

---

# Supplementary: LSB: Local Self-Balancing MCMC in Discrete Spaces

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Detailed Balance

We want to prove that  $p(\mathbf{x})T(\mathbf{x}'|\mathbf{x}) = p(\mathbf{x}')T(\mathbf{x}|\mathbf{x}')$  for all  $\mathbf{x}' \neq \mathbf{x}$ . We have that

$$\begin{aligned} p(\mathbf{x})A(\mathbf{x}', \mathbf{x}) \frac{g\left(\frac{\tilde{p}(\mathbf{x}')}{\tilde{p}(\mathbf{x})}\right)1[\mathbf{x}' \in N(\mathbf{x})]}{Z(\mathbf{x})} &= \\ p(\mathbf{x}')A(\mathbf{x}, \mathbf{x}') \frac{g\left(\frac{\tilde{p}(\mathbf{x})}{\tilde{p}(\mathbf{x}')} \right)1[\mathbf{x} \in N(\mathbf{x}')] }{Z(\mathbf{x}')} \end{aligned}$$

By observing that  $1(\mathbf{x}' \in N(\mathbf{x})) = 1(\mathbf{x} \in N(\mathbf{x}'))$  and using the balancing property, we can simplify previous equality to obtain the following relation:

$$\begin{aligned} p(\mathbf{x})A(\mathbf{x}', \mathbf{x}) \frac{g\left(\frac{\tilde{p}(\mathbf{x}')}{\tilde{p}(\mathbf{x})}\right)}{Z(\mathbf{x})} &= \\ p(\mathbf{x}')A(\mathbf{x}, \mathbf{x}') \frac{\frac{\tilde{p}(\mathbf{x})}{\tilde{p}(\mathbf{x}')} g\left(\frac{\tilde{p}(\mathbf{x}')}{\tilde{p}(\mathbf{x})}\right)}{Z(\mathbf{x}')} \end{aligned}$$

Therefore, we can apply standard algebra to simplify even more

$$\tilde{p}(\mathbf{x})A(\mathbf{x}', \mathbf{x}) = \tilde{p}(\mathbf{x}')A(\mathbf{x}, \mathbf{x}') \frac{Z(\mathbf{x})}{Z(\mathbf{x}')}$$

Finally, recall that for balancing functions  $A(\mathbf{x}, \mathbf{x}') = \min \left\{ 1, \frac{\tilde{p}(\mathbf{x})Q(\mathbf{x}'|\mathbf{x})}{\tilde{p}(\mathbf{x}')Q(\mathbf{x}|\mathbf{x}')} \right\} = \min \left\{ 1, \frac{Z(\mathbf{x}')}{Z(\mathbf{x})} \right\}$  and therefore previous equality becomes an identity, namely:

$$\tilde{p}(\mathbf{x})A(\mathbf{x}', \mathbf{x}) = \tilde{p}(\mathbf{x}')A(\mathbf{x}, \mathbf{x}')$$

thus proving detailed balance.

## 2 Ergodicity

Let's consider a Markov chain, namely

$$T(\mathbf{x}'|\mathbf{x}) = A(\mathbf{x}', \mathbf{x})Q(\mathbf{x}'|\mathbf{x}) + 1[\mathbf{x}' = \mathbf{x}] \sum_{\mathbf{x}'' \in \mathcal{X}} (1 - A(\mathbf{x}'', \mathbf{x}))Q(\mathbf{x}''|\mathbf{x}) \quad (1)$$

with a proposal of the following form:

$$Q(\mathbf{x}'|\mathbf{x}) = \frac{g\left(\frac{\tilde{p}(\mathbf{x}')}{\tilde{p}(\mathbf{x})}\right)1[\mathbf{x}' \in N(\mathbf{x})]}{Z(\mathbf{x})} \quad (2)$$

12 We can prove the ergodicity of the Markov chain for the case where the fixed-point distribution  
 13  $p(\mathbf{x}) > 0$  for every  $\mathbf{x} \in \mathcal{X}$  and then extend it to a general distribution  $p$ .

14 Now, assume that  $p(\mathbf{x}) > 0$  for any point  $\mathbf{x} \in \mathcal{X}$ ,  $g(t) > 0$  for any  $t > 0$  and  $\mathcal{X}$  is a  $d$ -dimensional  
 15 discrete space. Then, the Markov chain in Eq. 1 with proposal defined according to Eq. 2 can reach  
 16 any state  $\mathbf{x}'$  from any state  $\mathbf{x}$  in  $d$  steps with non-zero probability. More formally, we can construct a  
 17 new Markov chain by applying  $d$  times the original one and identify its transition probability with  
 18  $T^d(\mathbf{x}'|\mathbf{x})$ . We can easily check, thanks to our assumptions, that  $T^d(\mathbf{x}'|\mathbf{x}) > 0$  for any  $\mathbf{x}, \mathbf{x}'$ . In  
 19 other words, the original Markov chain is regular. This is sufficient to satisfy the assumptions of the  
 20 fundamental theorem of homogeneous Markov chains [1], thus proving ergodicity.

21 We can extend the previous result to any arbitrary  $p$  (namely considering cases where  $p(\mathbf{x}) = 0$  for  
 22 some  $\mathbf{x} \in \mathcal{X}$ ). This can be achieved by modifying our assumptions on  $g$ , namely considering that  
 23  $g(t) > 0$  for any  $t \geq 0$  and reusing the same proof strategy.

### 24 3 Monotonicity

25 We consider parametrization LSB 2, namely  $g_\theta(t) = \min\{h_\theta(t), th_\theta(1/t)\}$ , and we introduce a  
 26 regularizer for the learning objective defined in Section 3, to penalize violations to the condition  
 27  $h_\theta(1/t) - \frac{1}{t} \frac{dh_\theta(1/t)}{dt} \geq 0$ , or equivalently  $\frac{dth_\theta(1/t)}{dt} \geq 0$ :

$$\mathcal{L}_{reg}(\theta, t) = \max \left\{ \frac{th_\theta(1/t) - (t + \epsilon)h_\theta(1/(t + \epsilon))}{\epsilon}, 0 \right\}$$

28 where the left argument in the max operator corresponds to the finite difference approximation of  
 29  $-\frac{dth_\theta(1/t)}{dt}$ . The final objective used in the experiments (for LSB 2) is defined in the following way:

$$J(\theta) = E_{Q_{init}}\{\mathcal{J}(\theta, \mathbf{x})\} + E_{Q_{\theta_0}}\{\mathcal{J}(\theta, \mathbf{x})\} + E_{U(0,2)}\{\mathcal{L}_{reg}(\theta, t)\} \quad (3)$$

### 30 4 Derivation of the Objective Function

$$\begin{aligned} KL\{\tilde{T}_\theta(\mathbf{x}'|\mathbf{x})||p(\mathbf{x}')\} &= \sum_{\mathbf{x}' \in \mathcal{X}} \tilde{T}_\theta(\mathbf{x}'|\mathbf{x}) \log \frac{\tilde{T}_\theta(\mathbf{x}'|\mathbf{x})}{p(\mathbf{x}')} \\ &= \sum_{\mathbf{x}' \neq \mathbf{x}} \frac{T_\theta(\mathbf{x}'|\mathbf{x})}{Z_T} \log \frac{T_\theta(\mathbf{x}'|\mathbf{x})}{p(\mathbf{x}')Z_T} \\ &= \sum_{\mathbf{x}' \neq \mathbf{x}} \frac{T_\theta(\mathbf{x}'|\mathbf{x})}{Z_T} \log \frac{T_\theta(\mathbf{x}'|\mathbf{x})Z_p}{\tilde{p}(\mathbf{x}')Z_T} \\ &= \frac{1}{Z_T} \sum_{\mathbf{x}' \neq \mathbf{x}} A_\theta(\mathbf{x}', \mathbf{x}) Q_\theta(\mathbf{x}'|\mathbf{x}) \log \frac{A_\theta(\mathbf{x}', \mathbf{x}) Q_\theta(\mathbf{x}'|\mathbf{x}) Z_p}{\tilde{p}(\mathbf{x}') Z_T} \\ &= \frac{1}{Z_T} \sum_{\mathbf{x}' \neq \mathbf{x}} Q_\theta(\mathbf{x}'|\mathbf{x}) A_\theta(\mathbf{x}', \mathbf{x}) \log \frac{A_\theta(\mathbf{x}', \mathbf{x}) Q_\theta(\mathbf{x}'|\mathbf{x}) Z_p}{\tilde{p}(\mathbf{x}') Z_T} \\ &= \frac{1}{Z_T} \sum_{\mathbf{x}' \in \mathcal{X}} Q_\theta(\mathbf{x}'|\mathbf{x}) A_\theta(\mathbf{x}', \mathbf{x}) \log \frac{A_\theta(\mathbf{x}', \mathbf{x}) Q_\theta(\mathbf{x}'|\mathbf{x}) Z_p}{\tilde{p}(\mathbf{x}') Z_T} \\ &= \frac{1}{Z_T} E_{Q_\theta} \left\{ A_\theta(\mathbf{x}', \mathbf{x}) \log \frac{A_\theta(\mathbf{x}', \mathbf{x}) Q_\theta(\mathbf{x}'|\mathbf{x})}{\tilde{p}(\mathbf{x}')} \right\} - \frac{\log Z_T}{Z_T \log Z_p} \\ &\propto E_{Q_\theta} \left\{ A_\theta(\mathbf{x}', \mathbf{x}) \log \frac{A_\theta(\mathbf{x}', \mathbf{x}) Q_\theta(\mathbf{x}'|\mathbf{x})}{\tilde{p}(\mathbf{x}')} \right\} \\ &= E_{Q_{\theta_0}} \left\{ \omega_{\theta, \theta_0} A_\theta(\mathbf{x}', \mathbf{x}) \log \frac{A_\theta(\mathbf{x}', \mathbf{x}) Q_\theta(\mathbf{x}'|\mathbf{x})}{\tilde{p}(\mathbf{x}')} \right\} \doteq \mathcal{J}(\theta, \mathbf{x}) \end{aligned}$$

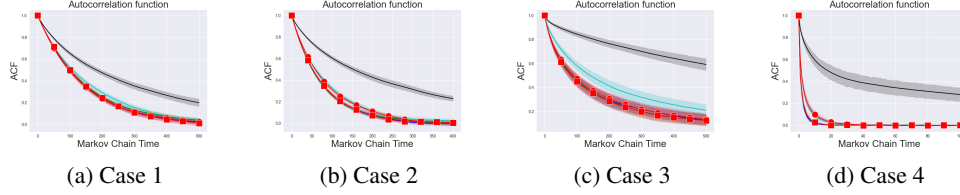


Figure 1: Samplers' performance on four cases of the Ising model ( $30 \times 30$ ) for the mixing phase. (a) Case 1: Independent-noisy, (b) case 2: Independent-clean, (c) case 3: Dependent-noisy, (d) case 4: Dependent-clean

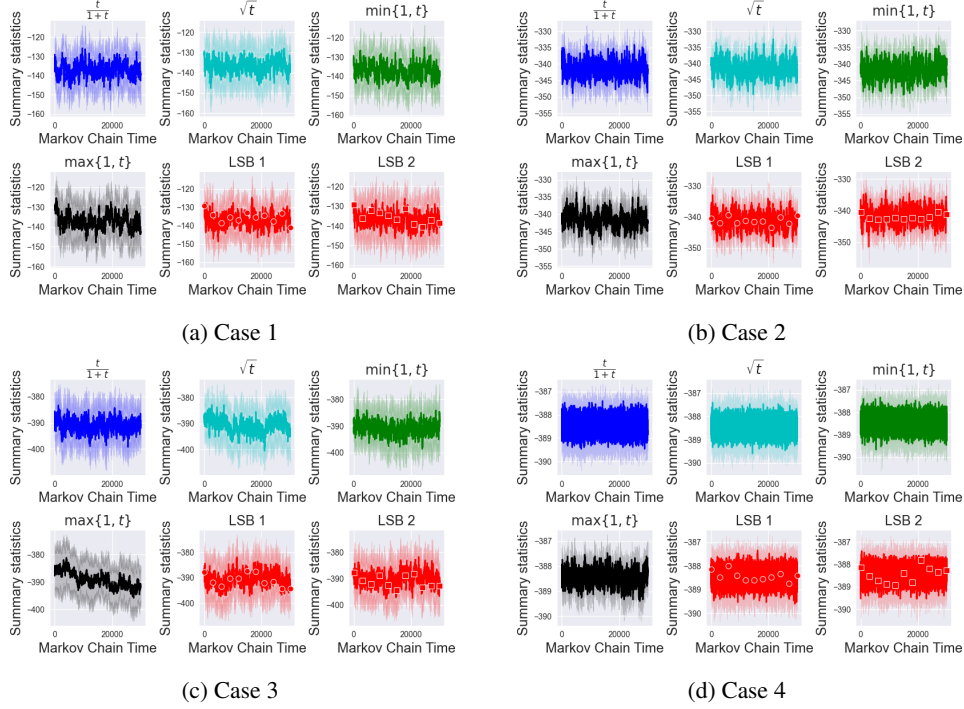


Figure 2: Traceplots on four cases of the Ising model ( $30 \times 30$ ) for the mixing phase. (a) Case 1: Independent-noisy, (b) case 2: Independent-clean, (c) case 3: Dependent-noisy, (d) case 4: Dependent-clean

## 5 Hyperparameters Used in the Experiments

- Learning rate  $\eta = 1e - 4$  for SGD optimizer with momentum.
- Burn-in iterations  $K = 2000$  (for Ising)  $K = 500$  (for UAI).
- Iterations for sampling 30000 (for Ising) 10000 (for UAI).
- Batch size  $N = 30$  (for Ising)  $N = 5$  (for UAI).
- Monotonic network 20 blocks of 20 neurons.

## 6 Further Results for Ising

See Figure 1 and Figure 2.

## 7 Further Results for UAI

See Figure 3.

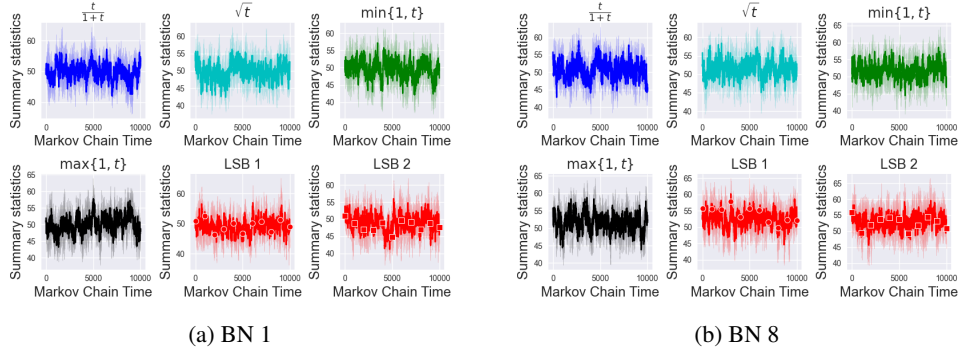


Figure 3: Traceplots on UAI benchmarks model (100 vars near-deterministic dependencies) for the mixing phase.

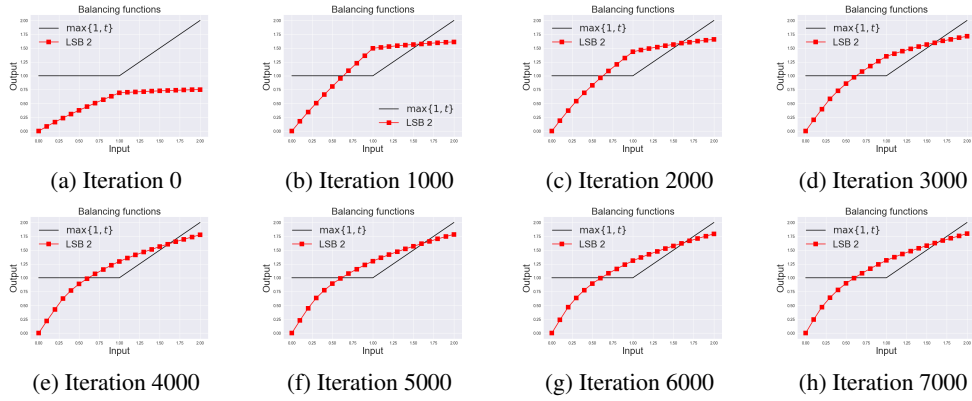


Figure 4: Training the monotonic network to match  $\max\{1, t\}$  balancing function.

## 8 Capacity of the Monotonic Network used in the Experiments

We analyze the capacity of the monotonic network used in our experiments, to see whether it can learn basic balancing functions. In particular, we train the monotonic network to match the  $\max\{1, t\}$  function using a L2 loss (we use SGD with learning rate  $1e - 2$ ). As we can see from Figure 4, the network has no sufficient capacity to well approximate the target function. That explains why in Figure 3(a) of the main manuscript, LSB 2 is performing slightly worse than  $\max\{1, t\}$ .

## References

- [1] R. Neal. *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. Department of Computer Science, University of Toronto Toronto, Ontario, Canada, 1993.