Ising-CF: A Pathbreaking Collaborative Filtering Method Through Efficient Ising Machine Learning

Zhuo Liu[1], Yunan Yang[2], Zhenyu Pan[1], Anshujit Sharma[1], Amit Hasan[3], Caiwen Ding[3], Ang Li[4], Michael Huang[1] and Tong Geng [1]

[1]University of Rochester, [2]ETH Zürich, [3]University of Connecticut, [4]Pacific Northwest National Laboratory

2023 60th ACM/IEEE Design Automation Conference (DAC) | 979-8-3503-2348-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/DAC56929.2023.10247860

***Abstract*—Due to the Ising model's strong expressivity and Ising machines' unique computational power, it is highly desired if Ising-based learning can be used in real-world applications. Unfortunately, the challenges in learning the Ising model and gaps between the practical accuracy of Ising machines and the theoretical accuracy of the Ising model impede the realization of Ising machines' potential. Hence, we propose an Ising Machine Learning framework, Ising-CF, for collaborative filtering, a widely-used recommendation method. Specifically, Ising-CF uses Linear Neural Networks with Besag's pseudo-likelihood and voltage polarization for fast, accurate Ising model learning and an Ising-specific logarithmic quantization for ns-level Ising machine inference with near-theoretical accuracy, 7.3% over SOTA.**

***Index Terms*—Ising Machine, Ising Model, Collaborative Filtering, Machine Learning**

I. Introduction

Collaborative Filtering (CF) is an increasingly critical technique of information filtering in this information explosion era and is playing a dominant role as a recommender in big data applications. After decades of development, CF techniques have been developed from latent linear models to complex deep non-linear models, sparking huge interest in both industry and academia. Amazon, YouTube, iTunes, Microsoft, and Netflix all use CF for commercial recommendations; Physicians use CF for drug recommendation, and Chemists use CF for chemical reactant recommendation [1].

Despite continuous efforts on the development of CF, there is, unfortunately, still no significant performance improvement in the past decade. As Fig. 1 shows, the recalls of current SOTA CFs are still around the ones 14 years ago, with massive space for improvement. Moreover, real-time CF recommendation is another challenge since almost all SOTA solutions are based on Matrix Multiplication in digital systems, which strictly bound the theoretical minimum latency far exceeding many real-world demands.

The stagnation of the development of CF is mainly due to the lack of information on users and items, making it hard to improve accuracy by simply increasing the model's complexity. Specifically, unlike other recommendation methods (e.g., META's DLRM), which work with adequate information, CF typically needs to learn highly sophisticated correlations among objects from very limited information [1]. Therefore, an efficient CF demands highly flexible logic interconnects among all users and items to enable both accurate and fast cascading passes of the limited user-item interplay information

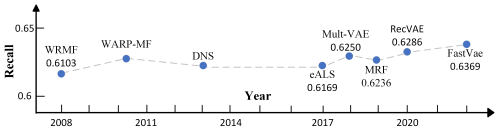

Fig. 1. Recall (%) of SOTA CFs with Movielens-100K in the past decade.

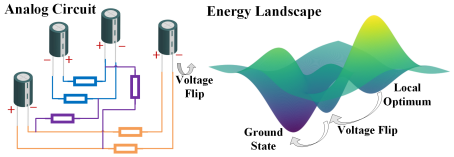

Fig. 2. Analog circuit and energy landscape of BRIM Ising machine.

where traditional methods applied to digital systems fall short. New CF methods are highly desired.

To this end, we investigate the potential of the Ising model and the Ising machine as the key to future CF for following reasons. (1) Ising models' strong expressivity: the Ising model is a probabilistic graphical model rooted in statistical physics (of magnetism) that has been widely used to describe complex systems. Ising models use complete graphs to embed the correlations among systems' objects providing strong connectivity, expressivity, and unique long-range cascading propagation of information. Solving the Ising model is to find the low-energy modes of systems' energy landscapes. (2) Ising machines' computation with "speed of electrons": Recently developed Ising machines (e.g. BRIM [2]) automatically and quickly find lowest-energy states via physics-*based* dynamical systems that naturally seek those states (e.g., via charging and discharging of nano-scale capacitors as shown in Fig. 2) much faster and more efficiently than can be done by the state-of-the-art, physics-*inspired* algorithm using traditional von Neumann computing. It has been shown that an mW-level CMOS-based BRIM solves properly formulated NP-complete optimization problems with orders of magnitude speedups over traditional processors. These observations suggest that if CF can be formulated as the Ising model with precisely learned parameters so that the low-energy states represent correct recommendations, the resulting Ising-based CF approach will enjoy both high expressivity from the Ising model and high performance from Ising machines.

Despite the above promising potential of Ising model-based approaches, real-world adoption of the Ising model and Ising machines come with prohibitive challenges. First, it is hard to construct the Ising model with precise parameters trained from historical data. On the one hand, globally optimizing

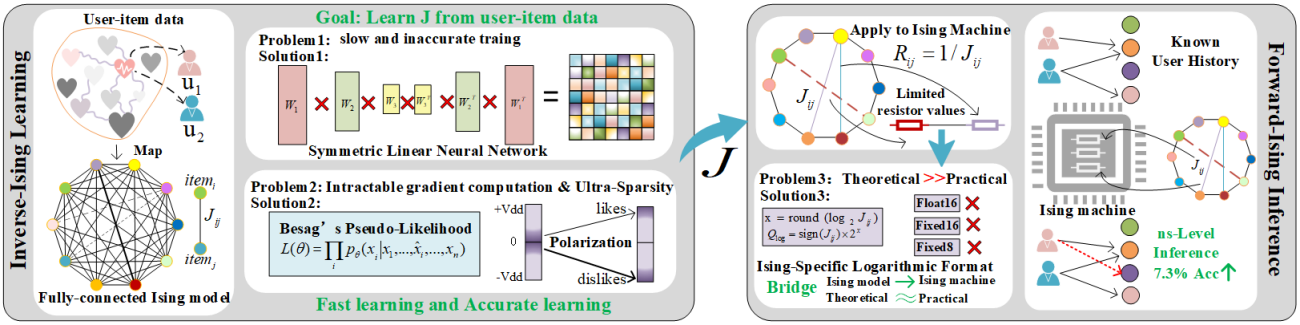

Fig. 3. An overview of the proposed Ising-CF framework including Inverse-Ising learning (left) and Forward-Ising inference (right).

n^2 parameters of the Ising model during training is infeasible with a large model size n. In comparison, local optimization of parameters reduces Ising models' expressivity but makes training easier. On the other hand, the exact maximum likelihood estimation for constructing the Ising model is intractable as exact gradient computation takes exponential time in n. Second, for the applications with significantly unbalanced positive and negative samples (e.g., 0.01% recommendation rate in CF), the positive voltages carried by the capacitors representing the items for recommendation are normally clustered around 0 volt, blurring the boundary of "like" and "dislike" and leading to accuracy degradation. Third, the physical Ising machines only support limited precision (e.g., 8 bits) of coupling strengths, resulting in the same limited choices of coupling parameters, which can cause accuracy gaps between theory and practice.

To address these challenges and unleash the potential of the Ising model, we propose an **Ising Machine Learning** framework (shown in Fig. 3), Ising-CF, taking CF as a case study. In particular, Ising-CF consists of two stages, Inverse-Ising learning and Forward-Ising inference. (1) Inverse-Ising learning formulates CF as the Ising model with precisely learned Ising parameters (J), creating energy landscapes where the local minima represent recommendations for different users. For accurate and fast training, Inverse-Ising uses symmetric linear neural networks to provide efficient global optimization of Ising parameters and Besag's pseudo-likelihood with voltage polarization to enlarge the voltage differences between recommendation and non-recommendation items with simplified gradient computation. (2) Forward-Ising inference maps the Ising model trained in (1) onto an Ising machine for inference and is equipped with an Ising-specific logarithmic quantization that better matches the distribution of the Ising parameters, further closing accuracy gaps between the theoretical Ising model obtained during training and the practical Ising model run on Ising machines. To the best of our knowledge, Ising-CF is the first work to use Ising methods to outperform SOTA solutions in the real world. Our contributions are as follows:

- We propose an Ising Machine Learning framework for CF, Ising-CF, a path-breaking recommendation method that realizes Ising methods' potential in the real world.
- We propose a novel Inverse-Ising learning method equipped with Linear Neural Network, pseudo-likelihood, and voltage polarization to fast and accurately construct the Ising model that matches real-world problems.
- We propose an Ising-specific logarithmic quantization technique that enables lossless theoretical Ising model mapping to practical Ising machines.
- Experiments show that Ising-CF provides, on average 8186× speedup & 7.3% recall improvement over 6 carefully selected SOTA solutions with 4 real-world datasets.

II. Background and Related Work

A. Ising Model and Ising Machine

The Ising model is typically shown as a graph with two-state spins as nodes and couplings among spins as edges. Given n spins in the system, the phase space $s = \{\sigma_i\}_{i=1}^n \in \{\pm 1\}^n$ is governed by the Hamiltonian (or the energy function) $H(s)$, forming an energy landscape throughout the entire phase space. Specifically, the energy function (Eq. 1) is composited by a real-valued coupling matrix $J \equiv \{J_{ij}\}_{i,j=1}^n$, describing the topology of the graph and the interaction strength of neighboring spins, and a vector $h \equiv \{h_i\}_{i=1}^n$ representing the reaction of each spin to an external field. Both J and h are Ising model parameters obtained during Inverse-Ising learning.

$$H = \sum_{i<j} J_{ij}\sigma_i\sigma_j + \mu \sum_i h_i\sigma_i. \quad (1)$$

An Ising machine is a physical implementation of the Ising model, which naturally tends to evolve towards the state $s = \{\sigma_1, \ldots, \sigma_n\}$ with lower energy $H(s)$. Therefore an Ising machine acts as an effective solver to an optimization problem in the Ising formulation. Many hardware prototypes or concepts of Ising machines have been developed, including D-Wave's quantum annealers, Coherent Ising Machines, Electronic Oscillator-based Ising Machines, and the recently proposed efficient CMOS-compatible BRIM [2]. In this paper, we use BRIM as a substrate for Ising-CF in Forward-Ising inference . More details can be found in [2, 3].

B. Related Work

Collaborative filtering methods: One-class collaborative filtering has been a hot topic recently, and numerous methods have been proposed. Classical neighborhood-based approaches are based on item-item (user-user) similarity matrices. Matrix factorization methods are further proposed by decomposing the rating matrix, e.g., WRMF, WAPR-MF [4]. Recently, variational autoencoders have been utilized to solve collaborative filtering and have become SOTA methods, e.g., FastVae [5].

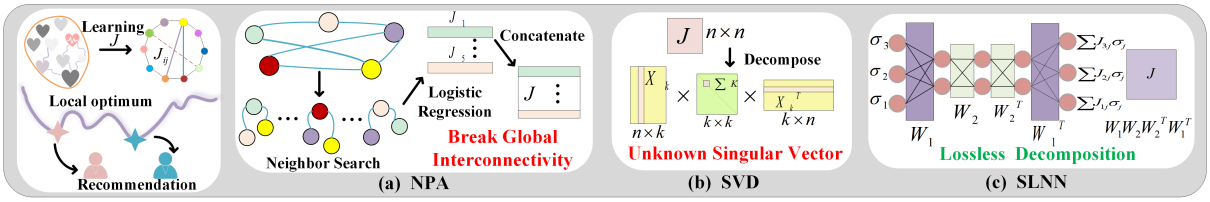

Fig. 4. Overview of Inverse-Ising learning (left) and illustration of 3 methods for coupling matrix decomposition including (a) NPA, (b) SVD, (c) SLNN.

Ising-based methods: The Ising machines have been increasingly popular in solving many optimization problems. [6, 7] uses a fully-connected Ising machine to solve traveling salesman problems. [2, 3] uses the Ising machine to solve graph cut problems. However, the potential of Ising machines in real-world applications has not been demonstrated before.

III. Methodology

A. Problem Definition and Framework Overview

As one of the most important techniques in recommendation systems, CF aims to recommend items to users with potential interests according to historical user preference (e.g., likes & dislikes). We use $\mathbb{U}$ and $\mathbb{I}$ to denote the set of all users and all items, respectively, with $m = |\mathbb{U}|$ and $n = |\mathbb{I}|$. Let $S = (s_1, \ldots, s_u, \ldots, s_m)$ denote the observation of user-item interactions of m users where the vector $s_u = (\sigma_{u,i}, \ldots, \sigma_{u,n})$ represents the user u's interaction with n items. For each user's observation (s_u), the Boltzmann distribution can be used to determine its likelihood:

$$P(s_u) = \frac{1}{Z} \exp\left\{ \sum_{i<j} J_{ij}\sigma_{u,i}\sigma_{u,j} + \sum_i h_i \sigma_{u,i} \right\}, \quad (2)$$

where Z is the partition function, J_{ij} represents the correlations of item i & j in recommendation, h_i represents the recommendation bias of item i, and $\sigma_{u,i}$ represents the interaction between user u and item i with $\sigma_{u,i}$ = 1 or -1 representing "like" or "dislike" respectively. Let $O = \{(u,i)|\sigma_{u,i} \in \{1,-1\}\}$ be a set of observed user-item interactions and $\hat{O} = \{(u,i)|\sigma_{u,i}\}$ be the set of unobserved user-item interests. The goal of CF is to predict the status of $\sigma_{u,i}$ in $\hat{O}$ given the observed interaction history in O.

We map the CF problem above onto an Ising model to employ the Ising machine in user-item recommendations. Specifically, we use spins of the Ising model to represent items and find shared Ising parameters (J_{ij}, h_i) that describe item correlations for the recommendation. We then construct a shared Ising model that applies to all users. To generate recommendations for a particular user, we fix the spins (items) whose interactions with this user are observed. These spins form a restricted region with an energy landscape, and we then employ Ising machines to find the lowest energy state in the space automatically. The spin configurations (lowest energy states) found by an Ising machine are the recommendations for this user. Different users have different observed interactions, so their recommendation results are located at different local minima of the shared landscape.

Ising-CF is a framework that can efficiently and accurately construct the shared Ising model through ***Inverse-Ising learning (S.III.B)*** and perform ultra-fast recommendation through ***Forward-Ising inference (S.III.C)*** using the Ising machine.

B. Inverse-Ising Learning

At the heart of many approaches to reconstructing the Ising model is the maximum likelihood framework. As mentioned above, $S = (s_1, \ldots, s_u, \ldots, s_m)$ follows the Boltzmann distribution of the Ising model and the observation s_u for user u corresponds to one sample drawn from this distribution. The so-called maximum likelihood estimator is as follows:

$$\{J_{ij}, h_i\}^{ML} = \text{argmax}_{J_{ij},h_i} P(s_1, s_2, ..., s_m | J_{ij}, h_i), \quad (3)$$

where P is given in (2) and $1 \le i, j \le n$. Hence, the goal of Inverse-Ising learning is to learn J (ignoring h for clarity) effectively from the observed data. To meet this goal, we use interconnection coupling matrix decomposition and Besag's pseudo-likelihood approximation.

1) Interconnection Coupling Matrix Decomposition

It is intractable to optimize n^2 parameters in J simultaneously considering the huge numbers of items (n) in real-world applications. We first investigate the neighborhood pursuit and singular value decomposition methods to reduce the size of the optimization problem to tackle this challenge.

(a) Neighborhood Pursuit Algorithm (NPA): The node-wise ℓ^1-regularized logistic regression approach, called neighborhood pursuit [8], can be used to estimate the Ising model. As Fig. 4 shows, its key idea is to break a whole Ising model into many sub-models. According to the pair-wise Markov property, an interaction $\sigma_{u,i}$ of user u with item i only depends on its neighbors $\sigma_{u,\setminus i}$. Thus, each of the n items is taken in turn as the dependent variable against the remaining items. The coupling interconnections for interaction i correspond to the vector $J_{i,.} = (J_{i,j}, \forall j \in \mathbb{I}_{\setminus i})$. Under this assumption, the estimation of the coefficient vector $J_{i,.}$ for item i is finally accomplished by addressing the following ℓ^1-regularized logistic regression problem:

$$\hat{J}_{i,.} = \text{argmax} \frac{1}{m} \sum_{u=1}^{m} \sum_{i=1}^{n} \log l(\sigma_i \in \mathbb{I}_{\setminus i}, \hat{J}_{i,.}) + \lambda \|\hat{J}_{i,.}\|_1, \quad (4)$$

where l is a logistic function. Hence, learning the coupling matrix J is feasible via a row-wise training strategy. However, NPA breaks the global connection of the Ising model where the high expressivity of the Ising model comes from, resulting in poor accuracy in real-world problems.

(b) Singular Value Decomposition (SVD): SVD is a widely-used matrix factorization method that enables low-rank approximation. For the Ising model, the coupling matrix J that is required to be symmetric due to Ising machines' architecture can be approximated by its truncated SVD: $J \approx X_k \Sigma X_k^T$, where X_k is an $n \times k$ orthogonal matrix and Σ is a $k \times k$ diagonal matrix whose elements are the singular values in decreasing order. The matrix J can also be approximated using the sum of the multiplication of the left and right singular

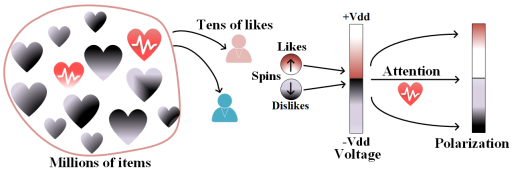

Fig. 5. Illustration of the Ising-voltage polarization.

vectors with large singular values (see Fig. 4 (b)). That is,

$$J \approx \sum_{i=1}^{k} e_i x_i^T x_i = e_1 x_1^T x_1 + ... + e_k x_k^T x_k, \tag{5}$$

where $\{x_i\}$ are column vectors of X_k. Although SVD can offer a low-rank parameterization of J, it still has three main shortcomings. First, not all real-world applications have low-rank features. Second, no prior knowledge is given about singular vectors for real-world applications. Finally, the SVD representation of J imposes extra constraints requiring matrices X_k and Σ to be orthogonal and diagonal respectively.

(c) Symmetric Linear Neural Network (SLNN): To address the problems of the previous two methods, we propose to use SLNNs to decompose and parameterize J, considering that the Hamiltonian of the Ising model is a linear map of spin configurations. In the SLNN, the dimension of the input and output layers are equal to the number of items (n) and the output is a linear transformation of the input spin configurations. The symmetry of the LNN guarantees matrix J to be symmetric meeting the requirement of Ising Machines. As Fig. 4 shows, J can be represented using the weight matrices of SLNNs:

$$J = \prod_{l=1}^{L} W_l, \tag{6}$$

where L is the number of layers and W_l is the weight matrix at the l-th layer. Compared with other methods, SLNN has the following strengths. First, it is a joint learning strategy without breaking the global interconnections of the Ising model, thus achieving higher accuracy. Second, unlike SVD, elements in J learned with SLNN can take any real values, which brings higher flexibility to the Ising model construction. Finally, SLNNs lead to faster Inverse-Ising learning by significantly reducing parameter dimensions.

2) Besag's Pseudo-Likelihood

After parameterizing J using SLNNs, we now apply maximum likelihood to learn the Ising model. The log-likelihood of the model parameters given the observed data S is

$$\begin{aligned} L_S(J) &= \frac{1}{m}\sum_{u=1}^{m} \log P(s_u|J) \\ &= \{\sum_{i<j} J_{ij} \langle \sigma_{u,i}\sigma_{u,j}\rangle + \sum_i J_{ii} \langle \sigma_{u,i}\rangle - \log Z(J)\}, \end{aligned} \tag{7}$$

where $Z(J)$ is the partition function. To apply the gradient descent optimization, we compute the gradients as follows:

$$\frac{\partial L_S(J)}{\partial J_{ij}} = \langle \sigma_{u,i}\sigma_{u,j}\rangle^S - \langle \sigma_{u,i}\sigma_{uj}\rangle^M, \quad i \neq j, \tag{8}$$

$$\frac{\partial L_S(J)}{\partial J_{ii}} = \langle \sigma_{u,i}\rangle^S - \langle \sigma_{u,i}\rangle^M, \tag{9}$$

where $\langle\cdot\rangle^S$ and $\langle\cdot\rangle^M$ are known as data-dependent and model-dependent expectations, respectively. To calculate these expectation values, one has to calculate the average of 2^n configurations, which is only feasible for ultra-small systems. Many approaches, e.g., MCMC and Gibbs samplings, have been applied to approximate this expectation, but they are still too expensive. To avoid these sampling steps in training, we use Besag's log pseudo-likelihood to approximate the loss function instead, which also yields asymptotically consistent estimates. The log pseudo-likelihood is defined as the sum of the conditional log-likelihoods of the items: $L(S|J) = \sum_{u=1}^{m}\sum_{i=1}^{n} \log p(\sigma_{u,i}|\mathbb{I}_{\setminus i}; J)$, where the conditional log-likelihood is derived as:

$$p(\sigma_{u,i}|\mathbb{I}_{\setminus i}; J) = \frac{\exp\left(2\sigma_{u,i}\sum_{j\in\mathbb{I}_{\setminus i}} J_{ij}\sigma_{u,j} + J_{ii}\sigma_{u,i}\right)}{\exp\left(2\sigma_{u,i}\sum_{j\in\mathbb{I}_{\setminus i}} J_{ij}\sigma_{u,j} + J_{ii}\sigma_{u,i}\right) + 1}. \tag{10}$$

By adopting Besag's pseudo-likelihood, our Inverse-Ising learning can avoid computationally expensive sampling-based gradient approximation and provide a tractable gradient calculation that achieves consistent parameter estimation.

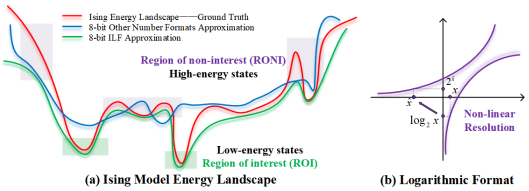

Fig. 6. Energy landscape and Ising-specific logarithmic quantization.

3) Ising-Voltage Polarization

As mentioned in Section I, an Ising machine uses capacitors to represent spins (items) in the Ising model (CF) and uses the voltage values carried by capacitors to make recommendations. Specifically, items with voltage from 0 to $+$Vdd are predicted as "like", while the ones from $-$Vdd to 0 are regarded as "dislike". With the Ising model trained with standard log-pseudo-likelihood, most of the spins that should have positive voltages end up with small voltages clustered around 0 volt, blurring the differences between "like" and "dislike" and further leading to low accuracy. To this end, we propose Ising-voltage polarization, which pushes the positive voltages on spins that should be recommended towards $+$Vdd and the negative ones down to $-$Vdd, better at differentiating "like" and "dislike" and ensuring the positive items can be recognized. Specifically, we augment log pseudo-likelihood to train the model with extra attention on the likelihood of positive interactions. Let r_i be the polarization factor for item i, and the new log pseudo-likelihood can be formulated as:

$$L = \sum_{(u,i)\in O^-} \log p(\sigma_{u,i}|\mathbb{I}_{\setminus i}; J) + r_i \sum_{(u,i)\in O^+} \log p(\sigma_{u,i}|\mathbb{I}_{\setminus i}; J), \tag{11}$$

where O^+ and O^- are sets of positive and negative user-item interactions respectively. By increasing r_i, voltages of more items will be pulled up closer to $+$Vdd and flipped to "$+1$", providing the support of adjustable recommendation rates.

C. Forward-Ising Inference

Forward-Ising inference is to map J matrix to the resistor network on an Ising machine, perform annealing with an Ising

machine, and read out spin states as recommendations. As the last two steps are automatically handled by BRIM [3], this section mainly discusses how to map the Ising model onto an Ising machine losslessly with the proposed Ising-specific logarithmic data format.

1) Ising-Specific Logarithmic Format (ILF)

To perform forward Ising inference on an Ising machine, we need to map theoretical values of elements in J learnt in Inverse-Ising learning to practical values of resistors on Ising machines. Unfortunately, due to the manufacturing limitations, real Ising machines can only be programmed with very limited numbers of resistor values (e.g., BRIMs only support 8-bit resistor values), making it hard to achieve high practical accuracy on the Ising machine with existing data formats. Specifically, Ising models learnt with high-precision data formats (e.g., Floating-Point/Fixed-Point 64/32) have high theoretical accuracy, however, the elements of their J matrices that use massive different values need to be mapped onto hundreds of resistor values, inevitably leading to unacceptable accuracy degradation. As for the models learnt with low-precision formats (e.g., fixed-point/integer 8/4), their J matrices have the potential to be accurately mapped onto resistors, however, the local minima of the energy landscape constructed with the low-precision J fail to represent the correct recommendation. To enable high accuracy on Ising machines, new Ising-specific data formats are required.

We observe that the practical accuracy of the Ising model is determined by how accurate the local minima of its energy landscape are. In another word, the Region of Interest (ROI) of the Ising model construction is at the low energy states. Furthermore, the J matrix trained through Inverse-Ising follows the logarithmic distribution. Motivated by these observations, we believe the non-linear data format can provide more efficient training of the Ising model with less number of bits. To this end, we propose a logarithmic representation friendly to Ising machines, ILF, and investigate its efficiency in the Ising model training. Our evaluation demonstrates that ILF needs less number of bits than other data formats to construct the Ising model with precise local minima and high accuracy. In particular, ILF only needs 4 bits to represent an accurate Ising model for CF, which can be easily mapped onto 8-bit resistor values losslessly.

ILF quantization function is defined as:

$$Q_{log}(\hat{x} \rightarrow x) = \text{sign}(x) \times s \times 2^{\frac{g}{N_I}\hat{x}}, \quad (12)$$

$$\hat{x} = \text{clamp}(\text{round}(\log_2(|x|/s) \times (N_I/g)), 0, 2^{b-1} - 1), \quad (13)$$

where g is used to adjust the distance of adjacent logarithmic representations, b is the number of bits, s is a scaling factor.

2) Mapping ILF-based Ising Model to Ising Machine

To map the Ising model trained through the Inverse-Ising process onto BRIM, we set resistor values as $R_{ij} = 1/J_{ij}$.

IV. Evaluation

A. Experimental Setup

Dataset: The proposed Ising-CF is evaluated with 4 real-world datasets widely used in recommendation system research, including Yahoo!Music-v1.0, MovieLens-100K, MovieLens-1M, and EachMovie. The user rating is converted to implicit feedback (like/dislike) with a threshold of 3.5.

TABLE I
Accuracy Comparison with Different Training Methodologies

Dataset	Yahoo-Music	MovieLens-100K	MovieLens-1M	Each-Movie
NPA	0.16325	0.32153	0.39562	0.32661
SLNN	0.21924	0.34151	0.42158	0.37434
SLNN+P	**0.36486**	**0.52948**	**0.53195**	**0.53251**

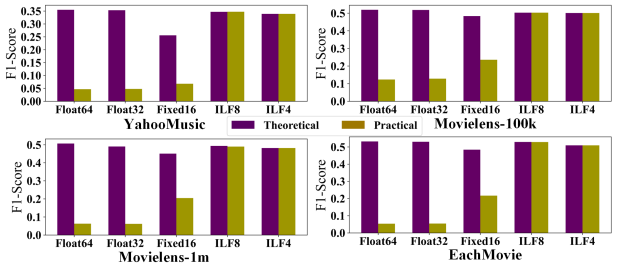

Fig. 7. Accuracy Comparison of Ising-CFs with Different Data Formats.

Experimental Metric: To evaluate the model performance, we use F1-score (**F1**), Recall (**Rec**), and Precision (**Pre**) as metrics. Note that different from image classification models that typically use Accuracy for model performance evaluation, recommendation models use **Rec** and **F1**. Following this tradition, we use F1, Rec, and Pre in the external accuracy comparison (Ising-CF vs. 6 SOTA methods) and use F1 only in the internal comparison among Ising-CF with 3 different design choices and 6 data formats for clarity.

Platforms: The Forward-Ising inference of Ising-CF is performed on a simulated BRIM system (20W) with 100 BRIM tiles (200mW for each) that work independently. The SOTA methods are performed on an Intel Gold 6330 CPU (205W) and a Nvidia A100-PCIe-40GB GPU (250W).

B. Evaluation of Accuracy

Ising-CFs with different design choices: Table I compares F1 accuracy of Ising-CFs that are constructed with the different training methods during Inverse-Ising learning including NPA, SLNN, and augmented SLNN with Ising-voltage polarization (SLNN+P). The results show that the proposed SLNN-based Inverse-Ising learning and Ising-voltage polarization significantly improve the accuracy of Ising-CF by 19%.

Ising-CFs (SLNN+P) with different data formats: Fig.7 compares the F1 accuracy of Ising-CFs that are quantized with different data formats for four datasets, including Floating-Point 64&32, Fixed-Point 16, and the proposed ILF 8&4. All Ising-CFs are trained with the SLNN+P method. For each data format, we report both the theoretical accuracy obtained from Inverse-Ising learning and the practical accuracy delivered by Ising machines after Ising parameters are mapped onto resistors. Results show that though Floating Points generally provide higher theoretical F1 accuracy, there are always significant accuracy drops after model mapping due to the limited choices of resistor values. In contrast, the models using ILFs with 4 bits achieve comparable accuracy to FP64 and can be mapped onto Ising machines with negligible accuracy drop.

TABLE II
OVERALL ACCURACY COMPARISON: PRACTICAL ACCURACY OF ISING-CFS DELIVERED BY BRIMS VS 6 SOTA CF METHODS.

Dataset	Yahoo-Music			MovieLens-100K			MovieLens-1M			Each-Movie		
Method	**F1**	**Rec**	**Pre**	**F1**	**Rec**	**Pre**	**F1**	**Rec**	**Pre**	**F1**	**Rec**	**Pre**
Eals-WRMF	0.1565	0.4162	0.0964	0.4096	0.6169	0.3066	0.2932	0.4815	0.2108	0.3818	0.5831	0.2839
Mult-VAE	0.3001	0.4217	0.2329	0.4750	0.6250	0.3831	0.4731	0.7673	0.3420	0.4488	0.6327	0.3478
Mult-DAE	0.3053	0.4203	0.2397	0.4730	0.6234	0.3811	0.4558	0.7413	0.3291	0.4581	0.6453	0.3551
MRF	0.2989	0.4096	0.2354	0.4761	0.6236	0.3851	0.4024	0.6482	0.2918	0.4499	0.6293	0.3501
RecVAE	0.2704	0.3788	0.2103	0.4817	0.6286	0.3904	0.4603	0.7451	0.3331	0.4788	0.6724	0.3718
FastVae	0.2831	0.4392	0.2089	0.4549	0.6369	0.3538	0.3708	0.6386	0.2613	0.4321	0.6674	0.3195
Ising-CF	**0.3393**	**0.4943**	**0.2583**	**0.5104**	**0.6619**	**0.4153**	**0.4941**	**0.7903**	**0.3594**	**0.5008**	**0.6926**	**0.3922**

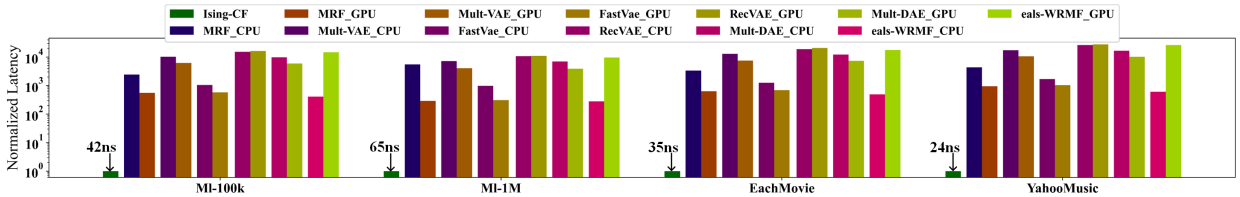

Fig. 8. Comparison of Normalized latency: Ising-CFs on 20W BRIM Ising System vs 6 SOTA solutions on 205W Gold 6330 CPU and 250W A100 GPU.

Overall Comparison with SOTA CFs: Table II compares the accuracy (F1, Rec, and Pre) of final Ising-CFs (i.e., practical accuracy with SLNN+P & ILF4) with 6 carefully selected SOTA CF methods. *eals-WRMF* uses a classical matrix factorization method based on element-wise alternating least squares. *Mult-VAE* and *Mult-DAE* are deep non-linear models trained with the multinomial likelihood. *MRF* is a probabilistic graphical model combined with autoencoders and neighbor-based approaches. *RecVAE* is a deep neural network collaborative filtering method. *FastVae* decomposes the inner-product-based softmax probability to improve the recommendation quality and efficiency. Results demonstrate that Ising-CFs outperform SOTA solutions in terms of accuracy, delivering more accurate recommendations. The average recall increase is 7.3%, a significant improvement compared with the one made in the past decade.

C. Cross-Platform Evaluation of Latency

Fig. 8 compares the recommendation latency of final Ising-CFs run on a BRIM system with 6 SOTA CF methods implemented on high-end server-level CPU and GPU. Results show that, due to the unique computational power brought by BRIM, Ising-CFs can generate recommendations (i.e., inference) within tens of nanoseconds (ns), deliver on average 8186× speedups over CPUs and GPUs, while the BRIM system consumes only 10% power.

V. CONCLUSION

This paper proposes a novel Ising Machine Learning framework, Ising-CF, that successfully realizes the potential of both Ising models' strong expressivity and Ising Machines' unique computational power in the real world. Taking CF as a case study, Ising-CF is the first to demonstrate that Ising methods can outperform traditional SOTA solutions of real-world applications. Experimental results show that Ising-CF delivers, on average 8186× speedups and 7.3% accuracy improvement over SOTA CF methods.

ACKNOWLEDGEMENT

This work was supported, in part, by DARPA under contract FA8650-23-C-7312, by NSF awards 2233378 and 2231036, and by the U.S. DOE Office of Science, Office of Advanced Scientific Computing Research, under award 78284: "ComPort: Rigorous Testing Methods to Safeguard Software Porting". In addition, Y. Yang acknowledges support from Dr. Max Rossler, the Walter Haefner Foundation, and the ETH Zurich Foundation.

REFERENCES

[1] Y. Koren, S. Rendle, and R. Bell, "Advances in collaborative filtering," *Recommender systems handbook*, 2022.

[2] R. Afoakwa, Y. Zhang, U. K. R. Vengalam, Z. Ignjatovic, and M. Huang, "BRIM: Bistable resistively-coupled Ising machine," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 749–760.

[3] A. Sharma, R. Afoakwa, Z. Ignjatovic, and M. Huang, "Increasing Ising machine capacity with multi-chip architectures," in *2022 International Symposium on Computer Architecture (ISCA '22)*.

[4] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua, "Fast matrix factorization for online recommendation with implicit feedback," in *Proceedings of the 39th International ACM SIGIR conference*, 2016, pp. 549–558.

[5] J. Chen, D. Lian, B. Jin, X. Huang, K. Zheng, and E. Chen, "Fast variational autoencoder with inverted multi-index for collaborative filtering," in *Proceedings of the ACM Web Conference 2022*.

[6] Q. Tao and J. Han, "Solving traveling salesman problems via a parallel fully connected Ising machine," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, 2022, pp. 1123–1128.

[7] T. Zhang and J. Han, "Efficient traveling salesman problem solvers using the Ising model with simulated bifurcation," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE.

[8] Z. Pan, A. Sharma, J. Y.-C. Hu, Z. Liu, A. Li, H. Liu, M. Huang, and T. T. Geng, "Ising-traffic: Using ising machine learning to predict traffic congestion under uncertainty," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.