Global Warming: Temperature Estimation in Annealers

Jack Raymond*, Sheir Yarkoni and Evgeny Andriyash

D-Wave Systems Inc., Burnaby, BC, Canada

Sampling from a Boltzmann distribution is NP-hard and so requires heuristic approaches. Quantum annealing is one promising candidate. The failure of annealing dynamics to equilibrate on practical time scales is a well understood limitation, but does not always prevent a heuristically useful distribution from being generated. In this paper, we evaluate several methods for determining a useful operational temperature range for annealers. We show that, even where distributions deviate from the Boltzmann distribution due to ergodicity breaking, these estimates can be useful. We introduce the concepts of local and global temperatures that are captured by different estimation methods. We argue that for practical application it often makes sense to analyze annealers that are subject to post-processing in order to isolate the macroscopic distribution deviations that are a practical barrier to their application.

Keywords: quantum annealing, thermal annealing, maximum likelihood estimation, pseudo-likelihood, ergodicity breaking

1. INTRODUCTION

Boltzmann distributions are important in many areas of science, and sampling from these distributions is a major bottleneck in many interesting applications. The tasks of uniform generation, approximate counting, and inference (e.g., estimation of marginal probabilities) are often NP-hard (Sinclair and Jerrum, 1989; Cooper, 1990; Long and Servedio, 2010). Heuristic samplers that sample approximately from a Boltzmann distributions are applied in practice to large scale problems [for example, in machine learning (Salakhutdinov and Hinton, 2009; Rolfe, 2016)].

One approach to heuristic sampling is to use an annealer. Whether thermal or quantum, an annealer generates independent samples by slowly transforming an easily prepared initial state into a random final state associated with a given objective function (Kirkpatrick et al., 1983; Kadowaki and Nishimori, 1998). In the case of the simulated thermal annealer (STA), an initial random sample is evolved through a schedule of decreasing temperature toward a specified terminal temperature (Kirkpatrick et al., 1983; Landau and Binder, 2005). In quantum annealing, the initial state is a ground state of some driver Hamiltonian (often a uniform superposition of states). During the annealing process, the state is evolved by slowly changing the Hamiltonian toward the target Hamiltonian (Kadowaki and Nishimori, 1998).

Although annealers have primarily been considered in the context of optimization, they can also be used as heuristic samplers of Boltzmann distributions. With sufficient resources, STA samples from a Boltzmann distribution (Kirkpatrick et al., 1983; Neal, 1993). However, the resources required per sample to achieve this are prohibitory in interesting applications, so that it is typically run as a heuristic without theoretical guarantees. Previous studies have also indicated that samples produced by the D-Wave quantum annealers may produce samples well described by finite temperature Boltzmann distributions (Bian et al., 2010; Denil and de Freitas, 2011; Amin, 2015; Benedetti et al., 2016; Dumoulin et al., 2015; Amin et al., 2016).

In this paper, we investigate several methods for determining how close sample distributions produced by annealers are to a family of Boltzmann distributions parameterized by inverse temperature β. These methods estimate the parameter β best describing samples drawn from an annealer, and also provide measures of closeness. The annealers we evaluate are the latest-model D-Wave[1] quantum annealer – the D-Wave 2X (DW2X) (Johnson et al., 2011; Denchev et al., 2016; King et al., 2016, 2015), and an implementation of simulated thermal annealing on a single CPU. We consider two knobs for each annealer that modify the heuristic distributions generated: rescaling the STA terminal temperature/DW2X terminal energy scale, or changing annealing time (either in the DW2X or STA). Sample quality shown here does not reflect performance of optimally tuned versions of these annealers, and are simply presented to compare various β estimation techniques.

We observe a significant difference between *local* (subspace) and *global* (full space) features of the annealer distributions. We find that even though samples are locally similar to a Boltzmann distribution, the global deviation can be large. This gives rise to a "global warming" effect: the fact that global distributional features indicate a higher temperature than local distributional features. We consider several estimators of inverse temperature and evaluate their efficacy. Some estimators are sensitive to details of the dynamics, and indicate a significant difference between the DW2X and STA. Other estimators are more sensitive to ergodicity breaking and macroscopic distribution features, where the DW2X and STA show a qualitatively similar behavior.

We treat our heuristic samplers as black-boxes and consider temperature estimation as the problem of determining the best fit among a single-parameter exponential family of models. This problem has a long history, and best practice is well established (Geyer and Thompson, 1992; Lehmann and Casella, 1998; Wainwright and Jordan, 2008). Inference of Ising models parameters under some systematic schemes is NP-hard (Wainwright and Jordan, 2008; Bresler et al., 2014). However, heuristic approaches, such as log-pseudo-likelihood are known to perform well in practice (Besag, 1975), and some schemes are provably convergent with reasonable resources (Bhattacharya and Mukherjee, 2015; Montanari, 2015). Bhattacharya and Mukherjee (2015) recently considered the log-pseudo-likelihood estimator for β and found that estimation based on only a single sample is possible; their focus was primarily on the convergence properties of this estimator. Multi-parameter estimation (estimation of couplings and fields) is more commonly studied, and is pertinent to the class of Ising models we study, though beyond the scope of this paper. In this context, efficient methods of estimation for strongly interacting models include pseudo-likelihood and variational approaches (Aurell and Ekeberg, 2012; Nguyen and Berg, 2012; Albert and Swendsen, 2014).

Many recent papers have shown that physical quantum annealers approximate Boltzmann distributions (Bian et al., 2010; Denil and de Freitas, 2011; Amin, 2015; Benedetti et al., 2016; Dumoulin et al., 2015; Amin et al., 2016; Perdomo-Ortiz et al., 2016). In some of these approaches, temperature estimators have been developed,

and these estimators have been effectively applied in correcting the annealer parameterization to produce the desired distribution. A significant focus has been the impact of noise, or systematic specification errors, in D-Wave processors. Remedies have been proposed to allow more effective sampling, but scaling is either poor or unproven; in some methods only a restricted set of problem classes are appropriate. An extension to the temperature estimation method of Benedetti et al. (2016) is discussed in supplementary materials,[2] but we prefer the more standard estimators presented in the main text. Some work considering closeness to quantum Boltzmann distributions has appeared (Amin, 2015; Amin et al., 2016).

Our paper evaluates several standard methods, but differs from previous studies in that it uses insight specific to annealers in the analysis of deviations and development of temperature estimators. Noise sources and quantum features in physical quantum annealers are discussed only briefly. Some estimators we evaluate have a firm theoretical basis, such as maximum likelihood, but where this is lacking we will not focus on formal properties, such as convergence, bias, and variance.

Qualitatively, the deviation of the STA distributions from the Boltzmann distribution for hard-to-sample Hamiltonians has been understood within physics and computer science since the idea of annealers was conceived (Kirkpatrick et al., 1983; Neal, 1993; Landau and Binder, 2005). As we modify the inverse temperature in the STA from its initial value to the terminal inverse temperature value (β_T), we move from a distribution where classical states are uniformly distributed to a distribution divided into disjoint subspaces of low energy (which can be identified qualitatively with modes of the probability distribution, or valleys in the free energy landscape). Under annealing dynamics, a state localized in one subspace cannot easily transition into another subspace that is separated by an energetic barrier once the inverse temperature becomes large (late in the annealing procedure) – this is called *ergodicity breaking* (Neal, 1993; Landau and Binder, 2005). In the case of the STA, we gradually decrease the annealing temperature. Temperature is in one-to-one correspondence with expected energy in a Boltzmann distribution, and equilibrated samples are characterized by tight energy ranges. These samples are partitioned into subspaces by the energy barriers as temperature decreases, at which point the samples in each subspace will evolve independently, and be characterized by a local distribution. Rare fluctuations do allow samples to cross barriers, but are exponentially suppressed in the height of the energy barrier later in the anneal. Therefore, the distribution between subspaces will reflect the distribution at the point in the anneal where dynamics between the subspaces became slow, rather than the equilibrated distribution associated with the terminal model. This effect is called "freeze-out." We provide a schematic in **Figure 1**.

Similarly, in the case of quantum annealing, we proceed through a sequence of quantum models of decreasing transverse field (and increasing classical energy scale). With respect to the terminal diagonal Hamiltonian, the energy is again decreasing throughout the anneal, and some characteristic mean energy
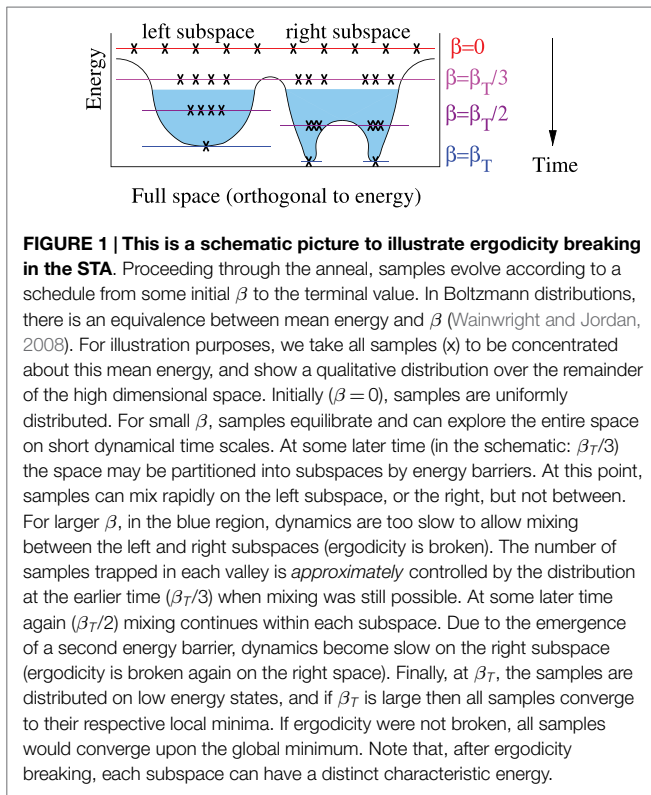

FIGURE 1 | This is a schematic picture to illustrate ergodicity breaking in the STA. Proceeding through the anneal, samples evolve according to a schedule from some initial β to the terminal value. In Boltzmann distributions, there is an equivalence between mean energy and β (Wainwright and Jordan, 2008). For illustration purposes, we take all samples (x) to be concentrated about this mean energy, and show a qualitative distribution over the remainder of the high dimensional space. Initially ($\beta = 0$), samples are uniformly distributed. For small β, samples equilibrate and can explore the entire space on short dynamical time scales. At some later time (in the schematic: $\beta_T/3$) the space may be partitioned into subspaces by energy barriers. At this point, samples can mix rapidly on the left subspace, or the right, but not between. For larger β, in the blue region, dynamics are too slow to allow mixing between the left and right subspaces (ergodicity is broken). The number of samples trapped in each valley is *approximately* controlled by the distribution at the earlier time ($\beta_T/3$) when mixing was still possible. At some later time again ($\beta_T/2$) mixing continues within each subspace. Due to the emergence of a second energy barrier, dynamics become slow on the right subspace (ergodicity is broken again on the right space). Finally, at β_T, the samples are distributed on low energy states, and if β_T is large then all samples converge to their respective local minima. If ergodicity were not broken, all samples would converge upon the global minimum. Note that, after ergodicity breaking, each subspace can have a distinct characteristic energy.

defines the sample set at intermediate stages. Energy barriers become impassable as the transverse field weakens and tunneling becomes slow, so that the process of ergodicity breaking is qualitatively similar (Kadowaki and Nishimori, 1998; Amin, 2015). Tunneling dynamics are affected by energy barriers in a different manner to thermal excitation dynamics, which is why there is some enthusiasm for quantum annealing; for some problems, it may not suffer the same dynamical slowdown that is true of STA (Denchev et al., 2016).

For many problem classes, the points of ergodicity breaking become well defined in the large system size limit, and can often be directly associated with thermodynamic phase transitions (Landau and Binder, 2005; Mezard and Montanari, 2009). This is true both of transitions related to symmetry breaking (such as ferromagnetic transitions) where domains are formed according to a simple symmetry of the problem, and those related to random (*a priori* unknown) problem structure (such as spin glass transitions). Our analysis will capture ergodicity breaking that relates only to the random problem structure, as this is a more practical barrier to heuristic sampling. This point is discussed further in the discussion section.

In this paper, we consider heuristic sampling from classical Ising spin models. The state x will consist of N spins, defined on $\{-1, +1\}^N$. The Hamiltonian is

$$H(x) = \sum_{ij} x_i J_{ij} x_j + \sum_i h_i x_i, \tag{1}$$

where J and h are unitless model parameters called couplings and fields, respectively. The Boltzmann distribution corresponding to

this Hamiltonian at inverse temperature β is

$$B_\beta(x) = \frac{1}{Z(\beta)} \exp(-\beta H(x)), \tag{2}$$

where Z is the partition function. Throughout the paper, we will use the standard, though improper, abbreviation in which x can denote both the random variable X, and its realization. We study two problem classes compatible with the Chimera topology described in Section 4.1, tailored to the D-Wave architecture (Katzgraber et al., 2014; Rønnow et al., 2014; King et al., 2015).

1.1. Outline

In Section 2, we introduce Kullback–Leibler divergence and mean square error on correlations as measures of closeness to the Boltzmann distribution, and from these develop standard estimators of inverse temperature.

In Section 3, we develop local self-consistent approximations for efficiently evaluating our inverse temperature estimators; we call these *locally consistent* inverse temperature estimators. We argue that in the context of annealers the approximation may determine an inverse temperature significantly different (typically larger) than that obtained by a full (computationally intensive) evaluation method. However, we will show that the estimate has meaning in that it captures local distribution features. Applying our approximation to maximum likelihood estimation, we recover in a special case the commonly used pseudo-log-likelihood approximation method.

In Section 3.2, we argue for evaluating post-processed distributions in place of raw distributions, for the purposes of removing superficial deviations in the heuristic distribution, and for determining the practical (as opposed to superficial) limitations of heuristic annealers.

We then present experimental results relating to our objectives and estimators in Section 4. We conclude in Section 5.

In supplementary materials, we present various supporting results to complement the main text, and additional results and methods (Raymond et al., 2016). In particular, we introduce a new *multi-canonical* approximation for β estimation inspired by Benedetti et al. (2016), we address issues related to practical usage of a DW2X, and we develop a method for calculation of KL-divergence. As part of our work we study the RAN1 and AC3 models of the main text, as well as two problem classes not tailored to the DW2X architecture (Bart, 1988; Douglass et al., 2015).

1.2. Practical Guidelines for Using Annealers in Sampling Applications

Based on the results in this paper, we offer the following advice for selecting and interpreting temperature estimation methods in the context of samples drawn from an annealer.

- It is important to define a suitable objective that is minimized by Boltzmann samples, and check that the objective is indeed small for the heuristic sampler for some temperature. It is not sufficient to find the best temperature, since there will always be a best temperature even for bad distributions. Comparisons of β estimates between heuristics are not meaningful in the absence of this analysis.

- Robust evaluation of a heuristic sampler will often require input from an independent (exact or heuristic) method, such as statistical estimates against which to compare. Attempting to quantify error in a locally self-consistent manner could be misleading.
- The DW2X rescaling parameter/STA terminal temperature, total annealing time, and post-processing, should be tuned to the sampling objective.
- A temperature can be estimated accurately and efficiently by standard methods, such as the log-pseudo-likelihood method. If ergodicity breaking is a weak effect, the log-pseudo-likelihood estimator is sufficient.
- It is valuable to consider several different types of estimator, since different estimators may be sensitive to different distribution features. Disagreement among estimators may reveal a pattern of ergodicity breaking, or imply a path to error mitigation.
- Efficient post-processing can move the distribution toward Boltzmann distribution by correcting local deviations. We are interested in the best practical heuristic, and so efficient post-processing should be applied.
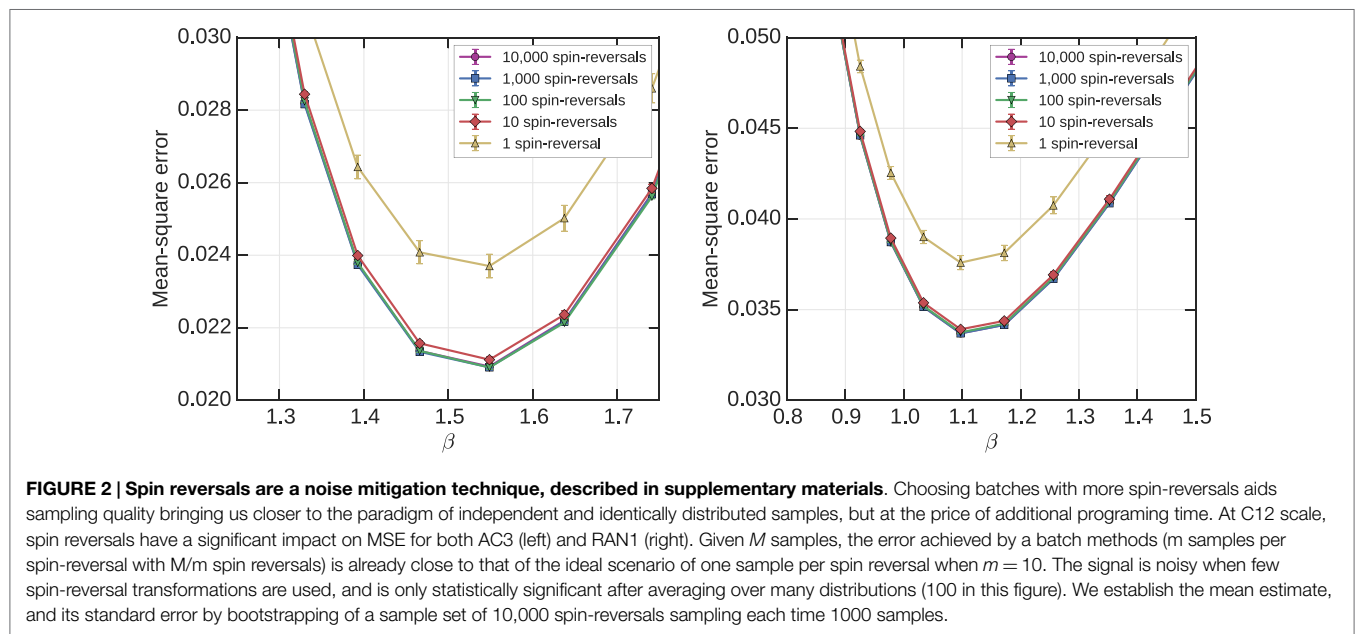
We argue in this paper, in line with previous literature and experimental results, that the temperature that best describes an annealer distribution is expected to be a function not only of the annealer parameterization, but also of the target Hamiltonian. There is no single parameter β that is optimal for all Hamiltonians. While this should be borne in mind, closely related Hamiltonians (e.g., those of a given class, created by a slow learning procedure, or otherwise of comparable statistical properties) do yield comparable estimates for temperature, so that it may be efficient to estimate temperature properties on a small subset of the problems of interest and effectively generalize.

2. ESTIMATORS FOR TEMPERATURE

We assume that annealers generate independent and identically distributed samples, according to a distribution P_A. For the STA, this is reasonable given powerful pseudo-random number generators. For the DW2X, correlated noise sources (discussed in Section 4.3) means this is an approximation that is more difficult to analyze. In **Figure 2**, we show evidence that these weak correlations in time do not strongly affect our results and conclusions. The experimental structure is demoted to supplementary materials.

We are interested in comparing these heuristic distributions to a family of Boltzmann distributions [equation (2)] parameterized by inverse temperature β. Amongst such models we wish to find the best fit, and measure its goodness. We will consider the best temperature to be that which minimizes some objective function. Since the distribution P_A is a heuristic distribution, and not Boltzmann, this temperature may vary between objectives. Given an objective that is minimized at some unique inverse temperature, we then need an estimator for this temperature working on the basis of finite sample sets. An effective estimator should be consistent, with low bias and variance (Geyer and Thompson, 1992; Lehmann and Casella, 1998; Shirts and Chodera, 2008; Bhattacharya and Mukherjee, 2015; Montanari, 2015). The estimators we study will be consistent, and in some cases optimal with respect to variance and bias [e.g., the Maximum Likelihood estimators (Geyer and Thompson, 1992; Lehmann and Casella, 1998)].

Either to evaluate the objective or to estimate temperature (i.e., minimize the objective) note that we must evaluate some statistics from the Boltzmann distribution, e.g., the mean energy, an energy gap, or marginal distribution. Inference for any of these quantities is NP-hard in the model classes we study. It is often in practice easier to evaluate the energy, and perhaps $\log(Z)$, than marginal statistics, but estimation of all these quantities is

FIGURE 2 | Spin reversals are a noise mitigation technique, described in supplementary materials. Choosing batches with more spin-reversals aids sampling quality bringing us closer to the paradigm of independent and identically distributed samples, but at the price of additional programing time. At C12 scale, spin reversals have a significant impact on MSE for both AC3 (left) and RAN1 (right). Given M samples, the error achieved by a batch methods (m samples per spin-reversal with M/m spin reversals) is already close to that of the ideal scenario of one sample per spin reversal when $m = 10$. The signal is noisy when few spin-reversal transformations are used, and is only statistically significant after averaging over many distributions (100 in this figure). We establish the mean estimate, and its standard error by bootstrapping of a sample set of 10,000 spin-reversals sampling each time 1000 samples.

slow in the worst case. For purposes of the models and temperatures explored, we are able to accurately estimate the mean energy, $\log(Z)$, or marginal expectations under the Boltzmann distribution by either dynamic programing or parallel tempering (Hukushima and Nemoto, 1996; Wainwright and Jordan, 2008; Selby, 2014). With these values in hand, we can efficiently evaluate our temperature estimators, and in most cases the objective (an exception, Kullback–Leibler divergence, is discussed in supplementary materials). However, a scalable estimator requires us to find effective approximation methods for these quantities or to define different estimators, and is the subject of Section 3.

2.1. Maximum Likelihood (Minimum Kullback-Leibler Divergence)

When comparing distributions, a natural objective function to minimize is the Kullback–Leibler (KL) divergence between the sampled distribution (from the annealer) P_A and the corresponding Boltzmann distribution B_β, as follows:

$$D_{KL}[P_A, B_\beta] = \sum_x P_A(x)\log\left(\frac{P_A(x)}{B_\beta(x)}\right). \quad (3)$$

The Kullback–Leibler divergence is an important information-theoretic quantity, which places various limitations on the efficacy of P_A for modeling B_β, and vice-versa (Wainwright and Jordan, 2008).[3]

P_A has no β dependence, so that at the minimum of this function with respect to β, we obtain an energy matching criterion $\mathrm{EM}(\beta) = 0$, where

$$\mathrm{EM}(\beta) = \sum_x P_A(x)H(x) - \sum_x B_\beta(x)H(x). \quad (4)$$

The energy matching criterion yields the maximum likelihood estimator for β – the likelihood that the annealed samples were drawn from a Boltzmann distribution. Maximum likelihood is perhaps the most well established of procedures for estimating model parameters from data – in this case, the data being the samples drawn from P_A (Geyer and Thompson, 1992; Lehmann and Casella, 1998; Wainwright and Jordan, 2008). Note that the Boltzmann distribution is an exponential model, and so it is natural to define the estimator in terms of expected energy, which is the sufficient statistic associated to the parameter β (Wainwright and Jordan, 2008).

2.2. Minimum Mean Square Error on Correlations

In the context of machine learning, an important potential application of annealers, the important feature of samples may be the quality of some statistics that are derived from them. In particular, a machine learning process may require accurate estimation of single variable expectations, and expectations for products of

variables (correlations). For this reason, we consider an alternative objective, the mean square error (MSE) on correlations:

$$\mathrm{MSE}[P_A, P_\beta] = \frac{1}{M}\sum_{ij:J_{ij}\neq 0}\left(\sum_x [P_A(x) - P_\beta(x)]x_i x_j\right)^2, \quad (5)$$

M is number of non-zero couplings. We consider specifically the mean error on correlations (excluding errors on single variable expectations) since the models we study experimentally are all zero-field problems so that $E[x_i] = 0$ for all β by symmetry. Unlike the KL-divergence, MSE is not a convex function of β in general, although intuitively this might be expected for many problem classes and reasonable heuristics. A derivative of equation (5) with respect to β will yield a criterion for local optimality. This is a complicated expression dependent on many statistics, but is straightforward to approximate numerically in our examples.

We will find that in application to annealers the minimum for this second objective [equation (5)] can disagree with the maximum likelihood (minimum KL-divergence) estimator [equation (4)], typically being more sensitive to ergodicity breaking in our experiments. Note that, once ergodicity breaking has occurred, the mean energy can be improved as samples settle toward their respective local minima. The maximum likelihood value can be larger than that implied at the point of ergodicity breaking. By contrast, the distribution between the now disconnected subspaces that determines the correlations cannot be much improved as samples settle toward their local minima. Therefore, the minimum MSE estimator will typically indicate a smaller value for β, better inline with the point of ergodicity breaking. We stress that MSE is not in any sense a special objective function in this regard; many variations are possible and should be chosen in an application orientated manner.

3. LOCAL APPROXIMATIONS FOR EVALUATING OBJECTIVES AND ESTIMATORS

The problem with the objectives and estimators outlined in Section 2 is that their use requires inference with respect to the Boltzmann distribution that is independent of the heuristic annealer, which is NP-hard to perform: estimation of either the expected energy, or correlations.

In this section, we show how, beginning from the annealed distribution, we can build a reasonable approximation to the Boltzmann distribution and thereby evaluate the estimator self-consistently. The estimators are motivated as approximations to those of Section 2. However, we will show that even in cases where the approximation is poor, the estimator can still reveal useful information about the distribution.

3.1. Statistics of the Heuristic Distribution, and of the Boltzmann Distribution, by Local Self-Consistency

Our estimators and objectives require us to evaluate statistics of the annealed distribution P_A. Estimates of mean energy or correlations based on P_A can be obtained by evaluating those

---

[3]The reverse form of the KL-divergence, or its symmetrized form, are also interesting. We choose this form as it allows for evaluation in the limit $P(x) \rightarrow 0$, among other technical factors. This is discussed further in supplementary materials.

statistics from the sample set $S = \{x\}$, or equivalently, evaluating their corresponding expressions using the plug-in estimator to P_A:

$$\hat{P}_A(x) = \frac{1}{|S|} \sum_{x' \in S} \delta_{x,x'}. \qquad (6)$$

The quality of estimates depends on variance and sample size. In experiments, we typically use sample sets of size 10^4 that are sufficient for temperature estimation and evaluation of the objectives. The evaluation of the KL-divergence is one exception: our approximation [equation (6)] is known to fail when applied to the entropy term $-\sum_x P_A(x) \log P_A(x)$ (Grassberger, 2003; Paninski, 2003). In the supplementary materials, we discuss why evaluation is problematic, and propose a mitigation strategy.

We must also evaluate energy, correlations, and $\log(Z)$ under the Boltzmann distribution, which is NP-hard. However, under the assumption that $P_A(x)$ is close to the Boltzmann distribution, we may make a *locally consistent* approximation. The approximation to $B_\beta(x)$ is

$$\hat{B}_\beta(x) \propto \sum_{x'} \hat{P}_A(x') W_\beta(x|x'), \qquad (7)$$

where W_β is a β-dependent kernel. It efficiently maps any state into a new state, with the property that the distribution is unchanged if it is a Boltzmann distribution

$$B_\beta(x) \propto \sum_{x'} B_\beta(x') W_\beta(x|x'). \qquad (8)$$

The transition kernels used in Markov chain Monte Carlo (MCMC) methods, either singly or iteratively, are suitable candidates for W_β (Landau and Binder, 2005; Mezard and Montanari, 2009). The blocked Gibbs MCMC method is described in Section 4.2. The simplest example of W_β is conditional resampling of a single variable, which is an element in the blocked Gibbs sampling procedure. All variables except i are unchanged, and i is resampled according to the conditional Boltzmann distribution $B_\beta(x_i | x \setminus x_i)$ given the neighboring values. We label this kernel by (i), indicating the updated variable

$$W_\beta^{(i)}(x|x') = B_\beta(x_i | x'(\neq x_i')) \prod_{j(\neq i)} \delta_{x_j, x_j'}. \qquad (9)$$

Applying the approximation [equation (7)] in combination with the kernel [equation (9)] to maximum likelihood estimation we obtain an energy matching criterion for (i). Each kernel i defines an energy matching criterion and an estimate for β, but it is not possible to simultaneously satisfy the criteria for all i. We can make a composite energy matching criterion by weighting each of the criteria equally: taking an average over EM(β) for each i. In this case, we recover the maximum log-pseudo-likelihood (MLPL) estimator (Besag, 1975; Bhattacharya and Mukherjee, 2015). The MLPL estimator is normally derived and motivated slightly differently. An alternative way to combine the kernels $\{W^{(i)}\}$ is to define a composite kernel as a sum of the individual kernels. We prefer the MLPL estimator in this paper due to its prevalence in the literature and well-established statistical properties. We discuss

this further in supplementary materials, where alternative locally self-consistent estimators are also examined.

In the case of the Hamiltonian equation (1), the MLPL estimate is the solution to EM$(\beta) = 0$, where

$$\mathrm{EM}(\beta) = \sum_{x \in S} \sum_i x_i \zeta_i(x) \frac{\exp(2\beta x_i \zeta_i(x))}{1 + \exp(2\beta x_i \zeta_i(x))}, \qquad (10)$$

where $\zeta_i(x) = [h_i + \sum_j (J_{ij} + J_{ji}) x_j]$ is the effective field. Note that $-2 x_i \zeta_i$ is the energy change of flipping the state of spin i. Provided there exists at least one $\zeta_i(x) > 0$, and one value $\zeta_i(x) < 0$ (at least one local excitation in some sample), then this equation has a unique finite solution which can be found, for example, by a bisection search method.

For our purposes, the MLPL estimator is a special case of a more general *locally consistent* estimator. We choose a kernel, approximate B_β [equation (7)], and then evaluate the energy matching criterion [equation (4)]. The locally consistent approach can also be applied straightforwardly to the MSE, and minimum MSE estimator, of Section 2.2. In the main text, the only locally consistent estimator for which we present results is the MLPL estimator. In the supplementary materials, we perform an experiment to demonstrate how the strength of the kernel impacts temperature estimation.

Consider the following interpretation for the role of the kernel: We take every sample that the annealer produces, and conditionally resample according to W_β. We then take this new set of samples as an approximation to Boltzmann samples drawn according to B_β. Since W_β obeys detailed balance, it necessarily brings the distribution toward the Boltzmann distribution. Consider again **Figure 1**, and note that resampling single spins, or doing some other efficient conditional resampling procedure (i.e., some short-run MCMC procedure) does not lead to a significant macroscopic redistribution of the samples, except in the high-temperature regime where fast dynamical exploration of the space is possible. Thus, the approximation [equation (7)] will typically inherit the macroscopic bias of the sampling distribution through $\hat{P}_A$, but correct local biases. The locally consistent estimator is, therefore, effective in capturing any *local* deviation in the distribution P_A not representative of B_β.

3.2. Toward Global Distribution Features

Objective functions, such as maximum likelihood and minimum mean square error on marginals, are influenced by a combination of local and global distribution features. We have already seen that locally consistent estimators, such as MLPL, can assign a meaningful temperature for local deviations from the Boltzmann distribution. However, these may fail to capture macroscopic features. It would be useful to have an objective, or estimator for temperature, that reflects only the macroscopic distribution features. One way to do this is to manipulate P_A so that the local distributional features are removed.

We have also not considered so far the practical application of annealers as heuristic samplers. By our definition, for an annealer to be useful, it must do well on the appropriate objective, and be fast. However, these two aims are typically in tension. A method that allows one to trade off these two goals is post-processing.

In post-processing, we take individual samples, or the set of all samples, and apply some additional procedures to generate an improved set of samples. This requires additional resources and can be heuristic, or employed in a manner guaranteed to improve the objective.

Those distribution features that can be manipulated by post-processing will be called local. Local, since it is assumed that efficient post-processing will not be so powerful as to manipulate the macroscopic distribution in interesting cases. Among the easiest local feature to correct in annealers is local relaxation: the tendency of states to decrease in energy toward their local minima at the end of the anneal as illustrated in **Figure 1**.

Post-processing has two uses considered in this paper: to extract macroscopic distribution features (by discounting local distortions), and to improve the heuristic distributions. A post-processed distribution can be represented as

$$P_{\beta,A}(x) = \sum_{x'} W_{\beta}(x|x')P_A(x'), \tag{11}$$

where W_{β} is again a kernel.

With post-processing, we now have three distributions of interest: $P_A(x)$, $P_{\beta,A}(x)$ and $B_{\beta}(x)$. Until this section, we were interested exclusively in the closeness of $P_A(x)$ and $B_{\beta}(x)$, and under the assumption that $B_{\beta} \approx P_{\beta,A}(x)$ we developed efficient approximations to maximum likelihood (or minimum MSE) estimation in Section 3.1. We can now present a different interpretation. If we are interested in local deviations in the distribution, we should compare $P_A(x)$, $P_{\beta,A}(x)$; whereas if we are interested in global deviations we should compare $P_{\beta,A}(x)$ and $B_{\beta}(x)$. For practical purposes, only the latter comparison makes sense, since we can efficiently correct local errors. However, the former is important in understanding why annealers fail, and how to correct their distributions. The locally self-consistent estimators (such as MLPL) minimize a divergence between $P_A(x)$ and $P_{\beta,A}(x)$, and so should be interpreted as local approximations (yielding a local temperature estimate). In the case that $P_A(x) = B_{\beta}(x)$, then the distinction between global and local is no longer relevant, and this local approximation is a consistent and low variance estimator for the unique β describing the distribution.

Implicit in our definition [equation (11)] is a restriction to "do no harm" post-processing, we post-process at the β that defines the Boltzmann distribution of interest (or that to which we wish to compare). The criterion [equation (8)] does no harm since it is guaranteed to move any distribution toward a Boltzmann distribution in some sense, and never away from it. In the case that W_{β} involves only conditional resampling according to B_{β} {rule [equation (9)] is one such case}, it is straightforward to show that the $D_{KL}[P_{\beta,A}, B_{\beta}] \leq D_{KL}[P_A, B_{\beta}]$. It is reasonable to expect, though not guaranteed, that other objectives will improve under do no harm kernels. Heuristic approaches without such guarantees may sometimes do better in practice, but carry risks.

For purposes of isolating macroscopic features of the distribution, it is ideal to apply enough post-processing to remove the local distortions; but leave the macroscopic features intact. This is a balancing act that strictly exists only as a concept, since the distinction between local and global is blurred except perhaps in the large system size limit, and there will typically be several relevant

scales not just two. In experiments, we present results for post-processing consisting of one sweep of blocked Gibbs sampling (described in Section 4.2), a weak form of post-processing. For purposes of improving the heuristic, W_{β} should be chosen powerful enough that the time-per-sample is not significantly impacted. One sweep of blocked Gibbs sampling meets the criterion of being a small overhead in time per-sample for the DW2X and STA under the operation conditions we examine, so we can infer something of the power of post-processing to efficiently correcting annealer non-idealities.

If the heuristic distribution is a function of β [equation (11)], we must take into consideration the dependence of $P_{\beta,A}$ on β in objective minimization. KL-divergence minimization becomes distinct from maximum likelihood in the case that samples are a function of β, and the energy matching criterion [equation (4)] is modified in the former case. This point is further discussed in supplementary materials.

In Section 3, we developed locally self-consistent estimators. We emphasize that with post-processing, these estimators can be made redundant, unless the post-processing method kernel [equation (11)] is significantly different from the kernel used in the local self-consistency trick [equation (7)]. The self-consistency trick [equation (11)] uses information about how samples are redistributed under the kernel to determine β – if this kernel matches the post-processing kernel, then it detects the effect of post-processing and very little else. Therefore, in the evaluation of post-processed distributions, care should be taken in applying and interpreting self-consistent approximations to β.

4. EXPERIMENTAL RESULTS

In Section 4.1, we present the two models we will study in the main text. We then describe in Section 4.2 a simple Markov Chain Monte Carlo procedure called blocked Gibbs sampling, and how we use this to create the STA. The blocked Gibbs method is also applied in our post-processing experiments. We then describe our usage of the DW2X as a sampler. Experimental results demonstrating the methods of Sections 2 and 3.2 are subsequently presented.

4.1. RAN1 and AC3 Models

In this main section, we consider only two models, RAN1 and AC3, which are spin-glass models compatible with the DW2X topology (Bunyk et al., 2014). This topology is described by a *Chimera* graph shown in **Figure 3**; each variable is a circle with an associated programmable field h_i, and each edge is associated with a programmable coupling J_{ij}. Qubits are arranged in unit cells, each a $K_{4,4}$ bipartite graph of 8 qubits. Due to manufacturing errors, some qubits and couplings are defective and cannot be programed.

The RAN1 and AC3 problems are defined on a Chimera graph with 1100 qubits across a 12×12 cell grid (C12). In experiments, we consider models that exploit all available couplings and qubits (C12), as well as models that use only 127 qubits on a 4×4 cell subgrid (C4), and 32 qubits on a 2×2 cells subgrid (C2).

RAN1 is a simple spin-glass model without fields ($h_i = 0$), and with independent and identically distributed couplings uniform

FIGURE 3 | Working graph of the DW2X used; the topology is called Chimera.
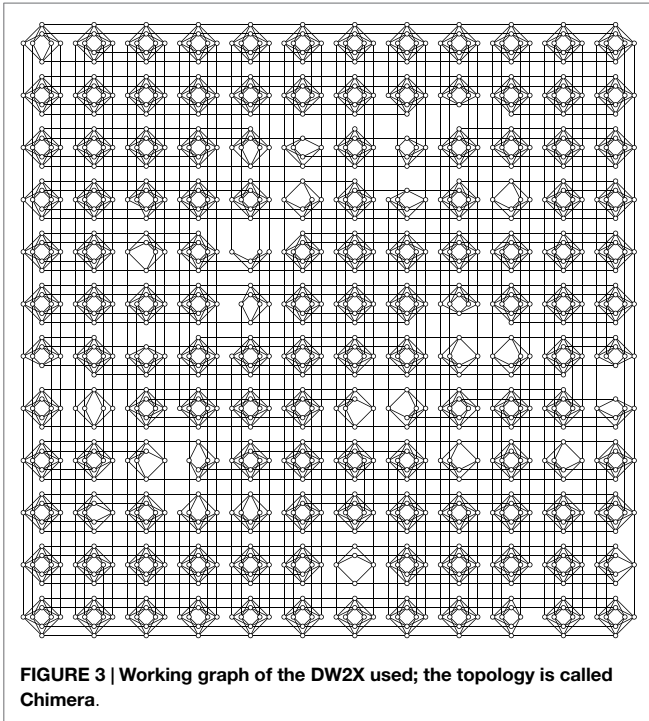
on $J_{ij} = \pm 1$ (King et al., 2015). Recent work has indicated that for some algorithms RAN1 may be a relatively easy problem in which to discover optima (Rønnow et al., 2014), and that asymptotically there is no finite temperature spin-glass phase transition (Katzgraber et al., 2014), making it questionable as a benchmark. However, the problem class demonstrates many interesting phenomena at intermediate scales and has become a well-understood benchmark for experimental analysis.

The AC3 model is a simple variation of the RAN1 class. Again we have no fields, but the intra-cell couplings' values (those between variables in the same cell) are sampled uniformly at random from $J_{ij} = \pm 1/3$, and inter-cell couplings set to $J_{ij} = -1$ (King et al., 2015).[4] By making couplings relatively stronger between cells, longer range interactions are induced through sequences of strongly correlated vertical, or horizontally, qubits. If we can consider the inter-cell couplings to dominate energetically, then the low energy solution space becomes compatible with that of a bipartite Sherrington Kirkpatrick model (Venturelli et al., 2015). We find that the AC3 problem is an interesting departure from RAN1 since the solution space is less dependent on local interactions, and also because we modify the precision of couplings (from an alphabet of ± 1, to an alphabet of $\pm 1/3$, -1).

4.2. Blocked Gibbs Sampling, and Simulated Thermal Annealing

Blocked Gibbs is a standard Markov chain Monte Carlo procedure closely related to the Metropolis algorithm procedure (Carreira-Perpiñán and Hinton, 2005; Landau and Binder, 2005). It is the

---

[4]We could equivalently assign the couplings between cells to ± 1 at random; due to a simple symmetry, the problem is not meaningfully changed.

basis for the STA in this paper, and also the post-processing results.

First note that because the Chimera graph is bipartite, it is 2-colorable. Given a coloring, the variables in a set of a given color are conditionally independent given the variables of the complementary set. Thus, it is possibility to simultaneously resample all states in one set, each according to the probability

$$P_\beta(x_i|x \setminus x_i) = \frac{\exp(-\beta \zeta_i(x)x_i)}{2\cosh(\beta \zeta_i(x))}, \quad (12)$$

where $\zeta(x) = (J + J^T)x + h$. We proceed through the colors in a fixed order, for each color resampling all variables. An update of all variables is called a sweep. This procedure can be iterated at a fixed temperature, and the distribution of samples is guaranteed to approach the Boltzmann distribution parameterized by β over sufficiently many sweeps. A graph coloring need not be optimal, and can always be found (given sufficiently many colors), so that this algorithm generalizes in an obvious manner to non-bipartite graphs. The blocked Gibbs sampling procedure at large β is not very efficient in sampling for multi-modal distributions, since samples are immediately trapped by the nearest modes (which may be of high energy), and escape only over a long timescale. A more effective strategy for multi-modal problems is blocked Gibbs annealing, in which β is slowly increased toward some terminal value β_T according to a schedule (a schedule assigns one temperature to each sweep of blocked Gibbs). In this paper, we consider an annealing schedule that is a linear interpolation between 0 and β_T. Given the restriction to a linear schedule the STA has two parameters: the total anneal time, and the terminal inverse temperature β_T. The setting of these parameters is discussed in Section 4.4.

4.3. Quantum Annealing with the D-Wave 2X

In the case of the DW2X, annealing is controlled by a time-dependent transverse field $\Delta(t)$ and an energy scale $E(t)$. These quantities are shown in **Figure 4** for the DW2X used in this paper. The physical temperature (T) of the system varies with time and load on the device and is difficult to estimate. We have experimentally observed a physical temperature which is 22.9 in the median, with quartiles of 22.0 and 25.6 mK, over the data collection period. We did not analyze the time scales associated with temperature fluctuations in depth, but much of the variation occurs on long time scales, so that in a single experiment we typically found a tighter range of temperatures applied. The unitless Hamiltonian operator in effect during the anneal is

$$\hat{H}(s) = \frac{h}{k_B T}\left(\frac{\Delta(s)}{2}\sum_i \hat{\sigma}_x^{(i)} + \frac{E(s)r}{2}\left[\sum_{ij}\frac{J_{ij}}{max(\{|J_{ij}|, |h_i|\})}\right.\right.$$

$$\left.\left.\times \hat{\sigma}_z^{(i)}\hat{\sigma}_z^{(j)} + \sum_i \frac{h_i}{max(\{|J_{ij}|, |h_i|\})}\hat{\sigma}_z^{(i)}\right]\right) \quad (13)$$

where $s = t/t_{max}$ is the rescaled time, the coefficient h is Planck's constant, and $\hat{\sigma}$ are the Pauli matrices. The Hamiltonian parameters can be considered rescaled to maximum value 1 (the function

FIGURE 4 | The DW2X operational energy scales during the anneal, the Hamiltonian is equation (13).
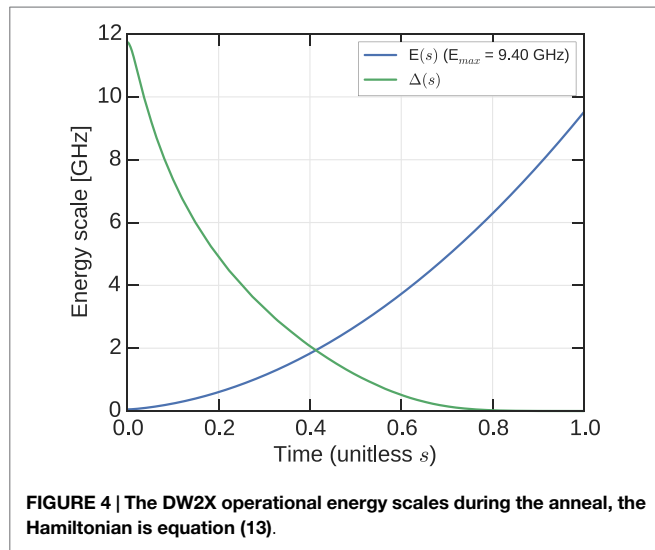
of the denominator max(·)). r and t_{max} are the rescaling parameter and the anneal time parameter, respectively. Throughout this paper, we adopt the convention that the Hamiltonian for the problem is fixed, and treat r as a parameter of the heuristic (DW2X). In current operation of the DW2X, this manipulation is achieved in practice by turning off autoscale, and manually rescaling the values of $\{J_{ij}, h_i\}$ submitted. We find the convention of modifying the quantum Hamiltonian equation (13) to be more intuitive than considering modification of the classical Hamiltonian that is submitted: when we rescale downwards, we are implying the use of smaller energy scales $E(t)$ in the annealer, and we anticipate that β is reduced.

4.4. Parameterization of the DW2X and STA in Experiments

Our criteria for selection of the default rescaling parameter r and the anneal time is minimization of the "Time to Solution" (TTS) (Rønnow et al., 2014; Isakov et al., 2015; King et al., 2015). TTS is the expected time required to see a ground state for the first time. For the DW2X, it is optimal to choose a minimal programmable anneal time (20 μs), and a maximum programmable energy scale $r = 1$. A second (and not uncorrelated) reason to use these parameters is that they are the default operation mode of the DW2X. Since our objective is not optimization, TTS is not the optimal way to use the DW2X, but represents a standard choice that allows for phenomena to be explored.

The STA parameters are chosen in simple manner in part according to TTS, and in part for convenient comparison with the DW2X. We choose the STA parameter β_T so that the distribution of local excitations in the DW2X and STA are comparable in the median case of 100 randomly generated C12 problems: by equating the *local* properties' differences between the annealers, we expose more interesting *global* features. To equate local properties β_T is chosen equal to the MLPL estimate of ta at full scale (C12) for the DW2X (see **Figure 5**). This implies $\beta_T = 3.54$ for RAN1, and $\beta_T = 4.82$ for AC3. In the case of RAN1, our choice $\beta_T = 3.54$ is not too dissimilar to the value ($\beta = 3$) that had been previously

used for optimization applications (Rønnow et al., 2014; Isakov et al., 2015; King et al., 2015).

We choose two sets for the number of sweeps of the STA, and show both in most figures. The first set is chosen to minimize TTS in the median instance. For both RAN1 and AC3, we find an approximately linear trend in the width the Chimera graph, so that 12,000, 4000, and 2000 sweeps were close to optimal for C12, C4, and C2 problems, respectively.

The second set was chosen to match the time per sample of the DW2X. A recent efficient implementation of the Metropolis algorithm for RAN1 problems achieves a rate of 6.65 spin flips per nano-second (Isakov et al., 2015). In C12 problems, we have 1100 active qubits, and so 20 μs would allow for $\frac{20000\ ns}{1100\ \text{spin flips}}$ * 6.65 spin flips/$ns \approx$ 120 sweeps (updates of all variables). For C4 and C2, we rescale linearly for simplicity (40 and 20, respectively).

In experiments, we evaluate the STA and DW2X on the basis of sample batches, each batch consisting of 10^4 samples. We approximate the samples as independent and identically distributed. In the case of the DW2X, this is an approximation and different programing procedures can impact quality of results. Our DW2X batches are generated by collecting 10^4 samples split across 10 programing cycles. This is a standard collection procedure that trades off quality of samples against timing considerations – including annealing time, programing time, and read-out time. The programing cycles exploit spin-reversals, a noise mitigating technique that strongly suppresses correlations between programing cycles (Boixo et al., 2013). The effect of this batch structure is presented in **Figure 2**.

4.5. Maximum Likelihood and Maximum Log-Pseudo-Likelihood Estimation

In this section, we consider the DW2X and STA without post-processing. The maximum log-pseudo-likelihood (MLPL) estimate can be interpreted as a form of locally self-consistent approximation to maximum likelihood, as described in 3.1, and here we compare it to the more computationally intensive maximum likelihood (ML) estimate of Section 2.1. MLPL and maximum likelihood estimators indicate different values of β, reflecting the failure of the locally self-consistent approximation [equation (8)]. MLPL captures a local temperature, consistent with the range over which the kernel [equation (9)] redistributes the sample. We consider the estimates with variation of the DW2X rescaling parameter r and the STA terminal temperature β_T relative to the default settings, for our annealing procedures on 100 randomly generated RAN1 problems at each of 3 sizes: C2 (32 variables), C4 (127 variables), and C12 (1100 variables). Results for AC3 are presented in supplementary materials.

We first consider the behavior of the STA with a range of terminal inverse temperatures between 0 and the default value for the problem class. Due to ergodicity breaking, we expect samples to fall out of equilibrium before the terminal temperature is reached, and so the distribution may be characterized by a value of β smaller than β_T, reflecting the range of inverse temperature for which dynamics slowed down. **Figure 6** shows the maximum likelihood and MLPL estimates based on the same sample sets. We see that the MLPL estimates follow a linear curve, which would indicate that the terminal model is indeed the best fit to the
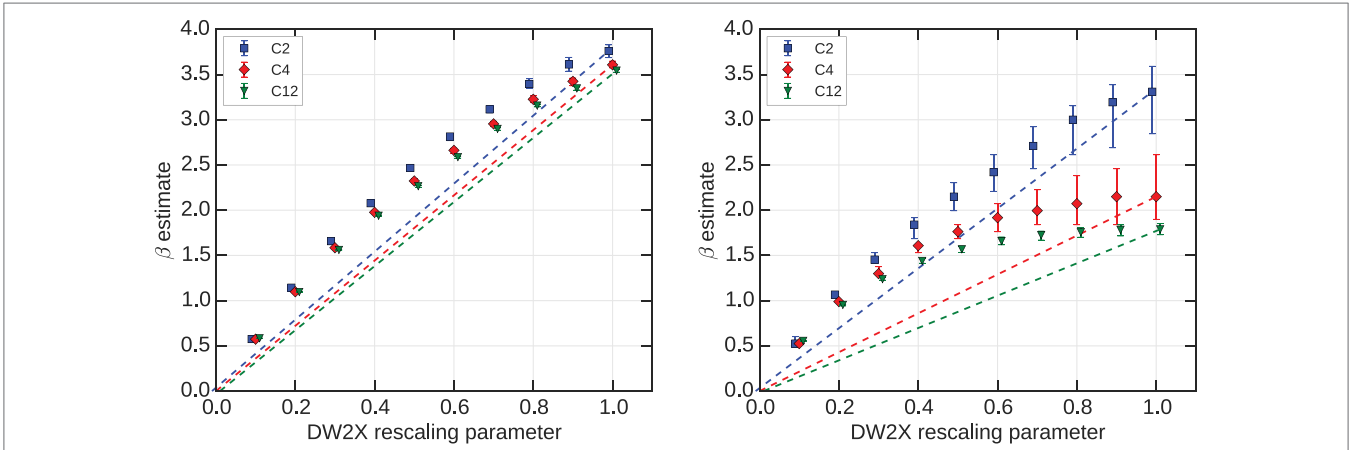
FIGURE 5 | As per Figure 6, but using samples from the DW2X. (Left) The MLPL method is non-linear, unlike the STA results. (Right) The maximum likelihood method shows qualitatively similar features to the STA. Note that the C12 result for the MLPL estimate at $r = 1$ is identical to that in **Figure 6**, the STA parameter β_T was chosen to meet this criteria as discussed in Section 4.2.
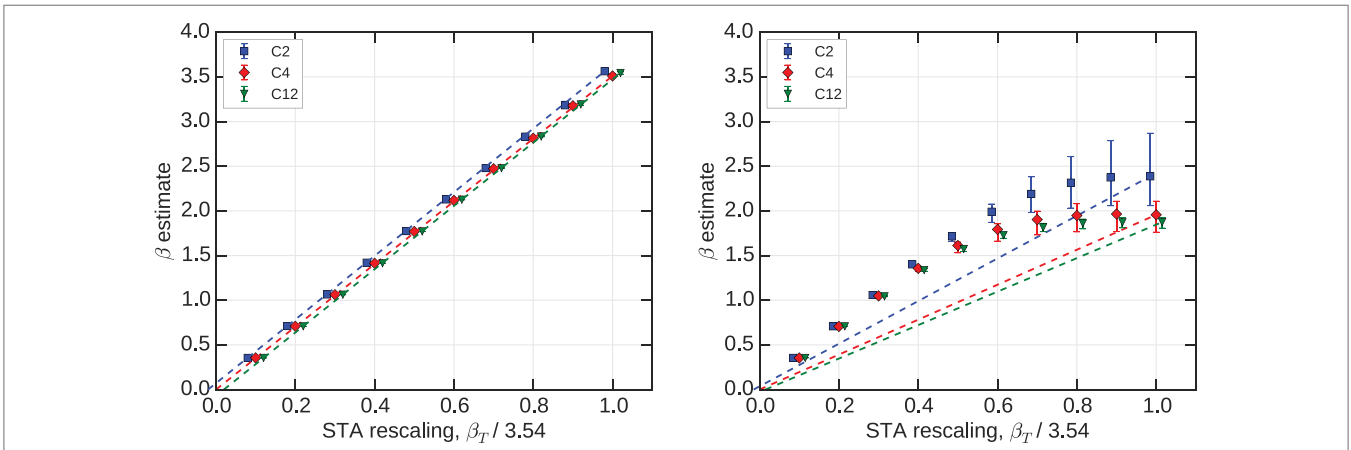

FIGURE 6 | Bars represent quartiles over 100 instances of RAN1 at each scale, estimates for β are in each case 10^4 samples generated by the STA. C2 and C12 data points are offset for visibility. (Left) Temperature estimates by the MLPL method. The MLPL estimate for thermal annealers matches the terminal model at all scales. (Right) Temperature estimates by ML. A non-linear trend is apparent due to ergodicity breaking, which is not captured by MLPL estimates.

samples, with no evidence of the ergodicity breaking we describe. By contrast, the maximum likelihood estimator is concave, with β significantly smaller than β_T.

The DW2X can also be manipulated by changing the rescaling parameter r on the interval $[0, 1]$. We thus repeat this experiment using sample sets from the DW2X. In **Figure 5** we see that both locally (MLPL) and globally (ML) the estimators are concave as a function of this rescaling. As with the STA, estimates are consistently larger for MLPL than maximum likelihood, and maximum likelihood estimates decrease for larger, more complicated, problems. Maximum likelihood exhibits some small decrease with system size. A naive interpretation of the rescaling parameter might lead to a general hypothesis that $\beta \propto r$, where the constant of proportionality can be determined by the terminal energy scale in annealing (see **Figure 4**). However, the physical dynamics of qubits are controlled by a *single qubit freeze-out* phenomenon that is discussed in supplementary material, with data summarized in **Figure 7**. The single qubit freeze-out figure implies that the

non-linearity of the MLPL curve in spin-glass problems is a function of the problem precision (granularity of the settings of J_{ij} and h_i), or more specifically, the pattern of energy gaps. Problems of higher precision, such as the AC3 problem class, have less pronounced non-linearity. The single qubit freeze-out phenomenon also anticipates the system size dependence seen in **Figure 5**, and in some other models not presented. AC3 results, and further discussion of this point, are in supplementary materials.

We believe the phenomenon underlying both MLPL and maximum likelihood to be easily understood, and similar in both the DW2X and STA, although the interaction of the dynamics with the energy landscape may be quite different. **Figure 1** provides a useful example to explain this. When ergodicity is broken, and the sample set is divided over subspaces, each subset relaxes toward the terminal model restricted to the subspace; the MLPL method effectively averages an estimate over these subspaces. By contrast, the maximum likelihood estimate accounts in part for the distribution between modes, determined at an early (higher

energy) stage of the anneal that is better described by smaller inverse temperature. Note that the equilibrated distribution at the ergodicity breaking point is a quantum one for the DW2X, unlike thermal annealers, so we ought to understand the deviation in the local and global temperatures in terms of the quantum parameterization ($\Delta(t)$, $E(t)$) (Amin, 2015). Still, the classical description in terms of β [equation (2)] provides the correct intuition.
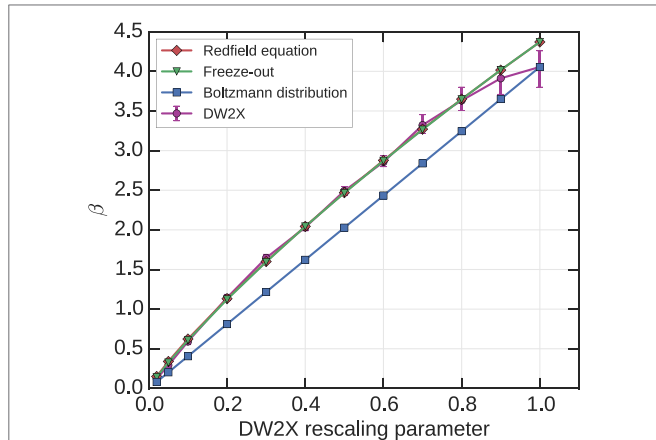

FIGURE 7 | An explanation for non-linearity of the MLPL estimator is possible through examination of single qubit dynamics in the DW2X. This study is described in supplementary materials. MLPL becomes equivalent to maximum likelihood in the limit of a single qubit. The single qubit β dependence on the rescaling parameter, for the single qubit Hamiltonian $H(x) = x_i$ is shown. We plot the DW2X median value together with 25–75 quantiles as error bars. Simulation of the physical dynamics gives the Redfield curve, and the "freeze-out" curve shows is based on the assumption of a single freeze-out point. The theory, thus, is in agreement with experiment. The DW2X is not equilibrated at the end of the anneal even for a single qubit model. By contrast, in a system that is locally or globally equilibrated at the end of the anneal a linear dependence would be expected, as is seen for the STA.

Response curves, such as these, can be used to choose a suitable parameterization of the terminal model – we can choose β_T so as to minimize KL-divergence. Similar curves could be constructed for any parameter, and with a distribution that is subject to post-processing. In the absence of the maximum likelihood curve (due to absence of approximates to the energy), a local information curve, such as MLPL curve, can be used as a compromise. The maximum likelihood curves, and MLPL curve for the DW2X, are problem dependent. Some temperature estimators require knowledge of, or assumptions about, the form of these curves – notably the method of Benedetti et al. (2016) and our multi-canonical method, both described in supplementary materials.

To judge quality of the approximation at the "best" β, it is appropriate to consider the quantity being minimized, KL-divergence. However, this is a difficult quantity to estimate from samples in the absence of parametric assumptions. We present a method for estimation of KL-divergence in supplementary materials, but find that for RAN1 and AC3 problems it is ineffective at C12 scales. This is an important reason to consider an alternative, such as the MSE estimator.

4.6. The Mean Square Error Estimator

The mean square errors on correlations associated with the DW2X and STA for a typical instance of RAN1, and a typical instance of AC3, are shown for the C12 (1100 variables) problem size in **Figure 8**. Though there is significant variation between the curves associated with different instances of these models, we chose among 100 random instances exemplars that are typical in the minimizing temperature and the mean square error (β, MSE) for DW2X.

Both the DW2X and STA performances are best characterized at intermediate β values, and variation from the default annealer settings is also shown to modify MSE. In the RAN1 exemplar, we demonstrate the effect of halving the terminal temperature or rescaling parameter, which improves MSE (except perhaps at very large values for β). In the case of the AC3 exemplar,
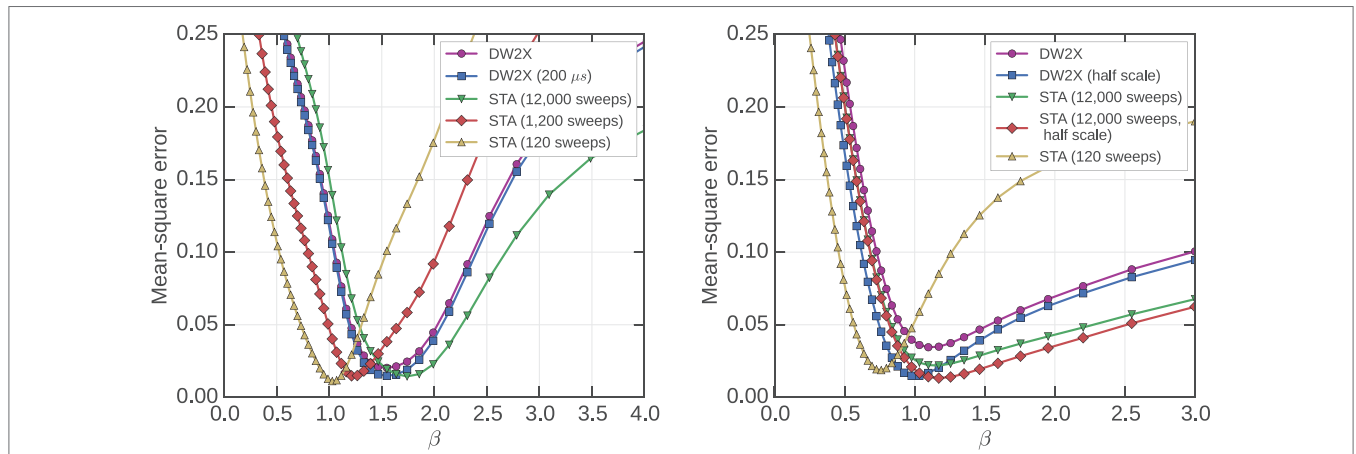

FIGURE 8 | Both the DW2X and STA can be used to sample from a RAN1 problem with small errors over an intermediate range of β. Objective performance is shown for two typical instances under several annealer operating conditions. (Left) Results for the AC3 exemplar. (Right) Results for the RAN1 exemplar. SEs determined by jack-knife methods are negligible compared to the marker size. To avoid clutter, we show only variation of the anneal time in the left figure, and only variation of the rescaling parameter ($r = 0.5$ in the DW2X, $\beta_T/2$ in the STA) in the right figure.

we demonstrate the effect of varying the anneal time. Results indicate that longer anneals tend to improve MSE at lower temperature. To prevent clutter, we have shown only one variation of a default parameter per model. Switching the parameter being varied results is qualitatively similar. Annealing for longer is expected to allow equilibration to lower temperatures, and so a better match is to be expected. Annealing with smaller r or β_T, concentrates annealing resources toward the initial part of the anneal where dynamics are effective (rather than at the end of the anneal where ergodicity breaking has already occurred and cannot be mitigated). It also reduces the tendency of samples to settle toward local minima that might distort the approximation for intermediate β. In the case of the DW2X, some noise sources and quantum mechanical phenomena can complicate this simple picture, but this is not obviously at play.

Figure 9 shows the mean square error achieved against the point it is minimized, for a large set of C12 (1100 variable) problem instances. Ideally, we wish for heuristics that are both sampling at large β and with small errors. **Figure 10** shows the minimum MSE estimator in comparison to the maximum likelihood estimator. These two estimates indicate different operational temperature ranges. The fact that the maximum likelihood estimator is significantly larger is to be expected in annealers. Late in the anneal samples sink toward their respective local minima, decreasing the mean energy significantly (see **Figure 1**). The mean energy is strongly dependent on the local relaxation, and hence the local inverse temperature, which as we saw in the previous section is large. In the case of MSE, the correlation error is for most models not strongly affected by this local decrease in energy. It is more sensitive to the distribution between modes, which is set only by the temperatures characterizing ergodicity breaking.

By annealing with modified β_T or r, improvements are made for sampling intermediate or small β. By contrast, it is relatively hard to sample effectively from large values of β by modifying these parameters; additional time resources are required to make

an impact. For this reason, we may argue that, generally, the larger the inverse temperature estimate, the more useful the annealer will be for hard sampling applications. However, it is important to note that an estimate for β, independent of the objective measure, may be risky or misleading. In **Figure 10** (right), the DW2X system at full scale indicates a lower minimum MSE β than the DW2X system operating at half scale. However, we can see that at this larger β value the half scale system is still more effective as a heuristic.

4.7. Effectiveness of Post-Processing

In Sections 4.5 and 4.6, we demonstrated how adjusting anneal duration, or the terminal temperature, can allow better objective outcomes. In this section, we consider briefly the effect of post-processing by one sweep of blocked Gibbs as discussed in Section 3.2. This post-processing changes dramatically the local distribution of samples, hence the MLPL estimate. However, the KL-divergence and mean square error, and the temperatures minimizing these objectives, are affected by a combination of the local and global distribution and so are modified in a non-trivial way by post-processing. Post-processing always strongly modifies local distribution properties, but only in the easy to sample regime (at small β) does it significantly impacts global distribution problems.

Figure 11 shows that MSE on correlations are, as expected, improved by post-processing. The improvements are dramatic in the regime $\beta \approx 0$, impressive over intermediate values of β, but almost negligible for larger β. After post-processing, the MSE curve has two minima. There is no guarantee and there should be two local minima working with arbitrary distributions. The first local minimum ($\beta = 0$) is evidence for the power of post-processing. Since one sweep of blocked Gibbs samples effectively at $\beta \approx 0$, independent of the initial condition P_A, $\beta = 0$ will be a global minimum for any heuristic distribution. The second local
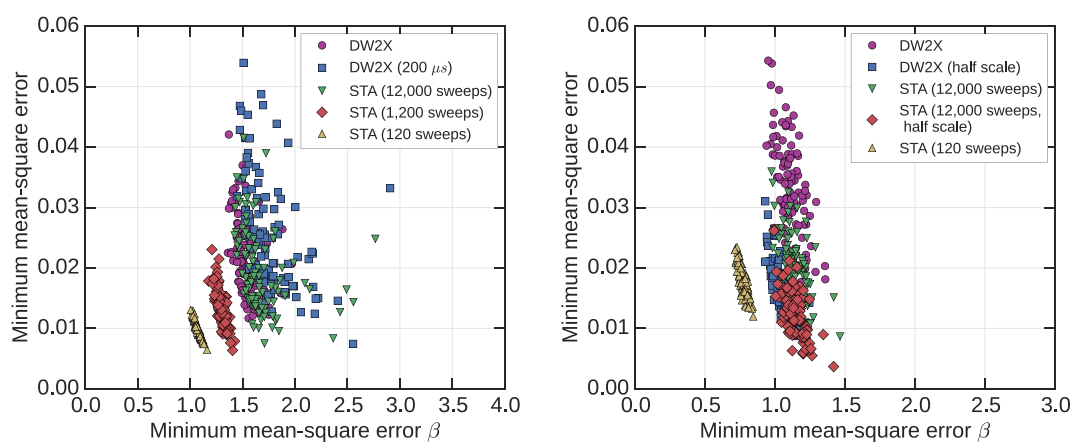

FIGURE 9 | Statistics over the set 100 AC3 (left) and RAN1 (right) problems at C12 scales. Small objective values at large inverse temperatures are difficult to obtain, and so desirable in a heuristic sampler. Sampling effectively at small inverse temperature is less valuable (e.g., 120 sweep STA). Modification of the annealing parameters significantly changes the distribution, allowing more effective emulation at some inverse temperatures. To avoid clutter, we show only variation of the anneal time in the left figure, and only variation of the terminal model rescaling (r, β_T) in the right figure; the effects are qualitatively similar in each of these models at C12 scale.
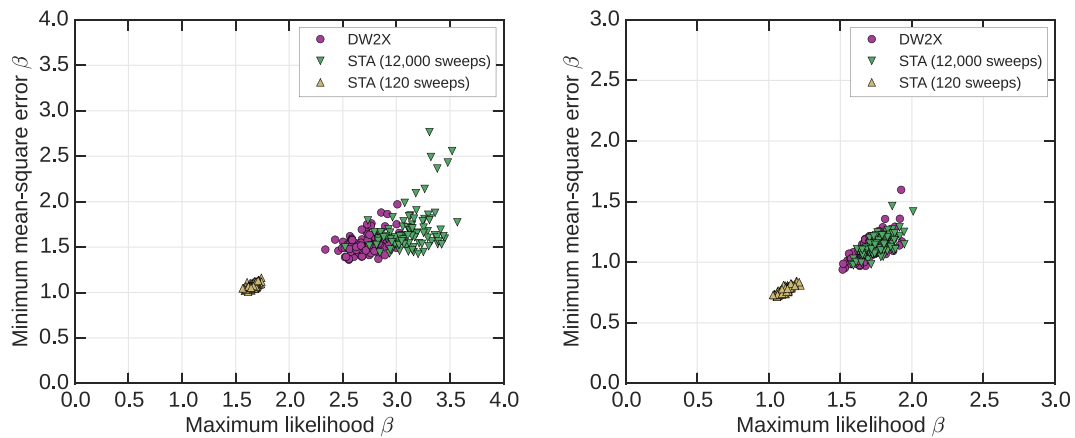
FIGURE 10 | Statistics over the set 100 AC3 (left) and RAN1 (right) problems at C12 scales, as per Figure 9. Minimum MSE and maximum likelihood estimators of temperature give different, but strongly correlated, results. The maximum likelihood estimate is typically larger, a partial explanation is the sinking of samples toward local minima late in the anneal, which through its impacts on mean energy has consequences for the maximum likelihood estimate.
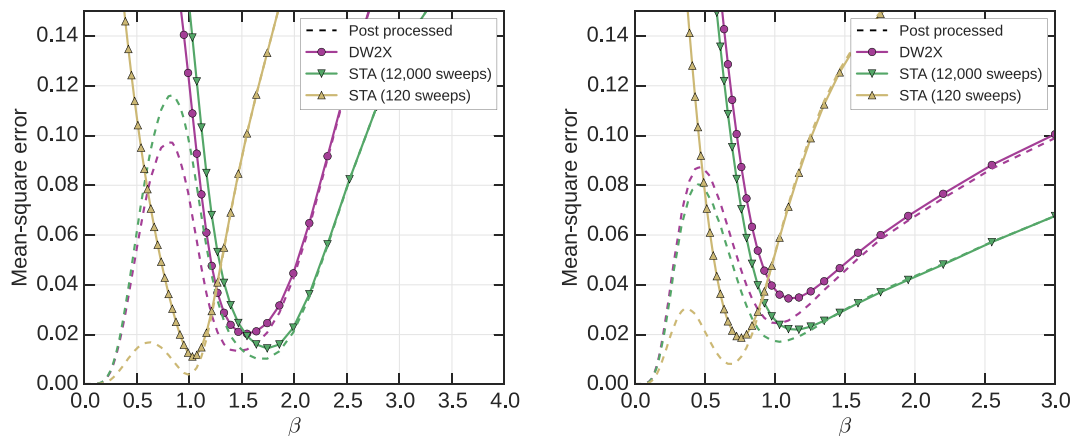

FIGURE 11 | Simple forms of post-processing can have a quantitatively large effect on objectives at small and intermediate temperatures. The C12 problem exemplars consistent with Figure 8 are shown. At low β a single sweep of blocked Gibbs can completely correct all errors. At high β, there is relatively little effect; however, the effect is significant in the intermediate range of inverse temperature where the annealers can be considered most effective.

minimum appears due to the closeness (at the macroscopic level) of the annealing distribution to some particular low-temperature Boltzmann distribution, a sweet spot of operation that may be of practical interest.

Post-processing reduces the error everywhere, but more so at smaller β where the time-scales for macroscopic redistribution are shorter. For this reason, we expect the minimizing value of β to move leftward with post-processing. If the post-processing allows global redistribution of samples, we may anticipate the disappearance of the local maximum separating the "easy for post-processing regime" from the "good for this annealer" regime; at which point a best operational regime for the annealer is less clear. However, we can assume that powerful post-processing of this kind is too expensive in the types of multi-modal problems where annealers are useful.

Figure 12 shows the statistics for the local minimum mean square error estimator, and its relation to the local maximum likelihood estimator, to be compared against Figure 10 that has

no post-processing. The local minimizer is the right most local minimum of the post-processed curve (see Figure 11), indicating the good operating regime. A global minimum is always at $\beta = 0$, but this is not of interest as it reflects the post-processor and not the heuristic.

In Figure 12 (left), we see that, relative to the distribution without post-processing, there is a slight shift leftwards in all distributions, and significant shift downwards that appears approximately proportional to the MSE without post-processing. The effect of local relaxation is to give the impression of better estimation at low temperature. Here, the effect is partially lifted to reveal that the macroscopic distribution may be characterized by slightly smaller β.

In Figure 12 (right), we see how the minimum MSE estimate decreases in all annealers, but by a less significant amount than the downward trend in the maximum likelihood estimator. The two remain strongly correlated under post-processing. It is natural to expect that the local relaxation, which shifts samples toward
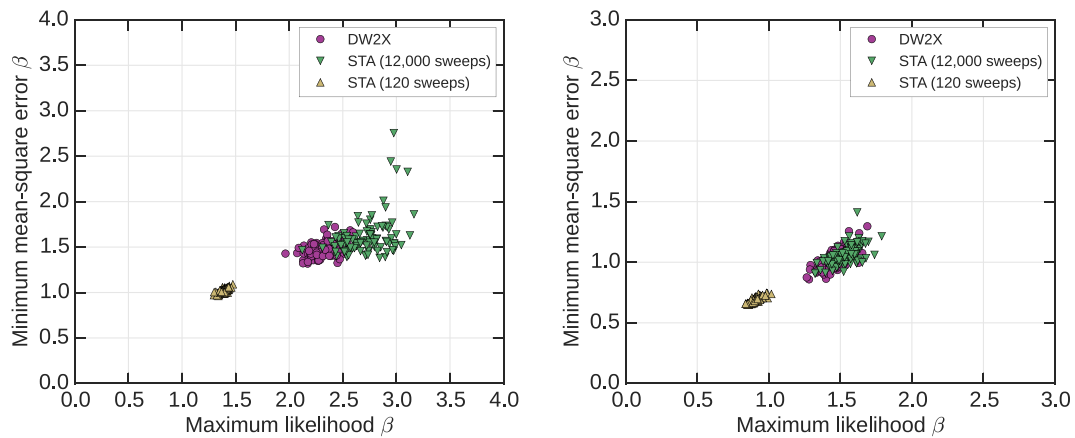
FIGURE 12 | As per Figure 10, but now all distributions are modified by one sweep of blocked Gibbs sampling. The estimators for inverse temperature are reduced, as the effect of the local distribution (characterized by larger inverse temperature) is partially removed.
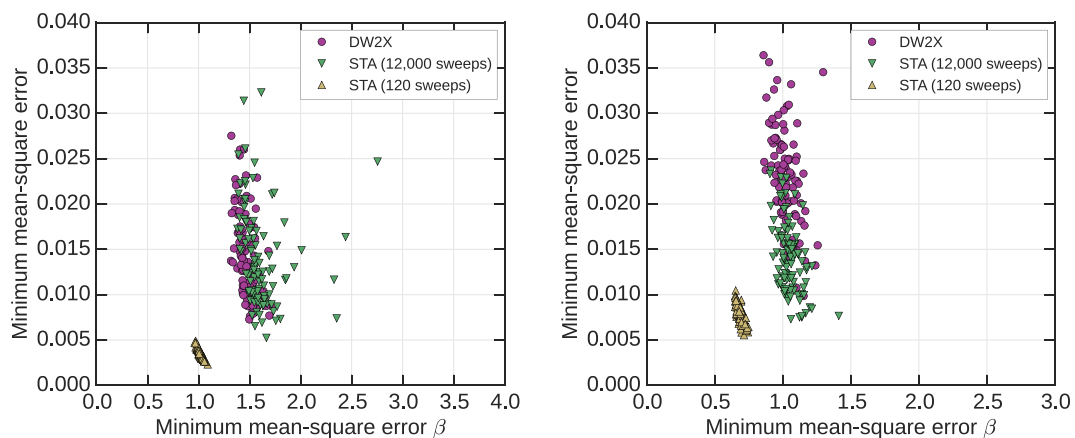

FIGURE 13 | As per Figure 9, but with distributions modified by one sweep of blocked Gibbs sampling. Objectives are improved everywhere very significantly, and by a comparable fraction across the different annealers.

their local minimum, may have a bigger impact on KL-divergence than MSE, because it is easy to raise the energy by post-processing that impacts maximum likelihood in a systematic manner, but it is difficult to redistribute samples macroscopically, which may be required to alter minimum MSE.

The particular effects demonstrated on the exemplar instances of **Figure 12** are reflected at the distribution level in **Figure 13**.

The estimation of KL-divergence is described in supplementary materials, and allows us to measure KL-divergence post-processed RAN1 and AC3 problems, with reasonable precision up to C4 scale. In **Figure 14**, we demonstrate results for an exemplar on a C4 graph. The pattern observed is qualitatively similar to **Figure 11** in that we see a global minimum at $\beta = 0$ reflecting the effectiveness of the post-processing, and a local minimum at intermediate β that reflects a promising region for application of the annealer as a heuristic. As discussed in the supplementary materials, bias can be a problem with our KL-divergence estimator. For this reason, we present a C4 (rather than C12) sized problem and only the post-processed (rather than unprocessed and post-processed)

estimates. To assess the bias (the sensitivity of the estimate to the finite sample set size), the jack knife bias-corrected estimator is also shown (Efron, 1982). At C4 scale, the bias does not significantly obscure the phenomena, particularly for the more powerful annealers (the STA with 4000 sweeps, and the DW2X).

5. DISCUSSION

In this paper, we have considered several temperature estimators applied in the context of a physical quantum annealer set up for optimization (DW2X), and a comparable simply parameterized simulated thermal annealer (STA). We have demonstrated how different objective measures of closeness to the Boltzmann distribution respond differently to local and global distribution features. An important phenomenon we observe is that in annealed distributions we have a range of temperature estimates according to the method employed. We have shown that estimators indicating larger temperature are those responsive to macroscopic (global) features of the distribution. From an estimator
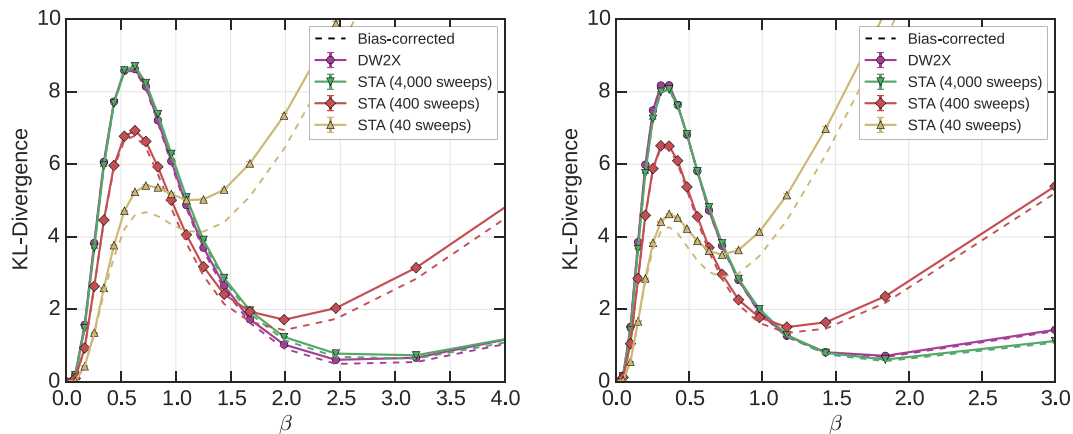
FIGURE 14 | KL-divergence results for two exemplar instances of AC3 (left) and RAN1 (right) each at C4 scale (127 variables). Full lines indicate the estimate, and the dashed lines indicate jack-knife bias-corrected estimates. The variance determined by the jack-knife method is negligible by comparison with symbol size. The bias is very large for 40-sweep annealing, indicating that we have insufficient samples to properly determine the KL-divergence. Elsewhere, we judge the bias not to significantly impact our conclusions.

perspective, we have *global warming*: the more effective the estimator is in capturing global distribution features, the higher the temperature that is typically indicated. Ergodicity breaking qualitatively explains the origin of this phenomenon in annealers, both the DW2X and STA. We have provided some practical guidelines in Section 1.1.

Local distributional features, which are well characterized by a temperature, are easy to estimate by self-consistent methods. We showed in the main text standard methods for temperature estimation from samples drawn from a single distribution. Self-consistent approaches are efficient and approximate the target distribution from the same samples by which the heuristic is evaluated. We presented a simple form for this, but the principle generalizes. However, the main problem with such methods is that they may indicate a good fit on the basis of incomplete (or biased) information about the target distribution. If a heuristic sample set fails to see a representative set of modes in the distribution, then the evaluation will inevitably skewed by the missing information.

The local approximation method is able to capture an important difference between quantum and simulated annealers related to the difference between dynamics of the DW2X and STA. This is realized in the non-linear dependence of the local temperature estimate to variation of the DW2X rescaling parameter. Quantum simulations on single qubits provide a qualitatively accurate explanation for this phenomenon, and are described in supplementary material (Raymond et al., 2016).

Describing the global distribution in terms of temperature(s) is more tricky; we proposed KL-divergence and MSE as measures of deviation from the Boltzmann distribution, and based on these objectives developed estimators for the best temperature. Each of these objectives is affected in slightly different ways by deviations locally and globally from the Boltzmann distribution. The maximum likelihood estimator is more strongly affected by the local distribution than the minimum MSE estimator, and indicates a larger estimate of inverse temperature. To remove the local distribution effects, we have proposed to take the initial distribution and apply local post-processing in order to isolate

the macroscopic distribution effects that are truly a limitation on practical performance. Applying some degree of post-processing may also be valuable in practice, in particular for the DW2X since the post-processing is strictly classical and complementary to the quantum dynamics.

We emphasize that because efficient post-processing allows significant manipulation of the local temperature, we consider this temperature not particularly important in practical applications. If we post-process, the post-processing temperature itself will be synonymous with the local temperature; the local temperature need not be measured.

Ergodicity breaking that relates to symmetry breaking, and ergodicity breaking without symmetry breaking should be distinguished. The problem classes we study in experimental sections are random, but there is a global sign symmetry, $P(x) = P(-x)$, other problems may exhibit different (or no) symmetries. Our annealer implementations (initial conditions, and dynamics) are chosen such that the heuristic distribution (P_A) also exhibits the same symmetry. If a symmetry is known, a well implemented sampler should be designed not to break such a symmetry. Symmetry implies that a mode at x' will imply a mode at $-x'$, and it may be that ergodicity breaking occurs between the two halves of the solution space (and be characterized by some temperature). However, this ergodicity breaking does not lead to a departure from the Boltzmann distribution, and so is of less practical interest than the ergodicity breaking that relates to random problem structure. The objectives we study are for this reason insensitive to ergodicity breaking (relying on energy and correlation statistics), and capture the temperature related only to the non-trivial ergodicity breaking.

Important ideas incidental to the main thread are discussed in supplementary materials (Raymond et al., 2016). These include the following: a description of single qubit experiments that explain the non-linearity of MLPL estimates in the DW2X; consideration of the effect of embedding on the distribution of samples; the development of an effective estimator for the Kullback–Leibler divergence; a consideration of spin-reversal transformations to

mitigate sampling error in the DW2X; and an experiment to test how the choice of kernel [equation (7)] affects the locally self-consistent temperature estimates.

At various points in this paper, we have included results both for the DW2X and STA. We have motivated a default parameterization for each algorithm in Section 4.4, and it should be clear that these choices are not aimed at, or appropriate for, a competitive comparison. It would be a complicated task to make a fair comparison, and it would also detract from the main theme of this paper since it would distort or disguise the phenomena we wish to highlight. We chose a default DW2X parameterization suitable for optimization, and two STA parameterizations that allow for qualitative comparison. It has been shown in experiments that both annealers can be improved with simple parameter modifications. It is interesting that, despite the fact that the DW2X annealer has been designed and tuned for optimization, it produces good statistics at intermediate temperature ranges, and that the STA with long anneal time shows a qualitatively similar behavior.

Given fixed time resources, optimal annealer operation is to a large extent determined by the schedule: the allocation of time resources over the anneal path (or modification of the anneal path itself, for multi-parameter paths as in the DW2X). The basic principle of schedule optimization is to allocate resources where dynamics are most effective (i.e., before, and close to, detrimental points of ergodicity breaking, allocating far fewer resources beyond that point) (Neal, 1993; Ghate and Smith, 2008). The STA schedule we implement is linear in temperature – this was simple to explain, and found to significantly outperformed a geometric schedule, over the temperature range implemented. The DW2X schedule by contrast could not be manipulated at this level of detail, the basic form of the schedule is fixed by engineering considerations.

When considering the value of annealers in inference problems, it is also important not to forget a variety of other powerful inference methods that may achieve a similar objective. In particular, simple variations on the STA, such as annealed (or population) importance sampling methods and other multi-canonical MCMC methods can often be tailored to the graphical structure of the problem under investigation (Neal, 2001; Hukushima and Iba, 2003; Hamze and de Freitas, 2004; Landau and Binder, 2005; Selby, 2014; Zhu et al., 2015).

From the perspective of both errors on correlations and KL-divergence, the balance of evidence certainly indicates that there is potentially a sweet spot for application of either the DW2X or STA to sampling. This sweet spot may be problem type dependent, but can be tuned to a degree, by modification of the annealing parameters, and more importantly, by post-processing. However, evaluation of this sweet spot is difficult to do self-consistently, and someone interested in applications may have to undertake hard work to discover (and have confidence in) annealer performance. Having available curves, such as those in Section 4.5, probably for some weakly post-processed distribution, would allow parameters of the annealer to be set optimally. It may seem computationally intensive (defeating the value of the heuristic) to evaluate the macroscopic distribution before using an annealer, but it is reasonable to assume that for some classes of problems at large scale, the local and global temperature properties will be common

across the class. In other time-dependent applications of annealers, the statistics of the distributions being learned change slowly, so that only periodic evaluations of the temperatures may be required.

Temperature plays an important role in describing annealers, even when they are implemented as optimizers. Temperature concepts might extend to some other heuristic optimization algorithms. In various optimization applications DW2X has been compared against walkSAT, HFS, and other heuristic solvers (Selby, 2014; Douglass et al., 2015; King et al., 2015). walkSAT and HFS optimizers are simple Markov Chain methods proposed to solve, respectively, the SAT, and chimera structured, optimization problems. They always return either global or local optima and in this sense the local temperature can be considered infinite (since, with respect to the sample neighborhood, it seems only ground states are returned). However, macroscopic dynamics are fundamentally similar to an annealer, with the same failure mode. In application to a multi-modal energy landscape, the process mixes across the space being weakly informed by the energy, before being trapped by modes as it falls below some energy threshold. In this sense, there is a global temperature with the same interpretation as proposed in this paper.

An important potential application of quantum annealers is in machine learning (Benedetti et al., 2016; Amin et al., 2016; Rolfe, 2016), where other heuristic samplers (not annealers) are prevalent. A common heuristic used in machine learning is called contrastive divergence (CD) (Hinton, 2002; Carreira-Perpiñán and Hinton, 2005). B_β is approximated in CD by taking the ground truths (a set of training examples) and evolving them by an MCMC procedure. This is an example of the post-processing scheme used and recommended in this paper, except that annealed samples are replaced by the ground truths. Like the annealing distribution, the distribution of training examples may be incorrect in both its local and global features, the effect of the MCMC procedure is to tidy up local distribution deviations. After post-processing, the distribution is used directly – in effect the post-processing temperature is taken to be correct,[5] without consideration of potential macroscopic deviations. The success of this algorithm in practice indicates that learning procedures may be quite tolerant of macroscopic deviations from the Boltzmann distribution in application provided the local temperature is correct. This would be good news since it may be computationally expensive to quantify macroscopic deviations, but it is easy to measure and manipulate local temperature in annealers.

One feature of D-Wave quantum annealers that might lead us to consider a different approach is the quantum part, as already discussed. The single qubit freeze-out, and dynamical slow-down at larger scales, is described by quantum models. The quantum Boltzmann distribution may be a better fit to sample sets drawn from the DW2X, and perhaps in "post-processing" we should think of the quantum space as the target, rather than the classical one. This is certainly a promising direction for future work.

---

[5]In machine learning, we can take the post-processing temperature to be $\beta = 1$, without loss of generality.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fict.2016.00023.

REFERENCES

Albert, J., and Swendsen, R. H. (2014). The inverse ising problem. *Phys. Procedia* 57, 99–103. doi:10.1016/j.phpro.2014.08.140

Amin, M. H. (2015). *Searching for Quantum Speedup in Quasistatic Quantum Annealers*. ArXiv:1503.04216.

Amin, M. H., Andriyash, E., Rolfe, J., Kulchytskyy, B., and Melko, R. (2016). *Quantum Boltzmann Machine*. ArXiv:1601.02036.

Aurell, E., and Ekeberg, M. (2012). Inverse ising inference using all the data. *Phys. Rev. Lett.* 108, 090201. doi:10.1103/PhysRevLett.108.090201

Bart, K. (1988). Bidirectional associative memories. *IEEE Trans. Syst. Man Cybern.* 18, 49–60. doi:10.1109/21.87054

Benedetti, M., Realpe-Gómez, J., Biswas, R., and Perdomo-Ortiz, A. (2016). Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applications in deep learning. *Phys. Rev. A.* 94:022308. doi:10.1103/PhysRevA.94.022308

Besag, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc. D Stat.* 24, 179–195.

Bhattacharya, B. B., and Mukherjee, S. (2015). *Inference in Ising Models*. Arxiv:1507.07055.

Bian, Z., Chudak, F., Macready, W. G., and Rose, G. (2010). *The Ising Model: Teaching An Old Problem New Tricks*. D-Wave Publications. Available at: http://www.dwavesys.com/

Boixo, S., Albash, T., Spedalieri, F. M., Chancellor, N., and Lidar, D. A. (2013). Experimental signature of programmable quantum annealing. *Nat. Commun.* 4, 2067. doi:10.1038/ncomms3067

Bresler, G., Gamarnik, D., and Shah, D. (2014). "Hardness of parameter estimation in graphical models," in *Advances in Neural Information Processing Systems 27*, eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc.), 1062–1070.

Bunyk, P., Hoskinson, E. M., Johnson, M. W., Tolkacheva, E., Altomare, F., Berkley, A. J., et al. (2014). Architectural considerations in the design of a superconducting quantum annealing processor. *IEEE Trans. Appl. Supercond.* 24, 1700110. doi:10.1109/TASC.2014.2318294

Carreira-Perpiñán, M. A., and Hinton, G. E. (2005). "On contrastive divergence learning," in *AISTATS05* eds R. G. Cowell, and Z. Ghahramani (NJ: Society for Artificial Intelligence and Statistics), 33–40. Available at: http://www.gatsby.ucl.ac.uk/aistats/

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artif. Intell.* 42, 393–405. doi:10.1016/0004-3702(90)90060-D

Denchev, V. S., Boixo, S., Isakov, S. V., Ding, N., Babbush, R., Smelyanskiy, V., et al. (2016). What is the computational value of finite-range tunneling? *Phys. Rev. X.* 6:031015. doi:10.1103/PhysRevX.6.031015

Denil, M., and de Freitas, N. (2011). "Toward the implementation of a quantum RBM," in *NIPS 2011 Deep Learning and Unsupervised Feature Learning Workshop*, Vol. 5 (Cambridge, MA: MIT Press).

Douglass, A., King, A. D., and Raymond, J. (2015). "Chapter constructing SAT filters with a quantum annealer," in *Theory and Applications of Satisfiability Testing – SAT 2015: Proceedings of the 18th International Conference, September 24-27, 2015* (Austin, TX: Springer International Publishing, Cham), 104–120.

Dumoulin, V., Goodfellow, I., Courville, A., and Bengio, Y. (2015). "On the challenges of physical implementations of RBMs," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence.* Arxiv:1312.5258v2.

Efron, B. (1982). "The Jackknife, the bootstrap and other resampling plans," in *CBMS-NSF Regional Conference Series in Applied Mathematics* (Philadelphia, PA: SIAM), 5–11. Lectures given at Bowling Green State Univ., June 1980.

Geyer, C. J., and Thompson, E. A. (1992). Constrained monte carlo maximum likelihood for dependent data. *J. R. Stat. Soc. Series B Stat. Methodol.* 54, 657–699.

Ghate, A., and Smith, R. L. (2008). A dynamic programming approach to efficient sampling from Boltzmann distributions. *Oper. Res. Lett.* 36, 665–668. doi:10.1016/j.orl.2008.07.009

Grassberger, P. (2003). *Entropy Estimates from Insufficient Samplings.* ArXiv Physics e-prints.

Hamze, F., and de Freitas, N. (2004). "From fields to trees," *Proceedings of the Twentieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)* (Arlington, VA: AUAI Press), 243–250.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14, 1771–1800. doi:10.1162/089976602760128018

Hukushima, K., and Iba, Y. (2003). Population annealing and its application to a spin glass. *AIP Conf. Proc.* 690, 200–206. doi:10.1063/1.1632130

Hukushima, K., and Nemoto, K. (1996). Exchange monte carlo method and application to spin glass simulations. *J. Phys. Soc. Japan* 65, 1604–1608. doi:10.1143/JPSJ.65.1604

Isakov, S. V., Zintchenko, I. N., Rønnow, T. F., and Troyer, M. (2015). Optimised simulated annealing for ising spin glasses. *Comput. Phys. Commun.* 192, 265–271. doi:10.1016/j.cpc.2015.02.015

Johnson, M. W., Amin, M. H. S., Gildert, S., Lanting, T., Hamze, F., Dickson, N., et al. (2011). Quantum annealing with manufactured spins. *Nature* 473, 194–198. doi:10.1038/nature10012

Kadowaki, T., and Nishimori, H. (1998). Quantum annealing in the transverse ising model. *Phys. Rev. E* 58, 5355–5363. doi:10.1103/PhysRevE.58.5355

Katzgraber, H. G., Hamze, F., and Andrist, R. S. (2014). Glassy chimeras could be blind to quantum speedup: designing better benchmarks for quantum annealing machines. *Phys. Rev. X* 4, 021008. doi:10.1103/PhysRevX.4.021008

King, A. D., Hoskinson, E., Lanting, T., Andriyash, E., and Amin, M. H. (2016). Degeneracy, degree, and heavy tails in quantum annealing. *Phys. Rev. A.* 93: 052320. doi:10.1103/PhysRevA.93.052320

King, J., Yarkoni, S., Nevisi, M. M., Hilton, J. P., and McGeoch, C. C. (2015). *Benchmarking a Quantum Annealing Processor with the Time-to-Target Metric.* Arxiv:1508.05087.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680. doi:10.1126/science.220.4598.671

Landau, D. P., and Binder, K. (2005). *A Guide to Monte Carlo Simulations in Statistical Physics*, 2nd Edn. Cambridge, UK: Cambridge University Press.

Lehmann, E. L., and Casella, G. (1998). *Theory of Point Estimation.* Springer Texts in Statistics, 2nd Edn, New York, NY: Springer.

Long, P. M., and Servedio, R. (2010). "Restricted Boltzmann machines are hard to approximately evaluate or simulate," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Madison, WI: Omnipress), 703–710.

Mezard, M., and Montanari, A. (2009). *Information, Physics, and Computation.* New York, NY, USA: Oxford University Press, Inc.

Montanari, A. (2015). Computational implications of reducing data to sufficient statistics. *Electron. J. Stat.* 9, 2370–2390. doi:10.1214/15-EJS1059

Neal, R. M. (1993). *Probabilistic Inference Using Markov Chain Monte Carlo Methods.* Toronto, Canada: Tech. rep., Dept. of Computer Science, University of Toronto. Technical Report CRG-TR-93-1.

Neal, R. M. (2001). Annealed importance sampling. *Stat. Comput.* 11, 125–139. doi:10.1023/A:1008923215028

Nguyen, H. C., and Berg, J. (2012). Mean-field theory for the inverse ising problem at low temperatures. *Phys. Rev. Lett.* 109, 050602. doi:10.1103/PhysRevLett.109.050602

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Comput.* 15, 1191–1253. doi:10.1162/089976603321780272

Perdomo-Ortiz, A., O'Gorman, B., Fluegemann, J., Biswas, R., and Smelyanskiy, V. N. (2016). Determination and correction of persistent biases in quantum annealers. *Sci. Rep.* 6, 18628. doi:10.1038/srep18628

Raymond, J., Yarkoni, S., and Andriyash, E. (2016). *Global Warming: Temperature Estimation in Annealers (Supplementary Materials).* ArXiv:1606.00919v4.

Rolfe, J. (2016). *Discrete Variational Autoencoders.* ArXiv:1609:02200.

Rønnow, T. F., Wang, Z., Job, J., Boixo, S., Isakov, S. V., Wecker, D., et al. (2014). Defining and detecting quantum speedup. *Science* 345, 420–424. doi:10.1126/science.1252319

Salakhutdinov, R., and Hinton, G. (2009). "Deep Boltzmann machines," in *JMLR Proceedings of the International Conference on Artificial Intelligence and Statistics,* Vol. 5 (Cambridge, MA: MIT Press), 448–455.

Selby, A. (2014). *Efficient Subgraph-Based Sampling of Ising-Type Models with Frustration.* Arxiv:1409.3934.

Shirts, M. R., and Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* 129, 124105. doi:10.1063/1.2978177

Sinclair, A., and Jerrum, M. (1989). Approximate counting, uniform generation and rapidly mixing markov chains. *Inform. Comput.* 82, 93–133. doi:10.1016/0890-5401(89)90067-9

Venturelli, D., Mandrà, S., Knysh, S., O'Gorman, B., Biswas, R., and Smelyanskiy, V. (2015). Quantum optimization of fully connected spin glasses. *Phys. Rev. X* 5, 031040. doi:10.1103/PhysRevX.5.031040

Wainwright, M. J., and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* 1, 1–305. doi:10.1561/2200000001

Zhu, Z., Ochoa, A. J., and Katzgraber, H. (2015). Efficient cluster algorithm for spin glasses in any space dimension. *Phys. Rev. Lett.* 115, 077201. doi:10.1103/PhysRevLett.115.077201