima. Thanks to permutation symmetry, the metastability of this model can be entirely studied in the magnetization space ($m(\sigma) = \frac{\sum_i \sigma_i}{n}$). One can construct strongly metastable laws by expanding the free energy Ψ around its positive minimizer $m_0 > 0$:

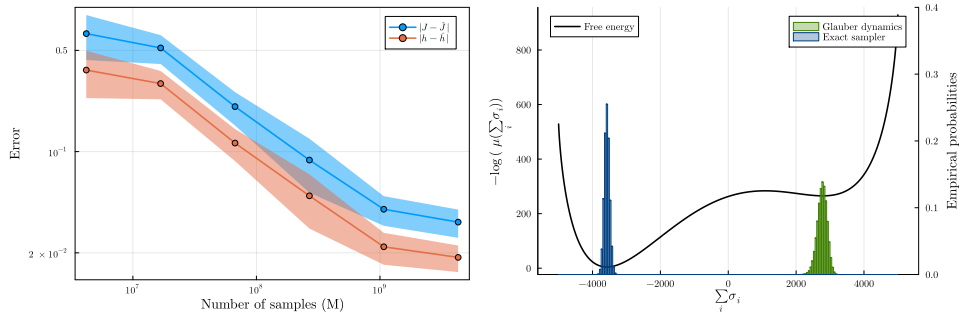
$$\Phi^{(K)}(m) = \sum_{k=2}^{K} \frac{\Psi^{(k)}(m_0)}{k!}(m - m_0)^k, \qquad \nu^{(K)}(\sigma) \propto \exp\{-n\Phi^{(K)}(m(\sigma))\}\binom{n}{\frac{n(1+m(\sigma))}{2}}.$$

Numerically we confirm that $\nu^{(4)}$ is $O(n^{-2})$-strongly metastable for $J > 1$ and small h, and that truncating Ψ in a narrow window around m_0 yields η decaying exponentially in n. Numerical results are given in Figure 1.

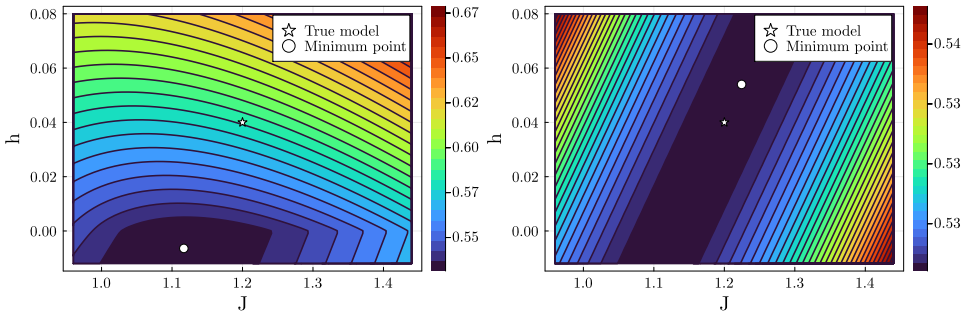
Pseudo-likelihood succeeds, MLE fails. We simulate Glauber dynamics in a regime where the true Gibbs measure strongly favors *negative* magnetization, but the chain started at all-plus remains stuck near the *positive* well. Training PL on these metastable samples accurately recovers the model parameters, whereas MLE (which matches global statistics of the data) yields a spurious solution, even flipping the sign preferred by the true magnetization. Concretely, for $(J, h) = (1.2, 0.04)$ on large n, PL trained on metastable data recovers parameters with small error, while the negative log-likelihood landscape attains its minimum far from the ground truth. This numerical experiment mirrors the theory: local conditionals are close, but global sufficient statistics are not. Details of these experiments are in Figure 2.

Discussion

The picture that emerges is simple: metastability hurts global-metric estimators but not conditional ones. For reversible chains, the *local* information carried by a strongly metastable law is already almost that of the equilibrium model, and PL exploits exactly this information. In slow-mixing regimes (small conductance), the resulting bias is tiny compared to sampling noise, so learning from "wrong" data is still possible, provided we learn from a metastable distribution that has the right local statistics.



(a) (b)



(c) Maximum likelihood loss (d) Psuedo-likelihood loss

Figure 2: **(a)** Error in learning the Curie-Weiss model on 5000 spins. Samples here are produced by Glauber dynamics "stuck" at the positive minima of the free energy. True parameters here are $J = 1.2, h = 0.04$. **(b)** The true distribution is highly biased towards towards negative magnetization as seen by free energy curve. There is a metastable distribution with positive magnetization that is highly suppressed in terms of probability. The empirical distributions of samples ($M = 4 \times 10^9$) drawn by an exact sampler and Glauber dynamics is overlaid on top of the free energy. This shows that the Markov chain is effectively stuck around the positive minima. **(c) and (d)** Comparison of the loss function landscape for the CW model with true parameters $J = 1.2, h = 0.04$. These are plotted with $M = 2^{32}$ samples produced by Glauber dynamics "stuck" at the positive minima of the free energy. Negative log-likelihood computed from this data clearly has it's minimum far from the true model. The sign of the magnetization is opposite of the true model. This is expected as maximum likelihood tries to match the sufficient statistics of the data to the model. PL loss function has the minima close to the true model and learns the magnetic field with the right sign.

References

[1] David A Levin, Malwina J Luczak, and Yuval Peres. Glauber dynamics for the mean-field ising model: cut-off, critical power law, and metastability. *Probability Theory and Related Fields*, 146:223–265, 2010.

[2] Andrey Y Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. Optimal structure and parameter learning of Ising models. *Science advances*, 4(3):e1700791, 2018.

[3] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[4] Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. In *Advances in Neural Information Processing Systems*, pages 8071–8081, 2019.