Adaptive Parallel Tempering

Masoud Mohseni[1, *]

[1] *Large Scale Integrated Photonics (LSIP), Hewlett Packard Labs, Santa Barbara, CA, USA*

(Dated: July 31, 2023)

In this note, we provide a simple but general approach for automatically finding optimal hyper-parameters for an important class of probabilistic solvers known as replica-exchange Monte Carlo or Parallel Tempering (PT).

Sampling low energy states of complex systems over discrete spaces are generally very hard. This task has diverse application in so-called ”forward combinatorial problems” such as solving constrained optimization, finding low-energy states of spin glasses, approximating partition functions, and probabilistic inference in graphical models. This problem is also at the core of many ”inverse combinatorial problem” where samples of low energy states are available and the goal is to find the underlying model, or problem, that can generate such states. These inverse problems are equivalent to the task of training certain classes of generative machine learning architectures, including structured probabilistic models, energy-based models, Bayesian networks, and binary neural networks. The computational difficulty in such problems originates from the existence of many local minima and/or high entropic barriers.

Conventional Monte Carlo sampling methods with local updates could get trapped within local minima and suffer from very long equilibration times. Cluster methods [1, 2] do not work for such models because of the existence of frustration at different length scales. Parallel tempering [3–5] is a generalization of the conventional Monte Carlo method. The intuition behind the PT algorithm is that for certain hard problems we should be able to elevate the temperatures to avoid getting stuck in suboptimal metastable states indefinitely. This is in contrast to single-replica Simulated Annealing (SA) where temperature is gradually reduced in a sequential fashion and there is no way of escaping a local minima once the temperature is already reduced below the barrier height.

The key idea of PT is to construct many identical replicas of a given graph or problem and operate each replica at a different (algorithmic) temperature. Occasionally the replicas compare their states with neighbouring replicas and if a higher temperature replica carries a lower energy state, this state is swapped with the lower temperature. Just as in simulated annealing, even an energetically unfavorable swap may be accepted by a rule that preserves the detailed-balance condition. Operationally all the replicas that live at different temperatures can be simulated in parallel. After a fixed number of conventional Monte Carlo sweeps, all the replica swaps can be performed simultaneously. This procedure is repeated many times until we have arrive at a steady state. Parallel tempering has generally much faster convergence to equilibrium than conventional Monte Carlo. The acceptance probability to swap two replicas at adjacent temperatures T_i and T_j satisfies detailed balance and can be written as:

$$p_{ij} = \min\{1, \exp[(\beta_i - \beta_j)(E(\beta_i) - E(\beta_j))]\}, \qquad (1)$$

where $\beta_i = 1/T_i$ is the inverse temperature and $E(\beta_i)$ is the configuration energy.

How can we develop an efficient way of implementing PT? Just as the β profile in time for SA need to be optimized, the β profile in space for PT needs to be carefully adjusted. For sufficiently complex problems, simple predetermined choices of β_i over replicas, such as constant or geometric temperature profiles, may cause inefficient swap attempts. This can create bottlenecks for information flow across those replicas that reside near a critical temperature (e.g., those near a spin-glass phase transition). Increasing the density of replica uniformly across entire temperature profile might in principle resolve such bottlenecks but will lead to significant computational overhead correspond to redundant replicas that are placed far from the critical temperature.

Adaptive Parallel Tempering (APT) is a powerful adaptive strategy that allows for a simple hyper-parameter optimization which finds (i) the optimal number of replicas (r), (ii) the optimal β_i profile (temperature spacing between replics), and (iii) minimum replica temperature all automatically in an instance-wise fashion. The main observation is that the inverse temperature spacings $\beta_i - \beta_j$ should be chosen in such a way that the swap probabilities are not too high and not too low (empirically a value from the range $(0.2, 0.3)$ is good enough). An over-simplified choice of a geometrical schedule ($\beta_i = r\beta_{i-1}$ with fixed r) is not necessarily efficient in this respect. A temperature schedule that maintains the fixed swap probability independent of the replica's temperature is much more efficient [6]. A number of methods have been developed to construct such a schedule adaptively [7–10]. There are also methods that maintain denser temperature spacings in the vicinity of simulation bottlenecks such as phase transitions [11].

Here, we present a simple version of our adaptive algorithm which maintains a *constant* acceptance probability, p_a, across each neighbouring pairs of replica independent of replicas' location in the temperature profile [12]. The basic intuition behind this algorithm is that the inverse

[*] mohseni@google.com

temperature spacings should be small in the regions of high energy fluctuations, i.e. in the regions with large specific heat:

We can measure the energy variance $\sigma(\beta_i)^2$ for each of replicas at temperature $T_i = 1/\beta_i$ from

$$\sigma(\beta_i)^2 = \left\langle E(\beta_i)^2 \right\rangle - \left\langle E(\beta_i) \right\rangle^2 \tag{2}$$

We note that the specific heat, c_i, for each replica i can be related to the energy variance as:

$$c_i = \frac{dE(\beta_i)}{dT_i} = \frac{1}{T_i^2}\left(\left\langle E(\beta_i)^2 \right\rangle - \left\langle E(\beta_i) \right\rangle^2\right) = \beta_i^2 \sigma(\beta_i)^2 \tag{3}$$

The energy difference over $d\beta_i$ can then be written as:

$$\frac{dE(\beta_i)}{d\beta_i} = -\frac{1}{\beta_i^2}\frac{dE(\beta_i)}{dT_i} = -\frac{1}{\beta_i^2} c_i = -\sigma(\beta_i)^2 \tag{4}$$

For sufficiently small $\Delta\beta_i$, we can write:

$$E(\beta_i + \Delta\beta_i) = E(\beta_i) + \frac{dE(\beta_i)}{d\beta_i}\Delta\beta_i = E(\beta_i) - \sigma(\beta_i)^2 \Delta\beta_i \tag{5}$$

Thus, the acceptance probability for each swap operation between two replicas can be obtained:

$$\begin{aligned} p_a &= \min\left\{1, \exp\left[\Delta\beta_i\left(E\left(\beta_i + \Delta\beta_i\right) - E\left(\beta_i\right)\right)\right]\right\} \\ &= \min\left\{1, \exp\left[-(\sigma(\beta_i)\Delta\beta_i)^2\right]\right\} \end{aligned} \tag{6}$$

Now, by choosing $\Delta\beta_i = \frac{\alpha}{\sigma(\beta_i)}$, where α is a constant hyperparameter, we arrive at:

$$p_a = \min\left\{1, \exp\left[-\alpha^2\right]\right\}, \tag{7}$$

which is independent of β_i.

Thus, for each instance of our problem, we can construct the adaptive temperature profile as a simple prepprocessing procedure:

Start with input parameters: maximum replica temperature T_0 (minimum inverse temperature β_0) and a fixed hyperparameter α which corresponds to a desired acceptance rate according to Eq. 7. We then perform a number of Monte-Carlo sweeps at a fixed inverse temperature β_i and calculate the next inverse temperatures β_{i+1} from

$$\beta_{i+1} = \beta_i + \frac{\alpha}{\sigma(\beta_i)}. \tag{8}$$

All other β_i values are subsequently determined iteratively until $\sigma(\beta_{\text{final}}) \leq \sigma_{min}$ for a specified σ_{min} at which point the temperature is low enough for this problem instance and we do not need any more replicas at lower temperatures and the preprocessing step is finished. This algorithm maintains the replica swap probability p_a that is more or less independent of temperature and are related to the parameter α via $p_a \approx e^{-\alpha^2}$. In this approach we keep $\Delta\beta_i$ small in the regions where energy variance $\sigma(\beta_i)^2$ is high and vice-versa. This resolves the issue of low acceptance rates near a spin-glass phase transition.

Decent values for α and σ_{min} are rather easy to find. We can setup ansatz values for σ_{min} by noting that the energy fluctuations shouldn't be smaller than the minimum energy gradient that can possibly induced after flipping variables that are connected by the minimum values of J_{ij}. Moreover, the appropriate range of α do not vary much among various problem classes. For example $\alpha = 1.1$ is usually good enough for most problems and optimal values are rarely outside $[0.85, 1.25]$.

[1] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **58**, 86 (1987).

[2] U. Wolff, Phys. Rev. Lett. **62**, 361 (1989).

[3] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).

[4] C. J. Geyer, in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (American Statistical Association, 1991) p. 156–163.

[5] K. Hukushima and K. Nemoto, Journal of the Physical Society of Japan **65**, 1604 (1996), https://doi.org/10.1143/JPSJ.65.1604.

[6] A. Kone and D. A. Kofke, The Journal of Chemical Physics **122**, 206101 (2005), https://doi.org/10.1063/1.1917749.

[7] D. A. Kofke, The Journal of Chemical Physics **117**, 6911 (2002), https://doi.org/10.1063/1.1507776.

[8] N. Rathore, M. Chopra, and J. J. de Pablo, The Journal of Chemical Physics **122**, 024111 (2005), https://doi.org/10.1063/1.1831273.

[9] C. Predescu, M. Predescu, and C. V. Ciobanu, The Journal of Chemical Physics **120**, 4119 (2004), https://doi.org/10.1063/1.1644093.

[10] C. Predescu, M. Predescu, and C. V. Ciobanu, The Journal of Physical Chemistry B **109**, 4189 (2005), pMID: 16851481, https://doi.org/10.1021/jp045073+.

[11] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, J. Stat. Mech. **2006**, P03018 (2006).

[12] M. Mohseni, D. Eppens, J. Strumpfer, R. Marino, V. Denchev, A. K. Ho, S. V. Isakov, S. Boixo, F. Ricci-Tersenghi, and H. Neven, arXiv:2111.13628 (2021).