Published in final edited form as:

Phys Rev E. 2020 March ; 101(3-1): 032107. doi:10.1103/PhysRevE.101.032107.

Erasure machine: Inverse Ising inference from re-weighting of observation frequencies

Junghyo Jo^{1,2,3,*}, Danh-Tai Hoang^{4,5}, Vipul Periwal^{4,†}

¹Department of Statistics, Keimyung University, Daegu 42601, Korea

²School of Computational Sciences, Korea Institute for Advanced Study, Seoul 02455, Korea

³Department of Physics Education, Seoul National University, Seoul 08826, Korea

⁴Laboratory of Biological Modeling, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

⁵Department of Natural Sciences, Quang Binh University, Dong Hoi, Quang Binh 510000, Vietnam

Abstract

Maximum Likelihood Estimation (MLE) is the bread and butter of system inference for stochastic systems. In some generality, MLE will converge to the correct model in the infinite data limit. In the context of physical approaches to system inference, such as Boltzmann machines, MLE requires the arduous computation of partition functions summing over all configurations, both observed and unobserved. We present a conceptually transparent data-driven inference computation based on a re-weighting of observed configuration frequencies that allows us to recast the inference problem as a simpler calculation. Modeling our approach on the high-temperature limit of statistical physics, we re-weight the frequencies of observed configurations by multiplying with reciprocals of Boltzmann weights and update the Boltzmann weights iteratively to make these products close to the high temperature limit of the Boltzmann weights. This converts the required partition function computation in the re-weighted MLE to a tractable leading order high temperature term. We show that this is a convex optimization at each step. Then, for systems with a large number of degrees of freedom where other approaches are intractable, we demonstrate that this data-driven algorithm gives accurate inference with both synthetic data and two real-world examples.

I. INTRODUCTION

Inferring underlying models from observed configurations is a general task for machine learning. Maximum Likelihood Estimation (MLE) is a mathematically rigorous approach to parameter estimation for stochastic systems. If a system is observed in configuration σ with frequency $n(\sigma)$ in a set of N observations, then MLE posits that an estimate of the true probabilities $p^*(\sigma)$ is

*Corresponding author: jojunghyo@snu.ac.kr. †Corresponding author: vipulp@mail.nih.gov. J.J. and D.-T.H. contributed equally to this work.

$$p^*(\sigma) = \arg \max_{p(\sigma)} \prod_{\sigma} p(\sigma)^{n(\sigma)}. \quad (1)$$

Taking the constraint $\sum_{\sigma} p(\sigma) = 1$ into account, one obtains the intuitive result

$$p^*(\sigma) = \frac{n(\sigma)}{N} \equiv f(\sigma). \quad (2)$$

When the number of observed configurations is much smaller than the set of possible configurations, Eq. (2) is useful only in the context of models which relate observed configuration frequencies to predictions of probabilities for the observation of other configurations.

Ising and Potts models, known as Markov random fields or undirected graphical models in the machine learning and statistical inference fields, are important classes of physical models to represent the probabilities $p(\sigma)$ of observed configurations σ . In particular, the models have been adopted to explain neural activities [1–3], gene expression levels [4], protein structures [5, 6], gene recombinations [7], bird interactions [8], financial markets [9, 10], and human interactions [11]. Searching in the space of graph structures encoding interactions between variables is an NP-hard problem [12].

As a concrete example, for a dataset comprised of N observed configurations of M binary variables $\sigma_i = \pm 1$, the binary variables are associated with Ising spins and the probability of observing a specific configuration $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_M)$ is assumed to be the normalized Boltzmann weight:

$$p(\sigma) = \frac{\exp(w^I O_I(\sigma))}{Z(w)} \text{ with } Z(w) \equiv \sum_{\sigma} \exp(w^I O_I(\sigma)), \quad (3)$$

using the Einstein summation convention between repeated raised and lowered indices, where $\{O_I\}$ is a set of operators appropriate for the problem of interest, for example the set of linear and quadratic terms, $\{\sigma_i\} \cup \{\sigma_i \sigma_j, i < j\}$. The inference problem is to determine the parameters w^I from the data. Applying MLE estimation, we wish to find $p^*(\sigma)$ that maximizes the likelihood $\mathcal{L} \equiv \prod_{\sigma} p(\sigma)^{f(\sigma)} = \prod_{\hat{\sigma}} p(\hat{\sigma})^{f(\hat{\sigma})}$ as the frequency $f(\sigma)$ of unobserved configurations is 0. Note that $\hat{\sigma}$ represents observed configurations in $\{\sigma\}$. Taking the logarithm of $\mathcal{L}$, we find

$$\frac{\partial \ln \mathcal{L}}{\partial w^I} = \langle O_I \rangle_f - \langle O_I \rangle_p. \quad (4)$$

Here, for any observable O_I defined on the set of all configurations, $\langle O_I \rangle_f \equiv \sum_{\hat{\sigma}} f(\hat{\sigma}) O_I(\hat{\sigma})$ is summed over the set of observed configurations, and the model prediction $\langle O_I \rangle_p \equiv \sum_{\sigma} p(\sigma) O_I(\sigma)$ is summed over all configurations. The dependence on the parameters w^I is entirely in $\langle O_I \rangle_p$. Gradient ascent using Eq. (4) to find w^I has the usual issues with local maxima for undersampled cases but the most computationally intensive part is the evaluation

of $\langle O \rangle_p$ for every step. No matter the size of the available data, N , this computation is a sum with 2^M terms.

The computational intractability of the partition function is well-known to physicists [13]. Due to the centrality of this inverse problem, many approximate solutions have been developed [14], including (resummed) mean-field methods [15, 16], Bethe approximations [17], and machine learning with variational autoregressive networks [18]. The Boltzmann machine is a representative method that approximates $\langle O \rangle_p$ in Eq. (4) by using Monte-Carlo samples [19]. Thus its solution becomes asymptotically exact as the number of samples approaches infinity. If the models have only linear and quadratic terms, first and second moments of the data are sufficient statistics to solve the problem, but attempts to use this information alone give inaccurate results for large numbers of spins, suggesting the use of higher moments to improve inference. The adaptive cluster expansion uses heuristics to truncate likelihood computations [20, 21], and the probabilistic flow method uses relaxation dynamics to aim for pre-specified analytically tractable target distributions, extracting information about the true distribution from reversed dynamics [22]. However, both approaches are computationally expensive and thus not applicable for large systems.

It has become clear over a decade of work that methods based on logistic regression perform much better for strongly coupled interactions than mean-field approaches. An excellent review and performance comparison between different approaches for inverse Ising inference are available [14, 23]. Many of these approaches use regularized pseudo-likelihood estimators assuming local interaction graphs with restricted connectivity [24, 25]. The pseudo-likelihood attempts to circumvent the difficulty of computing the exact partition function. The initial work of Ravikumar et al. [24] provides incorrect inference for large couplings but recent improvements [23] have found extensions to this regime as well and achieved very good performance on graphs with limited average degrees, necessary for the locality assumption underlying this regularized approach. Aurell and Ekeberg showed that the sparse regularization of Ravikumar et al. worked better with an explicit threshold for small couplings [25]. In brief, the weak connections below a chosen threshold were removed from a complete graph, and the regularized pseudo-likelihood inference was recomputed using neighborhoods obtained from the reduced graph. Later, Lokhov et al. refined the reduced graph inference step by removing the regularization parameter [23]. Decelle and Ricci-Tersenghi [26] showed that the local character of the pseudo-likelihood leads to inaccuracies for the interaction inferred between two spins if their local neighborhoods lead to very different estimates. They avoid this problem by using decimation and obtain excellent results for graphs with bounded degree distributions.

Of course, when we are faced with an inference problem, we have no way of knowing if the couplings are strong or weak, or if the interaction graphs of the spins have dense or sparse connectivity or have heavy-tailed degree distributions. Our aim in this study is to rethink model inference for such problems to simplify the calculation of $Z(w)$ using elementary considerations, instead of theoretical assumptions and/or sampling techniques.

The starting point of our approach is the observation that temperature is central to statistical physics and that the high temperature limit of partition functions is easy to compute. How

can we use temperature to make our computations simpler when the object of machine learning inference is to find its conjugate variable, the energy? Suppose we have a guess for the energy function. By re-weighting the frequencies of observed configurations with appropriate reciprocal powers of the Boltzmann weight, we can shift most of the energy dependence from the second ‘model’ term on the right hand side of Eq. (4) to the first ‘data’ term. The ‘data’ term involves easily computed sums over the observed data that are much smaller than computing the entire partition function. On the other hand, the partition function needed for the computation of the second term can now be computed as the leading order term in the high temperature expansion, and the parameters in the energy can be updated with Eq. (4) applied to the re-weighted frequencies. We give the details of this intuitive sketch in the following, and we show that this entirely data-driven algorithm is computationally very fast and accurate even in the hard inference regime of strong coupling with small numbers of samples.

In Sec. II, we explain our theory of re-weighting for simplifying the calculation of $Z(w)$ and show that each iteration step is a convex, indeed quadratic, minimization. Then, we demonstrate its inference performance with synthesized data and real data in Sec. III, and discuss our results in Sec. IV. Complete source code with documentation is available on GitHub [27].

II THEORY

The key idea is to re-weight the frequencies of observed configurations to make every configuration *almost* equally likely, as happens in the high-temperature limit of statistical physics. Suppose we re-weight the data distribution $f(\sigma)$ by multiplying with $p^{\epsilon-1}(\sigma)$ of an arbitrary model distribution $p(\sigma)$, which is almost the exact reciprocal of $p(\sigma)$ for ϵ small:

$$f_{\epsilon}(\sigma) \propto f(\sigma)p^{\epsilon-1}(\sigma) \quad (5)$$

for any ϵ with the normalization $\sum_{\sigma} f_{\epsilon}(\sigma) = 1$. The MLE solution for the re-weighted distribution is then $p_{\epsilon}^*(\sigma) = f_{\epsilon}(\sigma)$ following Eq. (2). Here, if we set the model distribution $p(\sigma) = p^*(\sigma) = f(\sigma)$, then we have $f_{\epsilon}(\sigma) \propto f^{\epsilon}(\sigma)$, which becomes almost flat for ϵ small. In other words, erasure of the frequency information in observed configurations with this re-weighting can only be accomplished if we know the true distribution $p^*(\sigma)$. Similarly, we define a re-weighted model probability:

$$p_{\epsilon}(\sigma) \propto p(\sigma)p^{\epsilon-1}(\sigma) = p^{\epsilon}(\sigma). \quad (6)$$

Therefore, we obtain

$$p_{\epsilon}(\sigma) = \frac{\exp(-\epsilon E(\sigma))}{Z_{\epsilon}} \text{ with } Z_{\epsilon} \equiv \sum_{\sigma} \exp(-\epsilon E(\sigma)), \quad (7)$$

where energy $E(\sigma) = -w^T O(\sigma)$ or more specifically $E(\sigma) = \sum_i h^i \sigma_i + \sum_{j < k} J^{jk} \sigma_j \sigma_k$ for the Ising model. Here it is tempting to interpret ϵ as the usual inverse temperature β in statistical

mechanics, because the final formula after the re-weighting procedure looks the same as $p_\epsilon(\sigma) \propto p(\sigma)p^{\epsilon-1}(\sigma) = p^\epsilon(\sigma)$ versus $p_\beta(\sigma) \propto p^\beta(\sigma)$. However, unlike the re-weighted model distribution $p_\epsilon(\sigma) \propto p^\epsilon(\sigma)$, the re-weighted data distribution does not satisfy $f_\epsilon(\sigma) \propto f^\epsilon(\sigma)$ until we update and iterate, as explained below, to obtain the true model distribution $p(\sigma) = p^*(\sigma)$ in Eq. (5). Therefore, we keep the symbol ϵ instead of the conventional symbol β to emphasize the re-weighting procedure which affects both the model and the observed frequencies of the data.

Now we explain the specific procedure for optimizing the model parameter $w = \{h^i, J^{jk}\}$ with the re-weighting idea. Since we do not know the model distribution $p(\sigma)$ at the beginning, we start with a random, possibly poor, re-weighting $p^{\epsilon-1}(\sigma)$ that defines $f_\epsilon(\sigma)$ in Eq. (5). Holding the re-weighted frequencies of the data $f_\epsilon(\sigma)$ fixed, we now want to find w that maximizes the likelihood $\mathcal{L}_\epsilon \equiv \prod_\sigma p_\epsilon(\sigma)^{f_\epsilon(\sigma)}$. For the optimization, note that we update the parameter dependence only in $p_\epsilon(\sigma)$. Keeping this in mind, we formulate the logarithm of $\mathcal{L}_\epsilon$:

$$\begin{aligned} \ln \mathcal{L}_\epsilon &= \sum_\sigma f_\epsilon(\sigma) \ln p_\epsilon(\sigma) \\ &= \sum_\sigma f_\epsilon(\sigma) \ln \frac{\exp(-\epsilon E(\sigma))}{\sum_{\sigma'} \exp(-\epsilon E(\sigma'))}, \end{aligned} \quad (8)$$

where the w -dependence is varied only in $E(\sigma) = -w^I O_I(\sigma)$, not in $f_\epsilon(\sigma)$. Then, the gradient of the logarithm of $\mathcal{L}_\epsilon$ is

$$\begin{aligned} \frac{\partial \ln \mathcal{L}_\epsilon}{\partial w^I} &= \sum_\sigma f_\epsilon(\sigma) O_I(\sigma) - \sum_\sigma p_\epsilon(\sigma) O_I(\sigma) \\ &= \langle O_I \rangle_{f_\epsilon} - \langle O_I \rangle_{p_\epsilon}. \end{aligned} \quad (9)$$

This modified gradient looks similar to the original gradient in Eq. (4), but the situation is dramatically changed. Now the two expectations in Eq. (9) are easily computable. The first expectation still needs to consider only observed configurations $\hat{\sigma}$:

$$\langle O_I \rangle_{f_\epsilon} = \frac{\sum_{\hat{\sigma}} f_\epsilon(\hat{\sigma}) O_I(\hat{\sigma})}{\sum_{\hat{\sigma}} f_\epsilon(\hat{\sigma})}, \quad (10)$$

because $f_\epsilon(\sigma) \propto f(\sigma)p^{\epsilon-1}(\sigma) = 0$ for unobserved configurations σ . In the presence of ϵ , the second expectation is

$$\langle O_I \rangle_{p_\epsilon} = \sum_\sigma p_\epsilon(\sigma) O_I(\sigma) = \frac{\sum_\sigma \exp(-\epsilon E(\sigma)) O_I(\sigma)}{Z_\epsilon} \quad (11)$$

with $E(\sigma) = -w^I O_I(\sigma)$ and $Z_\epsilon = \sum_\sigma \exp(\epsilon w^I O_I(\sigma))$. Here the re-weighted partition function Z_ϵ can be expanded as follows to expose ϵ dependence:

$$\begin{aligned}
Z_\epsilon &= \sum_{\sigma} \exp(-\epsilon E(\sigma)) = \sum_{\sigma} \exp\left(\sum_i \epsilon h^i \sigma_i + \sum_{j < k} \epsilon J^{jk} \sigma_j \sigma_k\right) \\
&= \sum_{\sigma} \prod_i \cosh(\epsilon h^i) [1 + \sigma_i \tanh(\epsilon h^i)] \prod_{j < k} \cosh(\epsilon J^{jk}) [1 + \sigma_j \sigma_k \tanh(\epsilon J^{jk})] \\
&= 2^M \prod_i \cosh(\epsilon h^i) \prod_{j < k} \cosh(\epsilon J^{jk}) \left[1 + \sum_{l < m} \tanh(\epsilon h^l) \tanh(\epsilon h^m) \tanh(\epsilon h^n) \right. \\
&\quad \left. + \sum_{l < m < n} \tanh(\epsilon J^{lm}) \tanh(\epsilon J^{ln}) \tanh(\epsilon J^{mn}) + \mathcal{O}(\epsilon^4)\right].
\end{aligned} \tag{12}$$

Now the logarithm of Z_ϵ is

$$\begin{aligned}
\ln Z_\epsilon &= M \ln 2 + \sum_i \ln \cosh(\epsilon h^i) + \sum_{j < k} \ln \cosh(\epsilon J^{jk}) + \mathcal{O}(\epsilon^3) \\
&= M \ln 2 + \sum_I \ln \cosh(\epsilon w^I) + \mathcal{O}(\epsilon^3),
\end{aligned} \tag{13}$$

where we used $\ln(1 + \mathcal{O}(\epsilon^3)) \approx \mathcal{O}(\epsilon^3)$ and $w = \{h^i, J^{jk}\}$. Finally, we obtain

$$\langle O_I \rangle_{p_\epsilon} = \frac{1}{\epsilon} \frac{\partial \ln Z_\epsilon}{\partial w^I} = \tanh(\epsilon w^I) \approx \epsilon w^I \tag{14}$$

by truncating at the first order of ϵ . Thus, in the small ϵ limit (strictly speaking, the small ϵw^I limit), the expectation $\langle O_I \rangle_{p_\epsilon}$, which usually requires expensive calculations to deal with every configuration, does not require any calculation at all because it is simply ϵw^I .

Finally, the gradient ascent of the logarithm of $\mathcal{Z}_\epsilon$ in Eq. (9) allows us to update the model parameter w^I :

$$\delta w^I = \alpha \frac{\partial \ln \mathcal{Z}_\epsilon}{\partial w^I} = \alpha \left(\langle O_I \rangle_{f_\epsilon} - \epsilon w^I \right) \tag{15}$$

with an arbitrary learning rate α . We use $\alpha = 0.1$ in this study. The first term is numerically evaluated with no expansion in ϵ from the data because we cannot assume that we are close to $p(\sigma) \approx f(\sigma)$. This update rule is interesting from two perspectives. First, $\epsilon = 1$ leads to the Hopfield solution of $w^I = \langle O_I \rangle_f$ at the maximum likelihood condition ($\delta w^I = 0$). Second, the second term $-\epsilon w^I$ works as a regularizer for constraining the amplitude of w^I , which contributes to make this algorithm stable. This term may look similar to ridge regression but it is conceptually and computationally completely different as we introduced no ad hoc regularization. Lastly, it is worth mentioning that decreasing $|\delta w^I|$ in Eq. (15) indeed guarantees increasing the re-weighted log-likelihood:

$$\begin{aligned}
\ln \mathcal{L}_\epsilon &= \sum_\sigma f_\epsilon(\sigma) \ln \frac{\exp(\epsilon w^I O_I(\sigma))}{Z_\epsilon} \\
&= \epsilon w^I \sum_\sigma f_\epsilon(\sigma) O_I(\sigma) - \ln Z_\epsilon \\
&= \epsilon w^I \langle O_I \rangle_{f_\epsilon} - \frac{(\epsilon w^I)^2}{2} - M \ln 2 \\
&= -\frac{1}{2} \left(\langle O_I \rangle_{f_\epsilon} - \epsilon w^I \right)^2 + \frac{\langle O_I \rangle_{f_\epsilon}^2}{2} - M \ln 2 \\
&= -\frac{(\delta w^I)^2}{2\alpha^2} + \frac{\langle O_I \rangle_{f_\epsilon}^2}{2} - M \ln 2,
\end{aligned} \tag{16}$$

where we used the approximation, $\ln Z_\epsilon \approx M \ln 2 + (\epsilon w^I)^2/2$, in the third line. This important property originates from the first-order truncation of ϵ in Eq. (14). Thus, as the gradient of $\ln \mathcal{L}_\epsilon$ vanishes ($\delta w^I \rightarrow 0$), and the re-weighted data distribution f_ϵ becomes almost flat, $\mathcal{L}_\epsilon$ converges to 2^{-M} , which corresponds to maximal uncertainty. We emphasize here that going to higher orders in the expansion of Z_ϵ is unnecessary, as the form of the leading order expansion already contains *all* the information about the complete space of configurations.

Summarizing the learning algorithm, we first re-weight the data distribution $f(\sigma)$ by multiplying with our present guess for the model distribution raised to a power $p^{\epsilon-1}(\sigma)$ so that it should almost completely erase the frequency information in observed configurations up to a distribution with Boltzmann weights $\propto \exp(-\epsilon E)$. Holding this re-weighted data distribution fixed, we then update the model distribution with ϵ -dependent weights. Complete erasure obviously implies knowing the true model distribution $p(\sigma)$. Thus, the re-weighting allows us to solve the inference problem using the leading order high-temperature expansion. This motivates us to name this algorithm the *erasure machine*.

The practical recipe of the erasure machine is as follows:

- i. compute $p(\sigma) \propto \exp(w^I O_I(\sigma))$ initially with random or vanishing w^I — note that the inference is insensitive to the initial w^I ;
- ii. re-weight $f_\epsilon(\sigma) = f(\sigma) p^{\epsilon-1}(\sigma) / \sum_{\sigma'} f(\sigma') p^{\epsilon-1}(\sigma')$ — note that this does not depend on the normalization of p in step (i);
- iii. obtain $\langle O_I \rangle_{f_\epsilon} = \sum_{\hat{\sigma}} O_I(\hat{\sigma}) f_\epsilon(\hat{\sigma})$;
- iv. update $w^I \rightarrow w^I + \alpha \left(\langle O_I \rangle_{f_\epsilon} - \epsilon w^I \right)$;
- v. iterate (i)-(iv).

To complete the algorithm, we should determine when to stop the iteration. Writing $\sum_{\hat{\sigma}} f(\hat{\sigma}) E(\hat{\sigma}) \equiv \langle E \rangle_f$, the mean energy $\langle E \rangle_f$ kept decreasing and finally saturated. Here $\langle E \rangle_f$ is easily computed as a sum over only observed configurations. Thus we adopted the mean energy as a proxy for monitoring the iterations, and we stopped the iterations at a saturation point.

The algorithm has a hyperparameter ϵ . Note that we did not, in fact, maximize the likelihood defined in Eq. (1). Our computation strictly required only the relative likelihoods of all observed configurations which do not depend on the absolute probability. These relative likelihoods depend only on the differences in energies of configurations, and it follows that finding the w^I parameters accurately will lead to more accurate relative probability predictions. Our experiments with synthesized data, explained later, showed that larger mean energy $\langle E \rangle_f$ implies larger energies $E(\hat{\sigma})$ of observed configurations $\hat{\sigma}$. The maximal value of $\langle E \rangle_f$ is reached at the value of ϵ that minimizes the mean squared error (MSE) between the true parameters and the inferred parameters. An intuition for this result is that the algorithm is balancing between making the observed configurations more likely and taking into account the strong coupling form of the re-weighted model partition function, which implicitly takes into account the entire universe of configurations, observed and unobserved. The former pushes the energies of observed configurations to lower values, increasing the relative likelihood of observing these configurations, and the latter pulls these energies to higher values, decreasing the probability of observing these configurations out of the universe of all possible configurations. For ϵ too small, the erasure machine is underdamped and takes the data into account too strongly, whereas for ϵ too large, the machine is overdamped and the first order form of the model partition function cannot be made to agree with the exact ϵ dependence of the data term. We run parallel erasure machines with different ϵ values, and choose optimal ϵ that results in a maximal mean energy if we want accurate inverse Ising inference.

Note the difference between inference and prediction. Using $-\ln(Z_1/Z_0) \approx \int_0^1 d\beta \langle E \rangle_{p_\beta}$, with $\langle E \rangle_{p_\beta} \equiv -\partial \ln Z_\beta / \partial \beta = \sum_{\hat{\sigma}} E(\hat{\sigma}) \exp(-\beta E(\hat{\sigma})) / Z_\beta$ and $Z_\beta \equiv \sum_{\hat{\sigma}} \exp(-\beta E(\hat{\sigma}))$, as an approximation for a free energy $F_\beta \equiv -\ln Z_\beta$, we find that the likelihood is maximized at a different value of ϵ that is relevant for the prediction problem, because it takes into account the more likely configurations weighted accordingly. This is different from the inverse Ising problem where only the relative probabilities of any two configurations are relevant.

III. RESULTS

Now we demonstrate that the erasure machine can efficiently infer model parameters for explaining the distribution of observed configurations, especially in the regime where standard MLE is intractable. For clear demonstration, we adopt an energy function of the Ising model, $E(\sigma) = \sum_i h^i \sigma_i + \sum_{j < k} J^{jk} \sigma_j \sigma_k$, which has $L = M + M(M-1)/2$ parameters $w = \{h^i, J^{jk}\}$. First, for simulating data, we randomly set the parameter values from a Gaussian distribution with zero mean and some variance that specifies bias and coupling strength, and define them as w_{true} . We then generate observations $\hat{\sigma}$ from the distribution $p(\sigma | w_{\text{true}})$. The

goal of the inverse problem is to find values of these parameters, i.e., to recover w_{true} , in order to make the model distribution $p(\sigma|w_{\text{true}}) = \exp(-E(\sigma))/Z$ close to the observed distribution $f(\sigma)$ by examining only data $\{\hat{\sigma}\}$.

The erasure machine can iteratively determine w by using Eq. (15). As iteration goes on, inferred w gets closer to w_{true} as quantified by the mean squared error,

$\text{MSE} = L^{-1} \sum_I (w^I - w_{\text{true}}^I)^2$ (Fig. 1a). For small ϵ , too many iterations sometimes lead to worse inference. However, for large ϵ , inference accuracy improves during the iteration, and finally becomes saturated (Fig. 1a). The final value of MSE after training depends on the value of ϵ (Fig. 1b). Too small ϵ does not sufficiently regularize for determining w^I in Eq. (15), whereas ϵ too large can result in an over-damped iteration because of an invalid small ϵ expansion ($|\epsilon w^I| < 1$) in Eq. (12). Therefore, it is important to find the optimal ϵ for the smallest MSE. Note that the MSE is analytic in ϵ when it is large enough to avoid w values diverging, as mentioned above. The flattening at the minimum of the MSE (Fig. 1b) then implies that the MSE is actually a constant in a range of ϵ . As the sample size N becomes larger, the inference error of MSE becomes almost negligible for a wider range of ϵ , suggesting that this method becomes exact in the limit of infinite sample size.

For real data, however, the MSE is not available as w_{true} is unknown, so we examined an alternative measure, the mean energy of observed configurations, $\langle E \rangle_f = \sum_{\hat{\sigma}} f(\hat{\sigma}) E(\hat{\sigma})$,

without re-weighting, which is also clearly analytic in ϵ for large enough ϵ . This mean energy decreased during iterations (Fig. 1c). While the final value of mean energy also depends on the value of ϵ (Fig. 1d), for a range of ϵ , just as MSE became a constant, $\langle E \rangle_f$ shows minimal changes: MSE becomes minimal, whereas $\langle E \rangle_f$ becomes maximal.

Therefore, the erasure machine first works in parallel and independently for a range of ϵ , then picks the optimal ϵ that maximizes $\langle E \rangle_f$. The stable range become larger as the number of samples increases (Fig. 1b and d). This implies that the erasure machine is not sensitive to the choice of hyperparameter ϵ given sufficient samples.

We now compare the performance of the erasure machine (EM) with existing methods (Fig. 2). For a small system ($M=20$), MLE can be used because the number of every possible configuration ($2^{10} \approx 10^6$) is computable. In the regime of weak interactions and large sample size, the EM performs as well as MLE and pseudo-likelihood estimation (PLE) [28] (Fig. 2a). On the other hand, the Hopfield solution (HF) is less accurate than the other three methods (Fig. 2c). However, in the regime of strong interactions and/or small sample size, EM outperforms HF, MLE and PLE (Fig. 2f). For a larger system ($M=40$), MLE is intractable because the partition function includes 2^{40} configurations. EM works as well as PLE in the limit of weak interaction and large sample (Fig. 2i). However, EM works significantly better than PLE in the limit of strong interactions and/or small sample size (Fig. 2l). Nonetheless, HF does not do well compared with EM and PLE. As the system size becomes much larger ($M=200$), only EM and HF can be used because PLE becomes intractable. EM still works well in this limit (Fig. 2mr). Furthermore, because EM only considers observed configurations, it significantly reduces computing time, compared with MLE and PLE. For $M=20$, EM takes approximately 25 times and 2 times less time than

MLE and PLE, respectively. For $M=40$, EM takes approximately 8 times less time than PLE (Fig. 3).

Given the demonstration with the synthetic data, we apply the erasure machine to solve real-world problems. First, we apply the erasure machine to recover missing values in data. We used an experimental dataset on smoking. The data set contains 1,536 samples of 38 binary variables including yes/no answers of interviewees on smoking restrictions in public places and tobacco advertisement bans, and their personal information such as marriage status, gender, educational level, and smoking history [29]. We first split the data into 460 training samples (Fig. 4a) and 1,076 test samples (Fig. 4d). The large portion (70%) of test samples was saved for concrete validation of our inference. By applying the erasure machine with the training samples, we inferred the local fields H^i for individual variables (Fig. 4b) and pairwise interactions J^{ik} between variables (Fig. 4c). Here the estimated local fields and interactions cannot be directly evaluated because their true values are unknown in this real-world problem. Instead, we can evaluate the inference through the prediction for the test samples. We randomly selected 14 variables ($\approx 37\%$ of total 38 variables) in test samples, and define them ($\sigma_i = 0$) as missing variables (Fig. 4e). To recover the missing values, we divided the entire variables $\sigma = (\sigma_m, \sigma_m^c)$ into missing variables σ_m and observed variables σ_m^c . Then, by maximizing $p(\sigma_m | \sigma_m^c) \propto p(\sigma_m, \sigma_m^c) \equiv p(\sigma)$, we can reconstruct σ_m (Fig. 4f). To estimate the reconstruction performance of EM, we computed the accuracy for varying numbers of missing variables and compared with the PLE method. Overall, it is clear that EM produces a better result than PLE. For instance, when each sample has 14 missing values, the accuracy of EM and PLE is 84.2% and 64.4%, respectively. Furthermore, while the accuracy of EM is quite stable as the number of missing variables increases and does not depend strongly on the value of the hyperparameter ϵ , the accuracy of PLE markedly decreases (Fig. 4g).

As shown in the previous sections, the PLE is intractable for large systems ($M \geq 100$), however the erasure machine works effectively for large systems. Here we emphasize this aspect by reconstructing missing pixels in real images. We use the MNIST images of handwritten digits [30]. The gray-scale values x_i of 28×28 pixel images are binarized, $\sigma_i = 1$ for $x_i > 1$ or $\sigma_i = -1$ otherwise (Fig. 5a). Given a test image, we randomly select 90 pixels ($>10\%$ of total 784 pixels), and define them ($\sigma_i = 0$) as missing pixels (Fig. 5b). Our goal is to reconstruct the missing pixels to recover the original image. Specifically, we use $N=5,851$ samples of the digit 8 in the MNIST training data. First, if the i -th pixel has a common value of σ_i for more than 80% of the training samples, the i -th missing pixel in the test image is simply reconstructed by the common σ_i value. However, the remaining $M=222$ pixels have a large sample variation. Therefore, similar to the previous example, we apply the erasure machine to obtain $p(\sigma)$ by inferring pixel bias and interactions, then reconstruct missing pixels (Fig. 5c).

IV. DISCUSSION

Inferring underlying models from observed configurations has become a cynosure with the present flood of big data. However, big data is not yet big enough to use most available

inference methods for large systems. In this study, we proposed a data-driven algorithm for solving the inverse Ising problem without any assumption on the connectivity (e.g., weak coupling, sparse networks, no cycles, etc.) of underlying systems. Unlike standard maximum likelihood estimation, our algorithm relies entirely on observed configurations with no need to sum over the vast number of unseen configurations. We systematically re-weighted the frequency of observed configurations to increase the entropy of the re-weighted observed configuration distribution, and in the process simplified the computation of the exact partition function to taking its leading non-trivial order approximation.

We showed that this lowest order gives a quadratic form to the re-weighted log-likelihood. Nevertheless, one may be tempted to consider that higher order approximations would provide better results. Here some statistical physics background is helpful. The usual square lattice Ising model with only nearest neighbor interactions has been shown to have a convergent high-temperature expansion for a range of temperatures, but the proof explicitly uses the finiteness of the number of neighbors for each spin. In the general setting of machine learning, there is no chance for a general convergence theorem. The real reason why the leading order term is enough for our purposes is that its form implicitly encodes the fact that all values of every spin have been summed over. This informs the optimization step that there are many more configurations of the system than just the observed configurations. The value of the mean energy of observed configurations decreases to a limiting value for any fixed value of the hyperparameter ϵ as the iteration proceeds, but the value of ϵ chosen is the one with the highest limiting value.

Since the erasure machine requires only the computation of expectation values of observables in the re-weighted observed ensemble, it is very fast. Furthermore, it gives more accurate inference results than state-of-the-art pseudo-likelihood methods in the difficult inference regime of limited sample size or strong coupling.

In this study, we adopted the Ising model with local fields and pairwise interactions to demonstrate our idea of re-weighting. It is straightforward to extend this idea to other models using the exponential family of distributions. The concept of flattening the observed distribution and simplifying the partition function can be further extended to consider hidden variables [31, 32], continuous variables [22, 33], nonequilibrium asymmetric couplings [34], and other inference problems where the computation of the partition function is unfeasible.

Acknowledgments

We thank Juyong Song for discussions during the initial stages of this study. This work was supported by Intramural Research Program of the National Institutes of Health, NIDDK (D.-T.H., V.P.), and by the New Faculty Startup Fund from Seoul National University and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1F1A1052916) (J.J.).

References

- [1]. Schneidman E, Berry II MJ, Segev R, and Bialek W, Nature 440, 1007 (2006). [PubMed: 16625187]
- [2]. Cocco S, Leibler S, and Monasson R, Proceedings of the National Academy of Sciences 106, 14058 (2009).

- [3]. Watanabe T, Hirose S, Wada H, Imai Y, Machida T, Shirouzu I, Konishi S, Miyashita Y, and Masuda N, *Nature communications*4, 1370 (2013).
- [4]. Lezon TR, Banavar JR, Cieplak M, Maritan A, and Fedoroff NV, *Proceedings of the National Academy of Sciences*103, 19033 (2006).
- [5]. Weigt M, White RA, Szurmant H, Hoch JA, and Hwa T, *Proceedings of the National Academy of Sciences*106, 67 (2009).
- [6]. Cocco S, Feinauer C, Figliuzzi M, Monasson R, and Weigt M, *Reports on Progress in Physics*81, 032601 (2018). [PubMed: 29120346]
- [7]. Mora T, Walczak AM, Bialek W, and Callan CG, *Proceedings of the National Academy of Sciences*107, 5405 (2010).
- [8]. Bialek W, Cavagna A, Giardina I, Mora T, Silvestri E, Viale M, and Walczak AM, *Proceedings of the National Academy of Sciences*109, 4786 (2012).
- [9]. Bury T, *Journal of Statistical Mechanics: Theory and Experiment*2013, P11004 (2013).
- [10]. Borysov SS, Roudi Y, and Balatsky AV, *The European Physical Journal B*88, 321 (2015).
- [11]. Eagle N, Pentland AS, and Lazer D, *Proceedings of the national academy of sciences*106, 15274 (2009).
- [12]. Chickering DM, in *Learning from data* (Springer, 1996), pp. 121–130.
- [13]. Welsh D and Welsh DJ, *Complexity: knots, colourings and countings*, vol. 186 (Cambridge university press, 1993).
- [14]. Nguyen HC, Zecchina R, and Berg J, *Advances in Physics*66, 197 (2017).
- [15]. Tanaka T, *Physical Review E*58, 2302 (1998).
- [16]. Jacquin H and Raçon A, *Physical Review E* 94, 042118 (2016). [PubMed: 27841631]
- [17]. Ricci-Tersenghi F, *Journal of Statistical Mechanics: Theory and Experiment*2012, P08015 (2012).
- [18]. Wu D, Wang L, and Zhang P, *Physical review letters*122, 080602 (2019). [PubMed: 30932595]
- [19]. Ackley DH, Hinton GE, and Sejnowski TJ, *Cognitive science*9, 147 (1985).
- [20]. Cocco S and Monasson R, *Physical review letters* 106, 090601 (2011). [PubMed: 21405611]
- [21]. Nguyen HC and Berg J, *Physical review letters* 109, 050602 (2012). [PubMed: 23006160]
- [22]. Sohl-Dickstein J, Battaglino PB, and DeWeese MR, *Physical review letters*107, 220601 (2011). [PubMed: 22182019]
- [23]. Lokhov AY, Vuffray M, Misra S, and Chertkov M, *Science advances*4, e1700791 (2018). [PubMed: 29556527]
- [24]. Ravikumar P, Wainwright MJ, Lafferty JD, et al., *The Annals of Statistics*38, 1287 (2010).
- [25]. Aurell E and Ekeberg M, *Physical review letters* 108, 090201 (2012). [PubMed: 22463617]
- [26]. Decelle A and Ricci-Tersenghi F, *Physical review letters* 112, 070603 (2014). [PubMed: 24579583]
- [27]. Jo J, Hoang D-T, and Periwal V, *Erasure Machine* (Accessed: September 4, 2019), <https://nihcompmed.github.io/e-machine/>.
- [28]. Lee ED and Daniels BC, *Journal of Open Research Software* 7 (2019).
- [29]. Bakhturidze G, Mittelmark M, Aarø L, and Peikrishvili N, *BMJ open*3, e003461 (2013).
- [30]. Lecun Y, Bottou L, Bengio Y, and Haffner P, *Proceedings of the IEEE*86, 2278 (1998), ISSN 0018–9219.
- [31]. Hoang D-T, Jo J, and Periwal V, *Physical Review E*99, 042114 (2019). [PubMed: 31108681]
- [32]. Roudi Y and Hertz J, *Physical review letters* 106, 048702 (2011). [PubMed: 21405370]
- [33]. Donner C and Oppen M, *Physical Review E* 96, 062104 (2017). [PubMed: 29347355]
- [34]. Hoang D-T, Song J, Periwal V, and Jo J, *Physical Review E*99, 023311 (2019). [PubMed: 30934224]

FIG. 1.

(Color online) Learning of the erasure machine. (a) Mean squared error (MSE) between actual and inferred interactions and (c) mean energy of observed configurations during iterations with $N = 5,000$ for different ϵ values: $\epsilon = 0.3$ (dashed black line), 0.6 (solid red line), and 0.9 (dotted blue line). The minimal values of (b) MSE and (d) mean energy during iterations depending on ϵ values. For the learning, we used a system with $M = 40$ spins, and two sample sizes with $N = 5,000$ (filled black circles) and $N = 10,000$ (empty blue squares).

FIG. 2.

(Color online) Inference performance of the erasure machine. Inferred interactions versus actual interactions for large (first column) and small sample sizes (second column), and for weak (odd row) and strong interactions (even row). Note that the true interaction weights w_{true} are sampled from a normal distribution $\mathcal{N}(0, g^2/M)$. We use $g = 0.5$ and 1.0 to generate weak and strong interactions, respectively. Mean squared errors (MSE) between inferred and actual interactions are plotted as a function of sample sizes (third column). Inference performances of Hopfield solution (HF, empty green triangles), maximum likelihood estimation (MLE, empty black inverse triangles), pseudo-likelihood estimation (PLE, empty

blue squares), and erasure machine (EM, filled red circles) are compared for various system sizes ($M = 20, 40, 200$). Here MLE is not available for a larger systems ($M = 40, 200$), and PLE is not available for the largest system ($M = 200$).

FIG. 3.

(Color online) Computing time of the erasure machine. Computing time (seconds) is compared between erasure machine (EM, filled red circles), maximum likelihood estimation (MLE, empty black triangles), and pseudo-likelihood estimation (PLE, empty blue squares) for different systems sizes given $N = 10,000$ samples.

FIG. 4.

(Color online) Inferring missing data with erasure machine. (a) Configurations of some training samples with black dots representing +1 values and white dots representing -1 values. (b) Inferred local fields for each variable and (c) pairwise interactions between variables from training samples. (d) Original configurations of some test samples. (e) Noisy test samples by randomly missing 14 variable values from each original test sample (gray dots). (f) Recovered test samples using the inference of the erasure machine. (g) Recovery accuracy of the erasure machine (EM, filled red circles $\epsilon = \epsilon^* = 0.84$, empty green triangles $\epsilon = 0.7$, empty teal inverted triangles $\epsilon = 0.98$) and pseudo-likelihood estimation (PLE, empty blue squares) for varying numbers of missing variables.

FIG. 5.

(Color online) Image reconstruction by the erasure machine. (a) An MNIST image. (b) 90 pixel values are missed (green pixels). (c) Recovered image with missing pixels reconstructed by the erasure machine.