TOWARDS UNDERSTANDING THE TRAINING OF GFLOWNETS: A THEORETICAL PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Flow Networks (GFlowNets) learn policies that sample discrete objects in proportion to a positive reward. They are trained by minimizing balance objectives, whose exact satisfaction guarantees correct sampling. However, a comprehensive theory linking nonzero balance loss and approximate stationarity to sampling accuracy, especially under evolving sampling distributions and backward policies, is still lacking. To bridge this gap, we develop a unified theory for Flow Matching, Detailed Balance, and Trajectory Balance in this regime. We first bound terminal symmetric KL by the square root of the expected balance loss. Then, we show that approximate stationarity controls this distribution error up to a tangent-space residual measuring parameterization incompleteness. Finally, we establish finite-time convergence under fixed or controlled sampling and fixed, balance-trained, or separately updated backward policies. Under standard smoothness assumptions, with fixed or controlled sampling and a fixed or balance-trained backward policy, the expected loss is $O(\bar{\epsilon}^2 + T^{-1})$ and expected terminal symmetric KL is $O(\bar{\epsilon} + T^{-1/2})$ under uniform coverage. For a separately updated backward policy, we characterize the sufficient update decay to approach the residual neighborhood and preserve these rates. These results connect balance objectives to sampling accuracy and clarify how parameterization, coverage, and moving components govern GFlowNet training.

1 INTRODUCTION

Generative Flow Networks (GFlowNets, Bengio et al. (2021; 2023)) have emerged as a promising framework for training generative models that sample objects with probabilities proportional to a positive reward. High-reward objects are therefore sampled frequently, while different high-reward regions retain probability mass. This combination of quality and diversity has made GFlowNets useful in various applications such as molecule generation (Zhu et al., 2023; Koziarski et al., 2024), combinatorial optimization (Zhang et al., 2023), probabilistic inference (Manta et al., 2023; Deleu et al., 2024), text generation (Hu et al., 2024), and image generation (Liu et al., 2025; Yun et al., 2025).

GFlowNets transform the sampling problem into a sequential generation process on a directed acyclic graph (DAG). Starting from an initial state, a forward policy constructs an object through a series of local decisions until it reaches a terminal state. The central idea is to view probability mass as flow through the DAG. If the flow is conserved at intermediate states and the terminal flow matches the reward, then the forward policy samples terminal states from the desired reward-proportional distribution. This global property can be enforced through local or trajectory-level balance conditions, leading to the standard Flow Matching (FM, Bengio et al. 2021), Detailed Balance (DB, Bengio et al. 2023), and Trajectory Balance (TB, Whitammer et al. 2022) objectives.

This elegant construction explains what happens when every balance condition is satisfied exactly, but it does not explain what happens during practical training. A parameterized model may attain only a small nonzero loss or approach a stationary point rather than a global minimizer. Moreover, the states, edges, or trajectories used for training may be drawn from an off-policy and time-varying distribution, while the backward policy may be fixed, trained jointly with the balance loss, or updated using a separate criterion. These choices make the objective seen by stochastic gradient descent (SGD) evolve over time. Consequently, to understand GFlowNet training, a systematic study is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

required to connect loss and stationarity to terminal accuracy while accounting for these moving components. This motivates three essential questions:

1. *Does a small balance loss guarantee that the learned terminal distribution is close to the reward-proportional target?*
2. *When optimization reaches an approximate stationary point, how good is the resulting sampler, and what limits its accuracy?*
3. *As sampling strategies and backward policies are adapted during training, when can learning still approach the target distribution, and at what rate?*

Existing studies provide complementary pieces of this picture. Flow-imbalance analyses (Silva et al., 2025a) examine how balance violations affect the generated distribution, while generalization analyses (Krichel et al., 2024; Silva et al., 2025b) study learning from limited observations. Connections to variational inference and entropy-regularized reinforcement learning (Whitammer et al., 2023; Tiapkin et al., 2024), together with advances in loss design (Hu et al., 2025; Fawkes & Hartford, 2026), relate training objectives to distribution discrepancies. However, these relationships do not by themselves connect balance objectives to sampling accuracy. Yu (2025) establishes gradient-convergence results for FM and DB, as well as sample-complexity guarantees for TB. However, the objectives in their analysis differ from the commonly used squared log-ratio loss and do not clarify the quality of parameterized stationary points or the behavior of stochastic training under varying sampling and backward policies. Very recently, Lei et al. (2026) derived total variation guarantees from a uniform residual bound, providing a quantitative measure of sampling accuracy based on balance loss. A comprehensive literature review is available in Appendix A.

To answer the above questions, we develop a unified theory for FM, DB, and TB that connects balance loss, stationary-point quality, and finite-time sampling accuracy. Our theoretical results are summarized as follows:

1. **Balance loss controls sampling accuracy.** For FM, DB, and TB, we bound the terminal symmetric KL by the square root of the expected balance loss, with explicit occupancy-mass and coverage factors. Thus a small loss is informative when the training samples cover the states, edges, or trajectories on which the forward and backward flows disagree.
2. **Stationarity implies approximate optimality.** We decompose the log-flow imbalance into a gradient-visible tangent component and an orthogonal parameterization residual. This yields an approximate gradient-domination property in which the terminal distribution error is controlled by the gradient norm and a precisely defined parameterization-incompleteness term. Thus, approximate stationarity leads to a nearly accurate sampler.
3. **Finite-time convergence with moving training components.** We analyze stochastic GFlowNet training when the sampling distribution is fixed or changes in a controlled manner and when the backward policy is fixed, jointly optimized by the balance loss, or updated by a general rule. With fixed or controlled sampling and a fixed or balance-trained backward policy, the expected balance loss is bounded by $O(\bar{\epsilon}^2 + T^{-1})$, which gives expected terminal symmetric KL of $O(\bar{\epsilon}+T^{-1/2})$ under uniform coverage. For a separately updated backward policy, we characterize how fast its update must decay to ensure convergence and to preserve these rates.

The remainder of the paper is organized as follows. Section 2 introduces FM, DB, and TB in a unified framework. Section 3 relates their expected balance losses to terminal-distribution mismatch. Section 4 establishes the stationarity-to-optimality guarantee, and Section 5 gives the finite-time convergence analysis. Additionally, we extend the analysis to Sub-Trajectory Balance in Appendix E.

2 PRELIMINARIES

Let $\mathcal{X}$ be a discrete space with positive reward $R : \mathcal{X} \to \mathbb{R}_{>0}$. GFlowNet aims to sample $x \in \mathcal{X}$ with probability $R(x)/Z$, where $Z := \sum_{x\in\mathcal{X}} R(x)$ is assumed to be finite. The generation takes place on a DAG $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ with a unique initial state s_0. Every state is reachable from s_0, and the set of terminal states coincides with $\mathcal{X}$. We write $\mathcal{T}$ for the set of complete trajectories $\tau = (s_0, s_1, \ldots, s_{n(\tau)})$, where $n(\tau)$ is the number of edges and $s_{n(\tau)} \in \mathcal{X}$.

A **flow network** is specified by a positive **edge flow** function $F : \mathcal{E} \to \mathbb{R}_{>0}$ satisfying flow conservation at every nonterminal state other than s_0:

$$\sum_{s':(s'\to s)\in\mathcal{E}} F(s' \to s) = \sum_{s'':(s\to s'')\in\mathcal{E}} F(s \to s'').$$

Its **state flow** $F(s)$ is defined as the total flow going through state s. The network is **balanced** when $F(x) = R(x)$ for every terminal state, in which case $F(s_0) = Z$ and the induced terminal distribution is $F(x)/F(s_0) = R(x)/Z$.

A flow network induces a **forward policy** and a **backward policy** through

$$P_F(s' \mid s) := \frac{F(s \to s')}{F(s)}, \qquad P_B(s \mid s') := \frac{F(s \to s')}{F(s')}.$$

GFlowNets have different variants, defined by how the flows are parameterized and the corresponding balance losses resulting from different balance conditions. In this paper, we focus on three typical types: Flow Matching (FM), Detailed Balance (DB), and Trajectory Balance (TB).

Throughout the discussion, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ collects the parameters optimized by the balance loss, while $\boldsymbol{\varphi}$ denotes the parameters of a backward policy when it is specified separately. For FM, both policies are induced by the edge flows and there is no separate backward block. For TB and DB, when the backward policy is jointly trained by the balance loss, its parameters are included in $\boldsymbol{\theta}$; every occurrence of $P_B(\cdot;\boldsymbol{\varphi})$ below is then read as $P_B(\cdot;\boldsymbol{\theta})$.

Flow Matching GFlowNets(Bengio et al., 2021). An FM-GFN uses a positive edge flow function $F(s \to s';\boldsymbol{\theta})$. For $s \in \mathcal{S} \setminus \{s_0\}$, define the forward flow and the backward flow as

$$F^{\to}(s) := \sum_{s':(s'\to s)\in\mathcal{E}} F(s' \to s;\boldsymbol{\theta}), \quad F^{\leftarrow}(s) := \begin{cases} R(s) & , s \in \mathcal{X} \\ \sum_{s'':(s\to s'')\in\mathcal{E}} F(s \to s'';\boldsymbol{\theta}) & , \text{otherwise} \end{cases},$$

By extending the definition of $F^{\leftarrow}(s)$ to the initial state s_0, these edge flows induce normalized policies even before the balance conditions hold:

$$P_F(s' \mid s) = \frac{F(s \to s';\boldsymbol{\theta})}{F^{\leftarrow}(s)}, \qquad P_B(s \mid s') = \frac{F(s \to s';\boldsymbol{\theta})}{F^{\to}(s')}.$$

FM enforces the balance condition $F^{\to}(s) = F^{\leftarrow}(s)$ on every non-initial state $s \neq s_0$ using the pointwise loss

$$L_{\mathrm{FM}}(s) := \frac{1}{2}\left(\log \frac{F^{\to}(s)}{F^{\leftarrow}(s)}\right)^2.$$

Detailed Balance GFlowNets(Bengio et al., 2023). A DB-GFN uses a positive state flow function $F(s;\boldsymbol{\theta})$ on nonterminal states, a forward policy $P_F(s' \mid s;\boldsymbol{\theta})$, and a backward policy $P_B(s \mid s';\boldsymbol{\varphi})$. For $(s \to s') \in \mathcal{E}$, define the forward flow and the backward flow as

$$F^{\to}(s \to s') := F(s;\boldsymbol{\theta})P_F(s'|s;\boldsymbol{\theta}), \quad F^{\leftarrow}(s \to s') := \begin{cases} R(s')P_B(s|s';\boldsymbol{\varphi}) & , s' \in \mathcal{X} \\ F(s';\boldsymbol{\theta})P_B(s|s';\boldsymbol{\varphi}) & , \text{otherwise} \end{cases},$$

DB trains their equality on every edge using the pointwise loss

$$L_{\mathrm{DB}}(s \to s') := \frac{1}{2}\left(\log \frac{F^{\to}(s \to s')}{F^{\leftarrow}(s \to s')}\right)^2.$$

Trajectory Balance GFlowNets(Whitammer et al., 2022). A TB-GFN uses a positive estimated total flow $\widehat{Z}(\boldsymbol{\theta})$, a forward policy $P_F(s' \mid s;\boldsymbol{\theta})$, and a backward policy $P_B(s \mid s';\boldsymbol{\varphi})$. For $\tau = (s_0, s_1, \ldots, s_{n(\tau)}) \in \mathcal{T}$, define the forward flow and the backward flow as

$$F^{\to}(\tau) := \widehat{Z}(\boldsymbol{\theta}) \prod_{t=1}^{n(\tau)} P_F(s_t|s_{t-1};\boldsymbol{\theta}), \quad F^{\leftarrow}(\tau) := R(s_{n(\tau)}) \prod_{t=1}^{n(\tau)} P_B(s_{t-1}|s_t;\boldsymbol{\varphi}),$$

TB enforces their equality on every complete trajectory using the pointwise loss

$$L_{\mathrm{TB}}(\tau) := \frac{1}{2}\left(\log \frac{F^{\to}(\tau)}{F^{\leftarrow}(\tau)}\right)^2.$$

Sampling and Resampling Strategies. In practice, the losses and their gradients are estimated from sampled and possibly reweighted states, edges, or trajectories. We denote the resulting distribution by $\mu_{\boldsymbol{\omega}}$, where $\boldsymbol{\omega}$ collects the related parameters. It may depend on the forward policy, the backward policy, or other hyperparameters and can therefore evolve during training.

Backward Policy. For DB-GFN and TB-GFN, the backward policy $P_B(s \mid s'; \boldsymbol{\varphi})$ can be chosen freely: any positive policy on the incoming edges gives compatible flows and a forward policy by propagating the terminal rewards backward. In practice, P_B is commonly fixed to be uniform or trained jointly using the TB or DB loss. Some recent work (Jang et al., 2024; Gritsaev et al., 2025) also studies separate backward-policy criteria designed to improve exploration or exploitation.

Unified Framework(Hu et al., 2025). For the unified analysis, let $\mathcal{O}$ denote the relevant training-object space: $\mathcal{S} \setminus \{s_0\}$ for FM, $\mathcal{E}$ for DB, and $\mathcal{T}$ for TB. Write $\mu = \mu_{\boldsymbol{\omega}}$ for a normalized training distribution on $\mathcal{O}$. Let $L_2(\mu)$ be the space of real-valued functions on $\mathcal{O}$ with finite second moment, identified when they agree μ-almost surely, with

$$\langle f, g\rangle_\mu := \sum_{o\in\mathcal{O}} \mu(o) f(o) g(o), \qquad \|f\|_\mu := \langle f, f\rangle_\mu^{1/2}.$$

Denote the **log-flow imbalance** as $h_{\boldsymbol{\theta}}(o) := \log \frac{F^{\rightarrow}(o)}{F^{\leftarrow}(o)}$. The overall objective can be written as

$$\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}) := \frac{1}{2}\,\|h_{\boldsymbol{\theta}}\|_\mu^2 = \frac{1}{2}\mathbb{E}_{o\sim\mu}[h_{\boldsymbol{\theta}}(o)^2].$$

For gradient-based analysis, $h_{\boldsymbol{\theta}}$ and its $\boldsymbol{\theta}$-derivatives are square-integrable under μ, and differentiation may be interchanged with the loss summations in each argument being differentiated; sufficient conditions are given in Appendix D.4.

3 From Balance Loss to Distribution Matching

This section connects the balance objectives to the ultimate sampling goal. We first lift the comparison between terminal distributions to complete trajectories and then express the resulting trajectory discrepancy as a weighted sum of the log-flow imbalance $h_{\boldsymbol{\theta}}$. This yields a bound in terms of the expected training loss and the coverage provided by μ. The proofs are provided in Appendix B.

For $\tau = (s_0, s_1, \ldots, s_{n(\tau)}) \in \mathcal{T}$, define the forward and backward trajectory distributions by

$$p_F(\tau) := \prod_{t=1}^{n(\tau)} P_F(s_t|s_{t-1}), \qquad p_B(\tau) := \frac{R(s_{n(\tau)})}{Z} \prod_{t=1}^{n(\tau)} P_B(s_{t-1}|s_t),$$

where parameter arguments are suppressed. The backward trajectory distribution first samples a terminal state from the reward-proportional target and then follows the backward policy to the initial state. To avoid an endless generation process, we assume that both $\mathbb{E}_{p_F}[n(\tau)]$ and $\mathbb{E}_{p_B}[n(\tau)]$ are finite. For FM and DB on infinite DAGs, we also assume that $\sum_{o\in\tau} |h_{\boldsymbol{\theta}}(o)|$ is integrable under both trajectory laws; see Appendix B. Their terminal marginals are

$$p_F^{\top}(x) := \sum_{\tau: s_{n(\tau)}=x} p_F(\tau), \qquad p_R(x) := \frac{R(x)}{Z} = \sum_{\tau: s_{n(\tau)}=x} p_B(\tau), \qquad x \in \mathcal{X}.$$

The chain rule of KL divergence implies the following lemma.

Lemma 3.1 *Let p_F, p_B, $p_F^{\top}$ and p_R be defined as above. We have*

$$\mathrm{KL}(p_F^{\top}\|p_R) \le \mathrm{KL}(p_F\|p_B), \quad \mathrm{KL}(p_R\|p_F^{\top}) \le \mathrm{KL}(p_B\|p_F).$$

For a training object $o \in \mathcal{O}$, define its forward and backward occupancy weights by

$$\nu_F(o) := \sum_{\tau\in\mathcal{T}: o\in\tau} p_F(\tau), \qquad \nu_B(o) := \sum_{\tau\in\mathcal{T}: o\in\tau} p_B(\tau),$$

where, for TB, $o \in \tau$ means $o = \tau$. Define the discrepancy distribution as

$$\nu(o) := \frac{|\nu_F(o) - \nu_B(o)|}{c_\nu}, \quad \text{where } c_\nu = \sum_{o \in \mathcal{O}} |\nu_F(o) - \nu_B(o)|.$$

Note that for state and edge objectives, the total mass of each occupancy measure is the expected number of corresponding objects along a trajectory, while for trajectories, the total mass is one. Thus c_ν is at most the sum of these two total masses.

To connect trajectory-level KL to balance objectives, we decompose the log-density ratio on trajectories into accumulated local imbalances plus a term that is constant across trajectories. Symmetrizing the KL divergence cancels this constant and leaves only the imbalances that the training objective controls, yielding the following theorem.

Theorem 3.2 *If $c_\nu = 0$, then $\mathrm{KL}(p_F^\top \| p_R) + \mathrm{KL}(p_R \| p_F^\top) = 0$. If $c_\nu > 0$, then*

$$\mathrm{KL}(p_F^\top \| p_R) + \mathrm{KL}(p_R \| p_F^\top) \le c_\nu \sqrt{2 \left\| \frac{\nu}{\mu} \right\|_\infty \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta})}, \quad \textit{where } \left\| \frac{\nu}{\mu} \right\|_\infty := \sup_{o \in \mathcal{O}} \frac{\nu(o)}{\mu(o)}.$$

The ratio $\|\nu/\mu\|_\infty$ represents the coverage of the discrepancy distribution by the sampling distribution. We set $\nu(o)/\mu(o) = +\infty$ if $\nu(o) > 0 = \mu(o)$ and set it to zero if $\nu(o) = \mu(o) = 0$. A finite ratio automatically implies that μ covers the support of ν; otherwise the bound is valid but vacuous. Although ν is infeasible in practice, the training objective is effective when μ provides good coverage. In particular, a bounded coverage ratio and an expected loss of δ yield an $O(\sqrt{\delta})$ bound on terminal symmetric KL.

Trajectory comparisons also underlie the connection between GFlowNets and variational inference (Whitammer et al., 2023). Our coverage factor parallels distribution-mismatch coefficients in RL (Agarwal et al., 2021), where the relevant measure is the difference between forward and backward occupancies. For approximate balance, Lei et al. (2026) derive total-variation guarantees from uniform residual bounds. Theorem 3.2 instead controls terminal symmetric KL through the expected squared log-ratio loss under the training measure. A detailed comparison is provided in Appendix A.1.

4 From Stationarity to Optimality

Theorem 3.2 connects distribution error to log-flow imbalance and hence the balance loss. To understand what remains when optimization only reaches a stationary point, we further show that this imbalance can be decomposed into two parts: its projection onto the local tangent space and an orthogonal residual. The loss gradient controls the projected component through an effective Fisher matrix. The orthogonal residual, however, is invisible to first-order optimization, which can be interpreted as a measure of parameterization incompleteness. Therefore, this decomposition offers an approximate stationarity-to-optimality guarantee whose irreducible error is the parameterization-incompleteness term. Below we present the main results, while the proofs are available in Appendix C.

For the differentiable imbalance family introduced in Section 2, define the imbalance gradient feature (called the score function below)

$$\psi_{\boldsymbol{\theta}}(o) := \nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}(o).$$

At a given $\boldsymbol{\theta}$, the score features span the local tangent space

$$\mathsf{T}_{\boldsymbol{\theta}} := \left\{ \boldsymbol{u}^\top \psi_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{u} \in \mathbb{R}^d \right\}.$$

Thus $h_{\boldsymbol{\theta}} + \mathsf{T}_{\boldsymbol{\theta}}$ is the local affine model of the imbalance. The distance from this affine space to zero tells us the residual after the best correction in this linearized model.

Definition 4.1 (Parameterization Incompleteness) *Let $\Pi^\mu_{\mathsf{T}_{\boldsymbol{\theta}}}$ denote the orthogonal projection in $L_2(\mu)$ onto $\mathsf{T}_{\boldsymbol{\theta}}$. Define the tangent residual by*

$$r_{\mu,\boldsymbol{\theta}} := \left(I - \Pi^\mu_{\mathsf{T}_{\boldsymbol{\theta}}}\right) h_{\boldsymbol{\theta}}.$$

Its parameterization incompleteness under the fixed training measure μ is

$$\epsilon_\mu := \sup_{\boldsymbol{\theta} \in \Theta} \|r_{\mu,\boldsymbol{\theta}}\|_\mu = \sup_{\boldsymbol{\theta} \in \Theta} \min_{\boldsymbol{u} \in \mathbb{R}^d} \left\| h_{\boldsymbol{\theta}} - \boldsymbol{u}^\top \psi_{\boldsymbol{\theta}} \right\|_\mu \in [0, +\infty].$$

The residual is orthogonal to every score feature and hence does not appear in the loss gradient. The quantity ϵ_μ bounds its norm over the parameter domain. A small value limits the imbalance that first-order optimization cannot detect under μ.

Proposition 4.2 (Tabular Zero Residual) *On a finite DAG, the standard positive tabular models of FM, DB, and TB satisfy*

$$h_{\boldsymbol{\theta}} \in \mathsf{T}_{\boldsymbol{\theta}} \quad \textit{for every } \boldsymbol{\theta} \in \Theta.$$

Consequently, for every sampling measure μ,

$$r_{\mu,\boldsymbol{\theta}} = 0 \quad \textit{in } L_2(\mu), \qquad \epsilon_\mu = 0.$$

Notes on forward-looking parameterizations. Forward-looking GFlowNets (FL-GFNs; Pan et al. 2023) incorporate a given intermediate energy into the balance equations. With the convention $R(x) = e^{-E(x)}$, their state-flow parameterization can be written as

$$F(s;\boldsymbol{\theta}) = e^{-E(s)}\widetilde{F}(s;\boldsymbol{\theta}), \qquad \widetilde{F}(x) = 1 \quad (x \in \mathcal{X}).$$

When parameterized independently by the log-flows, FL-GFNs differ only by a fixed translation, so Proposition 4.2 still applies.

With Proposition 4.2, ϵ_μ can be viewed as a reasonable measure of parameterization incompleteness. The remainder of the section concerns a general parameterized family. We require two regularity conditions to control the tangent-space component.

Assumption 4.3 (Bounded Score Function) *There exists $M > 0$ such that, for the discrepancy distribution ν from Theorem 3.2 and every $\boldsymbol{\theta} \in \Theta$ under consideration with $c_\nu > 0$,*

$$\mathbb{E}_{o\sim\nu}\left[\|\psi_{\boldsymbol{\theta}}(o)\|_2\right] \le M,$$

where ν is induced at the same $\boldsymbol{\theta}$.

Assumption 4.4 (Effective Fisher Conditioning) *Define the Fisher information as*

$$\mathcal{I}_\mu(\boldsymbol{\theta}) := \mathbb{E}_{o\sim\mu}[\psi_{\boldsymbol{\theta}}(o)\psi_{\boldsymbol{\theta}}(o)^\top].$$

There exists $\lambda > 0$ such that, for every $\boldsymbol{\theta} \in \Theta$ under consideration,

$$\left\|\mathcal{I}_\mu(\boldsymbol{\theta})^\dagger\right\|_{2\to 2} \le \frac{1}{\lambda},$$

where $\mathcal{I}_\mu(\boldsymbol{\theta})^\dagger$ is the Moore–Penrose pseudo-inverse.

Assumption 4.3 controls how large a tangent direction can be on the objects emphasized by ν. Assumption 4.4 requires the nonzero eigenvalues of the Fisher matrix to be uniformly separated from zero. Unlike strict positive definiteness, it permits null directions caused by redundant coordinates and therefore accommodates over-parameterized models. Only the effective parameter subspace matters because null directions have zero first-order effect on the imbalance in $L_2(\mu)$. They need not be invisible outside the support of μ, which is why coverage remains a separate condition.

Lemma 4.5 (Tangent Projection from the Gradient) *Let*

$$\boldsymbol{u}^*_\mu(\boldsymbol{\theta}) := \mathcal{I}_\mu(\boldsymbol{\theta})^\dagger \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}).$$

Under Assumption 4.4,

$$\Pi^\mu_{\mathsf{T}_{\boldsymbol{\theta}}} h_{\boldsymbol{\theta}} = \boldsymbol{u}^*_\mu(\boldsymbol{\theta})^\top \psi_{\boldsymbol{\theta}}, \qquad \|\boldsymbol{u}^*_\mu(\boldsymbol{\theta})\|_2 \le \frac{1}{\lambda}\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta})\|_2.$$

This projection parallels feature regression in compatible function approximation and natural policy gradient (Sutton et al., 1999; Kakade, 2001). Here the regression target is the imbalance $h_{\boldsymbol{\theta}}$ and the features are its derivatives $\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}$. Accordingly, the effective Fisher is the Gram matrix of imbalance features, and the residual measures local tangent approximation error. Refer to Appendix A.2 for further discussions. Together with conditioning and coverage, this geometry links balance-loss stationarity to terminal sampling accuracy in the following theorem.

Theorem 4.6 *Suppose Assumptions 4.3 and 4.4 hold. For every $\boldsymbol{\theta} \in \Theta$, if $c_\nu > 0$, then*

$$\mathrm{KL}(p_F^\top \| p_R) + \mathrm{KL}(p_R \| p_F^\top) \le c_\nu \left(\sqrt{\left\| \frac{\nu}{\mu} \right\|_\infty} \epsilon_\mu + \frac{M}{\lambda} \left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}) \right\|_2 \right).$$

If $c_\nu = 0$, the symmetric KL divergence on the left-hand side is zero. The coverage term is understood as $+\infty$ whenever $\|\nu/\mu\|_\infty = +\infty$.

This theorem bounds distribution mismatch by the gradient norm and a residual term. The factor M/λ describes the conditioning of the visible component, while the coverage ratio controls transfer from training objects to occupancy discrepancies. With finite coverage, the zero-residual tabular case has zero terminal symmetric KL at every stationary point satisfying the assumptions.

5 Convergence Rates of Stochastic GFlowNet Training

Section 4 explains when a small gradient implies a small imbalance. We now ask whether stochastic training reaches such a regime, and how quickly. We first fix the sampling distribution, so that tangent geometry directly turns SGD descent into loss contraction. We then allow the sampling distribution to change, and finally allow a separate backward-policy update. The stochastic-optimization, moving-objective, and actor–critic background for is summarized in Appendix A.2; detailed proofs and regularity checks appear in Appendix D.

5.1 Fixed sampling: from geometry to convergence

Recall that $\boldsymbol{\theta}$ contains all parameters trained by the balance loss, including the backward parameters for joint balance training. Make the external arguments of the same expected loss explicit:

$$\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, \boldsymbol{\omega}; \boldsymbol{\varphi}) := \frac{1}{2} \mathbb{E}_{o \sim \mu_{\boldsymbol{\omega}}} [h_{\boldsymbol{\theta}, \boldsymbol{\varphi}}(o)^2].$$

The separate argument $\boldsymbol{\varphi}$ is omitted for joint balance training and for FM; in those cases write $h_{\boldsymbol{\theta}}$ in place of $h_{\boldsymbol{\theta}, \boldsymbol{\varphi}}$. For now, $\boldsymbol{\omega}$ and any separate $\boldsymbol{\varphi}$ are fixed. Joint balance training simply enlarges the optimization block: its backward-policy update is part of the descent on the same objective.

Let $\mathcal{F}_t$ contain the history before iteration t and write $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. All current parameters are $\mathcal{F}_t$-measurable, and $\mathcal{L}_t := \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t; \boldsymbol{\varphi}_t)$ denotes their current loss. The balance update satisfies

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \boldsymbol{g}_t, \qquad \mathbb{E}_t \boldsymbol{g}_t = \nabla_{\boldsymbol{\theta}} \mathcal{L}_t, \qquad \mathbb{E}_t \|\boldsymbol{g}_t\|_2^2 \le \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_t\|_2^2 + \sigma^2. \tag{1}$$

This derivative holds the sampling distribution and any separate backward parameters fixed, including when they change between iterations below. We assume $\mathbb{E}\mathcal{L}_0 < \infty$, finite second moments of all updates, and finite score second moments. All constants are deterministic and independent of t.

Assumption 5.1 (Smooth stochastic objective) *The oracle satisfies equation 1. For some $L > 0$, the loss has an L-Lipschitz gradient jointly in the arguments that change, on a convex admissible region containing every actual update segment.*

Assumption 5.2 (Uniform tangent geometry) *For the residual and Fisher matrix defined in Section 4, there exist $\bar{\epsilon} \ge 0$ and $\lambda > 0$ such that, almost surely at every iterate,*

$$\|r_{\mu_{\boldsymbol{\omega}_t}, \boldsymbol{\theta}_t}\|_{\mu_{\boldsymbol{\omega}_t}} \le \bar{\epsilon}, \qquad \|\mathcal{I}_{\mu_{\boldsymbol{\omega}_t}}(\boldsymbol{\theta}_t)^\dagger\|_{2 \to 2} \le \frac{1}{\lambda}.$$

Both quantities use $\psi_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} h$, with any separate backward parameters fixed at their current values.

Here $\bar{\epsilon}$ bounds the residual along training, including changes in the external arguments. The projection identity gives an approximate Polyak–Łojasiewicz (PL) inequality:

Lemma 5.3 (Approximate PL geometry) *Under Assumption 5.2, at every iterate,*

$$2\mathcal{L}_t \le \bar{\epsilon}^2 + \frac{1}{\lambda} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_t\|_2^2. \tag{2}$$

Thus the gradient detects loss above the residual level. Note that for a sufficiently small step-size, smoothness gives expected descent of order $\eta_t \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_t\|_2^2$, up to an $O(\eta_t^2)$ noise term. The lemma therefore converts this into contraction of $\mathcal{L}_t$ outside an $O(\bar{\epsilon}^2)$ residual neighborhood.

Theorem 5.4 (SGD with fixed sampling) *Suppose Assumptions 5.1 and 5.2 hold. If μ is fixed and P_B is fixed or jointly balance-trained, there exist $a, t_0, C > 0$ such that $\eta_t = a/(t + t_0)$ yields*

$$\mathbb{E}\mathcal{L}_T \leq C\bar{\epsilon}^2 + \frac{C}{T + t_0}. \tag{3}$$

To convert loss convergence into terminal accuracy, we require:

Assumption 5.5 *Assume $0 < Z < \infty$. Almost surely at every iterate, policies are normalized and positive on allowed edges, and flow quantities in h are finite and positive. Forward trajectories reach a terminal state and backward trajectories reach s_0 almost surely. For FM and DB, trajectory sums of $|h|$ are integrable under both $p_{F,t}$ and $p_{B,t}$. For $\bar{B}, \bar{C}_\nu < \infty$, uniformly in t,*

$$c_{\nu,t} \leq \bar{B}, \qquad \left\| \frac{\nu_t}{\mu_{\boldsymbol{\omega}_t}} \right\|_\infty \leq \bar{C}_\nu \quad \textit{whenever } c_{\nu,t} > 0.$$

For TB, $c_{\nu,t} \leq 2$ automatically; finite DAGs also ensure the FM/DB mass bound and integrability.

Corollary 5.6 *Under Assumption 5.5, the loss bound (3) implies, for some $C > 0$,*

$$\mathbb{E}\left[\mathrm{KL}(p_{F,T}^\top \| p_R) + \mathrm{KL}(p_R \| p_{F,T}^\top)\right] \leq C\bar{\epsilon} + \frac{C}{\sqrt{T + t_0}}. \tag{4}$$

This follows from Theorem 3.2 and Theorem 5.4. See Appendix D.2 for a detailed discussion.

For a different loss function, Yu (2025) establishes best-iterate gradient convergence for FM and DB of $O(T^{-1/2})$ and $O(T^{-1/3})$, respectively, while our result provides $O(T^{-1/2})$ last-iterate global convergence for FM, DB and TB. A detailed comparison is provided in Appendix A.1.

5.2 Changing sampling: retaining net descent

When training samples follow the evolving policies, changing μ changes the weights in the objective. Even a successful balance step can therefore be partly offset by the next sampling update. Write $\Delta_t^\omega := \boldsymbol{\omega}_{t+1} - \boldsymbol{\omega}_t$. We make the following assumptions on the evolving sampling distribution.

Assumption 5.7 *For $C_\omega > 0$, almost surely at every iterate,*

$$\|\nabla_{\boldsymbol{\omega}} \mathcal{L}_t\|_2 \leq C_\omega \sqrt{2\mathcal{L}_t}.$$

Assumption 5.8 *For constants $K_\theta, K_2 \geq 0$, almost surely at every iterate,*

$$\|\mathbb{E}_t \Delta_t^\omega\|_2 \leq K_\theta \eta_t \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_t\|_2,$$
$$\mathbb{E}_t \|\Delta_t^\omega\|_2^2 \leq K_2 \eta_t^2 (\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_t\|_2^2 + \sigma^2).$$

For fixed sampling, these conditions are vacuous and the constants are zero.

Assumption 5.7 follows from bounded h and a bounded sampling score $\|\nabla_{\boldsymbol{\omega}} \log \mu_{\boldsymbol{\omega}}(o)\|_2$. Assumption 5.8 includes direct on-policy parameter copying, but modifications such as exploration mixtures, tempering, or delayed updates require their own sensitivity and update bounds. Appendix D.4 gives sufficient conditions for these two assumptions. Together with the PL inequality, these bounds leave net descent when

$$\rho := \frac{C_\omega K_\theta}{\sqrt{\lambda}} < 1.$$

Theorem 5.9 (SGD with controlled sampling) *Suppose Assumptions 5.1, 5.2, 5.7, and 5.8 hold and $\rho < 1$. If there is no separate backward update (FM, or TB/DB with fixed or jointly balance-trained P_B), then, for suitable $a, t_0 > 0$, $\eta_t = a/(t + t_0)$ gives the loss bound (3) and, under Assumption 5.5, the terminal bound (4).*

Practical interpretation of the drift condition. Theorem 5.9 identifies $\rho < 1$ as a sufficient condition for retaining the baseline rates under changing sampling. To obtain a smaller ρ, we can apply fixed exploration mixtures, tempering and experience replay to reduce C_ω, or incorporate damped updates or target updates to reduce K_θ. Refer to Appendix D.5 for a detailed discussion.

5.3 Separate backward updates: controlling the extra drift

For TB and DB, updating $\boldsymbol{\varphi}$ by another criterion changes the balance equations and need not decrease the balance loss. Define

$$z_t := \left(\mathbb{E}_t\|\boldsymbol{\varphi}_{t+1} - \boldsymbol{\varphi}_t\|_2^2\right)^{1/2}, \qquad d_t := (\mathbb{E}z_t^2)^{1/2}.$$

Similar to the sampling parameters, we make the following assumptions.

Assumption 5.10 *For $C_\varphi > 0$, almost surely at every iterate,*

$$\|\nabla_{\boldsymbol{\varphi}}\mathcal{L}_t\|_2 \le C_\varphi\sqrt{2\mathcal{L}_t}.$$

Assumption 5.11 *For constants $K_\theta, K_\varphi, K_2 \ge 0$, almost surely at every iterate,*

$$\|\mathbb{E}_t\Delta_t^\omega\|_2 \le K_\theta\eta_t\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_t\|_2 + K_\varphi z_t,$$
$$\mathbb{E}_t\|\Delta_t^\omega\|_2^2 \le K_2\left[\eta_t^2(\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_t\|_2^2 + \sigma^2) + z_t^2\right].$$

For fixed sampling, these conditions are vacuous and the constants are zero.

Theorem 5.12 (SGD with separate backward updates) *Suppose Assumptions 5.1,5.2, 5.7, 5.10, and 5.11 hold, with $\rho < 1$. For suitable $a, t_0, c, C > 0$, $\eta_t = a/(t + t_0)$ yields*

$$\mathbb{E}_t\mathcal{L}_{t+1} \le (1 - c\eta_t)\mathcal{L}_t + C\eta_t\bar{\epsilon}^2 + C\eta_t^2 + C\frac{z_t^2}{\eta_t}. \tag{5}$$

1. *If $d_t = o(\eta_t)$, then $\limsup_{t\to\infty}\mathbb{E}\mathcal{L}_t \le C\bar{\epsilon}^2$. In particular, $d_t = O(t^{-b})$ with any $b > 1$ suffices.*
2. *If $d_t = O(\eta_t^{3/2})$, equivalently $d_t = O(t^{-3/2})$, then $\mathbb{E}\mathcal{L}_T = O(\bar{\epsilon}^2 + T^{-1})$.*
3. *If $d_t = O(t^{-b})$ with $1 < b < 3/2$, then $\mathbb{E}\mathcal{L}_T = O(\bar{\epsilon}^2 + T^{-2(b-1)})$.*

Under Assumption 5.5, case (i) gives an $O(\bar{\epsilon})$ neighborhood for expected terminal symmetric KL. In cases (ii) and (iii), the expected terminal symmetric KL is bounded, respectively, by $O(\bar{\epsilon} + T^{-1/2})$ and $O(\bar{\epsilon} + T^{-(b-1)})$.

The three additive terms in equation 5 separate representation error, SGD noise, and backward-policy drift. After taking expectation, the last is Cd_t^2/η_t. For the expected loss to approach the residual neighborhood it suffices that this be $o(\eta_t)$; to retain the baseline rate it suffices that it be $O(\eta_t^2)$, the same order as SGD noise.

6 Conclusion

In this paper, we develop a unified theory that connects balance loss, approximate stationarity, and sampling accuracy for FM, DB, and TB. We bound terminal symmetric KL by the square root of the expected balance loss and characterize stationary-point quality through a tangent-space residual that vanishes for standard positive tabular models. Under suitable regularity, coverage, and conditioning assumptions, we establish finite-time convergence guarantees and quantify how sampling and backward-policy drift must be controlled to preserve them. Our findings clarify the conditions under which optimization progress leads to accurate reward-proportional sampling, marking a significant advancement in understanding the training of GFlowNets. This insight is valuable for further algorithmic development and real-world applications. Moreover, our analysis builds upon and enhances existing function-approximation and distribution-shift techniques from RL and optimization, indicating their potential applicability to a broader range of fields.

However, the guarantees outlined in this paper depend on adequate coverage, well-conditioned effective tangent geometry, and controlled drift. These factors require further verification in practical implementations, particularly when neural networks are involved. We recognize these limitations and view them as important directions for future research, despite the challenges they present.

AI USE STATEMENT

In this work, we used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to draft parts of a research paper and to edit a research paper to improve readability. We have reviewed all AI-assisted work. The mathematical claims and proofs are carefully read and checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021. URL https://jmlr.org/papers/v22/19-736.html.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27381–27394. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e614f646836aaed9f89ce58e837e2310-Paper.pdf.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *The Journal of Machine Learning Research*, 24(1):10006–10060, 2023. URL https://jmlr.org/papers/v24/22-0364.html.

Vivek S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5): 291–294, 1997. URL https://doi.org/10.1016/S0167-6911(97)90015-3.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. URL https://doi.org/10.1137/16M1080173.

Xuyang Chen and Lin Zhao. Finite-time analysis of single-timescale actor-critic. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/160adf2dc118a920e7858484b92a37d8-Abstract-Conference.html.

Tristan Deleu, Padideh Nouri, Nikolay Malkin, Doina Precup, and Yoshua Bengio. Discrete probabilistic inference as control in multi-path environments. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=3C69sU1YkK.

Tristan Deleu, Padideh Nouri, Yoshua Bengio, and Doina Precup. Relative trajectory balance is equivalent to trust-pcl, 2025. URL https://arxiv.org/abs/2509.01632.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentum-based policy gradient. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1910–1934. PMLR, 2022. URL https://proceedings.mlr.press/v151/ding22a.html.

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9827–9869. PMLR, 2023. URL https://proceedings.mlr.press/v202/fatkhullin23a.html.

Jake Fawkes and Jason Hartford. f-trajectory balance: A loss family for tuning GFlownets, generative models, and LLMs with off- and on-policy data. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sDfu0qN6Y6.

Jie Feng, Ke Wei, and Jinchi Chen. Global convergence of natural policy gradient with hessian-aided momentum variance reduction, 2024. URL https://arxiv.org/abs/2401.01084v2. Preprint, version 2.

GFNOrg. torchgfn: A PyTorch library for GFlowNets, version 2.4.1, 2026. URL https://github.com/GFNOrg/torchgfn/tree/v2.4.1. Source implementation, accessed September 21, 2026.

Timofei Gritsaev, Nikita Morozov, Sergey Samsonov, and Daniil Tiapkin. Optimizing backward policies in GFlownets via trajectory likelihood maximization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Xj66fkrlTk.

Alex Hernandez-Garcia, Nikita Saxena, Alexandra Volokhova, Michał Koziarski, Divya Sharma, Joseph D Viviano, Pierre Luc Carrier, and Victor Schmidt. gflownet, 2024. URL https://github.com/alexhernandezgarcia/gflownet.

Edward J Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Esmeralda S. Whitammer. Amortizing intractable inference in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ouj6p4ca60.

Rui Hu, Yifan Zhang, Zhuoran Li, and Longbo Huang. Beyond squared error: Exploring loss design for enhanced training of generative flow networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4NTrco82W0.

Rui Hu, Yu Chen, and Longbo Huang. Finite-time convergence analysis of actor-critic with evolving reward. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=hocdjAEAIm.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Hyosoon Jang, Yunhui Jang, Minsu Kim, Jinkyoo Park, and Sungsoo Ahn. Pessimistic backward policy for GFlownets. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L8Q21Qrjmd.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/hash/4b86abe48d358ecf194c56c69108433e-Abstract.html.

Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2144–2203. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/kaledin20a.html.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016. URL https://doi.org/10.1007/978-3-319-46128-1_50.

Vijay R. Konda and John N. Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability*, 14(2):796–819, 2004. URL https://doi.org/10.1214/105051604000000116.

MichałKoziarski, Andrei Rekesh, Dmytro Shevchuk, Almer van der Sloot, Piotr Gaiński, Yoshua Bengio, Cheng-Hao Liu, Mike Tyers, and Robert Batey. Rgfn: Synthesizable molecular generation using gflownets. In A. Globerson, L. Mackey, D. Belgrave, A. Fan,

U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 46908–46955. Curran Associates, Inc., 2024. doi: 10.52202/079017-1488. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/53704142f230054140418ecd8857f391-Paper-Conference.pdf.

Anas Krichel, Nikolay Malkin, Salem Lahlou, and Yoshua Bengio. On generalization for generative flow networks, 2024. URL https://arxiv.org/abs/2407.03105.

Salem Lahlou, Tristan Deleu, Pablo Lemos, Dinghuai Zhang, Alexandra Volokhova, Alex Hernández-García, Lena Nehale Ezzine, Yoshua Bengio, and Nikolay Malkin. A theory of continuous generative flow networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18269–18300. PMLR, 2023. URL https://proceedings.mlr.press/v202/lahlou23a.html.

Zengxiang Lei, Ananth Shreekumar, Jonathan Rosenthal, Ruoyu Song, Alvaro A. Cardenas, Daniel J. Fremont, Dongyan Xu, Satish Ukkusuri, and Z. Berkay Celik. Stable gflownets with tv monitoring and probabilistic guarantees, 2026. URL https://arxiv.org/abs/2605.01729.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous Control with Deep Reinforcement Learning. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1509.02971.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022. URL https://doi.org/10.1016/j.acha.2021.12.009.

Yanli Liu, Kaiqing Zhang, Tamer Başar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/56577889b3c1cd083b6d7b32d32f99d5-Abstract.html.

Zhen Liu, Tim Z. Xiao, Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. Efficient diversity-preserving diffusion alignment via gradient-informed GFlownets. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Aye5wL6TCn.

Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jain, Andrei Cristian Nica, Tom Bosc, Yoshua Bengio, and Nikolay Malkin. Learning GFlowNets from partial episodes for improved convergence and stability. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23467–23483. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/madan23a.html.

Idriss Malek, Aya Laajil, Abhijith Sharma, Eric Moulines, and Salem Lahlou. Loss-guided auxiliary agents for overcoming mode collapse in gflownets. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. URL https://doi.org/10.1609/aaai.v40i29.39613.

Dragos Cristian Manta, Edward J Hu, and Yoshua Bengio. GFlownets for causal discovery: an overview. In *ICML 2023 Workshop: Sampling and Optimization in Discrete Space*, 2023. URL https://openreview.net/forum?id=25D2NyVVpr.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015. URL https://www.nature.com/articles/nature14236.

Nikita Morozov, Ian Maksimov, Daniil Tiapkin, and Sergey Samsonov. Revisiting non-acyclic GFlowNets in discrete environments. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44887–44910. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/morozov25a.html.

Alex Olshevsky and Bahman Gharesifard. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2):980–1007, 2023. doi: 10.1137/22M1483335. URL https://doi.org/10.1137/22M1483335.

Ling Pan, Nikolay Malkin, Dinghuai Zhang, and Yoshua Bengio. Better training of GFlowNets with local credit and incomplete trajectories. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26878–26890. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/pan23c.html.

Yordan Raykov and Rodrigo Veiga. Information-geometric forward policy training in gflownets, 2026. URL https://arxiv.org/abs/2608.03967.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586. URL https://projecteuclid.org/euclid.aoms/1177729586.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05952.

Tiago Silva, Rodrigo Barreto Alves, Eliezer de Souza da Silva, Amauri H Souza, Vikas Garg, Samuel Kaski, and Diego Mesquita. When do GFlownets learn the right distribution? In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=9GsgCUJtic.

Tiago Silva, Amauri H Souza, Omar Rivasplata, Vikas Garg, Samuel Kaski, and Diego Mesquita. Generalization and distributed learning of GFlownets. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=PJNhZoCjLh.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Daniil Tiapkin, Nikita Morozov, Alexey Naumov, and Dmitry P Vetrov. Generative flow networks as entropy-regularized rl. In *International Conference on Artificial Intelligence and Statistics*, pp. 4213–4221. PMLR, 2024. URL https://proceedings.mlr.press/v238/tiapkin24a.html.

Esmeralda S Whitammer, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5955–5967. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/27b51baca8377a0cf109f6ecc15a0f70-Paper-Conference.pdf.

Esmeralda S. Whitammer, Salem Lahlou, Tristan Deleu, Xu Ji, Edward J Hu, Katie E Everett, Dinghuai Zhang, and Yoshua Bengio. GFlownets and variational inference. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.00580.

Tianshu Yu. Secrets of gflownets' learning behavior: A theoretical study, 2025. URL https://arxiv.org/abs/2505.02035.

Taeyoung Yun, Dinghuai Zhang, Jinkyoo Park, and Ling Pan. Learning to sample effective and diverse prompts for text-to-image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23625–23635. IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yun_Learning_to_Sample_Effective_and_Diverse_Prompts_for_Text-to-Image_Generation_CVPR_2025_paper.html.

Dinghuai Zhang, Hanjun Dai, Esmeralda S Whitammer, Aaron Courville, Yoshua Bengio, and Ling Pan. Let the flows tell: Solving graph combinatorial problems with gflownets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11952–11969. Curran Associates, Inc., 2023. doi: 10.52202/075280-0522. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/27571b74d6cd650b8eb6cf1837953ae8-Paper-Conference.pdf.

Yiheng Zhu, Jialu Wu, Chaowen Hu, Jiahuan Yan, kim hsieh, Tingjun Hou, and Jian Wu. Sample-efficient multi-objective molecular optimization with gflownets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79667–79684. Curran Associates, Inc., 2023. doi: 10.52202/075280-3487. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fbc9981dd6316378aee7fd5975250f21-Paper-Conference.pdf.

A Related Work and Technical Perspective

We review the theoretical development of GFlowNets, emphasizing results on sampling accuracy and learning, and then explain how our analysis connects to techniques in reinforcement learning and optimization.

A.1 Theoretical Progress in GFlowNets

Foundations and connections to other theories. The foundations of GFlowNets characterize reward-proportional sampling through consistent flows on a construction graph (Bengio et al., 2021; 2023). Flow Matching (FM) and Detailed Balance (DB) impose local consistency, while Trajectory Balance (TB) enforces consistency along complete generation paths (Whitammer et al., 2022). Under the corresponding support and reward-matching conditions, exact balance yields the target terminal distribution. Sub-Trajectory Balance (SubTB) interpolates between edge and complete-trajectory constraints, providing intermediate scopes for credit assignment (Madan et al., 2023). Appendix E extends our analysis to SubTB under its stated conditions. These results establish the correctness of balanced solutions; understanding an imperfectly trained, parameterized sampler requires additional analysis.

Connections to other inference and learning frameworks clarify these objectives. Whitammer et al. (2023) establish expected-gradient identities between balance objectives and variational inference under specified trajectory distributions and choices of the differentiated parameter block. Tiapkin et al. (2024) reformulate GFlowNet learning as entropy-regularized RL using a particular reward and regularizer structure. More recently, Deleu et al. (2025) identify an equivalence between Relative Trajectory Balance and the KL-regularized RL method Trust-PCL. Loss-design theory further connects regression penalties to distributional objectives: Hu et al. (2025) relate expected regression gradients to gradients of divergences between flows on minimal cuts under an appropriate sampling structure. Fawkes & Hartford (2026) develop an f-trajectory-balance family with corresponding on-policy f-divergence gradients and target-preserving off-policy minima under their conditions. Such identities illuminate optimization signals and loss design, but require further arguments to characterize parameterized stationary points or finite-time sampling accuracy.

Imbalance sensitivity and representational limits. Silva et al. (2025a) investigate how balance violations affect sampling accuracy. Their Theorem 1 bounds terminal total variation (TV) for a specified perturbation of an initially balanced network, obtained by adding and propagating flow. The bounds reveal dependence on the flow scale and target mass of downstream terminal states. This is a sensitivity result for the stated perturbation construction, rather than a general bound on arbitrary expected training loss. They also characterize expressiveness limitations of GFlowNets using 1-WL graph neural networks, propose more expressive look-ahead policies, and introduce flow consistency in subgraphs (FCS) as a distributional diagnostic with theoretical connections to TV. Their representational results concern which distributions a model class can express. Our tangent residual instead measures imbalance unexplained by local parameter derivatives at the current model.

Generalization, coverage, and statistical guarantees. Krichel et al. (2024) formalize generalization and reward-perturbation stability and empirically study generalization to longer trajectories. Their stability proposition assumes globally minimized TB, a fixed uniform backward policy, and reward functions with equal total mass; the length-generalization experiments do not establish an unrestricted extrapolation guarantee. Silva et al. (2025b) develop data-dependent PAC-Bayesian bounds using bounded FCS risk, together with oracle bounds for terminal KL and DB-based risk under their sampling and regularity assumptions. Their Proposition 4.2 also bounds terminal TV through an expected squared trajectory log-density ratio and a distribution-mismatch factor involving χ^2 divergence. Its expectation samples terminal states from a chosen distribution and then backward trajectories; the ratio uses normalized target probabilities. Our training measure can instead weight states, edges, or complete trajectories, with normalization terms eliminated by symmetrization. Thus, coverage-dependent links between average discrepancy and sampling error already exist. Their analysis motivates Subgraph Asynchronous Learning, which divides training across smaller subnetworks. Our contribution is a common occupancy-discrepancy analysis for FM, DB, and TB, and its connection to stationary-point quality and stochastic optimization. The population bounds

here complement statistical generalization guarantees: coverage alone does not justify substituting empirical loss for population loss.

Comparison with Yu (2025). Yu studies four aspects of learning behavior—convergence, sample complexity, implicit regularization, and robustness—using the following objectives[1] (interpreted under our notations):

$$\hat{L}_{\mathrm{FM}}(s) = \left(F^{\rightarrow}(s) - F^{\leftarrow}(s)\right)^2, \quad \hat{L}_{\mathrm{DB}}(s) = \left(\frac{F^{\rightarrow}(s)}{F^{\leftarrow}(s)} - 1\right)^2, \quad \hat{L}_{\mathrm{TB}}(s) = \left(\frac{F^{\rightarrow}(s)}{F^{\leftarrow}(s)} - 1\right)^2.$$

These definitions differ from the squared log-ratio objectives used in practical implementations(Hernandez-Garcia et al., 2024; GFNOrg, 2026) and in this paper. Different losses have different weighting and gradients, so their convergence statements are not interchangeable.

Under tabular parameterization, fixed sampling and fixed backward policies, Yu's Theorems 2 and 3 give $O(T^{-1/2})$ and $O(T^{-1/3})$ rates for the expected minimum squared gradient norm of FM and DB over all T iterates. In contrast, our Theorem 5.4 give $O(T^{-1/2})$ rate for the expected terminal symmetric KL of the T's iterate of FM, DB and TB.

Comparison with Lei et al. (2026). Lei et al. study loss sensitivity, showing that small terminal TV need not prevent large balance loss, and develop guarantees in the converse direction. Their Theorem 3.5 bounds TV by $1 - e^{-2c}$ when every squared TB residual is at most c^2; the corresponding FM/DB bound is $1 - e^{-2Lc}$ for maximum trajectory length L. Their Theorem 3.6 provides a probabilistic certificate using independent forward trajectories and backward trajectories initialized from the target distribution. Its global form requires target-terminal samples; restricted-target sampling yields a subgraph certificate. They also propose adaptive reference flows to stabilize training, analyzing the associated fidelity trade-off.

Relative to Lei et al., our distributional statement uses expected squared log-ratio loss under a general, possibly off-policy training measure and controls terminal symmetric KL through occupancy mass and coverage. It does not require their uniform pointwise residual condition, but it can be vacuous under poor coverage and is a population inequality rather than a finite-sample certificate. The two analyses therefore control different error metrics under different hypotheses; neither gives an unconditional tightness ordering.

Extensions of the underlying environment. Other theoretical work broadens the modeling setting. Lahlou et al. (2023) formulate generalized GFlowNets using measurable spaces, flow measures, and Markov kernels, extending balance conditions and sampling correctness to continuous and hybrid spaces under assumptions including finite absorption. Morozov et al. (2025) revisit discrete non-acyclic environments, clarifying flow definitions, the role of backward policies, and the relationship between total flow and expected trajectory length. They also analyze fixed-backward training and stability issues when the backward policy is learned. These extensions address structural questions beyond the discrete DAG setting used here; transferring our approximate-balance and optimization analysis to them requires revisiting the occupancy and trajectory assumptions.

A.2 Connections to Reinforcement Learning and Optimization

Compatible approximation, natural gradients, and approximate PL geometry. Compatible function approximation regresses advantages onto policy-score features, connecting least-squares coefficients to natural policy gradients (NPG) (Sutton et al., 1999; Kakade, 2001). For general policy classes, transferred approximation error evaluates this regression under a comparator policy's occupancy, making distribution shift part of the optimality analysis (Agarwal et al., 2021). Under Fisher nondegeneracy and bounded transferred error, Liu et al. (2020) analyze NPG through compatible regression and a comparator-policy KL potential. For momentum policy gradient, Ding et al. (2022) derive a bound on the optimality gap by an approximation floor plus a multiple of the gradient norm. Related analyses develop this route for stochastic policy gradients (Fatkhullin et al., 2023) and NPG with Hessian-aided momentum variance reduction (Feng et al., 2024).

[1] There is a typo in their definition of TB loss, where the backward probabilities are missing.

In the context of GFlowNets, the regression target is log-flow imbalance and the features are its parameter derivatives. The residual is local to the tangent space and measured under the training distribution, unlike transferred advantage error under a comparator occupancy. The feature Gram matrix is the Gauss–Newton matrix of the squared-imbalance loss with the training measure fixed, and need not be a policy Fisher. Raykov & Veiga (2026) use the forward trajectory Fisher to precondition GFlowNet balance-loss gradients. This Fisher coincides with the imbalance Gram matrix for the forward-only TB block under on-policy weighting with backward policy and source flow held fixed; their decomposition and approximation guarantees concern update geometry rather than approximation-dependent gradient domination. A lower bound on the nonzero Gram eigenvalues turns the projection identity into an approximate PL inequality: squared gradient norm controls loss above the tangent-residual floor. This also connects to nonlinear least-squares and tangent-kernel analyses (Karimi et al., 2016; Jacot et al., 2018; Liu et al., 2022).

Stochastic-gradient baselines and objective drift. The fixed-sampling part of Section 5 belongs to the classical stochastic-approximation lineage initiated by Robbins and Monro (Robbins & Monro, 1951) and developed for recursive algorithms and stochastic optimization in later treatments (Bottou et al., 2018). Conditional unbiasedness and a second-moment bound for the balance-gradient oracle, together with smoothness, give the usual expected-descent recursion. In our setting, the effective Fisher lower bound and the tangent residual turn this recursion into an approximate Polyak–Lojasiewicz inequality: the loss contracts until it reaches the parameterization residual floor, with an $O(T^{-1})$ loss rate under harmonic stepsizes. The terminal-distribution bound then takes a square root, giving the $O(T^{-1/2})$ rate in Corollary 5.6. This is a population result for the stated training measure, rather than a generic claim that every nonconvex SGD trajectory converges globally.

When the sampling law changes with the iterate, or the backward policy is updated seperately, the objective seen by the SGD oracle thus evolves during training. Two-time-scale stochastic approximation provides the classical limiting-ODE and finite-time perspectives for coupled updates (Borkar, 1997; Konda & Tsitsiklis, 2004; Kaledin et al., 2020). The controlled-gradient and controlled-sampling assumptions in Section 5.2 and Section 5.3 make the first- and second-moment effects of this drift explicit. This interpretation is related to small-gain analyses of single-timescale actor–critic methods (Olshevsky & Gharesifard, 2023; Chen & Zhao, 2023) and to finite-time analyses with an evolving reward or objective (Hu et al., 2026), but the constants here are defined for the GFlowNet sampling measure and its imbalance features.

B PROOFS FOR SECTION 3

Proof of Lemma 3.1. The chain rule for relative entropy, applied to the terminal-state map $\tau \mapsto s_{n(\tau)}$, gives

$$\begin{aligned}
\mathrm{KL}(p_F \| p_B) &= \mathrm{KL}(p_F^\top \| p_R) + \sum_{x \in \mathcal{X}} p_F^\top(x) \mathrm{KL}\big(p_F(\cdot \mid s_{n(\tau)} = x) \| p_B(\cdot \mid s_{n(\tau)} = x)\big) \\
&\geq \mathrm{KL}(p_F^\top \| p_R), \\
\mathrm{KL}(p_B \| p_F) &= \mathrm{KL}(p_R \| p_F^\top) + \sum_{x \in \mathcal{X}} p_R(x) \mathrm{KL}\big(p_B(\cdot \mid s_{n(\tau)} = x) \| p_F(\cdot \mid s_{n(\tau)} = x)\big) \\
&\geq \mathrm{KL}(p_R \| p_F^\top).
\end{aligned}$$

□

Proof of Theorem 3.2. For FM and DB, the integrability condition in Section 3 and Tonelli's theorem give

$$\sum_{o \in \mathcal{O}} (\nu_F(o) + \nu_B(o)) |h_{\boldsymbol{\theta}}(o)| = \mathbb{E}_{p_F} \sum_{o \in \tau} |h_{\boldsymbol{\theta}}(o)| + \mathbb{E}_{p_B} \sum_{o \in \tau} |h_{\boldsymbol{\theta}}(o)| < \infty.$$

This condition is automatic on a finite DAG. It justifies the signed sum interchange below by Fubini's theorem, since $|p_F(\tau) - p_B(\tau)| \leq p_F(\tau) + p_B(\tau)$. For TB, objects are complete trajectories, so no such interchange is needed. If $c_\nu = 0$, then $p_F = p_B$ and the TB claim is immediate. Otherwise, it suffices to consider finite coverage ratio and finite loss, since the claimed bound is vacuous in the remaining cases. Then

$$\sum_{\tau \in \mathcal{T}} |p_F(\tau) - p_B(\tau)| \, |h_{\boldsymbol{\theta}}(\tau)| = c_\nu \mathbb{E}_\nu |h_{\boldsymbol{\theta}}| \leq c_\nu \sqrt{2 \left\| \frac{\nu}{\mu} \right\|_\infty \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta})} < \infty,$$

by Cauchy–Schwarz and change of measure. Thus the signed sums below are absolutely convergent in every nonvacuous case.

According to Lemma 3.1, we have

$$\begin{aligned}
\mathrm{KL}(p_F^\top \| p_R) + \mathrm{KL}(p_R \| p_F^\top) &\leq \mathrm{KL}(p_F \| p_B) + \mathrm{KL}(p_B \| p_F) \\
&= \sum_{\tau \in \mathcal{T}} (p_F(\tau) - p_B(\tau)) \log \frac{p_F(\tau)}{p_B(\tau)}.
\end{aligned}$$

For FM-GFN, we have

$$\begin{aligned}
\log \frac{p_F(\tau)}{p_B(\tau)} &= \sum_{t=1}^{n(\tau)} \log \frac{F^{\rightarrow}(s_t)}{F^{\leftarrow}(s_{t-1})} + \log \frac{Z}{R(s_{n(\tau)})} \\
&= \sum_{t=1}^{n(\tau)} h_{\boldsymbol{\theta}}(s_t) + \log \frac{Z}{F^{\leftarrow}(s_0)},
\end{aligned}$$

where the last equality uses $F^{\leftarrow}(s_{n(\tau)}) = R(s_{n(\tau)})$.

For DB-GFN, we have

$$\log \frac{p_F(\tau)}{p_B(\tau)} = \sum_{t=1}^{n(\tau)} h_{\boldsymbol{\theta}}(s_{t-1} \to s_t) + \log \frac{Z}{F(s_0; \boldsymbol{\theta})}.$$

For TB-GFN, we have

$$\log \frac{p_F(\tau)}{p_B(\tau)} = h_{\boldsymbol{\theta}}(\tau) + \log \frac{Z}{\widehat{Z}(\boldsymbol{\theta})}.$$

In every case, the log-density ratio is the sum of the relevant imbalances along τ plus a term that does not depend on τ. Since $\sum_{\tau\in\mathcal{T}}(p_F(\tau)-p_B(\tau))=0$, this constant cancels in the symmetrized expression. Therefore,

$$\begin{aligned}
&\sum_{\tau\in\mathcal{T}}(p_F(\tau)-p_B(\tau))\log\frac{p_F(\tau)}{p_B(\tau)}\\
&\quad=\sum_{\tau\in\mathcal{T}}(p_F(\tau)-p_B(\tau))\sum_{o\in\tau}h_{\boldsymbol{\theta}}(o)\\
&\quad=\sum_{o\in\mathcal{O}}(\nu_F(o)-\nu_B(o))\,h_{\boldsymbol{\theta}}(o).
\end{aligned}$$

If $c_\nu=0$, the last display is zero, so the nonnegative symmetric KL divergence is also zero. Suppose $c_\nu>0$. If $\|\nu/\mu\|_\infty=+\infty$, the loss bound is immediate in the extended sense under the coverage convention preceding Theorem 3.2. Otherwise, finiteness of the ratio implies that μ covers the support of ν, and

$$\begin{aligned}
&\sum_{o\in\mathcal{O}}(\nu_F(o)-\nu_B(o))\,h_{\boldsymbol{\theta}}(o)\\
&\quad\le\sum_{o\in\mathcal{O}}|\nu_F(o)-\nu_B(o)|\,|h_{\boldsymbol{\theta}}(o)|\\
&\quad=c_\nu\sum_{o\in\mathcal{O}}\nu(o)|h_{\boldsymbol{\theta}}(o)|\\
&\quad\le c_\nu\sqrt{\sum_{o\in\mathcal{O}}\nu(o)h_{\boldsymbol{\theta}}(o)^2}\\
&\quad=c_\nu\sqrt{\sum_{o\in\mathcal{O}}\mu(o)\frac{\nu(o)}{\mu(o)}h_{\boldsymbol{\theta}}(o)^2}\\
&\quad\le c_\nu\sqrt{\left\|\frac{\nu}{\mu}\right\|_\infty\sum_{o\in\mathcal{O}}\mu(o)h_{\boldsymbol{\theta}}(o)^2}\\
&\quad=c_\nu\sqrt{2\left\|\frac{\nu}{\mu}\right\|_\infty\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta})}.
\end{aligned}$$

□

C Proofs for Section 4

All functions below are evaluated on their relevant supports, and all flows and policy probabilities are positive there.

Proof of Proposition 4.2. We show for each objective that the tabular tangent space contains the current imbalance pointwise. The claim for every μ then follows immediately. For DB and TB, first hold the backward policy fixed. If independent backward coordinates are included in $\boldsymbol{\theta}$, assign zero to their direction; the same construction remains in the enlarged tangent space.

FM. Give every edge e an independent coordinate $z_e = \log F(e)$. For a non-initial state s, the FM imbalance is

$$h(s) = \log F^{\rightarrow}(s) - \log F^{\leftarrow}(s).$$

For an incoming edge $e = (r \to s)$ and an outgoing edge $e = (s \to r)$, respectively,

$$\frac{\partial h(s)}{\partial z_e} = \frac{F(e)}{F^{\rightarrow}(s)} > 0, \qquad \frac{\partial h(s)}{\partial z_e} = -\frac{F(e)}{F^{\leftarrow}(s)},$$

where the outgoing-edge term is absent when s is terminal. Choose one incoming edge $e_s = (p(s) \to s)$ for every training state and order the states topologically. In the Jacobian submatrix formed by these columns, column e_s has a nonzero diagonal entry in row s and can have one additional nonzero entry only in the earlier row $p(s)$. This triangular submatrix has nonzero diagonal, so the Jacobian has full row rank. Its range is therefore all functions on the FM training states, including h.

DB. Use independent log-state-flow coordinates for nonterminal states and independent softmax logits for the forward policy at each nonterminal state. Let $h(s, s')$ be the DB imbalance. Set $V(x) = 0$ for terminal states and define, in reverse topological order,

$$V(s) = \mathbb{E}_{s' \sim P_F(\cdot|s)}[h(s, s') + V(s')], \tag{6}$$

$$A(s, s') = h(s, s') + V(s') - V(s). \tag{7}$$

Then

$$\mathbb{E}_{s' \sim P_F(\cdot|s)}[A(s, s')] = 0, \qquad h(s, s') = V(s) - V(s') + A(s, s').$$

A direction V in the log-state-flow coordinates contributes $V(s) - V(s')$, with $V(s') = 0$ for a terminal child. A logit direction A contributes

$$A(s, s') - \mathbb{E}_{r \sim P_F(\cdot|s)}[A(s, r)] = A(s, s').$$

The combined direction is exactly $h(s, s')$ on every edge, proving that the DB imbalance lies in the tabular tangent space.

TB. Use the coordinate $z = \log \widehat{Z}$ and independent softmax logits for the forward policy. For each edge, define

$$\ell(s, s') := \log \frac{P_F(s' \mid s)}{P_B(s \mid s')} - \begin{cases} \log R(s'), & s' \in \mathcal{X}, \\ 0, & s' \notin \mathcal{X}. \end{cases}$$

The reward term is evaluated only at terminal states. Every complete trajectory $\tau = (s_0, \ldots, s_{n(\tau)})$ has the additive imbalance

$$h(\tau) = z + \sum_{t=1}^{n(\tau)} \ell(s_{t-1}, s_t). \tag{8}$$

Set $V(x) = 0$ for every terminal state and define, in reverse topological order,

$$V(s) = \mathbb{E}_{s' \sim P_F(\cdot|s)}[\ell(s, s') + V(s')], \tag{9}$$

$$A(s, s') = \ell(s, s') + V(s') - V(s). \tag{10}$$

The advantage is centered under $P_F(\cdot \mid s)$. Since $\ell(s, s') = V(s) - V(s') + A(s, s')$, summing along any trajectory gives

$$h(\tau) = z + V(s_0) + \sum_{t=1}^{n(\tau)} A(s_{t-1}, s_t). \tag{11}$$

Choose the direction in the normalizer coordinate to be $z + V(s_0)$ and the logit direction on edge (s, s') to be $A(s, s')$. Centering makes the corresponding softmax-score direction equal to $A(s, s')$, so the resulting trajectory direction is exactly $h(\tau)$.

This argument does not require equal trajectory lengths. The value is defined once per state, which makes merged paths consistent, and assigning zero value to every terminal handles multiple terminal states. Thus the TB imbalance lies in the tabular tangent space even though the number of trajectories may exceed the number of parameters.

For FM, DB, and TB, we have proved $h_{\boldsymbol{\theta}} \in \mathsf{T}_{\boldsymbol{\theta}}$ pointwise for every $\boldsymbol{\theta}$. Hence its orthogonal residual is zero in every $L_2(\mu)$ and the supremum defining ϵ_μ is zero. □

Proof of Lemma 4.5. Fix $\boldsymbol{\theta}$ and define the linear operator

$$\Psi_{\boldsymbol{\theta}} : \mathbb{R}^d \to L_2(\mu), \qquad (\Psi_{\boldsymbol{\theta}} \boldsymbol{u})(o) = \boldsymbol{u}^\top \psi_{\boldsymbol{\theta}}(o).$$

By construction, $\mathrm{Range}(\Psi_{\boldsymbol{\theta}}) = \mathsf{T}_{\boldsymbol{\theta}}$. Its adjoint and Gram matrix are

$$\Psi^*_{\boldsymbol{\theta}} f = \mathbb{E}_\mu[\psi_{\boldsymbol{\theta}}(o) f(o)], \qquad \Psi^*_{\boldsymbol{\theta}} \Psi_{\boldsymbol{\theta}} = \mathcal{I}_\mu(\boldsymbol{\theta}).$$

The standard pseudoinverse formula for the orthogonal projection onto the range of a finite-dimensional linear operator gives

$$\Pi^\mu_{\mathsf{T}_{\boldsymbol{\theta}}} = \Psi_{\boldsymbol{\theta}} \mathcal{I}_\mu(\boldsymbol{\theta})^\dagger \Psi^*_{\boldsymbol{\theta}}. \tag{12}$$

By the square-integrability and differentiation conventions in Section 2, the operator above is bounded and differentiation may pass through the loss expectation, with μ and any separate $\boldsymbol{\varphi}$ fixed. Hence

$$\Psi^*_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}} = \mathbb{E}_\mu[\psi_{\boldsymbol{\theta}}(o) h_{\boldsymbol{\theta}}(o)] = \nabla_{\boldsymbol{\theta}} \frac{1}{2} \mathbb{E}_\mu[h_{\boldsymbol{\theta}}(o)^2] = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}).$$

Substituting this identity into Equation 12 yields

$$\Pi^\mu_{\mathsf{T}_{\boldsymbol{\theta}}} h_{\boldsymbol{\theta}} = \Psi_{\boldsymbol{\theta}} \boldsymbol{u}^*_\mu(\boldsymbol{\theta}) = \boldsymbol{u}^*_\mu(\boldsymbol{\theta})^\top \psi_{\boldsymbol{\theta}}.$$

The Moore–Penrose solution is the minimum-norm coefficient among all coefficients representing this projection. Assumption 4.4 then gives

$$\|\boldsymbol{u}^*_\mu(\boldsymbol{\theta})\|_2 \le \|\mathcal{I}_\mu(\boldsymbol{\theta})^\dagger\|_{2\to 2} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta})\|_2 \le \frac{1}{\lambda} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta})\|_2.$$

□

Proof of Theorem 4.6. If $c_\nu = 0$, then $\nu_F(o) = \nu_B(o)$ for every o, and the signed occupancy bound established in the proof of Theorem 3.2 in Appendix B implies that the symmetric KL divergence is zero. Suppose henceforth that $c_\nu > 0$, so the discrepancy distribution ν is well defined.

If $\|\nu/\mu\|_\infty = +\infty$, the claimed bound holds in the extended sense under the coverage convention preceding Theorem 3.2. Otherwise, finiteness of this ratio implies that μ covers the support of ν, so the changes of measure below are valid.

The same signed occupancy bound from Appendix B and the definition of ν give

$$\begin{aligned} &\mathrm{KL}(p_F^\top \| p_R) + \mathrm{KL}(p_R \| p_F^\top) \\ &\quad \le \sum_{o \in \mathcal{O}} \big(\nu_F(o) - \nu_B(o)\big) h_{\boldsymbol{\theta}}(o) \le c_\nu \mathbb{E}_{o\sim\nu} |h_{\boldsymbol{\theta}}(o)|. \end{aligned}$$

By Definition 4.1 and Lemma 4.5,

$$h_{\boldsymbol{\theta}}(o) = r_{\mu,\boldsymbol{\theta}}(o) + \boldsymbol{u}^*_\mu(\boldsymbol{\theta})^\top \psi_{\boldsymbol{\theta}}(o).$$

For the residual component, Cauchy–Schwarz and importance sampling yield

$$\begin{aligned}\mathbb{E}_\nu|r_{\mu,\boldsymbol{\theta}}| &\le \sqrt{\mathbb{E}_\nu[r_{\mu,\boldsymbol{\theta}}^2]} \\ &= \sqrt{\mathbb{E}_\mu\left[\frac{\nu}{\mu} r_{\mu,\boldsymbol{\theta}}^2\right]} \\ &\le \sqrt{\left\|\frac{\nu}{\mu}\right\|_\infty} \|r_{\mu,\boldsymbol{\theta}}\|_\mu \\ &\le \sqrt{\left\|\frac{\nu}{\mu}\right\|_\infty} \epsilon_\mu.\end{aligned}$$

For the tangent component, Assumption 4.3 and Lemma 4.5 imply

$$\begin{aligned}\mathbb{E}_\nu\left|\boldsymbol{u}_\mu^*(\boldsymbol{\theta})^\top \psi_{\boldsymbol{\theta}}(o)\right| &\le \|\boldsymbol{u}_\mu^*(\boldsymbol{\theta})\|_2 \mathbb{E}_\nu\|\psi_{\boldsymbol{\theta}}(o)\|_2 \\ &\le \frac{M}{\lambda}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta})\|_2.\end{aligned}$$

Combining the two component bounds proves

$$\mathrm{KL}(p_F^\top\|p_R) + \mathrm{KL}(p_R\|p_F^\top) \le c_\nu\left(\sqrt{\left\|\frac{\nu}{\mu}\right\|_\infty}\epsilon_\mu + \frac{M}{\lambda}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta})\|_2\right).$$

□

D PROOFS FOR SECTION 5

We use the parameter blocks and conditional expectations defined in Section 5. Constants denoted by C may increase between inequalities but are independent of t and $\bar{\epsilon}$.

D.1 GEOMETRY AND A COMMON DESCENT ESTIMATE

Proof of Lemma 5.3. The setup in Section 5 makes $\boldsymbol{u} \mapsto \boldsymbol{u}^\top \nabla_{\boldsymbol{\theta}} h$ a bounded map from $\mathbb{R}^d$ into $L_2(\mu)$. Its range is finite-dimensional and hence closed, even when $\mathcal{O}$ is infinite. The interchange of differentiation and summation assumed in Section 2 gives $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}} = \mathbb{E}_\mu[h \nabla_{\boldsymbol{\theta}} h]$, so the projection identity of Lemma 4.5 still applies. At the current parameters, write $\boldsymbol{G} = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}$, $I = \mathcal{I}_\mu(\boldsymbol{\theta})$, and let Π be the $L_2(\mu)$ tangent projection. Lemma 4.5 gives the coefficient $\boldsymbol{u}^* = I^\dagger \boldsymbol{G}$, so

$$\|\Pi h\|_\mu^2 = \boldsymbol{u}^{*\top} I \boldsymbol{u}^* = \boldsymbol{G}^\top I^\dagger \boldsymbol{G}.$$

Since $r = h - \Pi h$ is orthogonal to Πh, Pythagoras and Assumption 5.2 yield

$$2\mathcal{L}_{\mathcal{O}} = \|r\|_\mu^2 + \boldsymbol{G}^\top I^\dagger \boldsymbol{G} \le \bar{\epsilon}^2 + \frac{\|\boldsymbol{G}\|_2^2}{\lambda}.$$

□

A common one-step estimate. We prove a recursion under the conditions of Theorem 5.12. Without a separate backward update, set $z_t = 0$ and omit the backward sensitivity condition; the sampling bounds then reduce to Assumption 5.8. With fixed sampling, omit the sampling block and set its constants to zero. Write $x_t = \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_t\|_2$, $\Delta_t^\theta = \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t$, and $\Delta_t^\varphi = \boldsymbol{\varphi}_{t+1} - \boldsymbol{\varphi}_t$. Joint smoothness gives

$$\begin{aligned}\mathbb{E}_t \mathcal{L}_{t+1} \le \mathcal{L}_t - \eta_t x_t^2 + \langle \nabla_{\boldsymbol{\omega}} \mathcal{L}_t, \mathbb{E}_t \Delta_t^\omega \rangle + \langle \nabla_{\boldsymbol{\varphi}} \mathcal{L}_t, \mathbb{E}_t \Delta_t^\varphi \rangle \\ + \frac{L}{2} \mathbb{E}_t \big[\|\Delta_t^\theta\|_2^2 + \|\Delta_t^\omega\|_2^2 + \|\Delta_t^\varphi\|_2^2 \big]. \end{aligned} \tag{13}$$

All gradients on the right are $\mathcal{F}_t$-measurable; no independence between updates is needed. The sensitivity and update bounds, together with $\|\mathbb{E}_t \Delta_t^\varphi\|_2 \le z_t$, bound the two drift terms by

$$C_\omega K_\theta \eta_t \sqrt{2\mathcal{L}_t}\, x_t + (C_\omega K_\varphi + C_\varphi) \sqrt{2\mathcal{L}_t}\, z_t.$$

Lemma 5.3 implies $\sqrt{2\mathcal{L}_t} \le \bar{\epsilon} + x_t/\sqrt{\lambda}$. With $\delta = 1 - \rho > 0$, the one-step bound becomes

$$\mathbb{E}_t \mathcal{L}_{t+1} \le \mathcal{L}_t - \delta \eta_t x_t^2 + C \eta_t \bar{\epsilon} x_t + C(\bar{\epsilon} + x_t) z_t + C \eta_t^2 (x_t^2 + \sigma^2) + C z_t^2.$$

Young's inequality bounds the mixed terms as

$$C \eta_t \bar{\epsilon} x_t + C(\bar{\epsilon} + x_t) z_t \le \frac{\delta}{4} \eta_t x_t^2 + C \eta_t \bar{\epsilon}^2 + C \frac{z_t^2}{\eta_t}.$$

Choose $\eta_* > 0$ sufficiently small that $\eta_* \le 1$ and the quadratic term $C\eta_t^2 x_t^2$ is at most $\delta \eta_t x_t^2/4$ whenever $\eta_t \le \eta_*$. Absorbing $C z_t^2$ and the fixed noise constant gives

$$\mathbb{E}_t \mathcal{L}_{t+1} \le \mathcal{L}_t - \frac{\delta}{2} \eta_t x_t^2 + C \eta_t \bar{\epsilon}^2 + C \eta_t^2 + C \frac{z_t^2}{\eta_t}. \tag{14}$$

Using $x_t^2 \ge \lambda(2\mathcal{L}_t - \bar{\epsilon}^2)$ yields

$$\mathbb{E}_t \mathcal{L}_{t+1} \le (1 - c\eta_t) \mathcal{L}_t + C \eta_t \bar{\epsilon}^2 + C \eta_t^2 + C \frac{z_t^2}{\eta_t}, \qquad c = \lambda(1 - \rho) > 0. \tag{15}$$

Decrease η_* if needed so that $c\eta_* \le 1/2$. Finite second moments of the updates imply integrability of x_t^2 by conditional Jensen and (1), and of z_t^2 by definition. Starting from $\mathbb{E}\mathcal{L}_0 < \infty$, (15) therefore gives finite expected loss at each finite time and justifies taking total expectation.

D.2 Fixed and controlled sampling rates

A scalar comparison. Choose $a > 1/c$, $t_0 \geq \max\{1, a/\eta_*\}$, and $\eta_t = a/(t+t_0)$. For any $0 < \alpha \leq 1$, a nonnegative sequence satisfying

$$u_{t+1} \leq (1 - c\eta_t)u_t + \frac{C}{(t+t_0)^{1+\alpha}}$$

obeys $u_t \leq D/(t+t_0)^\alpha$ for sufficiently large D. Indeed, choose $D \geq t_0^\alpha u_0$ and $(ca - \alpha)D \geq C$. With $k = t + t_0$, induction and convexity of $k \mapsto k^{-\alpha}$ give

$$u_{t+1} \leq \frac{D}{k^\alpha} - \frac{caD - C}{k^{1+\alpha}} \leq \frac{D}{k^\alpha} - \frac{\alpha D}{k^{1+\alpha}} \leq \frac{D}{(k+1)^\alpha}.$$

The contraction factors are nonnegative by the choice of t_0.

Proofs of Theorems 5.4 and 5.9. In (15), set $z_t = 0$; fixed sampling also gives $\rho = 0$. Take total expectation and choose $A \geq C/c$ for the coefficient of $\eta_t \bar{\epsilon}^2$. Then $u_t = (\mathbb{E}\mathcal{L}_t - A\bar{\epsilon}^2)_+$ satisfies

$$u_{t+1} \leq (1 - c\eta_t)u_t + C\eta_t^2.$$

The scalar comparison with $\alpha = 1$ proves (3). FM is included because both policies are induced by $\boldsymbol{\theta}$; joint backward training is included because its parameters belong to the same descent block. These arguments require finite-dimensional parameters, not a finite object space. The tabular zero-residual specialization in Section 5 is restricted to finite DAGs. □

Conversion to terminal accuracy. We first justify the bound of Theorem 3.2 under Assumption 5.5 on a countable DAG. Fix the current parameters and suppress t. Almost-sure termination makes p_F, p_B probability distributions on finite complete trajectories, with backward terminal marginal p_R. For FM and DB the integrability condition is

$$\mathbb{E}_{\tau\sim p_F} \sum_{o\in\tau} |h(o)| + \mathbb{E}_{\tau\sim p_B} \sum_{o\in\tau} |h(o)| < \infty.$$

Tonelli's theorem identifies this quantity with $\sum_o (\nu_F(o) + \nu_B(o))|h(o)|$. It therefore justifies the signed sum interchange in Appendix B. The telescoping identities hold on each finite trajectory, and their trajectory-independent constants cancel because p_F, p_B are normalized. Thus the same signed bound holds for terminal symmetric KL.

For TB, there is no interchange between trajectory and object sums. The identity $\log(p_F/p_B) = h + \log(Z/\widehat{Z})$ and normalization give the signed bound directly. Its absolute convergence follows from

$$\sum_\tau |p_F(\tau) - p_B(\tau)|\,|h(\tau)| \leq c_\nu \sqrt{2\bar{C}_\nu \mathcal{L}_\mathcal{O}} < \infty$$

when $c_\nu > 0$; when $c_\nu = 0$, $p_F = p_B$. In all three cases, the change of measure and Cauchy–Schwarz argument of Theorem 3.2 is valid on the countable object space. With the deterministic bounds from Assumption 5.5, Jensen's inequality now gives

$$\begin{aligned}\mathbb{E}\big[\mathrm{KL}(p_{F,T}^\top \| p_R) + \mathrm{KL}(p_R \| p_{F,T}^\top)\big] &\leq \bar{B}\sqrt{2\bar{C}_\nu}\,\mathbb{E}\sqrt{\mathcal{L}_T} \\ &\leq \bar{B}\sqrt{2\bar{C}_\nu\,\mathbb{E}\mathcal{L}_T}.\end{aligned}$$

This also holds when $c_{\nu,T} = 0$, since the terminal discrepancy is then zero. Substituting (3) proves (4); the same conversion applies to the bounds below.

For TB, $c_{\nu,t} \leq 2$ follows from normalization alone. For FM and DB, Tonelli's theorem gives

$$c_{\nu,t} \leq \sum_o (\nu_{F,t}(o) + \nu_{B,t}(o)) = \mathbb{E}_{\tau\sim p_{F,t}} n(\tau) + \mathbb{E}_{\tau\sim p_{B,t}} n(\tau).$$

A deterministic uniform bound on this sum of expectations, conditional on the current parameters, suffices for $\bar{B}$. Finiteness at each iterate alone does not give a uniform constant. On a finite DAG of maximum path length H, one may take $\bar{B} = 2H$ for FM and DB; the absolute-imbalance integrability condition also holds automatically.

D.3 Separate backward updates

Proof of Theorem 5.12. Taking total expectation in (15) uses $d_t^2 = \mathbb{E}z_t^2 = \mathbb{E}\|\boldsymbol{\varphi}_{t+1} - \boldsymbol{\varphi}_t\|_2^2$. With the stepsize and residual shift from Appendix D.2,

$$u_{t+1} \le (1 - c\eta_t)u_t + C\eta_t^2 + C\frac{d_t^2}{\eta_t}, \qquad u_t = (\mathbb{E}\mathcal{L}_t - A\bar{\epsilon}^2)_+. \tag{16}$$

If $d_t = o(\eta_t)$, the additive term is $\eta_t\xi_t$ for a deterministic sequence $\xi_t \to 0$. For any $\varepsilon > 0$, eventually $\xi_t \le c\varepsilon$, so

$$(u_{t+1} - \varepsilon)_+ \le (1 - c\eta_t)(u_t - \varepsilon)_+.$$

Since $\sum_t \eta_t = \infty$, the product of contraction factors tends to zero. Hence $\limsup_t u_t \le \varepsilon$, and letting $\varepsilon \downarrow 0$ proves the first claim. If $d_t = O(\eta_t^{3/2})$, the additional forcing is $O(\eta_t^2)$; the scalar comparison with $\alpha = 1$ proves the second claim, absorbing any finite initial segment into the constant. The terminal bounds follow from Appendix D.2. □

Intermediate decay rates. Under the hypotheses of Theorem 5.12, if $d_t = O(t^{-b})$ with $1 < b < 3/2$, then

$$\mathbb{E}\mathcal{L}_T = O(\bar{\epsilon}^2 + T^{-2(b-1)}).$$

Indeed, the forcing in (16) is $O((t+t_0)^{-2} + (t+t_0)^{1-2b})$. The scalar comparison applies with $\alpha = 2(b-1) \in (0,1)$. Under Assumption 5.5, the expected terminal symmetric KL is consequently $O(\bar{\epsilon} + T^{-(b-1)})$.

D.4 Sufficient conditions for sensitivity, sampling, and coverage

The following examples verify individual conditions used above; they do not replace the remaining hypotheses of the theorems. The derivative identities use the interchange convention in Section 2, with the external arguments made explicit in Section 5. For a countable object space, a sufficient condition for that interchange is local domination of each derivative of the summand $\mu_{\boldsymbol{\omega}}(o)h_{\boldsymbol{\theta},\boldsymbol{\varphi}}(o)^2/2$ by a summable sequence on a parameter neighborhood, with finite loss at one point of that neighborhood. The mean value theorem and dominated convergence then justify differentiating the sum in each relevant coordinate. The square-integrability of h and its $\boldsymbol{\theta}$-derivatives also ensures that the tangent operator in Appendix C is bounded and that $\mathbb{E}_\mu[h\nabla_{\boldsymbol{\theta}}h]$ is finite by Cauchy–Schwarz. On a finite object space these conditions hold for continuously differentiable, finite imbalances and sampling probabilities on a parameter neighborhood.

Direct sensitivity bounds. Holding sampling fixed, Cauchy–Schwarz gives

$$\|\nabla_{\boldsymbol{\varphi}}\mathcal{L}_{\mathcal{O}}\|_2 = \|\mathbb{E}_\mu[h\nabla_{\boldsymbol{\varphi}}h]\|_2 \le \sqrt{2\mathcal{L}_{\mathcal{O}}}\left(\mathbb{E}_\mu\|\nabla_{\boldsymbol{\varphi}}h\|_2^2\right)^{1/2}.$$

A uniform bound on the second factor supplies C_φ. For DB, $\nabla_{\boldsymbol{\varphi}}h$ is the negative backward log-policy score on an edge; for TB it is the sum of these negative scores along the trajectory. Uniformly bounded edge scores suffice for DB. For TB they suffice if $\mathbb{E}_{\tau\sim\mu_{\boldsymbol{\omega}_t}}n(\tau)^2$ is uniformly bounded; this follows in particular from bounded trajectory lengths. This second moment under the training distribution is distinct from the occupancy mass condition under $p_{F,t}, p_{B,t}$. For neural policies, these bounds concern derivatives with respect to the network parameters and do not follow from softmax normalization alone.

For differentiable sampling probabilities on a fixed positive support, let $s_\omega(o) = \nabla_{\boldsymbol{\omega}}\log\mu_{\boldsymbol{\omega}}(o)$. If $\|s_\omega(o)\|_2 \le S_\omega$ and $|h(o)| \le B_h$ uniformly, then

$$\|\nabla_{\boldsymbol{\omega}}\mathcal{L}_{\mathcal{O}}\|_2 = \frac{1}{2}\|\mathbb{E}_\mu[h^2 s_\omega]\|_2 \le \frac{S_\omega B_h}{2}\mathbb{E}_\mu|h| \le \frac{S_\omega B_h}{2}\sqrt{2\mathcal{L}_{\mathcal{O}}}.$$

Thus $C_\omega = S_\omega B_h/2$ suffices. Bounded sampling scores alone do not ensure this square-root bound for unbounded imbalances.

Affine sampling updates. Suppose $\Delta_t^\omega = A_t\Delta_t^\theta + B_t\Delta_t^\varphi$, where A_t, B_t are $\mathcal{F}_t$-measurable and their operator norms are bounded by k_θ, k_φ. Conditional Jensen and the balance oracle give

$$\|\mathbb{E}_t\Delta_t^\omega\|_2 \le k_\theta\eta_t\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_t\|_2 + k_\varphi z_t,$$
$$\mathbb{E}_t\|\Delta_t^\omega\|_2^2 \le 2k_\theta^2\eta_t^2(\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_t\|_2^2 + \sigma^2) + 2k_\varphi^2 z_t^2.$$

Hence Assumption 5.11 holds with $K_\theta = k_\theta$, $K_\varphi = k_\varphi$, and $K_2 = 2\max\{k_\theta^2, k_\varphi^2\}$; omit the backward term when that block is absent. This includes parameter copying $\boldsymbol{\omega}_t = \boldsymbol{\theta}_t$, but the sensitivity bound and $\rho < 1$ remain separate requirements. A nonlinear parameter map must satisfy the conditional-mean bound directly: a Lipschitz increment bound alone controls second moments, not the mean's Taylor remainder.

D.5 Sampling stabilisation: exploration, replay, and stale targets

This subsection gives representative ways to verify the hypotheses of Theorem 5.9, relating familiar exploration, replay, and target-update mechanisms to its constants. Omit any separate backward update and write $x_t = \|\nabla_{\boldsymbol{\theta}}\mathcal{L}_t\|_2$. Collect lag, refresh, and perturbation terms in

$$\|\mathbb{E}_t\Delta_t^\omega\|_2 \le K_\theta\eta_t x_t + b_t, \tag{17}$$
$$\mathbb{E}_t\|\Delta_t^\omega\|_2^2 \le K_2\eta_t^2(x_t^2 + \sigma^2) + v_t. \tag{18}$$

When $b_t = v_t = 0$, these are exactly Assumption 5.8; otherwise they must be absorbed into its bounds; retaining them as forcing terms requires a corresponding extension of the descent analysis.

Sampling sensitivity. Uniform-random exploration is familiar from ε-greedy policies (Sutton & Barto, 2018); GFlowNet exploration also uses noisy policies and tempering (Malek et al., 2026). We analyse an ε-noisy mixture over complete training objects: choose a fixed exploration distribution q with probability ε_{exp}, and the current sampler otherwise. For $0 \le \varepsilon_{\text{exp}} < 1$, let

$$\mu_{\boldsymbol{\omega}}^{(\varepsilon)} = (1-\varepsilon_{\text{exp}})p_{\boldsymbol{\omega}} + \varepsilon_{\text{exp}}q, \qquad \mathcal{L}_\varepsilon = (1-\varepsilon_{\text{exp}})\mathcal{L}_{\text{fresh}} + \varepsilon_{\text{exp}}\mathcal{L}_q. \tag{19}$$

If $\|\nabla_{\boldsymbol{\omega}}\mathcal{L}_{\text{fresh}}\|_2 \le C_{\omega,\text{fresh}}\sqrt{2\mathcal{L}_{\text{fresh}}}$, then

$$\|\nabla_{\boldsymbol{\omega}}\mathcal{L}_\varepsilon\|_2 \le \sqrt{1-\varepsilon_{\text{exp}}}\, C_{\omega,\text{fresh}}\sqrt{2\mathcal{L}_\varepsilon}. \tag{20}$$

Thus fixed exploration changes C_ω, while the underlying update rule determines K_θ and K_2. For exploration mixed at each action, the induced trajectory law is a product of local mixtures; its sensitivity follows from the resulting trajectory score.

Temperature scaling, also called softmax or Boltzmann exploration (Sutton & Barto, 2018; Malek et al., 2026), flattens a sampling distribution at temperatures above one. For logits $\ell_{\boldsymbol{\omega}}(o)$ and fixed $\tau > 0$ such that $Z_\tau(\boldsymbol{\omega}) := \sum_o \exp(\ell_{\boldsymbol{\omega}}(o)/\tau) < \infty$ on the parameter region under consideration, and with differentiation permitted through this normalizing sum,

$$p_{\boldsymbol{\omega}}^{(\tau)}(o) = \frac{\exp(\ell_{\boldsymbol{\omega}}(o)/\tau)}{\sum_{o'}\exp(\ell_{\boldsymbol{\omega}}(o')/\tau)}, \qquad \nabla_{\boldsymbol{\omega}}\log p_{\boldsymbol{\omega}}^{(\tau)}(o) = \frac{1}{\tau}\left(\nabla_{\boldsymbol{\omega}}\ell_{\boldsymbol{\omega}}(o) - \mathbb{E}_{p_{\boldsymbol{\omega}}^{(\tau)}}\nabla_{\boldsymbol{\omega}}\ell_{\boldsymbol{\omega}}(o')\right). \tag{21}$$

Both requirements are automatic for finitely many continuously differentiable logits. On a countable object space, a uniform local bound $\|\nabla_{\boldsymbol{\omega}}\ell_{\boldsymbol{\omega}}(o)\|_2 \le G$ together with $Z_\tau(\boldsymbol{\omega}_0) < \infty$ at a reference point suffices locally: on a convex neighborhood of radius r about $\boldsymbol{\omega}_0$,

$$e^{\ell_{\boldsymbol{\omega}}(o)/\tau} \le e^{Gr/\tau}e^{\ell_{\boldsymbol{\omega}_0}(o)/\tau}, \qquad \left\|\nabla_{\boldsymbol{\omega}}e^{\ell_{\boldsymbol{\omega}}(o)/\tau}\right\|_2 \le \frac{G}{\tau}e^{Gr/\tau}e^{\ell_{\boldsymbol{\omega}_0}(o)/\tau}.$$

These summable bounds justify both normalization and its derivative. Finiteness at one temperature does not imply finiteness at a higher temperature: logits $\ell(n) = -2\log n$, $n \ge 1$, are normalizable at $\tau = 1$ but not at $\tau = 2$.

If $\|\nabla_{\boldsymbol{\omega}}\ell_{\boldsymbol{\omega}}(o)\|_2 \le G$ and $|h(o)| \le B_h$ uniformly, the score bound $2G/\tau$ and Appendix D.4 give $C_{\omega,\tau} = B_hG/\tau$. For actionwise tempering on trajectories of length at most H, summing local scores gives the bound B_hHG/τ. Fixed tempering leaves K_θ, K_2 determined by the update rule.

Experience replay reuses previously collected data, as in DQN (Mnih et al., 2015); prioritized experience replay further changes their sampling weights (Schaul et al., 2016). Freezing both buffer

contents and weights gives fixed sampling for the associated empirical objective. For a fixed mixture of replay and fresh samples, $\mu_{\boldsymbol{\omega}} = (1-\beta)q + \beta p_{\boldsymbol{\omega}}$, the same argument gives

$$C_{\omega,\mathrm{mix}} = \sqrt{\beta}\, C_{\omega,\mathrm{fresh}}, \qquad 0 \le \beta \le 1. \tag{22}$$

Coverage and the effective Fisher lower bound must be recomputed under the resulting distribution. Changing buffer contents, exploration weights, temperatures, or mixture weights adds drift. A differentiable parameterisation of these changes allows their increments to enter b_t, v_t; direct changes of the sampling measure can instead be bounded in the loss. For replay, the oracle is evaluated against the empirical or mixture objective actually used.

Update size and stale targets. If the sampler follows damped balance increments,

$$\boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}_t + \alpha(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t), \qquad 0 < \alpha \le 1,$$

then

$$\mathbb{E}_t \Delta_t^{\omega} = -\alpha\eta_t \nabla_{\boldsymbol{\theta}} \mathcal{L}_t, \qquad \mathbb{E}_t \|\Delta_t^{\omega}\|_2^2 \le \alpha^2 \eta_t^2 (x_t^2 + \sigma^2),$$

so $K_\theta = \alpha$ and $K_2 = \alpha^2$. This provides a direct way to limit sampler motion by scaling each new balance increment.

Soft target updates, often called Polyak or exponential moving average (EMA) updates, are used in DDPG (Lillicrap et al., 2016). Here we apply that update rule to a sampling copy; the cited RL target networks are used to form learning targets. With $\boldsymbol{\omega}_{t+1} = (1-\alpha)\boldsymbol{\omega}_t + \alpha\boldsymbol{\theta}_{t+1}$ and $\boldsymbol{e}_t = \boldsymbol{\theta}_t - \boldsymbol{\omega}_t$, the increment satisfies

$$\Delta_t^{\omega} = \alpha \boldsymbol{e}_t + \alpha \Delta_t^{\theta}, \qquad \|\mathbb{E}_t \Delta_t^{\omega}\|_2 \le \alpha \|\boldsymbol{e}_t\|_2 + \alpha\eta_t x_t. \tag{23}$$

The tracking condition $\|\boldsymbol{e}_t\|_2 \le \kappa\eta_t x_t$ reduces this to $K_\theta = \alpha(1+\kappa)$ and $K_2 = 2\alpha^2(1+\kappa^2)$. Polyak averaging alone does not provide this tracking condition.

Periodic hard target updates in DQN (Mnih et al., 2015) motivate a lazy sampling copy, held fixed between synchronisations. Let I_t indicate a copy step and $\boldsymbol{b}_t = \boldsymbol{\theta}_t - \boldsymbol{\omega}_t$. Then

$$\begin{aligned} \Delta_t^{\omega} &= I_t(\boldsymbol{b}_t + \Delta_t^{\theta}), \qquad \|\mathbb{E}_t \Delta_t^{\omega}\|_2 \le I_t(\|\boldsymbol{b}_t\|_2 + \eta_t x_t), \\ \mathbb{E}_t \|\Delta_t^{\omega}\|_2^2 &\le 2I_t \|\boldsymbol{b}_t\|_2^2 + 2I_t \eta_t^2 (x_t^2 + \sigma^2). \end{aligned} \tag{24}$$

For a fixed stale copy $\boldsymbol{\omega}_t = \boldsymbol{\theta}_{t-m}$, $\Delta_t^{\omega} = \Delta_{t-m}^{\theta}$. A bound on $\boldsymbol{b}_t$ or on the past update in terms of the current x_t is therefore required before these rules can be represented by finite K_θ, K_2. In particular, the period m does not imply $K_\theta = 1/m$. Conditionally zero-mean additive update noise preserves the first-order bound and can be accounted for by increasing K_2, v_t; biased noise also contributes to b_t.

E Extension to Sub-Trajectory Balance

Sub-Trajectory Balance (SubTB; Madan et al. 2023) interpolates between single-edge and complete-trajectory credit assignment. This appendix extends the distributional, stationarity, and convergence results of Sections 3–5 using the additive structure of its imbalance. Coverage, tangent geometry, and drift remain explicit assumptions. Throughout this appendix, the training-object space in the unified notation of Section 2 is $\mathcal{O} = \mathcal{O}_{\mathrm{sub}}$.

Objective and sampling measure. Let $\mathcal{O}_{\mathrm{sub}}$ be the set of nonempty finite directed subtrajectories, including single edges and complete trajectories. Use a positive state flow $F(s;\boldsymbol{\theta})$ on nonterminal states and set $F(x;\boldsymbol{\theta}) = R(x)$ on terminal states. For $\sigma = (s_i, \dots, s_j)$ with length $\ell(\sigma) = j - i \geq 1$, define

$$\begin{aligned} F^{\rightarrow}(\sigma;\boldsymbol{\theta}) &:= F(s_i;\boldsymbol{\theta}) \prod_{k=i}^{j-1} P_F(s_{k+1} \mid s_k;\boldsymbol{\theta}), \\ F^{\leftarrow}(\sigma;\boldsymbol{\theta},\boldsymbol{\varphi}) &:= F(s_j;\boldsymbol{\theta}) \prod_{k=i}^{j-1} P_B(s_k \mid s_{k+1};\boldsymbol{\varphi}), \\ h_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\sigma) &:= \log \frac{F^{\rightarrow}(\sigma;\boldsymbol{\theta})}{F^{\leftarrow}(\sigma;\boldsymbol{\theta},\boldsymbol{\varphi})} = \sum_{k=i}^{j-1} h^{\mathrm{DB}}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(s_k, s_{k+1}). \end{aligned} \tag{25}$$

The last equality follows by telescoping the intermediate state flows. A single edge gives DB, and a complete trajectory gives TB with $\widehat{Z} = F(s_0;\boldsymbol{\theta})$. As in the main text, independently specified backward parameters are held fixed under a $\boldsymbol{\theta}$ derivative, whereas jointly balance-trained backward parameters are part of $\boldsymbol{\theta}$. We suppress fixed parameter arguments below and write $h_{\boldsymbol{\theta}}$ and h when no ambiguity can arise.

For a complete trajectory $\tau = (s_0, \dots, s_n)$, normalized length weights can be written as

$$w^{\gamma}_{ij}(\tau) := \frac{\gamma^{j-i}}{\sum_{0 \leq a < b \leq n} \gamma^{b-a}}, \qquad 0 \leq i < j \leq n, \quad \gamma > 0.$$

Here γ denotes the SubTB weighting parameter, distinct from the effective Fisher lower bound λ. Given a trajectory sampling law $q_{\boldsymbol{\omega}}$, the induced normalized distribution on subtrajectories is

$$\mu_{\boldsymbol{\omega}}(\sigma) := \mathbb{E}_{\tau \sim q_{\boldsymbol{\omega}}} \sum_{0 \leq i < j \leq n(\tau)} w^{\gamma}_{ij}(\tau) \mathbf{1}\{\tau_{i:j} = \sigma\}. \tag{26}$$

Its loss is the unified balance objective $\mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}, \boldsymbol{\omega}; \boldsymbol{\varphi}) = \frac{1}{2}\mathbb{E}_{\mu_{\boldsymbol{\omega}}}[h^2_{\boldsymbol{\theta},\boldsymbol{\varphi}}]$, equivalently the expected weighted sum of subtrajectory losses. For each fixed finite trajectory, $\gamma \downarrow 0$ gives the uniform average of its edge losses and $\gamma \to \infty$ gives its complete-trajectory loss. The induced sampling weights, as well as the imbalance formula, matter for these DB/TB limits. Passing a limit through an expectation on an unbounded trajectory space requires an integrable dominating function. The results below also allow other normalized subtrajectory distributions, including those constructed from incomplete episodes. For a fixed sampling law, write $\mu = \mu_{\boldsymbol{\omega}}$ and suppress the fixed external arguments of $\mathcal{L}_{\mathcal{O}}$. Sampling is held fixed when differentiating the balance loss, as in Section 5.

E.1 From SubTB loss to terminal-distribution accuracy

Fix a deterministic, parameter-independent rule $\mathcal{P}$ that partitions every complete trajectory into consecutive nonempty blocks:

$$0 = i_0 < i_1 < \cdots < i_m = n(\tau), \qquad \mathcal{P}(\tau) = \{\tau_{i_{r-1}:i_r} : 1 \leq r \leq m\}.$$

This rule is used only to organize the proof; the training distribution may still average over all overlapping subtrajectories. For the forward and backward trajectory laws of Section 3, define the block occupancies

$$\nu_F(\sigma) := \sum_{\tau:\, \sigma \in \mathcal{P}(\tau)} p_F(\tau), \qquad \nu_B(\sigma) := \sum_{\tau:\, \sigma \in \mathcal{P}(\tau)} p_B(\tau).$$

In this subsection, ν_F, ν_B, and the quantities below implicitly refer to this fixed rule $\mathcal{P}$. Set

$$c_\nu := \sum_{\sigma \in \mathcal{O}} |\nu_F(\sigma) - \nu_B(\sigma)|, \qquad \nu(\sigma) := \frac{|\nu_F(\sigma) - \nu_B(\sigma)|}{c_\nu}$$

when $c_\nu > 0$. The zero-denominator conventions are those of Theorem 3.2. Since every block contains at least one edge,

$$c_\nu \le \mathbb{E}_{p_F}|\mathcal{P}(\tau)| + \mathbb{E}_{p_B}|\mathcal{P}(\tau)| \le \mathbb{E}_{p_F} n(\tau) + \mathbb{E}_{p_B} n(\tau). \tag{27}$$

Proposition E.1 (SubTB loss and terminal mismatch) *Suppose the positive forward and backward policies induce probability laws p_F, p_B on complete trajectories, $0 < Z < \infty$, and $F(s_0)$ is finite and positive. Assume finite expected block counts and integrability of $\sum_{\sigma \in \mathcal{P}(\tau)} |h(\sigma)|$ under both laws. If $c_\nu = 0$, then*

$$\mathrm{KL}(p_F^\top \| p_R) + \mathrm{KL}(p_R \| p_F^\top) = 0.$$

Otherwise, if $\|\nu/\mu\|_\infty < \infty$ and $\mathcal{L}_\mathcal{O} < \infty$, then

$$\begin{aligned} &\mathrm{KL}(p_F^\top \| p_R) + \mathrm{KL}(p_R \| p_F^\top) \\ &\quad \le \sum_{\sigma \in \mathcal{O}} (\nu_F(\sigma) - \nu_B(\sigma)) h(\sigma) \\ &\quad \le c_\nu \sqrt{2 \left\| \frac{\nu}{\mu} \right\|_\infty \mathcal{L}_\mathcal{O}}. \end{aligned} \tag{28}$$

Without finite coverage, the loss bound is understood as vacuous.

Proof The additive identity in equation 25 and the partition property give

$$\sum_{\sigma \in \mathcal{P}(\tau)} h(\sigma) = \log \frac{F(s_0) \prod_{k=0}^{n(\tau)-1} P_F(s_{k+1} \mid s_k)}{R(s_{n(\tau)}) \prod_{k=0}^{n(\tau)-1} P_B(s_k \mid s_{k+1})}.$$

Thus $\log(p_F(\tau)/p_B(\tau))$ is this sum plus $\log(Z/F(s_0))$. The latter is constant in τ and cancels in the symmetric trajectory KL. Integrability permits interchanging trajectory and block sums, so

$$\mathrm{KL}(p_F \| p_B) + \mathrm{KL}(p_B \| p_F) = \sum_{\sigma \in \mathcal{O}} (\nu_F(\sigma) - \nu_B(\sigma)) h(\sigma).$$

Lemma 3.1 gives the first inequality and the case $c_\nu = 0$. For $c_\nu > 0$, the last sum is at most $c_\nu \mathbb{E}_\nu |h|$. Cauchy–Schwarz and change of measure give the second inequality. □

Single-edge partitions recover the DB occupancy construction, while a single block per trajectory recovers the TB construction and gives $c_\nu \le 2$. Intermediate partitions give the same identity. Summing all overlapping subtrajectories would count an edge (s_k, s_{k+1}) exactly $(k+1)(n-k)$ times, so those occurrence counts cannot replace block occupancies in this proof. The training objective may nevertheless average all overlapping subtrajectories through equation 26. Different fixed partitions can have different coverage ratios. The single-edge partition is available for positive all-length weights, subject to coverage by the trajectory sampling law; the complete-trajectory partition is natural at the TB endpoint. For restricted or incomplete sampling, the proposition is informative only when the chosen partition has finite discrepancy coverage.

E.2 From SubTB stationarity to optimality

Use the score, tangent space, residual, and effective Fisher matrix from Section 4, now with the SubTB imbalance h:

$$\begin{aligned} \psi_{\boldsymbol{\theta}}(\sigma) &:= \nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta}}(\sigma), & \mathsf{T}_{\boldsymbol{\theta}} &:= \{\boldsymbol{u}^\top \psi_{\boldsymbol{\theta}}(\cdot) : \boldsymbol{u} \in \mathbb{R}^d\}, \\ r_{\mu,\boldsymbol{\theta}} &:= (I - \Pi^\mu_{\mathsf{T}_{\boldsymbol{\theta}}}) h_{\boldsymbol{\theta}}, & \mathcal{I}_\mu(\boldsymbol{\theta}) &:= \mathbb{E}_\mu[\psi_{\boldsymbol{\theta}} \psi_{\boldsymbol{\theta}}^\top]. \end{aligned}$$

Let $\epsilon_\mu := \sup_{\boldsymbol{\theta} \in \Theta} \|r_{\mu,\boldsymbol{\theta}}\|_\mu$. Assume finite loss and score second moments and justify differentiation under the expectation, for example by local summable domination on a countable object space. These requirements hold pointwise for finite positive smooth tabular models.

Proposition E.2 (SubTB stationarity bound) *Under the hypotheses of Proposition E.1, suppose* $c_\nu > 0$, $\|\nu/\mu\|_\infty < \infty$, *and*

$$\mathbb{E}_\nu \|\psi_{\boldsymbol{\theta}}(\sigma)\|_2 \le M, \qquad \|\mathcal{I}_\mu(\boldsymbol{\theta})^\dagger\|_{2\to 2} \le \frac{1}{\lambda}.$$

Then

$$\mathrm{KL}(p_F^\top \| p_R) + \mathrm{KL}(p_R \| p_F^\top) \le c_\nu \left(\sqrt{\left\| \frac{\nu}{\mu} \right\|_\infty} \epsilon_\mu + \frac{M}{\lambda} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}\|_2 \right). \tag{29}$$

Proof As in Lemma 4.5, the normal equations give

$$\boldsymbol{u}_\mu^*(\boldsymbol{\theta}) = \mathcal{I}_\mu(\boldsymbol{\theta})^\dagger \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}, \qquad \|\boldsymbol{u}_\mu^*(\boldsymbol{\theta})\|_2 \le \lambda^{-1} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{O}}\|_2.$$

The projection identity is $\Pi_{\mathsf{T}_{\boldsymbol{\theta}}}^{\mu} h_{\boldsymbol{\theta}} = (\boldsymbol{u}_\mu^*)^\top \psi_{\boldsymbol{\theta}}$, so $h_{\boldsymbol{\theta}} = r_{\mu,\boldsymbol{\theta}} + (\boldsymbol{u}_\mu^*)^\top \psi_{\boldsymbol{\theta}}$. Finite coverage transfers the residual from $L_2(\mu)$ to the occupancy discrepancy. Applying the signed bound in Proposition E.1 and bounding the two components separately gives

$$\begin{aligned} \mathrm{KL}(p_F^\top \| p_R) + \mathrm{KL}(p_R \| p_F^\top) &\le c_\nu \mathbb{E}_\nu |h_{\boldsymbol{\theta}}| \\ &\le c_\nu \left(\sqrt{\left\| \frac{\nu}{\mu} \right\|_\infty} \|r_{\mu,\boldsymbol{\theta}}\|_\mu + M \|\boldsymbol{u}_\mu^*\|_2 \right). \end{aligned}$$

The claim follows from the definition of ϵ_μ. If $c_\nu = 0$, Proposition E.1 already gives zero terminal mismatch. □

Proposition E.3 (Tabular zero residual for SubTB) *On a finite DAG, independent nonterminal log-state-flows and independent forward softmax logits give* $h_{\boldsymbol{\theta}} \in \mathsf{T}_{\boldsymbol{\theta}}$ *pointwise at every positive parameter setting. Therefore* $\epsilon_\mu = 0$ *for every sampling measure* μ. *The statement also holds with independent backward coordinates jointly trained by the balance loss.*

Proof For the current DB edge imbalances, the construction in Appendix C defines $V(x) = 0$ on terminal states and

$$V(s) = \mathbb{E}_{s' \sim P_F(\cdot|s)}[h^{\mathrm{DB}}(s, s') + V(s')], \qquad A(s, s') = h^{\mathrm{DB}}(s, s') + V(s') - V(s).$$

The advantage has zero conditional mean under P_F. Give the log-state-flow coordinates direction V and the forward-logit coordinates direction A. This single direction $\boldsymbol{u}$ satisfies $\boldsymbol{u}^\top \nabla h^{\mathrm{DB}}(s, s') = h^{\mathrm{DB}}(s, s')$ on every edge. Summing along any subtrajectory gives

$$h(\sigma) = V(s_i) - V(s_j) + \sum_{k=i}^{j-1} A(s_k, s_{k+1})$$

and, by equation 25,

$$\boldsymbol{u}^\top \nabla h(\sigma) = \sum_{k=i}^{j-1} \boldsymbol{u}^\top \nabla h^{\mathrm{DB}}(s_k, s_{k+1}) = h(\sigma).$$

The endpoint term $V(s_j)$ remains when the subtrajectory ends at an internal state. The same state-wise construction works for overlapping segments and different lengths. Independent backward coordinates, if present, receive direction zero. Thus $h_{\boldsymbol{\theta}}$ belongs pointwise to $\mathsf{T}_{\boldsymbol{\theta}}$ and the projection residual vanishes under every μ. □

Forward-looking SubTB. For the fixed-energy parameterization of Pan et al. (2023), write $F(s; \boldsymbol{\theta}) = e^{-E(s)} \widetilde{F}(s; \boldsymbol{\theta})$, with $E(x) = -\log R(x)$ and $\widetilde{F}(x) = 1$ on terminals. Direct substitution gives

$$h(\sigma) = \log \frac{\widetilde{F}(s_i; \boldsymbol{\theta}) \prod_{k=i}^{j-1} P_F(s_{k+1} \mid s_k)}{\widetilde{F}(s_j; \boldsymbol{\theta}) \prod_{k=i}^{j-1} P_B(s_k \mid s_{k+1})} + E(s_j) - E(s_i). \tag{30}$$

For a single edge this is FL-DB. Along a longer segment the intermediate energy evaluations cancel, leaving its endpoint difference. The effective terminal flow remains $R(x)$, so the distributional proof

is unchanged. With E fixed, the independent coordinates $\log F(s) = \log \widetilde{F}(s) - E(s)$ differ by a translation with identity derivative, and Proposition E.3 applies to the translated tabular model.

For a neural log-flow $\widetilde{f}_{\boldsymbol{\theta}}(s) = \log \widetilde{F}(s; \boldsymbol{\theta})$, the energy term contributes no parameter derivative. If only the log-flow output is shifted, with the policy maps and μ fixed, the score features, tangent space, and Fisher are unchanged at a fixed parameter. However, the imbalance being projected changes and need not lie in that tangent space. Along actual training, the iterates and sampling law can change, so the residual and uniform conditioning assumptions must still be checked for $\widetilde{f}_{\boldsymbol{\theta}} - E$. The given energy is distinct from our imbalance-dependent tangent witness V. Learned or externally changing energies add derivative or drift terms and require a separate analysis.

E.3 Convergence of stochastic SubTB training

Let $\mathcal{L}_t = \mathcal{L}_{\mathcal{O}}(\boldsymbol{\theta}_t, \boldsymbol{\omega}_t; \boldsymbol{\varphi}_t)$. The balance update differentiates the pointwise SubTB loss while holding sampling and any separate backward parameters fixed. The following conditions specify exactly which parts of Section 5 must be checked for this objective:

1. The SubTB oracle satisfies equation 1; the initial expected loss is finite, and the loss is jointly smooth on the admissible update region as in Assumption 5.1. Loss derivatives can be interchanged with expectations, and the required update and score second moments are finite.
2. Almost surely along training, $\|r_{\mu_{\boldsymbol{\omega}_t}, \boldsymbol{\theta}_t}\|_{\mu_{\boldsymbol{\omega}_t}} \le \bar{\epsilon}$ and $\|\mathcal{I}_{\mu_{\boldsymbol{\omega}_t}}(\boldsymbol{\theta}_t)^\dagger\|_{2\to 2} \le 1/\lambda$, with deterministic uniform constants.
3. For changing sampling, the SubTB loss satisfies Assumptions 5.7 and 5.8, with $C_\omega K_\theta/\sqrt{\lambda} < 1$. With a separate backward update, also require Assumption 5.10 and use Assumption 5.11 instead of Assumption 5.8. Fixed sampling omits the sampling block and has $\rho = 0$.
4. For terminal-distribution conclusions, the hypotheses of Proposition E.1 hold at every iterate for one fixed partition rule, and $c_{\nu,t} \le \bar{B}$, $\|\nu_t/\mu_{\boldsymbol{\omega}_t}\|_\infty \le \bar{C}_\nu$ whenever $c_{\nu,t} > 0$, with deterministic finite bounds.

In particular, the last item includes integrability of the absolute block-imbalance sum under both current trajectory laws. Finite DAGs ensure finite block counts and pointwise integrability, but do not automatically give a uniform coverage or conditioning bound.

Corollary E.4 (SubTB training rates) *Under items 1–3, with fixed or jointly balance-trained P_B, suitable $a, t_0 > 0$ and $\eta_t = a/(t + t_0)$ give*

$$\mathbb{E}\mathcal{L}_T = O(\bar{\epsilon}^2 + T^{-1}).$$

If item 4 also holds, then

$$\mathbb{E}[\mathrm{KL}(p_{F,T}^\top \| p_R) + \mathrm{KL}(p_R \| p_{F,T}^\top)] = O(\bar{\epsilon} + T^{-1/2}). \tag{31}$$

For a separately updated backward policy, define z_t, d_t as in Section 5. Under items 1–3,

$$\mathbb{E}_t \mathcal{L}_{t+1} \le (1 - c\eta_t)\mathcal{L}_t + C\eta_t \bar{\epsilon}^2 + C\eta_t^2 + C\frac{z_t^2}{\eta_t} \tag{32}$$

for suitable $c, C > 0$. Consequently $d_t = o(\eta_t)$ gives $\limsup_t \mathbb{E}\mathcal{L}_t = O(\bar{\epsilon}^2)$; $d_t = O(\eta_t^{3/2})$ preserves the $O(\bar{\epsilon}^2 + T^{-1})$ loss bound; and $d_t = O(t^{-b})$, $1 < b < 3/2$, gives $O(\bar{\epsilon}^2 + T^{-2(b-1)})$. Under item 4, these imply, respectively, an $O(\bar{\epsilon})$ terminal neighborhood, an $O(\bar{\epsilon} + T^{-1/2})$ terminal bound, and an $O(\bar{\epsilon} + T^{-(b-1)})$ terminal bound.

Proof Projection and Pythagoras give

$$2\mathcal{L}_t = \|r_{\mu_{\boldsymbol{\omega}_t}, \boldsymbol{\theta}_t}\|_{\mu_{\boldsymbol{\omega}_t}}^2 + (\nabla_{\boldsymbol{\theta}} \mathcal{L}_t)^\top \mathcal{I}_{\mu_{\boldsymbol{\omega}_t}}(\boldsymbol{\theta}_t)^\dagger \nabla_{\boldsymbol{\theta}} \mathcal{L}_t \le \bar{\epsilon}^2 + \lambda^{-1} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_t\|_2^2.$$

The common descent proof in Appendix D.1 uses only this inequality, the unbiased oracle, joint smoothness, and the specified sensitivity and update bounds. Applying it to $\mathcal{L}_{\mathcal{O}}$ proves equation 32, with $z_t = 0$ when no separate backward update is present. The scalar recursion arguments in Appendix D give the stated rates with the same stepsize choices. Finally, Proposition E.1 and Jensen's inequality yield $\mathbb{E}[\mathrm{KL}(p_{F,T}^\top \| p_R) + \mathrm{KL}(p_R \| p_{F,T}^\top)] \le \bar{B}\sqrt{2\bar{C}_\nu \mathbb{E}\mathcal{L}_T}$. □

Length moments and the limits of the extension. For a separately parameterized backward policy,

$$\nabla_{\boldsymbol{\varphi}} h(\sigma) = -\sum_{k=i}^{j-1} \nabla_{\boldsymbol{\varphi}} \log P_B(s_k \mid s_{k+1}; \boldsymbol{\varphi}).$$

If each edge score has norm at most S_B, then $\mathbb{E}_{\mu_{\boldsymbol{\omega}_t}} \|\nabla_{\boldsymbol{\varphi}} h\|_2^2 \le S_B^2 \mathbb{E}_{\mu_{\boldsymbol{\omega}_t}} \ell(\sigma)^2$. A uniform second-moment bound on sampled subtrajectory lengths therefore supplies the backward sensitivity constant through Cauchy–Schwarz. Analogous bounded endpoint log-flow scores and forward edge scores control the balance-score second moment. These moments concern the training law $\mu_{\boldsymbol{\omega}_t}$; they are distinct from block-count and integrability conditions under $p_{F,t}, p_{B,t}$. Bounded imbalance and sampling scores give the sampling sensitivity bound as in Appendix D.4. On countable spaces, local summable domination of derivatives is one sufficient way to justify the required interchanges of differentiation and expectation.

Changing γ affects coverage, conditioning, and noise through $\mu_{\boldsymbol{\omega}_t}$; these constants need not be uniform in γ. Tabular zero residual alone does not ensure convergence. Fixed-energy FL-SubTB in equation 30 obeys the same conditional guarantees, and replay and Polyak targets retain the qualifications of Appendix D.5.