Convergences guarantees of GFlowNets

Anonymous Author(s)
Affiliation
Address
email

Abstract

Although they were introduced to approximate complex distributions defined up to normalization, Generative Flow Networks (GFlowNets) only provide strong guarantees once idealized conditions are matched. However, these conditions are never satisfied exactly in practice when they are trained using gradient-based methods. In this paper, we prove that minimizing the Trajectory Balance loss, a popular GFlowNet objective, does lead to an induced distribution getting closer to the target distribution of interest, confirming theoretically this long-standing intuition from the GFlowNet literature. We ultimately show that the KL divergence between both distributions is upper-bounded by the quantity being minimized, and we further verify this theoretical statement on a simple sampling task.

1 Introduction

Unlike more standard variational methods that minimize some divergence between an approximation and the target distribution of interest (typically, the reverse KL-divergence [Jordan et al., 1999]), *Generative Flow Networks* [GFlowNets; Bengio et al., 2021, 2023] instead minimize an objective quantifying "inconsistencies" with some conditions which, if they were to be satisfied, would ensure that we can sample from the target distribution. This provides strong guarantees in the idealized setting where those conditions are fully satisfied, but those are never met in practice since models are inevitably bound by their finite capacity and optimization. In this paper, we show that the divergence between the distribution induced by the GFlowNet and the target distribution can be theoretically bounded, confirming the intuition in this community that minimizing losses derived from the ideal conditions does yield better approximations. This sheds light on the behavior of GFlowNets in those inexact settings encountered in practice. We also validate these theoretical results empirically on some basic generative tasks.

2 Background

We first recall some basic notions of Generative Flow Networks (GFlowNets), and refer readers to [Bengio et al., 2023] for a more comprehensive treatment. GFlowNets were introduced as a framework for sampling from complex distributions over combinatorial objects [Bengio et al., 2021], where the sample space is often so large that computing the normalizing constant (partition function) is intractable. They cast the generation process as a sequential decision-making problem in a directed acyclic graph (DAG) of states $\overline{\mathcal{S}} = \mathcal{S} \cup \perp$, starting from a unique source s_0 and always terminating at a designated sink $\perp$. The objects of interest are the terminating states $x \in \mathcal{X} \subseteq \mathcal{S}$, defined as the states encountered immediately before reaching $\perp$.

A GFlowNet is parameterized by forward transition probabilities $P_F(s' \mid s)$, which specify how states are chosen along trajectories from s_0 to $\perp$. This induces a distribution over terminating states,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

called the *terminating state distribution*:

$$P_F^{\top}(x) \triangleq \sum_{\tau : s_0 \rightsquigarrow x} P_F(\tau) = \sum_{\tau : s_0 \rightsquigarrow x} \prod_{t=0}^{T_\tau} P_F(s_{t+1} \mid s_t), \tag{1}$$

where $s_0 \rightsquigarrow x$ denotes all the trajectories terminating at x, with $s_{T_\tau} = x$ and $s_{T_\tau+1} = \perp$. Moreover, we also define a positive reward function $R(x)$, quantifying a notion of preference over terminating states x. The objective of a GFlowNet is to find a forward transition probability such that $P_F^{\top}(x)$ matches the target distribution:

$$P^\star(x) = \frac{R(x)}{Z^\star}, \tag{2}$$

Among the various conditions introduced in the GFlowNet literature to characterize this P_F [Bengio et al., 2021, 2023, Madan et al., 2023, Zhang et al., 2023], the *Trajectory Balance* (TB) conditions [Malkin et al., 2022] stand out as being the most widely used in practice. It states that if there exists a positive scalar $Z > 0$ and a backward transition probability P_B (*i.e.*, $P_B(\cdot \mid s)$ is a distribution over the parents of s) such that for all trajectories $\tau = (s_0, s_1, \ldots, s_T, \perp)$,

$$Z \prod_{t=0}^{T} P_F(s_{t+1} \mid s_t) = R(s_T) \prod_{t=1}^{T} P_B(s_{t-1} \mid s_t), \tag{3}$$

then the terminating state distribution $P_F^{\top}(x) \propto R(x)$, and $Z \equiv Z^\star$. To learn this forward transition probability, the TB loss has been introduced as a way to quantify the "mismatch" in these conditions. It is a non-linear least-square objective of the form $\mathcal{L}_{\mathrm{TB}}(\phi, \psi) = \frac{1}{2}\mathbb{E}_{\pi_b}[\Delta^2_{\mathrm{TB}}(\tau; \phi)]$, where π_b is an arbitrary distribution over trajectories τ, and the residual is

$$\Delta_{\mathrm{TB}}(\tau; \phi) = \log \frac{Z_\phi \prod_{t=0}^{T} P_F^\phi(s_{t+1} \mid s_t)}{R(s_T) \prod_{t=1}^{T} P_B^\phi(s_{t-1} \mid s_t)}, \tag{4}$$

where Z_ϕ and P_B^ϕ are learned alongside the forward transition probability P_F^ϕ. It is clear that if this loss reaches its global minimum (leaving aside discussions of expressivity of P_F^ϕ), then the residuals are all zeros and (3) are satisfied, leading to the guarantees about ${P_F^{\phi^\star}}^{\top}(x)$.

Related to our work, Malkin et al. [2023] showed that the KL divergence between $P_F^{\phi\top}(x)$ and the target distribution can be bounded using the following data processing inequality [Zhang et al., 2019]

$$\mathrm{KL}\big(P_F^{\phi\top}(x) \,\|\, P^\star(x)\big) \leq \mathrm{KL}\big(P_F^\phi(\tau) \,\|\, P_B^\phi(\tau)\big), \tag{5}$$

with appropriately defined distributions over trajectories $P_F^\phi(\tau)$ and $P_B^\phi(\tau)$. However, the latter still remains intractable, since it depends on $P^\star(x)$ itself.

3 Convergence guarantees of GFlowNets

The approximation of (3) by minimizing the TB loss inevitably leads to errors, the source of which is typically two-fold: (1) due to the choice of the variational family to parametrize P_F^ϕ, which may not be expressive enough, and (2) due to the finite nature of optimization. In those cases, that are always encountered in practice, we have no guarantee as to what $P_F^{\phi\top}(x)$ could be, or how close to the target distribution $P^\star(x)$ it is. The objective of this section is to build towards establishing a connection between the mismatch quantified by (4) and the divergence between these two distributions. All the proofs of the propositions in this section are available in Appendix A.

3.1 Estimation of the partition function

One advantage of TB compared to other GFlowNets conditions is that it provides an approximation of the partition function with the learned scalar Z_ϕ, in addition to an approximate sampler for $P^\star$. However, it has been shown empirically that Z_ϕ alone may yield an unreliable estimation of the partition function [Malkin et al., 2022], which may be detrimental if used for model comparison. The following proposition offers a more precise relation between the true partition function $Z^\star$, its learned counterpart Z_ϕ, and the residuals (4).

Proposition 1 (Estimation of the partition function)**.** *The log-partition function* $\log Z^\star$ *of the target distribution (2) is related to* $\log Z_\phi$ *via*

$$\log Z^\star = \log Z_\phi + \log\Big(\mathbb{E}_{\tau\sim P_F^\phi}\big[\exp(-\Delta_{\mathrm{TB}}(\tau;\phi))\big]\Big). \tag{6}$$

This result is reminiscent of the estimation of the log-marginal likelihood via a variational lower-bound in [Zhou et al., 2024], which can be obtained by applying Jensen's inequality to (6). This proposition plays a key role in proving the results of the next section.

3.2 Convergence of the terminating state distribution

Using a similar argument as the one used to proved Proposition 1, we can show that the difference between the log-terminating state probability $\log P_F^{\phi\top}(x)$ and the target log-probability $\log P^\star(x)$ is also directly controlled by the residual (4) being minimized. This is the first evidence supporting the general wisdom in the GFlowNet literature that minimizing the TB loss leads to a more accurate approximation *at a particular* x.

Proposition 2. *For any* $x \in \mathcal{X}$*, the error between the log-terminating state probability associated with* P_F^ϕ *and the target log-probability is bounded by*

$$|\log P_F^{\phi\top}(x) - \log P^\star(x)| \le \max_{\tau\in\mathcal{T}} |\Delta_{\mathrm{TB}}(\tau;\phi)| + \max_{\tau:s_0\rightsquigarrow x} |\Delta_{\mathrm{TB}}(\tau;\phi)|. \tag{7}$$

We used $\mathcal{T}$ to denote the set of all trajectories. As an immediate consequence of this proposition, we can get a simpler (albeit looser, because it becomes independent of the terminating state x) bound

$$|\log P_F^{\phi\top}(x) - \log P^\star(x)| \le 2\max_{\tau\in\mathcal{T}} |\Delta_{\mathrm{TB}}(\tau;\phi)|. \tag{8}$$

The first term in the bound of Proposition 2, which is independent of x, has an intuitive interpretation. We could allocate all our training capacity to learn parameters ϕ so that the TB conditions match almost perfectly for all the trajectories leading to a certain terminating state x only (*e.g.*, using a behavior policy that focuses almost exclusively on those trajectories). In that case, we would have $\max_{\tau:s_0\rightsquigarrow x} |\Delta_{\mathrm{TB}}(\tau;\phi)| \approx 0$. However, that does not mean that we perfectly recovered $\log P^\star(x)$, because we have to take into account all the other terminating states as well in the computation of its normalization constant; this is precisely what $\max_{\tau\in\mathcal{T}} |\Delta_{\mathrm{TB}}(\tau;\phi)|$ controls for.

Instead of considering what happens at the level of a single terminating state x, we can also get more global guarantees on the divergence between these two distributions, similar to what typical variational methods would minimize. The following proposition gives a bound on their KL-divergence that still depends exclusively on $\Delta_{\mathrm{TB}}(\tau;\phi)$.

Proposition 3. *Let* ϕ *be the parameters of the forward transition probabilities* P_F^ϕ*, the backward transition probabilities* P_B^ϕ*, and the total flow* Z_ϕ *in the trajectory balance loss (4). The KL-divergence between the terminating state probability distribution associated with* P_F^ϕ *and the target distribution* $P^\star$ *is bounded by*

$$\mathrm{KL}\big(P_F^{\phi\top}(x)\,\|\,P^\star(x)\big) \le \mathbb{E}_{x\sim P_F^{\phi\top}}\Big[\max_{\tau:s_0\rightsquigarrow x} \Delta_{\mathrm{TB}}(\tau;\phi)\Big] - \min_{\tau\in\mathcal{T}} \Delta_{\mathrm{TB}}(\tau;\phi). \tag{9}$$

We should emphasize that since the bounds of Propositions 2 & 3 both involve the maximum error being incurred for any complete trajectory, they are often impractical for quantifying how close the distribution induced by the GFlowNet is to the target. Indeed, the number of trajectories is often much larger than the number of states, which itself is already combinatorially large. Another caveat is that the KL divergence in (9) does not measure if $P_F^{\phi\top}$ misses some of the modes of $P^\star$, potentially only giving a partial view on the accuracy of the approximation.

Nevertheless, the goal of these results is to confirm the intuition that the approximation found by the GFlowNet gets better as the TB loss is minimized, which was missing in the community as little was known about the behavior of an "inexact" GFlowNet that doesn't satisfy the TB conditions exactly. In Appendix C, we show that this is not limited to the TB loss, and a similar bound can be derived for another GFlowNet objective called the *Detailed Balance* (DB) loss [Bengio et al., 2023].

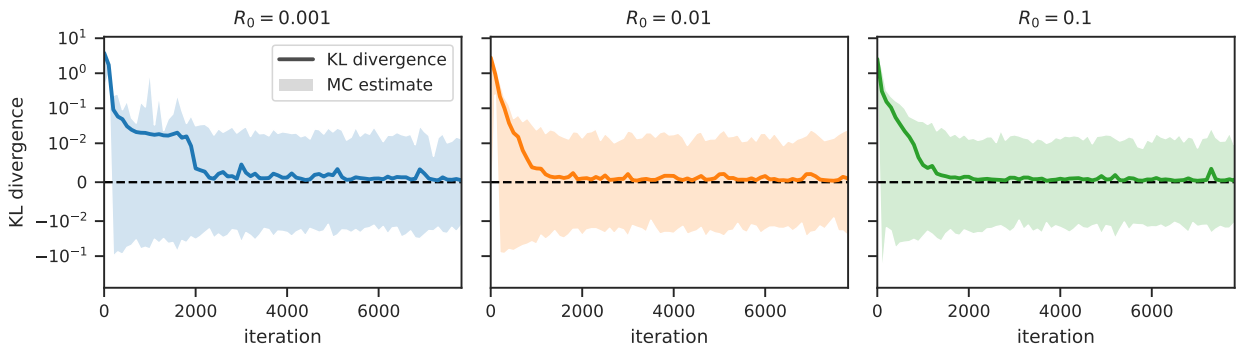

Figure 1: Empirical validation of Proposition 3 on the $8 \times 8 \times 8 \times 8$ hypergrid, with varying difficulty (controlled by R_0), during training. The true $\mathrm{KL}(P_F^{\phi\top}(x) \,\|\, P^\star(x))$ eventually decreases to 0, and always fall in the confidence interval given by estimates of (10) (shaded area). The confidence interval is obtained with bootstrapping. See Figure 2 for similar results on the 64×64 hypergrid.

4 Empirical validation

The bounds established in Propositions 2 & 3 both depend on the maximum error across all complete trajectories $\tau \in \mathcal{T}$. This dependence on worst-case trajectories makes them less practical for accurately assessing how closely the terminating state distribution approximates the target, even in relatively simple generative tasks. Instead, these results serve primarily to confirm the intuition that minimizing the TB loss moves the learned distribution closer to the target.

To provide a more practical validation of the theoretical results from the previous section, we turn to the following equality derived in the proof of Proposition 3, which connects the KL divergence to the residuals (the proposition being simply an upper bound of the RHS):

$$\mathrm{KL}\big(P_F^{\phi\top}(x) \,\|\, P^\star(x)\big) = \log\Big(\mathbb{E}_{\tau\sim P_F^\phi}\big[\exp(-\Delta_{\mathrm{TB}}(\tau;\phi))\big]\Big) + \mathbb{E}_{x\sim P_F^{\phi\top}}\Big[\log\Big(\mathbb{E}_{\tau\sim P_B^\phi(\cdot|x)}\big[\exp(\Delta_{\mathrm{TB}}(\tau;\phi))\big]\Big)\Big]. \tag{10}$$

Here, $\tau \sim P_B^\phi(\cdot \mid x)$ denotes trajectories sampled according to the backward transition probability P_B^ϕ, starting from the terminating state x. To corroborate the theoretical result of Proposition 3, we monitor whether the true KL divergence computed throughout training falls within the confidence interval induced by Monte Carlo estimates of (10) (see Algorithm 1 for details), in cases where both $P_F^{\phi\top}(x)$ and $P^\star(x)$ can be computed analytically.

We consider the hypergrid task introduced by Bengio et al. [2021], which is a simple and widely studied environment in the GFlowNet literature. In this environment, the state space is organized as an N-dimensional grid where all the states are terminating. The initial state is $(0, \ldots, 0)$, and transitions correspond to incrementing one of the coordinates by one, up to a maximum size H. The reward function in defined is such a way that it peaks around the corners of the hypergrid, with valleys in between controlled by a parameter R_0 determining the relative difficulty of finding all the modes of the target distribution (hence, the difficulty of approximating it); see Appendix B for details. Both $P^\star(x)$ and $P_F^{\phi\top}(x)$ can be evaluated in closed form here (the former by exhaustive enumeration of the terminating states and their rewards, the latter using dynamic programming [Malkin et al., 2023]).

In Figure 1, we show the evolution of $\mathrm{KL}(P_F^{\phi\top}(x) \,\|\, P^\star(x))$ throughout training on a 4-dimensional grid of size 8, with varying levels of difficulty, along with the confidence intervals derived from estimating (10). We observe that the true KL divergence goes down to 0, showing that the approximation found by the GFlowNet gets better, and this divergence stays steadily within the confidence interval during training, confirming our theoretical result in Proposition 3. Note that even on this small task, enumerating exhaustively all the trajectories in $\mathcal{T}$ is infeasible, making the exact evaluation of the bound (9) impossible.

References

E. Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Y. Bengio, S. Lahlou, T. Deleu, E. J. Hu, M. Tiwari, and E. Bengio. GFlowNet Foundations. *Journal of Machine Learning Research (JMLR)*, 2023.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine learning*, 1999.

K. Madan, J. Rector-Brooks, M. Korablyov, E. Bengio, M. Jain, A. Nica, T. Bosc, Y. Bengio, and N. Malkin. Learning GFlowNets from partial episodes for improved convergence and stability. *International Conference on Machine Learning (ICML)*, 2023.

N. Malkin, M. Jain, E. Bengio, C. Sun, and Y. Bengio. Trajectory Balance: Improved Credit Assignment in GFlowNets. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

N. Malkin, S. Lahlou, T. Deleu, X. Ji, E. Hu, K. Everett, D. Zhang, and Y. Bengio. GFlowNets and variational inference. *International Conference on Learning Representations (ICLR)*, 2023.

D. Zhang, C. Rainone, M. Peschl, and R. Bondesan. Robust Scheduling with GFlowNets. *International Conference on Learning Representations (ICLR)*, 2023.

M. Zhang, T. Bird, R. Habib, T. Xu, and D. Barber. Variational f-divergence Minimization. *Information Theory and Machine Learning Workshop (NeurIPS)*, 2019.

M. Zhou, Z. Yan, E. Layne, N. Malkin, D. Zhang, M. Jain, M. Blanchette, and Y. Bengio. PhyloGFN: Phylogenetic inference with generative flow networks. *International Conference on Learning Representations (ICLR)*, 2024.

Appendix

A Proofs of the convergence guarantees for the TB loss

Before proving the proposition from the main text, we will first state a lemma that will be useful in the proofs later

Lemma 1 (Bound on log-expectation-exp)**.** *Let $\boldsymbol{p} = (p_1, \dots, p_n)$ be a vector of probabilities (i.e., $p_i \geq 0$ and $\sum_i p_i = 1$), and $\boldsymbol{x} = (x_1, \dots, x_n)$ be an arbitrary vector. Then*

$$\left| \log \sum_{i=1}^{n} p_i \exp(x_i) \right| \leq \max_i |x_i|. \tag{11}$$

Proof. For any vector $\boldsymbol{y} = (y_1, \dots, y_n)$, and since $p_i \geq 0$, we have the following inequalities

$$\sum_{i=1}^{n} p_i \Big(\min_j y_j \Big) \leq \sum_{i=1}^{n} p_i y_i \leq \sum_{i=1}^{n} p_i \Big(\max_j y_j \Big). \tag{12}$$

Since $\sum_i p_i = 1$, both sides of these inequalities can be further simplified, only involving the minimum and maximum of $\boldsymbol{y}$. We can apply these inequalities to $y_i = \exp(x_i)$, and observe that $\min_j \exp(x_j) = \exp(\min_j x_j)$ (and similarly for the $\max$), because the exponential is a monotonically increasing function

$$\exp\Big(\min_j x_j \Big) \leq \sum_{i=1}^{n} p_i \exp(x_i) \leq \exp\Big(\max_j x_j \Big). \tag{13}$$

Taking the logarithm of the inequalities above, and again using the fact that $\log$ is monotonically increasing

$$\min_j x_j \leq \log \sum_{i=1}^{n} p_i \exp(x_i) \leq \max_j x_j. \tag{14}$$

Another way to write (14) is

$$\log \sum_{i=1}^{n} p_i \exp(x_i) \leq \max_i x_i \leq \max_i |x_i| \tag{15}$$

$$-\log \sum_{i=1}^{n} p_i \exp(x_i) \leq -\min_i x_i \leq \max_i |x_i|, \tag{16}$$

which concludes the proof. □

Proposition 1 (Estimation of the partition function)**.** *The log-partition function $\log Z^\star$ of the target distribution (2) is related to $\log Z_\phi$ via*

$$\log Z^\star = \log Z_\phi + \log \Big(\mathbb{E}_{\tau \sim P_F^\phi} \big[\exp(-\Delta_{\mathrm{TB}}(\tau; \phi)) \big] \Big). \tag{6}$$

Proof. Based on (4), for some complete trajectory $\tau = (s_0, s_1, \dots, s_T, \perp)$, we can find an expression for the product of the reward at the terminating state s_T and the backward probability of τ as a function of the residual $\Delta_{\mathrm{TB}}(\tau; \phi)$

$$R(s_T) P_B^\phi(\tau \mid s_T) = R(s_T) \prod_{t=1}^{T} P_B^\phi(s_{t-1} \mid s_t) = Z_\phi \exp(-\Delta_{\mathrm{TB}}(\tau; \phi)) \prod_{t=0}^{T} P_F^\phi(s_{t+1} \mid s_t), \tag{17}$$

where we used the notation $P_B^\phi(\tau \mid s_T) = \prod_{t=1}^{T} P_B^\phi(s_{t-1} \mid s_t)$. Since $P_B^\phi(\cdot \mid x)$ is a distribution over the complete trajectories terminating at x [Bengio et al., 2023], the reward $R(x)$ is given by

$$R(x) = \sum_{\tau : s_0 \rightsquigarrow x} R(x) P_B^\phi(\tau \mid x) = Z_\phi \sum_{\tau : s_0 \rightsquigarrow x} P_F^\phi(\tau) \exp(-\Delta_{\mathrm{TB}}(\tau; \phi)) \tag{18}$$

By definition of the partition function Z, and using the fact that P_F^ϕ induces a probability distribution over all the complete trajectories [Bengio et al., 2023]

$$Z^\star = \sum_{x \in \mathcal{X}} R(x) \tag{19}$$

$$= Z_\phi \sum_{x \in \mathcal{X}} \sum_{\tau : s_0 \rightsquigarrow x} P_F^\phi(\tau) \exp(-\Delta_{\mathrm{TB}}(\tau; \phi)) \tag{20}$$

$$= Z_\phi \sum_{\tau \in \mathcal{T}} P_F^\phi(\tau) \exp(-\Delta_{\mathrm{TB}}(\tau; \phi)) \tag{21}$$

$$= Z_\phi \mathbb{E}_{\tau \sim P_F^\phi} \big[\exp(-\Delta_{\mathrm{TB}}(\tau; \phi)) \big]. \tag{22}$$

We can conclude by taking the $\log$ of the equality above. □

Proposition 2. *For any $x \in \mathcal{X}$, the error between the log-terminating state probability associated with P_F^ϕ and the target log-probability is bounded by*

$$|\log P_F^{\phi\top}(x) - \log P^\star(x)| \leq \max_{\tau \in \mathcal{T}} |\Delta_{\mathrm{TB}}(\tau; \phi)| + \max_{\tau : s_0 \rightsquigarrow x} |\Delta_{\mathrm{TB}}(\tau; \phi)|. \tag{7}$$

Proof. We can use (4) to find an expression for $P_F^\phi(\tau)$ as a function of the residual $\Delta_{\mathrm{TB}}(\tau; \phi)$:

$$\prod_{t=0}^{T} P_F^\phi(s_{t+1} \mid s_t) = \exp(\Delta_{\mathrm{TB}}(\tau; \phi)) \frac{R(s_T)}{Z_\phi} \prod_{t=1}^{T} P_B^\phi(s_{t-1} \mid s_t). \tag{23}$$

For any $x \in \mathcal{X}$, using the notation $P_B^\phi(\tau \mid x) = \prod_{t=1}^{T} P_B^\phi(s_{t-1} \mid s_t)$ (with $s_T = x$), which is a properly defined distribution over the complete trajectories terminating in x, and by (1) of the terminating state probability distribution associated with P_F^ϕ, we get

$$P_F^{\phi\top}(x) = \sum_{\tau : s_0 \rightsquigarrow x} \prod_{t=0}^{T_\tau} P_F^\phi(s_{t+1} \mid s_t) \tag{24}$$

$$= \frac{R(x)}{Z_\phi} \sum_{\tau : s_0 \rightsquigarrow x} \exp(\Delta_{\mathrm{TB}}(\tau; \phi)) P_B^\phi(\tau \mid x) \tag{25}$$

$$= \frac{R(x)}{Z^\star} \frac{Z^\star}{Z_\phi} \mathbb{E}_{\tau \sim P_B^\phi(\cdot \mid x)} \big[\exp(\Delta_{\mathrm{TB}}(\tau; \phi)) \big] \tag{26}$$

$$= P^\star(x) \frac{Z^\star}{Z_\phi} \mathbb{E}_{\tau \sim P_B^\phi(\cdot \mid x)} \big[\exp(\Delta_{\mathrm{TB}}(\tau; \phi)) \big] \tag{27}$$

Although the normalization constant $Z^\star$ of the target distribution is still unknown, we can fortunately write it as a function of the residual as well thanks to Proposition 1. Combining it with (27), we get an expression of the difference in log-probabilities as a function of $\Delta_{\mathrm{TB}}(\tau; \phi)$ only, where the expectations over complete trajectories are respectively taken wrt. P_F^ϕ and $P_B^\phi(\cdot \mid x)$

$$\log P_F^{\phi\top}(x) - \log P^\star(x) = \log \Big(\mathbb{E}_{\tau \sim P_F^\phi} \big[\exp(-\Delta_{\mathrm{TB}}(\tau; \phi)) \big] \Big) + \log \Big(\mathbb{E}_{\tau \sim P_B^\phi(\cdot \mid x)} \big[\exp(\Delta_{\mathrm{TB}}(\tau; \phi)) \big] \Big). \tag{28}$$

Using the triangle inequality, we can conclude by applying Lemma 1 to both terms of the RHS

$$\Big| \log \Big(\mathbb{E}_{\tau \sim P_F^\phi} \big[\exp(-\Delta_{\mathrm{TB}}(\tau; \phi)) \big] \Big) \Big| \leq \max_{\tau \in \mathcal{T}} |\Delta_{\mathrm{TB}}(\tau; \phi)| \tag{29}$$

$$\Big| \log \Big(\mathbb{E}_{\tau \sim P_B^\phi(\cdot \mid x)} \big[\exp(\Delta_{\mathrm{TB}}(\tau; \phi)) \big] \Big) \Big| \leq \max_{\tau : s_0 \rightsquigarrow x} |\Delta_{\mathrm{TB}}(\tau; \phi)|. \tag{30}$$

□

Proposition 3. *Let ϕ be the parameters of the forward transition probabilities P_F^ϕ, the backward transition probabilities P_B^ϕ, and the total flow Z_ϕ in the trajectory balance loss (4). The KL-divergence between the terminating state probability distribution associated with P_F^ϕ and the target distribution $P^\star$ is bounded by*

$$\mathrm{KL}\big(P_F^{\phi\top}(x) \,\|\, P^\star(x)\big) \leq \mathbb{E}_{x \sim P_F^{\phi\top}} \Big[\max_{\tau : s_0 \rightsquigarrow x} \Delta_{\mathrm{TB}}(\tau; \phi) \Big] - \min_{\tau \in \mathcal{T}} \Delta_{\mathrm{TB}}(\tau; \phi). \tag{9}$$

Algorithm 1 KL-divergence estimation

1: **for** $m = 1, \ldots, M$ **do**
2: Sample a trajectory τ_m following P_F^ϕ, terminating at state x_m
3: Starting at x_m, sample a trajectory τ'_m following P_B^ϕ
4: Construct a Monte Carlo estimate of (10)

$$\widehat{\mathrm{KL}} = \log\left(\frac{1}{M}\sum_{m=1}^{M}\exp(-\Delta_{\mathrm{TB}}(\tau_m;\phi))\right) + \frac{1}{M}\sum_{m=1}^{M}\Delta_{\mathrm{TB}}(\tau'_m;\phi)$$

5: **return** the estimate $\widehat{\mathrm{KL}}$

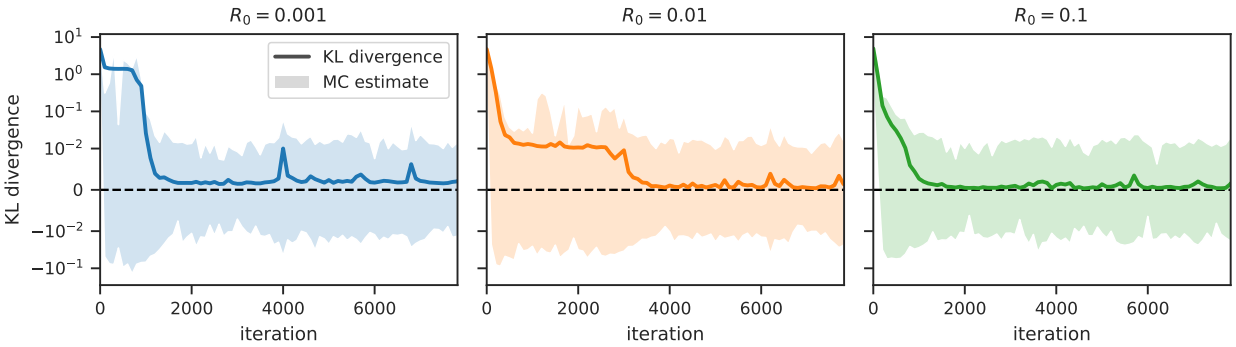

Figure 2: Empirical validation of Proposition 3 on the 64×64 hypergrid, with varying difficulty (controlled by R_0), during training. The true $\mathrm{KL}(P_F^{\phi\top}(x) \,\|\, P^\star(x))$ eventually decreases to 0, and always fall in the confidence interval given by estimates of (10) (shaded area). The confidence interval is obtained with bootstrapping.

Proof. This is a direct consequence of (28) in the proof of Proposition 2 above:

$$\begin{aligned}
\mathrm{KL}\big(P_F^{\phi\top}(x) \,\|\, P^\star(x)\big) &= \mathbb{E}_{x\sim P_F^{\phi\top}}\big[\log P_F^{\phi\top}(x) - \log P^\star(x)\big] && (31)\\
&= \log\Big(\mathbb{E}_{\tau\sim P_F^\phi}\big[\exp(-\Delta_{\mathrm{TB}}(\tau;\phi))\big]\Big) + \mathbb{E}_{x\sim P_F^{\phi\top}}\Big[\log\Big(\mathbb{E}_{\tau\sim P_B^\phi(\cdot|x)}\big[\exp(\Delta_{\mathrm{TB}}(\tau;\phi))\big]\Big)\Big] && (32)\\
&\le \Big[\max_{\tau\in\mathcal{T}} -\Delta_{\mathrm{TB}}(\tau;\phi)\Big] + \mathbb{E}_{x\sim P_F^{\phi\top}}\Big[\max_{\tau:s_0\rightsquigarrow x}\Delta_{\mathrm{TB}}(\tau;\phi)\Big] && (33)\\
&= \mathbb{E}_{x\sim P_F^{\phi\top}}\Big[\max_{\tau:s_0\rightsquigarrow x}\Delta_{\mathrm{TB}}(\tau;\phi)\Big] - \min_{\tau\in\mathcal{T}}\Delta_{\mathrm{TB}}(\tau;\phi), && (34)
\end{aligned}$$

where the inequality in (33) is a consequence of (14) in the proof of Lemma 1. □

B Details of the empirical evaluation

The reward function of the hypergrid environment is defined as

$$R(x) = R_0 + \frac{1}{2}\prod_{n=1}^{N}\mathbf{1}\left(\left|\frac{x_n}{H-1} - 0.5\right| \in (0.25, 0.5]\right) + 2\prod_{n=1}^{N}\mathbf{1}\left(\left|\frac{x_n}{H-1} - 0.5\right| \in (0.3, 0.4)\right) \tag{35}$$

In Algorithm 1 we describe a way to obtain a Monte Carlo estimate of the KL divergence using (10). In Figure 2, we show a similar experiment on the 2-dimensional hypergrid of size 64; the conclusions we made in the main text largely hold in this new setting.

C Proofs of the convergence guarantees for the DB loss

Bengio et al. [2023] showed that if P_F, P_B, and a "flow" function $F(s)$ defined over the all the states in the state space satisfy the following *Detailed Balance* (DB) conditions at every transition $s \to s'$ where $s' \neq \perp$

$$F(s)P_F(s' \mid s) = F(s')P_B(s \mid s'), \tag{36}$$

with the boundary condition $F(x)P_F(\perp \mid x) = R(x)$ at any terminating state $x \in \mathcal{X}$, then $P_F^\top(x) \propto R(x)$. Taking inspiration from Section 2, we can convert this condition into a loss that can be minimized of the form $\mathcal{L}_{\text{DB}}(\phi) = \frac{1}{2}\mathbb{E}_{\pi_b}[\Delta^2_{\text{DB}}(s, s'; \phi)]$, where

$$\Delta_{\text{DB}}(s, s'; \phi) = \log \frac{F_\phi(s)P_F^\phi(s' \mid s)}{F_\phi(s')P_B^\phi(s \mid s')} \qquad \Delta_{\text{DB}}(x, \perp; \phi) = \log \frac{F_\phi(x)P_F^\phi(\perp \mid x)}{R(x)} \tag{37}$$

Throughout this section, for a trajectory $\tau = (s_0, s_1, \ldots, s_T, \perp)$, we will use the notations $s_{T+1} = \perp$, so that we can write $\Delta_{\text{DB}}(s_T, s_{T+1}; \phi)$. The following proposition is the counterpart of Proposition 2 in the case of Detailed Balance.

Proposition 4. *For any $x \in \mathcal{X}$, the error between the log-terminating state probability associated with P_F^ϕ and the target log-probability is bounded by*

$$|\log P_F^{\phi\top}(x) - \log P^\star(x)| \leq \max_{\tau \in \mathcal{T}} \left| \sum_{t=0}^{T_\tau} \Delta_{\text{DB}}(s_t, s_{t+1}; \phi) \right| + \max_{\tau : s_0 \rightsquigarrow x} \left| \sum_{t=0}^{T_\tau} \Delta_{\text{DB}}(s_t, s_{t+1}; \phi) \right|. \tag{38}$$

Proof. The proof is similar to the proof of Proposition 2. From (37), we can derive that

$$P_F^\phi(s' \mid s) = P_B^\phi(s \mid s')\frac{F_\phi(s')}{F_\phi(s)}\exp(\Delta_{\text{DB}}(s, s'; \phi)) \tag{39}$$

$$P_F^\phi(s, \perp) = \frac{R(x)}{F_\phi(x)}\exp(\Delta_{\text{DB}}(x, \perp; \phi)) \tag{40}$$

For some trajectory $\tau = (s_0, s_1, \ldots, s_T, \perp)$,

$$P_F^\phi(\tau) = \prod_{t=0}^{T} P_F^\phi(s_{t+1} \mid s_t) = P_F^\phi(\perp \mid s_T) \prod_{t=0}^{T-1} P_F^\phi(s_{t+1} \mid s_t) \tag{41}$$

$$= R(s_T)\frac{1}{F_\phi(s_T)}\exp(\Delta_{\text{DB}}(x, \perp; \phi)) \prod_{t=0}^{T-1} P_B^\phi(s_t \mid s_{t+1})\frac{F_\phi(s_{t+1})}{F_\phi(s_t)}\exp(\Delta_{\text{DB}}(s_t, s_{t+1}; \phi)) \tag{42}$$

$$= \frac{R(s_T)}{F_\phi(s_0)}\exp\left[\sum_{t=0}^{T} \Delta_{\text{DB}}(s_t, s_{t+1}; \phi)\right] P_B^\phi(\tau \mid s_T) \tag{43}$$

We can also prove a similar result as Proposition 1, and show that we can write the true partition function $Z^\star$ as a function of $F_\phi(s_0)$ and the residuals:

$$Z^\star = \sum_x R(x) = F_\phi(s_0)\mathbb{E}_{\tau \sim P_F^\phi}\left[\exp\left[-\sum_{t=0}^{T_\tau} \Delta_{\text{DB}}(s_t, s_{t+1}; \phi)\right]\right] \tag{44}$$

Using (43) & (44), we can write the terminating state distribution as

$$P_F^{\phi\top}(x) = \sum_{\tau : s_0 \rightsquigarrow x} \frac{R(x)}{F_\phi(s_0)} P_B^\phi(\tau \mid x)\exp\left[\sum_{t=0}^{T_\tau} \Delta_{\text{DB}}(s_t, s_{t+1}; \phi)\right] \tag{45}$$

$$= \frac{R(x)}{Z^\star}\frac{Z^\star}{F_\phi(s_0)}\mathbb{E}_{\tau \sim P_B^\phi(\cdot \mid x)}\left[\exp\left[\sum_{t=0}^{T_\tau} \Delta_{\text{DB}}(s_t, s_{t+1}; \phi)\right]\right] \tag{46}$$

Writing this in terms of the difference of log-probabilities, we obtain

$$\log P_F^{\phi\top}(x) - \log P^\star(x) = \log\left(\mathbb{E}_{\tau\sim P_B^\phi(\cdot|x)}\left[\exp\left[\sum_{t=0}^{T_\tau}\Delta_{\mathrm{DB}}(s_t, s_{t+1};\phi)\right]\right]\right) + \log\left(\mathbb{E}_{\tau\sim P_F^\phi}\left[\exp\left[-\sum_{t=0}^{T_\tau}\Delta_{\mathrm{DB}}(s_t, s_{t+1};\phi)\right]\right]\right) \tag{47}$$

We can conclude using Lemma 1. □