Toward sampling from undirected probabilistic graphical models using a D-Wave quantum annealer

Yaroslav Koshka1 · M. A. Novotny2

Received: 21 January 2020 / Accepted: 19 July 2020 / Published online: 22 September 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
A D-Wave quantum annealer (QA) having a 2048 qubit lattice, with no missing qubits and couplings, allowed embedding of a complete graph of a restricted Boltzmann machine (RBM). A handwritten digit OptDigits dataset having 8 × 7 pixels of visible units was used to train the RBM using classical contrastive divergence. Embedding of the classically trained RBM into the D-Wave lattice was used to demonstrate that the QA offers a high-efficiency alternative to the classical Markov chain Monte Carlo (MCMC) for reconstructing missing labels of the test images as well as a generative model on a classically trained RBM. At any training iteration, the D-Wave-based classification had classification error less than half of that in MCMC. The main goal of this study was to investigate the quality of the QA sample from the RBM model probability distribution and compare it to a classical MCMC during the Gibbs sampling traditionally used in RBM training. For the OptDigits dataset, the states in the D-Wave sample belonged to about twice as many local valleys (LVs) as in the MCMC sample. All the lowest energy (the highest joint probability), local minima in the MCMC sample were also found by the D-Wave. The D-Wave missed many of the higher-energy LVs, while also finding many “new” LVs consistently missed by the MCMC. It was established that the “new” LVs that the D-Wave finds are important for the model distribution in terms of the energy of the corresponding local minima, the width of the LVs and the height of the escape barrier.

Keywords Adiabatic quantum annealer · Boltzmann machine · Local extrema · Monte Carlo · Undirected graphical model · Sampling · Simulated annealing · Simulated warming

* Yaroslav Koshka
ykoshka@ece.msstate.edu

M. A. Novotny
man40@msstate.edu

1 Department of Electrical and Computer Engineering, HPC2 Center for Computational Sciences, Mississippi State University, Mississippi State, MS 39762, USA

2 Department of Physics and Astronomy, HPC2 Center for Computational Sciences, Mississippi State University, Mississippi State, MS 39762, USA
1 Introduction

Generative models based on undirected probabilistic graphical models (Markov random fields) [1–3] remain very interesting, in part due to their ability to learn and model very complex probability distributions, as well as their suitability for deep unsupervised learning of the distributions represented by training data deprived of classification labels. However, the now classical training algorithms for undirected models, which include Markov chain Monte Carlo (MCMC) sampling and other approximations, lately surrendered a big part of their popularity to other models utilizing backpropagation. Those models proved to be more scalable and therefore were strongly empowered by the improvements in the computational power in the past decades. Besides the inferior scalability of the learning approaches used for undirected models, there still remain concerns about not only how fast but also how close \( P_m(\mathbf{v}) \) can approach \( P(\mathbf{v}) \), where \( P_m(\mathbf{v}) \) is the model distribution of Markov random fields and \( P(\mathbf{v}) \) is the target distribution represented by the training patterns. Limitations of classical sampling approaches make training generative models very difficult [4].

Backpropagation has been used as a powerful addition to deep Boltzmann machines (DBM) [3]. It can be applied as the last step of training, following the DBM pretraining by one of the classical techniques aimed at minimizing the Kullback–Leibler divergence (KL divergence). However, there are hopes for other approaches to improve learning, besides supplementing the undirected learning algorithm with backpropagation. Significant improvements to the gradient-descent-based log-likelihood maximization could come from more efficient and more precise sampling techniques that would deliver a better representation of the state of the model distribution under optimization.

Quantum annealers (QAs) are designed to solve a task that at a first glance has nothing to do with generating a sample from a model distribution. The main purpose of a QA is to find the global minimum or the ground state (GS) of the energy function of an Ising spin glass. While the Ising spin glass does represent the family of models to which Boltzmann machines (BM), including DBM, belong, the task of finding the GS is fundamentally different from sampling from a BM model distribution at a value of temperature parameter \( T = 1 \). However, due to the probabilistic nature of finding the GS by a QA, the solution procedure involves a very fast repetition of a large number of solution attempts (usually from 1000 to 10,000 repetitions). Many of those solutions end up to be not the GS but one of many excited states of the Ising spin glass energy function. The fact of a very fast generation of this “sample” (1000 results within tens of milliseconds) stimulates an interest in investigating the possibility of using this set of D-Wave solutions to generate or improve a sample from a BM model distribution. There is an expectation that the sample may not only follow the Boltzmann distribution at \( T = 1 \) but would be more representative of the model distribution compared to samples produced with help of classical sampling techniques (e.g., Gibbs sampling).

The promise and difficulties of applying QA hardware to BM problems have been investigated theoretically [5]. The actual quantum hardware was employed in
training deep neural networks, and different ways of embedding neural networks into the QA lattice and sampling from the D-Wave have been reported [6–11]. Besides using QAs for sampling, new models and training algorithms such as a Quantum BM have been proposed [12] and applied on the D-Wave [13]. More recently, RBM training with datasets approaching those of practical importance has been pursued by multiple groups [14–16]. In Ref. [14], a three-layer neural network used the D-Wave for the middle hidden layer. It was suggested that the D-Wave can be utilized in a multi-layer neural network, with each layer of multiple times the size of the D-Wave. In Ref. [15], a hybrid system combined a classical deep neural network autoencoder with a quantum annealing RBM on the D-Wave. Also, a promise of training complex graphs (e.g., having intralayer connections) has been demonstrated [16]. However, the origin of the improved in some cases (e.g., Ref [7].) trainability with a QA-based sampling, compared to some of the classical techniques, is yet to be fully understood and quantified, which was one of the motivations for this work.

In our previous work [17, 18], we have made a first attempt to compare classical samples generated using the Gibbs MCMC sampling and a “sample” produced with a help of the D-Wave QA. Statistical comparison of samples obtained by the Gibbs technique versus 1000 sample states generated from the D-Wave revealed significant differences in the observed outcomes. The D-Wave samples were insufficiently representative of the model distribution, specifically by missing many local valleys (LVs) of the configuration space found by the Gibbs sampling. On the positive side, the D-Wave demonstrated an ability to find LVs that were consistently missed by the Gibbs sampling, which could potentially be very interesting for sampling applications. However, a few important questions remained unanswered. Specifically, with respect to some LV parameters, how important may be those “new” LVs found by the D-Wave but missed by the classical MCMC? The parameters to consider may include the RBM energy (i.e., the probability of the corresponding states), the width of the LVs (and, possibly, the density of the high-probability states) and the height of the escape barrier.

Another (secondary) motivation for this work was that our results in Ref. [17] were obtained for a relatively simple (and a rather ideal in many ways) training and testing dataset—a toy problem of 8×8 bars-and-stripes (BAS) [19]. It remained unclear if the preliminary trends are applicable to other, more practical and more realistic, training cases, and if those training cases might reveal any additional differences between the results of sampling using the D-Wave versus the classical MCMC.

Finally, there was a concern that the less-than-perfect embedding of the RBM into the D-Wave lattice (caused by the D-Wave hardware limitations [10]) could have been partially responsible for some of the trends observed in Ref. [17].

In this work, a systematic investigation of those critical issues is undertaken. A much more realistic and popular in ML OptDigits dataset [20], with 8 labels (classes) selected, scaled down to 8x7 pixels, is used. Furthermore, we were lucky to have temporary access to D-Wave hardware that was free from some of the limitations responsible for the non-perfect embedding reported by us in Ref. [10]. We rely on a flawed but useful criterion for comparing the samples. The criterion is based on monitoring how
many and what kind of LVs of the RBM energy function the generated samples belong to. Convincing evidence is obtained in support of an observation that the D-Wave “sample,” while missing many important LVs represented in the Gibbs sample, also finds many very important LVs that are systematically missed by the MCMC. Discussed are opportunities for using the discovered property of the QA in order to achieve the ultimate goal—bring the model distribution $P_m(\mathbf{v})$ of Markov random fields much closer to the target distribution $P(\mathbf{v})$ compared to what was possible before, while doing that with much higher computational speed compared to the classical sampling approaches.

## 2 Probabilistic graphical models and QA

### 2.1 Requirements to sampling used for log-likelihood maximization

KL divergence is the measure of the difference between the target distribution $P(\mathbf{v})$ represented by the training patterns and the model distribution $P_m(\mathbf{v})$ that attempts to match $P(\mathbf{v})$ during the training.

$$\text{KL}(P(\mathbf{v})||P_m(\mathbf{v})) \equiv \left\langle \ln \left( \frac{P(\mathbf{v})}{P_m(\mathbf{v})} \right) \right\rangle_p$$

$$= \sum_{D} \left( P(\mathbf{v}) \ln(P(\mathbf{v})) - P(\mathbf{v}) \ln(P_m(\mathbf{v})) \right)$$

(1)

where $\langle \ldots \rangle_p$ represents an expectation value with respect to the target distribution. The first term does not depend on the model. When $D$ represents the data composed of $K$ patterns $\{D = \{\mathbf{v}_T(1), \mathbf{v}_T(2), \ldots, \mathbf{v}_T(K)\}\}$, and $M$ is the graphical model under the optimization, the second term is the expected log-likelihood $\ln \mathcal{L}(D : M)$. Let us see how the second term is optimized during the model learning. The following is an approximated form when $P_m(\mathbf{v})$ is estimated only for the training patterns appearing with equally high probability, while all the other states are expected to have a negligible probability in $P(\mathbf{v})$:

$$\ln \mathcal{L}(D : M) \approx -\frac{1}{K} \sum_k \ln \left( P_m(\mathbf{v}_T(k)) \right)$$

(2)

The approximation contains probabilities specifically for each of the training patterns (the probabilities that we want to maximize during the log-likelihood optimization). It may be useful to look at these probabilities to see how the probabilities for states other than the training patterns influence $\ln \mathcal{L}(D : M)$.

Specifically, for the energy-based models:

$$P_m(\mathbf{v}) = \frac{e^{-E_{\text{free},m}(\mathbf{v})}}{Z_m}$$

(3)

where $\mathbf{v}$ and $\mathbf{h}$ are vectors of the visible and hidden units of the model, respectively, $E_{\text{free},m}(\mathbf{v}) = -\ln \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$, and $Z_m$ is the standard partition function. Substituting (3) into (2), we get the following approximation for the log-likelihood:
\[
\ln \mathcal{L}(D : M) \approx \frac{1}{K} \sum_{k}^{K} \ln \sum_{h} e^{-E(\mathbf{v}_T(k), h)} - \ln \sum_{\mathbf{v}, h} e^{-E(\mathbf{v}, h)}
\] (4)

We can see that maximizing \( \ln \mathcal{L}(D : M) \) means not only maximizing the quantity \( \ln P_m(\mathbf{v}_T(k)) \), the log marginal probability of the training data vectors \( \mathbf{v}_T(k) \) under the model (i.e., minimizing the energy \( E(\mathbf{v}_T(k), h) \) of the data under the model). More importantly, maximizing \( \ln \mathcal{L}(D : M) \) also means reducing the marginal probability (increasing the energy) of all the other states in the state space.

While it is not feasible to scan through all the \( \mathbf{v}, h \) combinations in the state space, sampling techniques can give an adequate estimate when optimizing the term containing the partition function [the second term in Eq. (4)]. For that, the sampling must reflect the most important LVs, those that include states that contribute the most to the partition function. The term “local valley” in this work means a local valley in the space of the energy function \( E(\mathbf{v}, h) \) [see Eq. (5), in which the general state vector \( s \) comprises both \( \mathbf{v} \) and \( h \) for the case of RBM]. A LV includes all the states surrounding a particular local minimum of \( E(\mathbf{v}, h) \) that have the following property. When performing a classical Monte Carlo from each of those states at zero temperature (which means a deterministic transition to a next step having a lower energy), those transitions should end up in the same local minimum. With this definition, the LVs that are important for the quality of the sample must have a high density of states and low state energies \( E(\mathbf{v}, h) \). While there is no useful way to mathematically introduce a general concept of a LV width, a “width-related parameter” is introduced in this work. It is intended to provide some non-direct measure (in addition to the LV depth) influencing the number of states belonging to a particular LV.

The optimization of the log-likelihood is conducted using the gradient descent of the log-likelihood, which utilizes derivatives of the \( \ln \mathcal{L}(D : M) \) with respect to coefficients \( \theta \) that are used to parametrize the model. The comment above about the required quality of the sample from the model distribution equally applies to the gradient of the log-likelihood during the gradient ascent. For \( K' \) total samples, sampled vectors \( \mathbf{v}_{\text{Sampled}}(k') \) are explicitly included in the approximation of the log-likelihood gradient below:

\[
\frac{\partial \ln \mathcal{L}(D : M)}{\partial \theta} = -\frac{1}{K} \sum_{k}^{K} \ln \sum_{h} p(h | \mathbf{v}_T(k)) \frac{\partial E(\mathbf{v}_T(k), h)}{\partial \theta} + \sum_{\mathbf{v}, h} p(\mathbf{v}) p(h | \mathbf{v}) \frac{\partial E(\mathbf{v}, h)}{\partial \theta} \\
\approx -\frac{1}{K} \sum_{k}^{K} \ln \sum_{h} p(h | \mathbf{v}_T(k)) \frac{\partial E(\mathbf{v}_T(k), h)}{\partial \theta} + \frac{1}{K'} \sum_{k'}^{K'} \sum_{h} p(h | \mathbf{v}_{\text{Sampled}}(k')) \frac{\partial E(\mathbf{v}_{\text{Sampled}}(k'), h)}{\partial \theta}
\]

Specifically for the energy function \( E(\mathbf{v}, h) \) of the RBM used in this work, when we go back to the general expression for the log-likelihood, it further simplifies. However, summation over all possible values of the visible vector \( \mathbf{v} \) remains in the second term. For the gradients of the log-likelihood for one training pattern (i.e., the expression under the sum over all the training patterns above), we get:
When selecting RBM case: approximated expressions for the gradient of the log-likelihood are used. For the \(\upsilon\) values of the visible vector \(\upsilon\) sampled possible if the samples the summation disappear from the expression used in the iterative algorithm. Once

\[
\frac{\partial \ln \mathcal{L}(\theta|\upsilon(k))}{\partial \omega_{ij}} = [2p(H_i = 1|\upsilon(k)) - 1] \upsilon_j(k) - \sum_{d} p(d) [2p(H_i = 1|d) - 1] \upsilon_j(k)
\]

\[
\frac{\partial \ln \mathcal{L}(\theta|\upsilon(k))}{\partial b_j} = \upsilon_j(k) - \sum_{d} p(d) \upsilon_j(k)
\]

\[
\frac{\partial \ln \mathcal{L}(\theta|\upsilon(k))}{\partial c_i} = [2p(H_i = 1|\upsilon(k)) - 1] - \sum_{d} p(d) [2p(H_i = 1|\upsilon(k)) - 1]
\]

Ideally, calculation of these gradients would require summing over all possible values of the visible vector \(\upsilon\), which is computationally not feasible. Instead, the approximated expressions for the gradient of the log-likelihood are used. For the RBM case:

\[
\frac{\partial \ln \mathcal{L}(\theta|\upsilon(k))}{\partial \omega_{ij}} = [2p(H_i = 1|\upsilon(k)) - 1] \upsilon_j(k) - \frac{1}{K'} \sum_{k'}^{K'} [2p(H_i = 1|\upsilon_{\text{sampled}}(k')) - 1] \upsilon_{\text{sampled}}(k')
\]

\[
\frac{\partial \ln \mathcal{L}(\theta|\upsilon(k))}{\partial b_j} = \upsilon_j(k) - \frac{1}{K'} \sum_{k'}^{K'} \upsilon_{\text{sampled}}(k)\]

\[
\frac{\partial \ln \mathcal{L}(\theta|\upsilon(k))}{\partial c_i} = [2p(H_i = 1|\upsilon(k)) - 1] - \frac{1}{K'} \sum_{k'}^{K'} [2p(H_i = 1|\upsilon_{\text{sampled}}(k')) - 1]
\]

When selecting \(K'\) equal to the number of the training patterns \(K\), the \(1/K'\) factor and the summation disappear from the expression used in the iterative algorithm. Once again, in the three expressions above, an adequate estimate of the second terms is possible if the samples \(\upsilon_{\text{sampled}}(k')\) represent at least the most important LVs. For example, consider a sample missing (not finding) any of the values of \(\upsilon\) belonging to one or a few LVs in the current model distribution. Let us consider missed LVs that contribute significantly to \(Z_\text{m} (\text{i.e., valleys having a high density of states and low state energies } E(\upsilon, h))\). Consistently missing those LVs by the sample will dictate a wrong trajectory of the log-likelihood gradient descent, resulting in the model distribution after training not maximizing the log-likelihood given the training data and assigning high probabilities to wrong states.

With respect to the Monte Carlo sampling, the height of the lowest portion of the escape barrier for the LV is relevant. It influences the probability of a Monte Carlo chain escaping from the given LV, thereby possibly reducing the chance of states in the given LV being sampled. The height of the lower portion of the escape barrier in this work is defined as the lowest energy difference from the bottom of the LV that must be acquired for escaping into the neighboring valley during MCMC.

When the classical Gibbs technique is used for sampling, each variable is updated sequentially based on its conditional distribution given the state of all the other variables. In RBM, all hidden units are connected only to visible units and vice versa. Therefore, the sequential updates at time \(t\) look as the following:

\[
h_i^{(t)} \sim p(h_i|\upsilon^{(t)}), v_i^{(t+1)} \sim p(v_j|h^{(t)})
\]
This means that each hidden unit $h_i$ is updated using the probability conditioned on the state of all the visible units (i.e., the vector $\mathbf{v}$), and vice versa.

Specifically, for the RBM energy function:

$$p\left( H_i^{(t)} = 1 \mid \mathbf{v}^{(t)} \right) = \sigma \left( 2 \left( \sum_{j=1}^{m} w_{ij} \mathbf{v}_j^{(t)} + c_i \right) \right),$$

$$p\left( V_j^{(t+1)} = 1 \mid \mathbf{h}^{(t)} \right) = \sigma \left( 2 \left( \sum_{i=1}^{m} w_{ij} h_i^{(t)} + b_j \right) \right),$$

where $\sigma(x) = 1/(1 + e^{-x/T})$, and the “temperature” parameter $T$ is equal to 1 in machine learning applications.

Limitations of classical sampling approaches make training undirected graphical models very difficult [4]. During Monte Carlo jumps at the time $(t+1)$ from the previous state at $(t)$, high-probability states in the model distribution should be revealed by the sampling thanks to the higher probability of those jumps that lead toward the states of higher conditional probability. It means that not only do states belonging to wider and deeper LVs have a higher chance of being sampled (which would be reflective of the true probability distribution). This also means that of high importance is the sign of the gradient (the slope) of $E(\mathbf{v}, \mathbf{h})$ with respect to the coordinates in the space of the random variables $\mathbf{v}$ and $\mathbf{h}$ in the higher-energy areas surrounding particular LVs (the so-called basin of attraction). When the corresponding states have low probability, the $E(\mathbf{v}, \mathbf{h})$ gradient in these areas may have little consequence for the true KL divergence between the target distribution $P(\mathbf{v})$ represented by the training patterns and the model distribution $P_m(\mathbf{v})$. But this slope may be critical for how well the particular sampling techniques reflect this divergence. In other words, in the absence of a sufficient negative slope of $E(\mathbf{v}, \mathbf{h})$ in a wide basin of attraction of some LVs, the particular LVs may be missed by the Gibbs sample. This would happen despite some of those LVs possibly having a high density of high-probability states (i.e., deep and wide LVs that we do not want to be missed in the sample).

### 2.2 Adiabatic quantum annealing

Adiabatic QA can be considered as an alternative to the classical SA method for finding the global energy minimum of a spin glass [21]. In a QA, the search for the GS utilizes not only thermal fluctuations over barriers (as in the classical SA algorithm), but also quantum mechanical tunneling through the barriers surrounding local minima. The D-Wave was the first commercial quantum annealer (QA) [22]. The D-Wave architecture is an implementation of an Ising spin glass model [23]. The qubits correspond to an optimization variables $s_j$. There are couplings between some of the qubits. The weights of the couplings $J_{ij}$ and the bias fields $h_i$ are defined to represent a particular spin glass problem as in Eq. (5) for which the ground state is to be found. The ground state of the Ising spin glass is the state of the Ising spins that minimizes the energy function Eq. (5) [24]. The physics of the ground state determination by the D-Wave QA has been extensively investigated [25–27].
While the main job of a QA is to find the GS, the statistical nature of the search for the GS ensures that a big sample from the underlying model distribution may be produced. This sample usually contains not only the GS but also a very large number of excited states, and it may be generated very fast (1000 results within tens of milliseconds). While it is known that the D-Wave sample does not follow the Boltzmann distribution at the desirable value of the temperature parameter $T=1$, there are efficient ways of converting it to the desirable form of the distribution [9]. For that reason, investigation of fundamental properties of the samples produced by the D-Wave is one of the main goals of this work.

3 Methods

3.1 RBM embedding within the D-Wave lattice

The architecture used in the D-Wave hardware is the so-called Chimera graph shown in Fig. 1a, b. The quantum bits (qubits) are located at vertices of the Chimera graph. Some of the qubits are connected with couplings having weights $J_{ij}$. In addition, the value of the optimization variable is influenced by the bias fields $h_j$. The values of $J_{ij}$ and $h_j$ are assigned to represent a particular spin glass problem described by the following expression for $N$ qubits:

$$E(s) = -\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}s_is_j - \sum_{j=1}^{N} h_js_j$$

(5)

The Chimera lattice provides at most six connections for each qubit, or less when some qubit neighbors or couplings are missing in the particular version of the hardware. A common approach to increasing the connectivity of the graph is to combine (clone) multiple qubits to represent one logical unit by ensuring a high probability

Fig. 1 a A magnified view of D-Wave lattice. The six available couplings for the first qubit in the left column of the unit cell are highlighted with the bold lines. b The same as (a) highlighting the couplings for the first qubit in the right column of the unit cell. c The RBM embedding into the Chimera lattice used in this work. The bold lines highlight “ferromagnetic” couplings that combine (clone) qubits into a single logical RBM unit in order to increase the number of connections for each RBM unit.
of their alignment via providing strong ferromagnetic bonds in a classical sense (Fig. 1c). In this case, the number of connections for the given logical unit is equal to the sum of the connections for each qubit combined into the given unit. The cloning is conducted by assigning the maximum allowed value to the couplings $J_{ij}$ to ensure that all (or at least the majority) of the combined qubits have the same value. Those couplings approximate the ferromagnetic bonds and are shown with bold lines in Fig. 1c. The vertical bold lines in Fig. 1c combine qubits into visible RBM units, while hidden units are achieved by combining qubits using bold horizontal lines. For clarity of the figure, only some of the qubits are shown to be combined with bold lines.

In contrast to our previous work [17], the lattice of 16 × 16 unit cells of the D-Wave hardware used in this work had no missing qubits or couplings. The same embedding approach enabled 64 (16 × 4) visible and 64 hidden RBM units, each connected to all units of the opposite kind. Therefore, this version of the hardware allowed embedding of a complete RBM architecture.

### 3.2 RBM training by classical CD

As described in the previous section, the price for the complete RBM connectivity is a relatively small number of RBM units that can be embedded within the current D-Wave hardware. In our previous work, for a similar RBM embedding, a toy problem of 8 × 8 bars-and-stripes (BAS) [17, 18] was selected. That dataset was employed also at the first stage of this work for comparison. The main focus of the current investigation, however, was a much more realistic dataset, a popular in ML OptDigits dataset. It was scaled down to 8 × 7 pixels, binarized, and converted from the QUBO {0,1} to Ising {-1,1} format. The optimal embedding provided us with 64 visible units, which restricted us to using not all 10 but only 8 RBM units for labels. Therefore, 8 handwritten digit classes (from “0” to “7”) were used for the training and reconstruction. The RBM was trained with 1024 different training patterns by a classical algorithm (without the D-Wave) and then embedded into the D-Wave lattice. During that training, each of the 1024 training patterns was used to initiate a five-step Gibbs chain. Patterns obtained at the end of the Gibbs chain were used as a sample from the model distribution for the contrastive divergence (CD) training [28]. We followed an approach that was demonstrated in the previous work to be very effective. The values for the weight-decay parameter and the power term in the weight-decay expression were optimized by trial and error and applied to keep the absolute value of the maximum weights below 0.5, while also discouraging too small values of weights.

Bigger weights in excess of 0.5 would be too close to 1, and 1 is the maximum value allowed by the D-Wave hardware. The value of 1 was used in the RBM embedding into the D-Wave to establish a “ferromagnetic” bond between qubits combined to represent a single RBM unit. If some RBM weights also approach this value after training, some of the “ferromagnetic” bonds might be violated, compromising the validity of the embedding. On the other hand, too small values for $w_{nm}$ were discouraged to avoid the risk of having smaller weights dropping below the D-Wave
sensitivity limit. This often happens after performing an additional scaling with a scale factor for the D-Wave, which contributes to the classification error (Fig. 2). Encouraging most of the weights to be as large as possible (while smaller than 0.5) helped us minimize this problem.

3.3 Reconstruction of incomplete images and missing labels using the RBM embedding within the D-Wave lattice

After the RBM training using the classical CD, the D-Wave and MCMC were applied independently to reconstruct missing portions of incomplete input images and the missing labels. The procedure developed in the previous work [17] was used. Incomplete test images were drawn from the test set of 440 handwritten digits. For each image, all visible units corresponding to the available portion of the image were “clamped” to the value of the corresponding pixel. The clamping in the D-Wave hardware was achieved by providing the maximum value to the bias field $h_j$ of the corresponding qubits. The maximum value of $h_j$ was $-2$ or $2$, depending on the value at the pixel of the test image.

The D-Wave GS was sought as a representation of the most energetically favorable combination of the remaining qubits (and therefore the corresponding visible units and labels) for the given clamped incomplete input image. A critical advantage of using the D-Wave is that the standard MCMC image reconstruction method takes a large number of MCMC steps to obtain the reconstructed image and/or label. In contrast, one call to the D-Wave machine provides the immediate result for the image reconstruction and classification attempt.

From the 1000 solution repetitions provided by the D-Wave in a single call, the lowest energy state was used. When clamping the values of the qubits corresponding to the available portion of the image, this state gives the highest probability to the qubits representing the sought visible RBM units corresponding to the labels and missing pixels. From this state, a majority vote was taken among the values of the qubits combined to represent the reconstructed RBM units. The obtained majority vote represented the sought values for the missing image pixels and labels.

![Fig. 2](image.png) The percentage of the classification errors for the handwritten digits (bottom) and the percentage of the errors of incomplete image reconstruction (top) achieved by the D-Wave, shown as a function of the scale factor applied to the couplings and bias fields of the RBM embedding into the D-Wave.
The values of $J_{ij}$ obtained from the values of $w_{nm}$ may be not sufficiently smaller in magnitude than the $-1$ value of the ferromagnetic couplings [10]. This lack of an overwhelming comparative strength of some ferromagnetic bonds may cause violation of the requirement that all qubits combined to represent a given RBM unit must have the same values. As a result, the lowest energy state found by the D-Wave would be the GS of a different graph, not the lowest energy state of the trained RBM.

In order to further mitigate this problem, the values of $J_{ij}$ and $h_j$ were further reduced using the scale factor for the D-Wave, which was experimentally adjusted to achieve the best results in terms of both (1) the classification error and (2) the correspondence between the GSs found by the D-Wave and the classical SA [10]. The scale factor was optimized to obtain from the D-Wave the lowest classification error for the 440 handwritten digits. The results are shown in Fig. 2. Reasonably large values of the scale factor allow keeping the values of $J_{ij}$ corresponding to RBM weights sufficiently below 1. Thanks to that, most of the combined qubits are ensured to have the same value, providing a correct embedding. As a result, an optimal for the given RBM classification error by the D-Wave can be achieved (e.g., in Fig. 2 it corresponds to the scale factor equal to 4). However, a scale factor that is too large may lead to the smallest values of the coupling dropping below the sensitivity limit of the D-Wave hardware. This would hurt the quality of the embedding and would cause the deterioration of the classification error (Fig. 2).

### 3.4 Determination of the local valleys in the energy function

In Ref. [17], we introduced a conservative approach to assessing if the D-Wave has a potential to offer important advantages for sampling from the RBM model distribution. The approach is further applied in this work to the RBM trained with the handwritten digit dataset. A comprehensive statistical analysis is introduced to provide a meaningful comparison of the D-Wave and the Gibbs samples. The approach is based on comparing the D-Wave and the Gibbs samples from the point of view of what LVs the states in the samples belong to. Further, the relative importance of the LVs reflected in the D-Wave and the Gibbs samples was analyzed by comparing the RBM energy of the corresponding local minima (Eq. (5)), the depth of the LVs (i.e., the height of the escape barrier), and a width-related parameter of the valley.

The classical search for the local minima of the RBM model distribution utilized Gibbs sampling. It was performed on the original RBM graph (i.e., the graph free from the representation of the RBM units by combining qubits). Each of the 1024 training patterns was used as the initial state in the Gibbs chain. A specified number of Gibbs sampling steps was conducted at an initial temperature of $T = 1$. Following that, the value of $T$ was set to zero in order to perform relaxation to the bottom of the current LV. Up to 1000 Monte Carlo sweeps were performed, until no new local minima were generated at the end of the relaxation at $T = 0$.

The D-Wave was run 10,000 times in each D-Wave call to generate 10,000 attempts to find the lowest energy solutions for the embedding of the trained RBM (described in Section III.A). The 10,000 D-Wave anneals returned 10,000 solution attempts. Some of the same solutions were found multiple times, thereby reducing
the number of distinct sampled states from the given D-Wave call. Next, each distinct D-Wave solution was used as an initial state (i.e., the vector of the visible units). A MCMC was ran for each of those states, followed by a relaxation (a SA at $T=0$) to the bottom of the LV. The parameters of the MCMC and relaxation sequence were the same as in the case of the local minima determination using Gibbs sampling described above. Each of the found local minima can be used as an identification for the corresponding LV. To identify which LVs found by the Gibbs sampling and by the D-Wave are the same and which are different, we conducted a bit-by-bit comparison of the local minima states.

### 3.5 Simulated warming

Simulated warming (SW) was implemented by conducting MCMC starting from a bottom of each LV while gradually increasing the temperature from the initial value of $T=0$. In this version of the SW, the samples were collected and evaluated after each MCMC jump. A big enough number of distinct sampled states that remain in the same LV could be used as a comparative estimate of the size of the LV and, indirectly, the density of states. To include only those sampled states that belong to the same initial LV after SW, a relaxation from each of the sampled states (a SA at $T=0$) was conducted. The minimum of the LV to which the sampled state belonged was found, which was used to establish if the local minimum indeed corresponds to the initial LV of the D-Wave state, or if MCMC escaped into another LV. SW at different temperatures was also used to find the escape frequency. The escape frequency for a given LV was estimated by counting and taking an inverse of the number of MCMC jumps before the next state finds itself outside of the LV under consideration.

The height of the escape barrier $\Delta E_{esc}$ is:

$$\Delta E_{esc} = E_{S.P.} - E_{LM}$$

where $E_{LM}$ and $E_{SP}$ are the values of the RBM energy $E(\mathbf{\upsilon}, \mathbf{h})$ at the local minimum and at the saddle point of the LV, respectively. SW was used to find the activation energy $E_{act}$ of the escape frequency using the standard Arrhenius temperature dependence formula. The found value of $E_{act}$ was used as an estimate for $\Delta E_{esc}$.

### 4 Results and discussion

Classification, reconstruction and pattern generation results reported in the first part of this section primarily aim at verifying the D-Wave’s ability to correctly determine the lowest energy state of a complex probability distribution under a variety of constrains (e.g., clamped visible units and clamped labels). Having this reassurance about the validity of the embedding, the main novel results of this work are reported in Section IV, parts B and C.
4.1 Using D-Wave for classification on an RBM graph trained by classical CD

Figure 3 shows the classification error versus the training epoch, obtained (A) using the classical MCMC and (B) using the D-Wave reconstructing the qubits corresponding to the unknown labels of the test image. Here, the term epoch means a full cycle of iterations, with each training pattern participating only once. At any training iteration, the D-Wave gives at least two times better classification error. It was suggested by us in Ref. [17] that this superiority of the QA may be due to a lower chance of the labels getting stuck in a wrong local minimum during the reconstruction of the highest probability state. In Fig. 3, the advantage of the more than two times improvement of the classification error on the toy BAS dataset in Ref. [17] is reproduced here for the more difficult task of classifying 8 classes of handwritten digits.

A few examples of correct classifications of incomplete input images of various handwritten digits using the D-Wave are shown in Fig. 4. This classification was conducted after only 10 training epochs.

4.2 D-Wave embedding of the classically trained RBM used as a generative model

Further, the RBM was fully trained with 2000 epochs on the same handwritten digit dataset. Only the qubits representing the labels (i.e., combined with “ferromagnetic” bonds to represent RBM units of a label) were clamped to the desirable values of the label. The 1000 solution repetitions returned by the D-Wave provided 1000 generated images, some of which may repeat. The five lowest energy solutions for each of the 8 labels are shown in Fig. 5 as examples. Figure 5 is the illustration of the D-Wave’s ability to serve as a generative model.
Fig. 4 Examples of correct classifications of incomplete input images of various handwritten digits using the D-Wave, after 10 training epochs. Left: the original test images. Center: input images without labels “clamped” in the D-Wave call. Right: the reconstructions (i.e., the GSs) found by the D-Wave for the missing units and labels. Each row represents one example for each of the digit classes recognized by the D-Wave.
4.3 Comparison of D-Wave and Gibbs samples from the RBM model distribution

As discussed in the Introduction, our result of Ref. [17] was obtained with a D-Wave hardware in which some of the qubits and couplings were missing. This was one of the reasons for the imperfect RBM embedding [10]. The results of this work,
obtained when using the embedding into the D-Wave lattice with no missing qubits and couplings, were qualitatively similar. The D-Wave sample from the BAS-trained RBM includes a smaller number of the LVs compared to the classical sample (Fig. 6a). This fact alone indicates that the sample may be not sufficiently representative. Furthermore, more than 98% of the classically found LVs are missed by the D-Wave for the BAS dataset (Fig. 6b). On the positive side, the D-Wave sample includes LVs that were not present in the classical sample.

For the handwritten digit dataset, the comparison results were somewhat different. The D-Wave sample represented a larger number of the LVs than the number of

![Fig. 6 RBM trained with BAS. (a) The total number of the found local minima as a function of the training epoch. The two dependencies in a correspond to (triangles) the local minima found by SA at $T=0$ from each of the 1000 repetitions of the D-Wave solutions, and (circles) the local minima found by SA at $T=0$ after one classical Gibbs step from each of the training patterns. The number of the found local minima is normalized to the total number of the training patterns. (b) The percentage of the LVs in a found by the classical MCMC that are missed by the D-Wave sample](image-url)
the LVs in the classical sample (Fig. 7a). It could have been further expected from these results that the set of the LVs present in the D-Wave sample should include at least all the LVs from the Gibbs sample plus additional LVs that were not present in the Gibbs sample. The reality was more disappointing. There is only a relatively small overlap between the two sets of the LVs. For the case of the handwritten digit dataset, ~60–80% of the classically found LVs are missed by the D-Wave (Fig. 7b). On the other hand, this small overlap also means that the D-Wave sample
includes even more (compared to the BAS case) LVs that were missed by the classical sample.

Our further analysis targeted exclusively the handwritten digits dataset as a more realistic example of a useful training case.

During the RBM training, the marginal probability is maximized for the states having visible units corresponding to the training patterns, which also leads to a higher joint probability and a lower RBM energy. Lower-energy states contribute the most to the partition function in Eq. (4). Obviously, a hypothetical sample missing LVs with minima having low energy $E(\mathbf{u}, \mathbf{h})$ would be a more serious problem than a sample missing higher-energy LVs in the model distribution. The next step in the investigation aimed at establishing what kind of LVs (with respect to the energy of the local minimum) present in the classical MCMC sample are missed by the D-Wave sample.

Histograms of the RBM energies of the minima of the LVs present in the classical MCMC sample are shown in Fig. 8 as white bars. The dark bars in Fig. 8 are for those LVs in the MCMC sample that are also found by the D-Wave sample. The light bars are for those MCMC-found LVs that were not found by the D-Wave.

All the lowest energy local minima (the highest joint probability states in the model distribution) in the classical sample are also found by the D-Wave sample. With the increase in the RBM energy (lower probability states), an increasingly higher percentage of the classically found local minima are not present in the D-Wave sample. The result holds for different training epochs. Three different training epochs are illustrated in Fig. 8.

The next question was specifically about those “new” LVs that were found by the D-Wave but missed by the classical MCMC sample. The corresponding histograms for the RBM energies of the minima of the LVs present in the D-Wave sample are shown in Fig. 9. The lowest energy LVs found by the D-Wave are the same as what MCMC finds. However, the main interest now is in the “new” LVs found by the D-Wave sample but missed by the classical technique. It follows from Fig. 9 that those “new” LVs are distributed across a wide range of the RBM energy, dominating

![Fig. 8](image-url)  
**Fig. 8** Histograms of the RBM energies of the minima of the LVs present in the classical MCMC sample (found by SA at $T=0$ from each of the 1024 training patterns). Histograms are plotted for three different training epochs. The white bars are for all the local minima in the MCMC sample. The dark bars are for the minima of those LVs in the MCMC sample that are also represented by the D-Wave sample (found by SA at $T=0$ from each of the 1000 repetitions of the D-Wave solutions). The light bars are for those local minima that were not found in the D-Wave sample.
at higher energies (lower probability states) but also present in a significant amount at medium-to-low energies. This result indicates that at least some of the newly found LVs could be an important contribution to the quality of the sample. Finding or missing them could significantly influence how good or bad the model RBM distribution approaches the target distribution.

Further insight into what kind of the LVs are more likely to be represented by the D-Wave sample while missed in the MCMC sample was obtained by looking at the height of the lowest portion of the escape barrier for the LV. This lowest escape barrier corresponds to the lowest energy from the bottom of the LV that must be acquired for escaping into the neighboring valley. Figure 10 shows histograms of the

**Fig. 9** Histograms of the RBM energies of the minima of the LVs present in the D-Wave sample (found by SA at $T=0$ from each of the 1000 repetitions of the D-Wave solutions), plotted for three different training epochs. The white bars are for all the local minima in the D-Wave sample. The dark bars are for those local minima in the D-Wave sample that coincide with the local minima in the classical MCMC sample (found by SA at $T=0$ from each of the training patterns). The light bars are for those local minima that are new and not present in the classical sample

**Fig. 10** Histograms of the activation energy $E_{\text{act}}$ of the escape frequency from the D-Wave-found LVs. The dark bars are for those LVs in the D-Wave sample that coincide with the LVs in the classical MCMC sample. The light bars are for those LVs that are new and not present in the classical sample. The inset in (a) illustrates the Arrhenius plot for one of the LVs.
activation energy $E_{\text{act}}$ of the escape frequency from the D-Wave-found LVs during the SW.

While different heights of the escape barrier dominate at high versus low temperatures, $E_{\text{act}}$ used in (a) reflects the lowest portion of the escape barrier found from the lower-temperatures part of the Arrhenius curve. This estimate of the barrier height was independently confirmed by sampling the energy of the last state of the MCMC before escaping the valley.

Four of the D-Wave-found LVs in Fig. 10 having the highest $E_{\text{act}}$ have been also found in the classical MCMC sample. Except for those four found by both techniques, most of the “new” LVs found by the D-Wave but missed by the MCMC sample have the escape barrier height in a wide range of energies. In other words, among the “new” LVs, there are valleys of the variety of depths from very shallow (i.e., less important for the quality of the sample) to very deep (i.e., more important).

Finally, histograms of a width-related parameter of the D-Wave-found LVs are shown in Fig. 11. This parameter was estimated using a very rough approximation of a LV with a square well, which allows to assume that the “size” of the well and therefore the number of states inside is proportional to the product of its height and width. With this approximation, the width-related parameter was estimated by dividing the total number of the states inside the well (i.e., having the values of energy below $E_{\text{act}}$) by the depth of the well $E_{\text{act}}$. A qualitatively similar result was obtained using the intercept of the Arrhenius plot with the vertical axis as the width estimate.

With respect to the width-related parameter of the LVs, there is no clear trend for the “new” LVs (i.e., those found by the D-Wave while missed by the MCMC). Among the “new” LVs, there are valleys of the variety of widths from very narrow to very wide. However, at least for this number of epochs, the fraction of “new” LVs
missed by MCMC seems to somewhat increase for higher LV widths. The four widest D-Wave-found LVs happened to be all missed by MCMC.

Understanding what properties of the LV make it easier to be found by the D-Wave but missed by classical MCMC is beyond the scope of this paper. However, an important conclusion can be suggested from Figs. 8, 9 and 10. It appears that the D-Wave’s sample reflects many important LVs missed by MCMC, which contain high-probability states, a high density of states and high barriers for escaping from the LV.

5 Conclusion

Use of the D-Wave QA for classification of handwritten digits and the demonstration of the successful use of the D-Wave as a generative model is an important milestone in evaluating capabilities of QAs for machine learning. Both in the previous work dealing with a BAS dataset [17] and in this work, which focused on a much more complex dataset, the D-Wave gives at least two times better error for incomplete image classification compared to MCMC, with much faster performance (1000 D-Wave states found within tens of milliseconds were used to find the GS for reconstructing the missing visible units and labels).

Analysis of QA and MCMC samples from the RBM model distribution was conducted while using D-Wave hardware with no missing qubits and couplings. It allowed us to eliminate the concern that the (usual) non-idealities of the QA hardware could be responsible for some of the trends in Ref. [17].

Statistical analysis of the distribution of the “new” LVs represented in the D-Wave sample but missed by the classical MCMC sample was conducted. It revealed that the D-Wave sample normally contains some important LVs often missed by classical MCMC, many of which include high-probability states, a high density of states and high barriers for escaping from the LV.

The fact that the D-Wave is also missing many important LVs present in the classical sample suggests that further improvements are needed before it is possible to fully replace MCMC with a D-Wave sampling. However, we believe that the obtained results point out a strong potential of using QAs for achieving drastic improvements in the sampling speed as well as producing a much better sample for the gradient-descent-based log-likelihood maximization. The sample may be more representative of the model distribution compared to samples obtained with classical MCMC, Gibbs sampling, etc. Such improvements could further empower the broad family of generative models based on undirected probabilistic graphs, including deep architectures such as deep Boltzmann machines. This work points to the possibility that as QA hardware becomes more ubiquitous, a combination of QA and MCMC for ML may be the most efficient and effective approach. Investigation of the training using a combined Gibbs and D-Wave sample is the subject of our future work.
Acknowledgements The authors thank D-Wave Systems for access to their 2000 Q machine. This material is based on research sponsored by the Air Force Research Laboratory under agreement number FA8750-18-1-0096. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of the Air Force Research Laboratory (AFRL) or the US Government.

REFERENCES

1. Yoshua, B.: Learning deep architectures for AI. Found. Trends Mach. Learn. 2(1), 1–127 (2009). https://doi.org/10.1561/2200000006
2. Salakhutdinov, R.: Learning deep generative models. Annu. Rev. Stat. Appl. 2, 361 (2015)
3. Salakhutdinov, R.R., Hinton, G.E.: An efficient learning procedure for deep Boltzmann machines. Neural Comput. 24(8), 1967–2006 (2012). https://doi.org/10.1162/NECO_a_00311. Epub 2012 Apr 17
4. Frigessi, A., Martinelli, F., Stander, J.: Computational complexity of Markov Chain Monte Carlo methods for finite markov random fields. Biometrika 84, 1 (1997)
5. Dumoulin, V., Goodfellow, I.J., Courville, A.C., Bengio, Y.: (2014). On the challenges of physical implementations of RBMs. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, 2014, Quebec City, Quebec, Canada., pp. 1199–1205
6. Rose, G.: First ever DBM trained using a quantum computer (2014). https://dwave.wordpress.com/2014/01/06/first-ever-dbm-trained-using-a-quantum-computer/
7. Adachi, S.H., Henderson, M.P.: Application of quantum annealing to training of deep neural networks (2015). arXiv:1510.06356
8. Perdomo-Ortiz, A., O’Gorman, B., Fluegemann, J., Biswas, R., Smelyanskiy, V.N.: Determination and correction of persistent biases in quantum annealers (2015). arXiv:1503.05679v1
9. Benedetti, M., Reaple-Gómez, J., Biswas, R., Perdomo-Ortiz, A.: Estimation of effective temperatures in a quantum annealer and its impact in sampling applications: a case study towards deep learning applications. Phys. Rev. A 94, 022308 (2015)
10. Koshka, Y., Perera, D., Hall, S., Novotny, M.A.: Determination of the lowest-energy states for the model distribution of trained restricted Boltzmann machines using a 1000 Qubit D-Wave 2X quantum computer. Neural Comput. 29, 1815–1837 (2017)
11. Koshka, Y., Perera, D., Hall, S., Novotny, M.A.: Empirical investigation of the low temperature energy function of the Restricted Boltzmann Machine using a 1000 qubit D-Wave 2X. In: Proceedings of 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 1948–1954. https://doi.org/10.1109/ijcnn.2016.7727438
12. Amin, M.H., Andriyash, E., Rolfe, J., Kulchytskyy, B., Melko, R.: Quantum Boltzmann machine. arXiv:1601.02036
13. Benedetti, M., Realpe-Gómez, J., Biswas, R., Perdomo-Ortiz, A.: Quantum-assisted learning of hardware-embedded probabilistic graphical models. Physical Review X 7(4), 041052 (2017)
14. Dorband, J.E.: A Boltzmann machine implementation for the D-wave. In: 2015 12th International Conference on Information Technology—New Generations, Las Vegas, NV, 2015, pp. 703–707. https://doi.org/10.1109/ITNG.2015.118
15. Sleeman, J., Dorband, J., Halem, M.: A hybrid quantum enabled RBM advantage: convolutional autoencoders for quantum image compression and generative learning. arXiv:2001.11946
16. Liu, J., Spedalieri, F.M., Yao, K.T., Potok, T.E., Schuman, C., Young, S., Patton, R., Rose, G.S., Chankam, G.: Adiabatic quantum computation applied to deep learning networks. Entropy. 20(5), 380 (2018). https://doi.org/10.3390/e20050380
17. Koshka, Y., Novotny, M.A.: Comparison of use of a 2000 Qubit D-wave quantum annealer and MCMC for sampling, image reconstruction, and classification. IEEE Trans. Emerg. Top. Computat. Intell. (2000). https://doi.org/10.1109/TETCI.2018.2871466
18. Koshka, Y., Novotny, M.A.: 2000 Qubit D-wave quantum computer replacing MCMC for RBM image reconstruction and classification. In: 2018 International Joint Conference on Neural Networks (IJCNN): Rio de Janeiro, Brazil, July 2018, pp. 1–8 (2018). https://doi.org/10.1109/ijcnn.2018.8489746

Springer
19. MacKay, D.J.C.: Information Theory, Inference & Learning Algorithms. Cambridge University Press, Cambridge (2002)
20. Lichman, M.: UCI Machine Learning Repository (2013)
21. Santoro, G.E., Tosatti, E.: Topical review: optimization using quantum mechanics: quantum annealing through adiabatic evolution. J. Phys. A: Math. Gen. 39, R393–R431 (2006)
22. D-Wave Systems, Inc. http://www.dwavesys.com
23. Binder, K., Young, A.P.: Spin glasses: Experimental facts, theoretical concepts and open questions. Rev. Mod. Phys. 58, 801 (1986)
24. Stein, D.L., Newman, C.M.: Spin Glasses and Complexity. Princeton University Press, Princeton, NJ (2013)
25. Boixo, S., Rønnow, T.F., Isakov, S.V., Wang, Z., Wecker, D., Lidar, D.A., Martinis, J.M., Troyer, M.: Evidence for quantum annealing with more than one hundred qubits. Nat. Phys. 10(3), 218–224 (2014)
26. Trummer, I., Koch, C.: Multiple query optimization on the D-Wave 2X adiabatic quantum computer (2015). arXiv:1510.06437
27. Novotny, M.A., Hobl, L., Hall, J.S., Michielsen, J.S.: Spanning tree calculations on D-Wave 2 machines. In: Journal of Physics: Conference Series, vol. 681, 012005. International Conference on Computer Simulation in Physics and Beyond (CSP 2015), 6–10 Sept. 2015, Moscow, Russia, IOP Publishing Ltd. (2016). https://iopscience.iop.org/article/10.1088/1742-6596/681/1/012005/meta
28. Fischer, A., Igel, C.: Training restricted Boltzmann machines: an introduction. Pattern Recognit 47(1), 25–39 (2014)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.