基于部分回合评估 GFlowNet，以实现稳定灵活的策略训练

Puhua Niu德州农工大学，大学城，德克萨斯州 77843，美国
niupuhua.123@tamu.edu

Shili Wu德州农工大学，大学城，德克萨斯州 77843，美国
shiliwu@tamu.edu

Xiaoning Qian
德克萨斯州A&M大学，大学城，TX 77843，美国 布鲁克海文国家实验室 厄普顿，NY 11973，美国 xqian@tamu.edu

arXiv:2603.01047v1 [cs.LG] 1 Mar 2026

摘要

生成流网络（GFlowNets）旨在通过将生成过程解释为有向无环图中的轨迹，学习能够高效采样组合候选项的策略。在基于价值的训练流程中，目标是使学习策略的流与目标策略的估计流在部分回合上达到平衡，从而隐式地鼓励最小化策略散度。基于策略的策略通过交替估计策略散度和更新策略来进行训练，但在有向无环图下可靠地估计散度仍是一项重大挑战。本文通过证明流平衡也能提供一种有原则的策略评估器来衡量散度，从而衔接了这两种观点；此外，本文提出了一种基于部分回合的评估平衡目标，用于学习该评估器。合成任务和现实世界任务上的实验表明，评估平衡不仅提高了基于策略的训练的可靠性，还通过无缝支持参数化后向策略并允许集成离线数据收集技术，增强了其灵活性。我们的代码可在 github.com/niupuhua1234/Sub-EB 获取。

1 引言

生成流网络（GFlowNets）是组合空间 $\mathcal{X}$上的生成模型，例如以特定方式组织节点和边所形成的图，或按特定顺序由字母表构成的字符串。GFlowNets 旨在采样 $x \in \mathcal{X}$，其概率为 $\propto R(x)$，其中 $R(x)$是非负评分函数。该任务具有挑战性，因为 $|\mathcal{X}|$ 可能过大而无法计算归一化常数 $Z^* := \sum_{x\in\mathcal{X}} R(x)$，并且由于组合特性，分布模态可能高度孤立，不利于高效探索。在 GFlowNets（Bengio 等人，2021；2023年）中，生成或采样 $x \in \mathcal{X}$被分解为增量轨迹（轨迹），这些轨迹从空状态出发，经过中间状态，并在 x 处结束，将其作为期望的终止状态。这些轨迹 $\tau \in \mathcal{T}$可视为有向无环图（DAG）上的路径，状态为$s \in \mathcal{S}$，边为$(s \to s') \in \mathcal{E}$。轨迹的正测度（未归一化概率）被视为沿该有向无环图的流 的数量，而 $R(x)$是终止于 x的轨迹的总流量，因此采样轨迹将以概率 $\propto R(x)$终止于 x。

GFlowNet 训练问题的核心可以理解为：最小化前向轨迹分布 $P_F(\tau)$（由前向策略 $\pi_F(s'|s)$ 诱导）相对于后向轨迹分布 $P_B(\tau)$ 的差异：$= P_B(\tau|x)R(x)/Z^*$，该分布由给定的后向策略 $\pi_B(s|s')$ 和 $R(x)$ 诱导（ Bengio 等人，2023 年；Malkin 等人，2022b ）。其动机在于，在实际应用中，序列生成必须以正向方式进行。此外，边缘化-

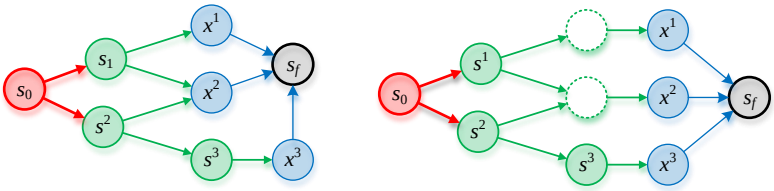


图 1：有向无环图（左）及其分层版本（右）的示意图。虚线圆圈表示将图转换为分级有向无环图时添加的虚拟状态。

$P^*(x) = \sum_{\tau|x} P_B(\tau)$ 对所有 $x \in \mathcal{X}$ 显然成立，因此任何后向策略都可用于定义目标后向轨迹分布。由于评估 Z^* 以及由此得到的归一化分布 $P_B(\tau)$ 被认为难以处理，直接优化 $P_F(\tau)$ 与 $P_B(\tau)$ 之间的某种分布散度并不可行。为规避这一问题，基于价值的方法将分布匹配问题重新表述为流匹配问题，并利用流值的平衡条件来推导训练目标。基本条件 $\mathbb{E}_{P_\mathcal{D}(s)}[\log \pi_F(s'|s)F(s)] = \mathbb{E}_{P_\mathcal{D}(s)}[\log \pi_B(s|s')F(s')]$ 意味着边流值 $F(s \to s') = \pi_F(s'|s)F(s)$ 与反向的目标边流值相匹配，其中 $P_\mathcal{D}(s)$ 是由数据收集策略 $\pi_\mathcal{D}(s'|s)$ 诱导的边缘状态分布。这一基本条件引出了多种训练目标，这些目标隐式地促进前向分布与后向分布之间的对齐。这些目标涵盖从边级表述到子轨迹级变体，提供了对目标流的多样化估计以及对流不平衡的量化（Bengio 等人，2023年；Madan 等，2023；Malkin等人，2022a）。或者，基于策略的方法（Malkin 等人，2022b；Niu 等，2024；Zimmermann 等人，2022）引入评估函数 $V(s)$，用于前向策略 $\pi_F(\cdot|s)$，该函数近似从状态 s出发的前向子轨迹（部分回合）分布与终止于 s的后向子轨迹分布之间的 Kullback-Leibler（KL）散度。然后，$V(s)$ 被用于更新前向策略 $\pi_F(\cdot|s)$。尽管基于 $V(s)$ 的策略更新已得到充分研究，如何可靠地学习 $V(s)$ 仍是一个开放性问题。

鉴于利用流平衡条件的基于价值的方法取得了成功，并且状态流函数 $F(s)$ 表示经过状态 s 的轨迹流总量，而评估函数 $V(s)$ 则表示状态 s 处的分布散度，这两者之间存在类比，因此探究这两个量之间的关系颇有意义。已知，最优前向策略和状态流函数可以由流平衡条件$\mathbb{E}_{\pi_F}[\log F(s)\pi_F(s'|s)] = \mathbb{E}_{\pi_F}[\log F(s')\pi_B(s|s')]$唯一确定。我们发现，对于任意固定的 π_F，满足该条件的状态流解与从状态 s 出发的精确 KL 散度一致，也就是真实评估函数 π_F。这为推导基于平衡条件的目标函数奠定了基础，从而可靠地学习 $V(s)$。我们的贡献如下：

- 我们建立了关于状态流函数 F与评估函数 V的平衡条件之间的联系。固定 π_F后，log F的期望平衡条件可直接导出 V的一个充分条件，我们称之为子轨迹评估平衡（Sub-EB）条件。
- 我们提出了 Sub-EB 目标，以可靠地学习评估函数 V，其中子轨迹（部分回合）作为流平衡的基本单位。
- 在模拟和真实世界数据集上的实验结果，包括超网格建模、生物与分子序列设计以及贝叶斯网络结构学习，证明了 Sub-EB 用于策略评估的有效性和可靠性。

2 预备知识

我们将 GFlowNets 的有向无环图限制为分层版本，因为任何有向无环图都可以通过添加虚拟非终止状态等价地转换为分层版本（Malkin 等人，2022b）。在有向无环图 $\mathcal{G} := (\mathcal{S}, \mathcal{E})$中，元素 $s \in \mathcal{S}$表示一个状态，元素 $(s \to s') \in \mathcal{E}(\subseteq \mathcal{S} \times \mathcal{S})$表示一条有向边。我们将时间范围的索引集定义为 $[H] := \{0, \ldots, H\}$。无环意味着状态空间 $\mathcal{S}$

可以划分为互不相交的子空间：$\mathcal{S}_0, \ldots, \mathcal{S}_{H+1}$，其中 $\mathcal{S}_h$ 的每个元素表示为 s_h，其中 $h \in [H+1]$。分层还意味着动作只能从 $\mathcal{S}_h$ 转移到 $\mathcal{S}_{h+1}$。初始状态和最终状态都是唯一的，即 $S_0 = \{s_0\}$ 和 $S_{H+1} = \{s_f\}$。终止状态集合 $S_H := \mathcal{X}$ 与评分函数 $R : \mathcal{X} \to \mathbb{R}^+$相关联。此外，（完整）轨迹集合定义为 $\mathcal{T} := \{\tau = (s_0 \to \cdots \to s_f) | \forall (s \to s') \in \tau : (s \to s') \in \mathcal{E}\}$。对于 $i < j \in [H+1]$，$\tau_{i:j}$、$\tau_{i:}$ 和 $\tau_{i:}$ 分别表示从 s_i 到 s_j、从 s_i 到 s_f以及从 s_0 到s_j的子轨迹（部分回合）。将 $\mathcal{T}$上的自然 σ-代数记为 $\sigma(\mathcal{T})$，对于任意事件$\Gamma \in \sigma(\mathcal{T})$，有 $P(\Gamma) := \sum_{\tau \in \Gamma} P(\tau)$。图 1 展示了一个示例性有向无环图及其分层版本。

GFlowNet 训练旨在使前向轨迹分布 $P_F(\tau) := \prod_{i=0}^{H} \pi_F(s_{h+1}|s_h)$ 与后向轨迹分布 $P_B(\tau) := \prod_{i=0}^{H} \pi_B(s_h|s_{h+1})$ 保持一致，其中 π_F和 π_B 分别是前向和后向马尔可夫策略（转移分布），并假设对于任意 $\tau \in \mathcal{T}$，它们满足 $P_F(\tau) > 0$和 $P_B(\tau) > 0$。相应地，$P_F(\tau_{i,j}|s_i) = \prod_{h=i}^{j-1} \pi_F(s_{h+1}|s_h)$，且 $P_B(\tau_{i,j}|s_j) = \prod_{h=i}^{j-1} \pi_B(s_h|s_{h+1})$。除了对于任意 $x \in \mathcal{X}$，$\pi_B(x|s_f) := R(x)/Z^*$之外，中间边上的 $\pi_B(s|s')$可以是任意的，因为 $P_B(x) = \sum_{\tau|x} P_B(\tau) = R(x)/Z^*$ 始终成立。因此，满足对所有 $\tau \in \mathcal{T}$ 均有 $P_F^*(\tau) = P_B(\tau)$ 的最优前向策略 π_F^*，会诱导其边缘分布$P_{F^*}(x) = R(x)/Z^*$，从而实现 GFlowNet 的最终目标。此外，在马尔可夫假设下（Bengio 等人，2023），与 π_B 对应的最优前向策略 π_F^* 是唯一的。

2.1 基于价值的训练

传统的变分推断（VI）方法通过最小化 KL 散度 $\mathcal{L}_{KL}(\theta) := D_{\mathrm{KL}}(P_F(\tau;\theta) \| P_B(\tau))$，但由于 Z^* 被认为难以处理，因此这种方法不可行（Malkin 等人，2022b）。为绕开难以处理的 Z^*，基于价值的方法引入了一个未归一化的马尔可夫分布 $F(\tau) := P_F(\tau)Z$，称为（轨迹）流（Bengio 等人，2023年），并引入总流量Z。因此，对于任意事件 $\Gamma \in \sigma(\mathcal{T})$，其流量为 $F(\Gamma) := \sum_{\tau \in \Gamma} P(\tau)$。特别地，边流量和状态流量分别为 $F(s \to s') := \sum_{\tau \ni (s \to s')} F(\tau)$ 和 $F(s) := \sum_{s'} F(s \to s')$。此外，$F(s_0) = F(s_f) = Z$，且 $\pi_F(s'|s) = F(s \to s')/F(s)$。类似地，最优流量为$F^*(\tau) := P_B(\tau)Z^* = P(\tau|x)R(x)$。因此，GFlowNets 的最优性条件可转化为 $\forall \tau \in \mathcal{T}: F(\tau) = F^*(\tau)$。得益于马尔可夫假设，该条件可以放宽并重新表述为子轨迹平衡（Sub-TB）条件（Madan 等，2023年）。它要求$\forall i < j \in [H+1]$:

$$\mathbb{E}_{P_{\mathcal{D}}(\tau_{i:j})}\left[\log\left(F(s_i)P_F(\tau_{i:j}|s_i)\right)\right] = \mathbb{E}_{P_{\mathcal{D}}(\tau_{i:j})}\left[\log\left(F(s_j)P_B(\tau_{i:j}|s_j)\right)\right], \tag{1}$$

其中，对于 $x \in \mathcal{X}$，$F(s_f)P_B(x|s_f) := R(x)$，而 $P_{\mathcal{D}}(\tau_{i,j}) = \sum_{\tau \ni \tau_{i,j}} P_{\mathcal{D}}(\tau)$ 由离线数据收集策略 $\pi_{\mathcal{D}}$ 诱导。假设对于任意 $\tau \in \mathcal{T}$，均有 $P_{\mathcal{D}}(\tau) > 0$ (Malkin 等人，2022b)。若不作此假设，一种更严格的替代方案是像 Madan 等（2023）那样，对每条单独的子轨迹强制施加流平衡。上述条件可以解释为：一条子轨迹的流量值应与反向上的目标流量值相匹配，后者表示期望的（近似）流量值。

利用平衡条件，基于价值的方法通常使用 Sub-TB 目标，将参数化策略 $\pi_F(s'|s;\theta)$ 和状态流量 $F(s;\theta)$ 优化至其最优解 $\pi_F^*(s'|s)$ 和 $F^*(s)$。目标函数定义如下：

$$\mathcal{L}_F(\theta) := \mathbb{E}_{P_{\mathcal{D}}(\tau)}\Big[\sum_{\tau_{i:j}} w_{j-i}\left(\delta_F(\tau_{i:j};\theta)\right)^2\Big], \quad \delta_F(\tau_{i:j};\theta) := \log \frac{P_F(\tau_{i:j}|s_i;\theta)F(s_i;\theta)}{P_B(\tau_{i:j}|s_j;\theta)F(s_j;\theta)}, \tag{2}$$

其中，w_{j-i} 表示由 $j - i$ 条边组成的子轨迹所对应的非零权重系数。为便于实际进行基于梯度的优化，$\mathcal{L}_F(\theta)$ 近似为 $\mathcal{L}_F(\theta)$: $\frac{1}{K}\sum_{\tau \in \mathcal{D}}[\sum_{\tau_{i:j} \in \tau} w_{j-i}\delta_F\ \tau_{i:j}\ \theta\ ]$, $\mathcal{D} = \{\tau^k\ \ \tau^k \sim P_{\mathcal{D}}\ \tau\ \}_{k=1}^{K}$ (;),

其中::()为样本集。

2.2 基于策略的训练

GFlowNets 的基于策略的训练方法（Niu 等，2024年）类似于强化学习（RL）中的策略梯度算法，旨在像传统的 VI 方法那样最小化 KL 散度 $\mathcal{L}_{KL}(\theta)$。这些方法遵循**演员-评论家**框架（Sutton 和 Barto，2018年；Haarnoja 等，2018年）。

评论器 在每次训练迭代中，首先计算行动者 π_F 的评论器（真实评估函数） $V^\dagger$，以捕捉子轨迹上的 KL 散度所体现的策略差异，从而便于计算整体轨迹散度。对于任意 $h \in [H]$， $V^\dagger(s_h;\theta)$ 定义如下：[1]

$$\mathbb{E}_{P_F(\tau_{h:}|s_h;\theta)}\Big[\sum_{i=h}^{H} R(s_i \to s_{i+1};\theta)\Big] = \log F^*(s_h) - D_{\mathrm{KL}}(P_F(\tau_{h:}|s_h;\theta)\|P_B(\tau_{h:}|s_h)) \tag{3}$$

其中 $R(s_i \to s_{i+1};\theta) := \log \frac{\widetilde{\pi}_B(s_i|s_{i+1})\,\pi_F(s_{i+1}|s_i;\theta)}{\widetilde{\pi}_B(x|s_f) := R(x)}$。如附录中的（21）所示，这一等式很容易验证（Niu 等，2024）。由于 $|\mathcal{T}|$ 可能非常庞大，通常无法对 $V^\dagger$ 进行解析计算，因此使用带有可学习参数的评估函数 V 对 $V^\dagger$ 进行建模。下一节将讨论 V 的学习目标。

行动者为最小化 $\mathcal{L}_{KL}(\theta)$，注意到 $\nabla_\theta V^\dagger(s_0;\theta) = -\nabla_\theta \mathcal{L}_{KL}(\theta)$，因为 $F^*(s_0)(= Z^*)$ 是一个常数。因此， $-V^\dagger(s_0;\theta)$ 可以作为 KL 散度的替代项。进一步应用强化学习中的策略梯度定理（Agarwal 等人，2021）， $\nabla_\theta V^\dagger(s_0;\theta)$ 可以简化，并用 V 表示为近似形式：

$$\mathbb{E}_{P_F(\tau)}\Big[\sum_{h=0}^{H} A^\gamma(s_h \to s_{h+1})\nabla_\theta \log \pi_F(s_{h+1}|s_h;\theta)\Big], \tag{4}$$

其中 $A^\gamma(s_h \to s_{h+1}) = \sum_{i=h}^{H} \gamma^{i-h}(R(s_i \to s_{i+1}) + V(s_{i+1}) - V(s_i))$ 。当超参数 γ 从 0 增加到 1 时，$A^\gamma(s_h \to s_{h+1})$ 会从 $R(s_h \to s_{h+1}) + V(s_{h+1}) - V(s_h)$ 变化为 $\sum_{i=h}^{H} R(s_i \to s_{i+1}) - V(s_h)$。因此， γ 使随机梯度下降（Schulman 等人，2016年；2017年）能够在 **方差-偏差** 之间进行权衡，这与 Sub-TB 方法中的 w 类似。

3 子轨迹评估平衡

如前几节所述，评估函数 V 在基于策略的方法中发挥着核心作用。本节首先介绍 Sub-EB 条件，该条件刻画了 π_F 与 V 之间的关系，与 π_F 和 F 之间的子轨迹平衡条件非常相似。随后，我们给出用于学习 V 的 Sub-EB 目标。此外，我们还将这些结果扩展到 π_B，并给出相应的条件和目标，以及反向评估函数 W。

3.1 前向策略的平衡条件与目标

给定一个前向策略 π_F，使得对于所有 $\tau \in \mathcal{T}$， $P_F(\tau) > 0$，其对应评估函数 V 的子轨迹评估平衡（Sub-EB）条件可写为 $\forall i < j \in [H+1]$:

$$\mathbb{E}_{P_F(\tau_{i:j})}\Big[\log\big(P_F(\tau_{i:j}|s_i)\exp V(s_i)\big)\Big] = \mathbb{E}_{P_F(\tau_{i:j})}\Big[\log\big(P_B(\tau_{i:j}|s_j)\exp V(s_j)\big)\Big], \tag{5}$$

其中，对于 $x \in \mathcal{X}$， $P_B(x|s_f)\exp V(s_f) := R(x)$。该条件可重新表述为：$\mathbb{E}_{P_F(s_i,s_j)}[V(s_i) - V(s_j)] = \mathbb{E}_{P_F(\tau_{i:j})}\Big[\log \frac{P_B(\tau_{i:j}|s_j)\,P_F(\tau_{i:j}|s_i)}{}\Big] = \mathcal{H}(P_F(s_i)) - \mathcal{H}(P_F(s_j)) - \mathbb{E}_{P_F(s_j)}[D_{\mathrm{KL}}(P_F(\tau_{i,j}|s_j)\|P_B(\tau_{i,j}|s_j))]$,，其中 $\mathcal{H}(\cdot)$ 表示分布的熵。因此，直观而言，该条件要求 s_i 和 s_j 处学习到的散度之差，与这两个状态之间子轨迹上的真实散度相匹配。

定理 3.1。设 V 是定义在 $\mathcal{S}$上的任意评估函数， F^* 是由后向策略 π_B诱导的最优流量。给定一个前向策略 π_F，

$$\forall h \in [H] : V(s_h) = \log F^*(s_h) - D_{\mathrm{KL}}(P_F(\tau_{h:}|s_h)\|P_B(\tau_{h:}|s_h)), \tag{6}$$

当且仅当 V满足 ***Sub-EB*** 条件 ***(5)***。

相应的证明见附录 A.1。

定理 3.2。设 F 是定义在 $\mathcal{S}$上的任意状态流函数， π_F 是任意前向策略，并且$(F^*、\ \pi_F^*)$ 是由后向策略 π_B诱导的最优流量和前向策略。那么，

$$\forall h \in [H] : \log F(s_h) = \log F^*(s_h) - D_{\mathrm{KL}}(P_F^*(\tau_{h:}|s_h)\|P_B(\tau_{h:}|s_h)) \tag{7}$$

且当且仅当 $(F、\ \pi_F)$ 满足 ***Sub-TB*** 条件 ***(1)*** 时， $\pi_F = \pi_F^*$

[1] Niu 等人（2024）考虑使用一个额外的总流估计器 $Z(\theta)$来缩小 $V^\dagger$的幅度。为简化符号表示，我们将把 Z 整合到我们的方法中的讨论推迟至附录 A.4。

相应的证明见附录 A.2。尽管先前的研究（Bengio 等人，2023年；Malkin等人，2022a）已探讨过 Sub-TB 条件的充分性和必要性，但定理 3.2 提供了另一种视角，更清晰地阐明了流函数与评估函数之间的联系。需要注意的是，上述 KL 项的最小值为零，因为 $P_F(\tau)$ 和 $P_B(\tau)$ 都是马尔可夫的。此外，根据 Bengio 等人（2023年）中的命题 23，实现 KL 项为零的轨迹流 $F(\tau) = P_F(\tau)Z$ 是唯一的。

利用平衡条件，我们定义 Sub-EB 目标，以优化参数化评估函数 $V(\cdot\,;\phi)$:

$$\mathcal{L}_V(\phi) := \mathbb{E}_{P_F(\tau)}\Big[\sum_{\tau_{i:j}} w_{j-i}\left(\delta_V(\tau_{i:j};\phi)\right)^2\Big],\ \ \delta_V(\tau_{i:j};\phi) = \log\frac{P_F(\tau_{i:j}|s_i)\exp V(s_i;\phi)}{P_B(\tau_{i:j}|s_j;\phi)\exp V(s_j;\phi)}, \quad (8)$$

其中，w_{j-i} 是对包含 $j-i$ 条边的子轨迹赋予的权重，与 Sub-TB 目标中的做法相同。传统的 λ-时间差分（TD）目标（54）详见附录 A.5（Niu 等，2024年；Schulman 等人，2016年），它仅利用从步骤 h 开始的事件以及逐边失配 $\delta_V(s \to s')$ 来学习 $V(s_h;\phi)$；而 Sub-EB 目标则纳入发生在 h 之前和之后的事件，并利用逐子轨迹失配，从而使 $V(s_h;\phi)$ 的学习更加均衡。此外，Sub-EB 支持自由选择权重方案，而 λ-TD 的权重方案则受限于 λ-衰减形式。附录 A.5 对我们的 Sub-EB 和 λ-TD 目标进行了详细比较。

需要指出的是，Sub-EB 目标专为估计 $V^\dagger$ of π_F而设计。在每次优化迭代中，计算其关于 ϕ的梯度，以更新 $V(\cdot\,;\phi)$，同时冻结 θ。相比之下，Sub-TB 目标 2 用于联合更新 $\pi_F(\cdot|\cdot\,;\theta)$ 和 $\log F(\cdot\,;\theta)$。[2]我们在算法 1 中总结了基于策略的方法进行 GFlowNet 训练的工作流程，其中$\widehat{\mathcal{L}}_V(\phi) := \frac{1}{K}\sum_{\tau\in\mathcal{D}}[\sum_{\tau_{i:j}\in\tau} w_{j-i}\delta_V(\tau_{i:j};\phi)]$ 以及 $\widehat{\nabla}_\theta V^\dagger(s_0;\theta)$:
$= \frac{1}{K}\sum_{\tau\in\mathcal{D}}[\sum_{h=0}^{H} A^\gamma(s_h \to s_{h+1})\nabla_\theta \log \pi_F(s_{h+1}|s_h;\theta)]$。

3.2 参数化反向策略

在本节及后续各节中，我们进一步讨论 Sub-EB 条件所带来的两个关键优势，这些优势提升了基于策略的训练的灵活性。λ-TD目标（54）要求 π_B 在整个优化过程中保持固定，因为 V^λ 相对于 ϕ被视为常数。为解决这一局限，Niu 等人（2024）提出了一种两阶段算法，其中每次训练迭代都包含一个正向阶段和一个反向阶段。在正向阶段，我们采样 $\mathcal{D} \sim P_F(\tau)$，基于 $\mathcal{D}$更新 V，并基于 V 和 $\mathcal{D}$更新 π_F。在反向阶段，我们采样$\mathcal{D}' \sim P_B(\tau|x)P_F(x)$，更新反向评估函数 W，使其逼近 π_B的真实反向评估函数 $W^\dagger$，并基于 W更新 π_B。 $W^\dagger$和 W的定义将在第 3.3 节中详述。与此同时，Gritsaev 等；Jang 等人（2024年）在正向阶段内为 π_B 引入了一个额外目标。当应用于基于策略的框架时，他们的方法首先采样一批 $\mathcal{D} \sim P_F(\tau)$，并通过最大化其对数似然$\sum_{\tau\in\mathcal{D}} \log P_B(\tau)$来更新 π_B。然后，在 π_B保持固定的情况下，使用同一批 $\mathcal{D}$，按照上述步骤更新 V和 π_F。

相比之下，Sub-TB（2）和 Sub-EB（8）目标都允许更新参数化的 π_B，而无需引入单独的反向阶段或额外目标。具体而言，在 Sub-EB 目标下，π_B与 V联合更新；在 Sub-TB 目标下，则与（π_F、$\log F$）联合更新。这使训练过程更加简洁高效，同时让后向策略能够在优化过程中动态调整。此外，除基本的策略梯度方法外，Sub-EB 目标还支持在更高级的基于策略的方法中使用参数化的 π_B，例如信赖域策略优化（TRPO）方法（Niu 等，2024）；在这种方法中，按照 λ-TD 目标，π_B 此前必须保持固定。

3.3 离线基于策略的训练

Niu 等人（2024）提出的单阶段和两阶段算法都以**在线**方式运行，因此训练过程中无法使用数据收集策略 $\pi_\mathcal{D} \neq \pi_F$。为克服这一限制，我们引入了一种**离线**基于策略的方法，这得益于 Sub-EB

[2]此处，θ 和 ϕ 作为独立的参数集引入。θ 对应于与π_F联合更新的函数，而 ϕ 则对应于不与其联合更新的函数。

算法 1 在线策略工作流

需要： $\pi_F(\cdot\,|\,\cdot\,;\ \theta)$，$\pi_B(\cdot\,|\,\cdot\,;\ \phi)$，$V(\cdot\,;\ \phi)$，批量大小 K，迭代总次数 N

对于 $n = 1, \ldots, N$ **执行**

$\mathcal{D} \leftarrow \{\tau^k :\ \tau^k \sim P_F\ (\tau)\ \}_{k=1}^{K}$

基于 $\mathcal{D}$，通过 ∇_ϕ 更新 ϕ $\widehat{\mathcal{L}}_V(\phi)$

基于 $\mathcal{D}$ 以及 V，通过 θ 更新 $-\widehat{\nabla}_\theta V^\dagger(s_0;\ \theta)$

结束循环

返回 $\pi_F(\cdot\,|\,\cdot\,;\ \theta)$、$\pi_B(\cdot\,|\,\cdot\,;\ \phi)$、$V(\cdot\,;\ \phi)$

算法 2 基于离线策略的工作流

要求： $\pi_B(\cdot\,|\,\cdot;\ \phi)$，$\pi_F(\cdot\,|\,\cdot\,;\ \theta)$，$W(\cdot\,;\ \theta)$，的——离线策略 $\pi_{\mathcal{D}}$，K，N

对于 $n = 1, \cdots\cdots, N$ **do**

$\mathcal{D}^\top \leftarrow \{x^k :\ x^k \in \tau^k,\ \tau^k \sim P_{\mathcal{D}}\ (\tau)\ \}_{k=1}^{K}$

$\mathcal{D} \leftarrow \{\tau^k : x^k \in \mathcal{D}^\top,\ \tau^k \sim P_B(\tau|x^k)\}_{k=1}^{K}$

基于 $\mathcal{D}$，将 θ 更新 $\nabla_\theta \mathcal{L}_W\ (\theta)$

基于 $\mathcal{D}$ and W，更新 ϕ，方式为 $-\nabla_\phi \mathbb{E}_{P_{\mathcal{D}}\ (x)}\ [W^\dagger\ (x;\ \phi)\ ]$

结束循环

返回 $\pi_B(\cdot\,|\,\cdot\,;\ \phi)$，$\pi_F(\cdot\,|\,\cdot\,;\ \theta)$，$W(\cdot\,;\ \theta)$

目标的灵活性。我们定义 $W^\dagger(s_0) := \log\ F(s_0)$。对于任意 $h \in [H]\backslash\{0\}$，$W^\dagger(s_h;\ \phi)$ 定义如下：

$$\mathbb{E}_{P_B(\tau_{:h}|s_h;\phi)}\Big[\sum_{i=0}^{h-1} R(s_i \leftarrow s_{i+1};\phi)\Big] = \log F(s_h) - D_{\mathrm{KL}}(P_B(\tau_{:h}|s_h;\phi)\|P_F(\tau_{:h}|s_h)), \tag{9}$$

在 $R(s_i \leftarrow s_{i+1};\ \phi)$ 的条件下：$= \log \frac{\pi_F(s_{i+1}|s_i)}{\pi_B(s_i|s_{i+1};\phi)}$。等式可轻易验证，如附录中的式（45）所示。众所周知，最小化 $\mathbb{E}_{P_{\mathcal{D}}(x)}[\mathcal{L}_{KL}(\phi|x)] := \mathbb{E}_{P_{\mathcal{D}}(x)}[D_{\mathrm{KL}}(P_B(\tau|x;\ \phi)\|P_F(\tau|x))]$也能减小 $P_B(\tau)$ 与 $P_F(\tau)$ 之间的差异（Niu 等，2024）。由于 $\nabla_\phi \log\ F(x) = 0$且 $-\nabla_\phi \mathbb{E}_{P_{\mathcal{D}}(x)}[W^\dagger(x;\ \phi)\ ] = \nabla_\phi \mathbb{E}_{P_{\mathcal{D}(x)}}\ [\mathcal{L}_{KL}(\phi|x)]$,，因此 $-\mathbb{E}_{P_{\mathcal{D}}(x)}[W^\dagger(x;\ \phi)]$ 可作为预期 KL 散度的替代指标。类似于前向情形，我们使用参数化评估函数 W来近似 $W^\dagger$，并按如下方式计算 $\nabla_\phi \mathbb{E}_{P_{\mathcal{D}}(x)}[W^\dagger(x;\ \phi)]$：

$$\mathbb{E}_{P_B^{\mathcal{D}}(\tau)}\Big[\sum_{h=0}^{H-1} A^\gamma(s_h \leftarrow s_{h+1})\nabla_\phi \log \pi_B(s_h|s_{h+1};\phi)\Big] \tag{10}$$

其中 $A^\gamma(s_h \leftarrow s_{h+1}) = \sum_{i=0}^{h} \gamma^{h-i}(R(s_i \leftarrow s_{i+1}) + W(s_i) - W(s_{i+1}))$ 和 $P_B^{\mathcal{D}}(\tau)\ :\ = P_B(\tau|x)P_{\mathcal{D}}(x)$。给定后向策略 π_B，W 的 Sub-EB 条件可写为 $\forall i < j \in [H+1]$：

$$\mathbb{E}_{P_B^{\mathcal{D}}(\tau_{i:j})}\Big[\log\big(P_F(\tau_{i:j}|s_i)\exp W(s_i)\big)\Big] = \mathbb{E}_{P_B^{\mathcal{D}}(\tau_{i:j})}\Big[\log\big(P_B(\tau_{i:j}|s_j)\exp W(s_j)\big)\Big], \tag{11}$$

其中，对于 $x \in \mathcal{X}$，$P_B(x|s_f)\ \exp\ W(s_f)\ :\ =\ R(x)$。该条件可改写为$\mathbb{E}_{P_B^{\mathcal{D}}(s_i,\ s_j)}\ [W(s_i) - W(s_j)] = \mathbb{E}_{P_B^{\mathcal{D}}(\tau_{i:j})}\ \Big[\log\ \frac{P_{B_{(\tau_{i:j}|s_j)}} P_F(\tau_{i:j}|s_i)}{}\Big]$，促使学习得到的散度差在 s_i 和 s_j 处与二者之间的真实散度相匹配。

定理 3.3。假设 W 是定义在 $\mathcal{S}\backslash\{s_f\}$ 上的任意评估函数，且 F 是与前向策略 π_F 相关联的流。给定后向策略 π_B，

$$\forall h \in [H-1] : W(s_{h+1}) = \log F(s_{h+1}) - D_{\mathrm{KL}}(P_B(\tau_{:h+1}|s_{h+1})\|P_F(\tau_{:h+1}|s_{h+1})), \tag{12}$$

当且仅当 W满足反向 *Sub-EB* 条件（*11*）时，W（x）$=$对数 R（x）成立。

相应的证明见附录A.3。基于反向 Sub-EB 条件，我们给出 $W(\cdot\,;\ \theta)$ 的后向Sub-EB目标如下：

$$\mathcal{L}_W(\theta) := \mathbb{E}_{P_B^{\mathcal{D}}(\tau)}\Big[\sum_{\tau_{i:j}} w_{j-i}\left(\delta_W(\tau_{i:j};\theta)\right)^2\Big],\ \ \delta_W(\tau_{i:j};\theta) = \log\frac{P_F(\tau_{i:j}|s_i;\theta)\exp W(s_i;\theta)}{P_B(\tau_{i:j}|s_j)\exp W(s_j;\theta)} \tag{13}$$

当 π_B 和 π_F 达到各自的最优值时，W 表达式中的 KL 项为零。因此，$F(x) = R(x)$，且 $P_F(x) = F(x)/\sum_x F(x) = R(x)/Z^*$，从而实现 GFlowNet 训练的目标。我们用于 GFlowNet 训练的离线基于策略的方法的工作流程见算法 2，其中 $\widehat{\mathcal{L}}_W(\theta) := \frac{1}{K}\sum_{\tau\in\mathcal{D}}[\sum_{\tau_{i:j}\in\tau} w_{j-i}\delta_W(\tau_{i:j};\ \theta)]$ 以及 $\widehat{\nabla}_\phi \mathbb{E}_{P_{\mathcal{D}}(x)}[W^\dagger(x;\ \phi)] := \frac{1}{K}\sum_{\tau\in\mathcal{D}}[\sum_{h=0}^{H-1} A^\gamma(s_h \leftarrow s_{h+1})\nabla_\phi \log\ \pi_B(s_h|s_{h+1}|\phi)]$。

4 相关工作

基于价值的 GFlowNet 训练 现有基于价值的 GFlowNet 训练工作可分为两个方向。第一个方向侧重于设计训练目标，以刻画目标流值，从而改进对流不平衡的估计。详细平衡（DB）目标（Bengio 等人，2023年）旨在减小前向边流量 $\pi_F(s'|s)F(s)$ 与后向目标边流 $\pi_B(s|s')F(s')$ 之间的对数失配。目标项 $F(s')P_B(s|s')$ 提供有偏但低方差的反馈，因为 F 编码了关于任务的已学习局部知识。轨迹平衡（TB）目标（Malkin等人，2022a）优化前向轨迹流 $P_F(\tau)Z$ 与真实后向轨迹流 $P_B(\tau|x)R(x) = F^*(\tau)$ 之间的对数失配，由此产生无偏但高方差的反馈。Sub-TB 目标（Madan 等，2023）通过最小化不同长度子轨迹的流对数失配，推广了 DB 和 TB。通过按子轨迹长度重新加权，该方法实现了偏差-方差权衡，从而取得更好的性能。在 Sub-TB 目标的基础上，人们还提出了各种改进。Kim 等人（2023a）引入了温度条件目标，其目标变为使 $P_F(x) \propto R^\beta(x)$，其中正标量$\beta < 1$。将 R 作为指数 β 可降低其尖锐程度，使其更容易匹配。由于此时 $P_F(x)$ 与 $R(x)$ 不匹配，这种表示仅在关注模式寻优时特别有用。对于上述所有目标，任意中间边（$s \to s'$）的 $\pi_B(s|s')$ 均可自由选择，而只有 $\pi(x|s_f)$ 固定为 $R(x)$。因此，中间边或子轨迹的目标流值并不能直接反映奖励函数 $\mathcal{R}$的真实信息。然而，在某些特殊情况下，如果 $\mathcal{R}$ 所覆盖的对象空间可以从 $\mathcal{X}$扩展到 $\mathcal{S}$，Pan 等人（2023年）和 Jang 等人（2023年）改进了 DB 和 Sub-TB 目标的形式，将关于 $\mathcal{R}$ 的局部知识直接传播到中间边。鉴于 Sub-TB 目标与我们的 Sub-EB 目标相似，这些改进也很容易适用于 Sub-EB，从而促进对 V的学习。

基于价值的方法的一个关键特征是，$\pi_\mathcal{D}$ 可以采用离策略方式，也就是说，它与 π_F不同，因此人们投入了大量精力来设计 $\pi_\mathcal{D}$（Kim 等，2023b；Rector-Brooks 等，2023）。其目标是在有效识别先于高奖励终止状态的边（探索）的同时，允许重新访问已发现的、能够带来高奖励的边（利用）。应用最广泛的方法是 α贪心设计，它以 α为比例将 π_F 与均匀策略相结合。然而，这些方法无法实现**深度探索**；深度探索不仅要求考虑即时信息增益，还要考虑状态转移（边）对未来学习的长期影响（Os-band 等，2019）。尽管从强化学习的角度出发，在高效探索设计方面已有许多理论进展（Azar 等人，2017年；Jin 等人，2018年；M´enard 等人，2021年），但其高昂的计算成本限制了它们在实际问题中的适用性。

基于策略的GFlowNet训练 如上所述，设计一种兼具计算效率并能够**深度探索** 的数据收集策略，仍是一项重大挑战。此外，提高训练效率真正需要的是识别流不平衡较大的边，而非能够带来较高终止奖励的边。然而，流不平衡与 π_F 密切相关，并会随着 π_F 的更新而变化。经验证据表明，在许多条件下，同策略训练（即 $\pi_\mathcal{D} = \pi_F$）能够实现更快的收敛（Atanackovic与Bengio，2024）。因此，基于策略的GFlowNet方法（Malkin 等人，2022b；Niu 等人，2024；Zimmermann等人，2022）采用同策略训练，通常对应于优化 $P_F(\tau)$ 与未归一化分布$P_B(\tau|x)R(x)(= P_B(\tau)Z^*)$之间的 KL 散度；该优化与 $P_F(\tau)$ 和$P_B(\tau)$之间散度的梯度等价。由于无需再设计数据收集策略，基于策略的方法面临的主要挑战转为稳健地估计散度及其梯度，也就是平衡估计器的方差与偏差。Malkin 等人（2022b）和 Silva等人（2024）通过经验构造估计器，并要求固定的 π_B。在基于梯度的优化过程中，这些估计器仅根据当前迭代中采样的训练数据计算。此类估计器通常偏差较低，但方差较高。基于强化学习中策略梯度的理论，Niu 等人（2024）提出了基于参数化评估函数 V的估计器，从而能够利用此前所有迭代中采样的数据。这些估计器通常偏差较高，但方差较低。Niu 等人（2024）结合经验方法与参数化方法，引入一个超参数来显式控制**方差—偏差** 权衡，从而显著提升了性能。由于这些基于策略的方法的有效性在很大程度上取决于 V 的学习方式，本文将继续探索这一挑战。

GFlowNet训练与强化学习 基于价值的方法（如 DB 和 Sub-TB），以及受强化学习启发的 DB 变体，包括 Munchausen DQN（记为 Q-Much）（Tiapkin 等，2024）和修正流迭代（RFI）（He 等，2025），都遵循 Q 学习框架（Haarnoja 等，2017）。相比之下，基于策略梯度的基于策略的方法则采用演员-评论家框架（Sutton 和 Barto，2018；Haarnoja 等，2018）。附录 A.6 对这两个框架在理论上的区别进行了详细讨论。为进一步说明 Sub-EB 目标和策略梯度如何在演员-评论家框架中发挥作用，我们在附录 A.7 中推导了适用于 GFlowNet 的基本软行动者–评论家算法（算法 3），以及一个修改后的变体（算法 4）。

5 实验

我们考虑基于树策略的方法：Silva 等人（2024）的经验方法，该方法在梯度估计期间使用控制变量（CV）技术来缩减方差；以及两种策略梯度方法（Niu 等，2024），它们通过 λ-TD 目标或提出的 Sub-EB 目标来学习 V。我们分别将这些方法称为 CV、强化学习和 Sub-EB。由于 Sub-TB（Madan 等，2023）（Tiapkin 等，2024）与 Sub-EB 密切相关，我们也将其纳入基于价值的方法的代表性基线。对于受强化学习启发的基于价值的方法变体，我们考虑 Q-Much（Tiapkin 等，2024）和 RFI（He 等，2025）。不过，鉴于 RFI 可能存在计算可扩展性问题（见附录 A.6），我们的基准实验中仅纳入 Q-Much。为了设计 Sub-TB 方法中的数据收集策略 $\pi_{\mathcal{D}}$，我们采用一种常见选择：以 $(1-\alpha)$ 的概率令 $\pi_{\mathcal{D}}$ 等于 π_F，以 α 的概率采用均匀策略（Shen 等，2023；Rector-Brooks 等，2023）。π_B 的常见设置可以是均匀策略或参数化策略。默认情况下，我们采用均匀策略，因为参数化策略不包含关于奖励函数的真实信息。我们还为 Sub-TB 和离线 Sub-EB（算法 2）加入局部搜索技术（Kim 等，2023b）来设计 $P_{\mathcal{D}}$，以便在轨迹采样期间明确促进对高奖励终止状态的探索。这些变体分别记为 Sub-TB-B 和 Sub-EB-B。如第 4 节所述，且下文的实验结果也证实，明确鼓励探索的离策略技术可能无益于分布建模。不过，如果重点是发现终止状态的众数，这些技术可能会很有价值。对于 Sub-EB、Sub-TB 及其变体，按照 Madan 等（2023）的做法，$i<j\in[H+1]$ 对应的权重 w_{j-i} 设为 $\lambda^{j-i}/\sum_{i<j\in[H+1]}\lambda^{j-i}$。
我们没有采用文献中常用的平均 l_1 距离，而是选择总变差 D_{TV} 和詹森-香农散度 D_{JSD} 作为衡量指标，用于比较竞争方法之间的性能，其中比较对象为 $P_F(x)$ 和 $P^*(x)$。其理由及正式定义详见附录 B。我们进行了三组实验。第一组在模拟环境中开展，称为“超网格”。第二组聚焦于使用真实世界数据集进行生物与分子序列设计任务。第三组涉及 GFlowNet 在贝叶斯网络（BN）结构学习中的真实世界应用。更多实验细节，例如模型配置和超参数选择，可见附录 B。我们的部分实现代码改编自 *torchgfn* 软件包（Lahlou 等人，2023年）。

超网格 超网格实验广泛用于测试 GFlowNet 的性能（Malkin 等人，2022b；Niu 等，2024年）。在此，状态是高度为 H 的 D 维超立方网格的坐标元组。生成过程的详细描述见附录 B.1。我们在 256×256、$128\times128\times128$ 和 $64\times64\times64$ 网格上进行实验。我们使用动态规划（Malkin等人，2022a）显式计算学习得到的 π_F 的 $P_F(x)$，并精确计算 $P_F(x)$ 与 $P^*(x)$ 之间的 D_{TV} 和 D_{JSD}。实验结果分别见图 2 中的 D_{TV}，以及附录 B.1 图 4 中的 D_{JSD}。

在 256×256 网格上可以观察到，将用于学习 V 的 λ-TD 目标替换为所提出的 Sub-EB 目标，显著提高了策略梯度方法的稳定性和收敛率。尽管两种策略梯度方法（RL 和 Sub-EB）的最终表现相近，但它们都优于 Sub-TB 和 CV。这些实验结果有力地证明，Sub-EB 目标有助于更可靠地学习评估函数V，从而提高基于策略的 GFlowNet 训练的稳定性并加快收敛。仅根据当前训练批次构造的经验梯度估计量，无法为基于策略的训练提供可靠的指导。在 $128\times128\times128$ 网格上，稳定性和收敛

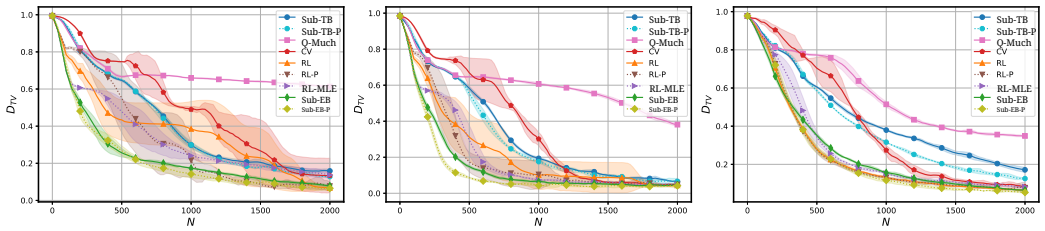


图 2：基于每种方法的五次随机初始化运行，展示不同训练方法在采用参数化 π_B和均匀 π_B时，于 256×256（左）、 128×128（中）和 $64\times64\times64$（右）网格上的 D_{TV} 均值及标准差（以阴影区域表示）。**默认情况下**，本文所有绘制曲线的指标值每 20 次迭代记录一次，在 $N = 2000$训练迭代期间记录，并使用长度为 5 的滑动窗口进行平滑处理。

率方面，Sub-EB 再次远胜 RL。尽管 CV、RL、Sub-EB 和 Sub-TB 的最终表现相近，但在收敛率方面，RL 和 Sub-EB 均优于 Sub-TB 和 CV。这些发现进一步验证了 Sub-EB 目标的有效性。最后，在 $64\times64\times64$ 网格上，RL 和 Sub-EB 均优于 Sub-TB 和 CV，但 RL 与 Sub-EB 的表现非常接近。这可能是因为 RL 在这项实验中的稳定性已经足够好，因此 Sub-TB 目标带来的优势并不明显。此外，网格高度 N 对建模难度的影响可能大于网格维度 D，因为超网格在各个维度上是齐次的，且众数之间的最小距离仅取决于网格高度 N（Niu 等，2024）。默认训练批次大小设为 128，而 Tiapkin 等人（2024）的超网格实验采用了 16 的批量大小和 20 的网格高度。为确保比较公平并评估这些差异的影响，我们还在 128×128 和 20×20 网格上评估了 Sub-TB 和 Q-Much 的表现。详细讨论和结果见附录 B.1。

关于 π_B 和 λ 的消融研究。为展示 Sub-EB 目标能够自然地适配参数化的 π_B，我们比较了不同方法在 256×256、 128×128和 $64\times64\times64$ 网格上采用参数化和均匀 π_B 时的性能。我们用 Sub-TB-P 和 Sub-EB-P 表示采用参数化 π_B的相应方法。此外，RL-P 和 RL-MLE 分别表示 RL 的两个变体，它们针对参数化 π_B分别采用 Niu 等人（2024）提出的两阶段算法和 Gritsaev 等人提出的方法。由于 CV 要求 π_B固定，因此不支持参数化的 π_B，故不纳入本消融研究。如图 2 及附录 B.1 中的图 4 所示，在所有考虑的方法中，Sub-EB-P 的性能和训练稳定性最佳。这证实了 Sub-EB 目标能够很好地适配反向策略，这些策略与评估函数一同参数化并联合更新。对于128×128 网格，我们还对 Sub-EB 权重w_{j-i}的超参数 λ 进行了消融研究（$=\lambda^{j-i}/\sum_{i<j\in[H+1]}\lambda^{j-i}$）。详细讨论和结果见附录 B.1。

序列设计 在这组实验中，我们使用 GFlowNets 生成长度为 D 的生物和分子序列，这些序列由 M 个构建块组成（Shen 等，2023）。我们使用了 Shen 等人（2023）中的核苷酸序列数据集（SIX6 和 PHO4）以及分子序列数据集（QM9 和 sEH）。生成过程和实验结果的详细说明见附录 B.2。强化学习、Sub-TB、Sub-EB、Sub-TB-B 和 Sub-EB-B 在模式发现和分布建模任务上的实验结果，验证了所提出的 Sub-EB 目标能够将离线采样技术整合到基于策略的方法中。

贝叶斯网络结构学习 在本组实验中，我们聚焦于贝叶斯网络（BN）结构学习的真实世界研究（Malkin 等人，2022b；Niu 等，2024）。此处，对象空间 $\mathcal{X}$对应于 BN 结构空间。生成过程的详细描述见附录 B.3。我们考虑节点数分别为 5、10 和 15 的三种情形，其中 $\mathcal{X}$ 的规模分别约为 2.95×10^4、 4.18×10^{18}和 2.38×10^{35}。对于大规模情形， P_F（x）或 P^*（x）在计算上均不可行。因此，我们报告训练过程中发现的前 100 个唯一图的平均奖励。由于有效的分布建模性能不仅意味着最优性，也意味着生成的候选项的多样性，我们

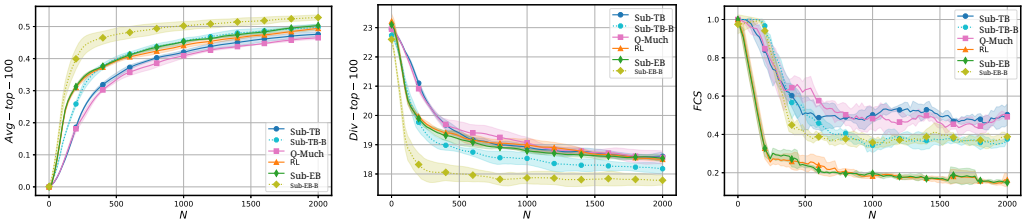


图 3：基于每种方法 5 次随机初始化运行结果，绘制 10 个节点上前 100 个不同候选图的平均奖励（左）、多样性（右）和 FCS（右）的均值与标准差（以阴影区域表示）。

还计算这 100 个图之间的平均成对汉明距离，作为多样性的度量。需要注意的是，**多样性既不宜过高也不宜过低**：前者对应于近乎随机的生成，而后者表明生成陷入有限的一组结构。由于 D_{TV} 的精确计算不可行，我们使用子图流一致性（FCS）（Silva 等，2025）作为 D_{TV}的无偏估计。对于 FCS，我们随机采样了 32 批终止状态，每批大小为 128，以构建蒙特卡洛估计量。

下面我们讨论两个大规模案例，并将小规模案例留到附录 B.3 中介绍。相应的实验结果见图 3 以及附录 B.3 中的图 12。在 RL、Sub-EB、Sub-TB 和 Q-Much 中，结果表明 Sub-EB 获得了最高的平均奖励。四种方法取得了相近的多样性。值得注意的是，RL 和 Sub-EB 比 Sub-TB 和 Q-Much 收敛更快，并且只有 RL 和 Sub-EB 在 FCS 指标所衡量的分布建模性能方面表现出色。综合这些发现，我们可以得出结论：所有方法都实现了适当的分布建模，而 Sub-EB 的表现最佳。这表明，Sub-EB 不仅能够实现可靠的基于策略的训练，还能有效扩展到大规模组合空间，同时提供高质量且多样化的解。对于两种变体 Sub-TB-B 和 Sub-EB-B，可以观察到，Sub-EB-B 在全部六种方法中取得了最高的平均奖励，但多样性下降得更为明显。由于局部搜索组件明确优先探索解空间中的高奖励区域，因此这种权衡——以降低多样性为代价显著提升奖励——是意料之中的。相比之下，Sub-TB-B 相较于其未增强版本（Sub-TB）确实取得了更高的平均奖励，同时多样性仅适度下降。然而，这种权衡明显弱得多。在没有局部搜索技术的情况下，Sub-EB 已经取得了与 Sub-TB-B 相当的平均奖励以及更高的多样性。总体而言，Sub-EB-B 比 Sub-TB-B 更有效，这与我们对最优性—多样性权衡的预期高度一致。这一发现进一步支持了基于策略的方法的优越性，并验证了 Sub-EB 目标能够在基于策略的框架中整合离线技术。

分子图设计在这组实验中，我们考虑了 Bengio 等人（2021）所述的分子图设计任务（使用 $|\mathcal{X}| \approx 10^{16}$）。基于 sEH 数据集（使用 $|\mathcal{X}| \approx 3.4 \times 10^{7}$）的序列设计任务（Shen 等，2023）是对这些任务的简化。附录 B.4 对这些图设计任务的生成过程及相应实验结果进行了详细描述。实验结果表明，Sub-EB 在大规模分子图设计方面整体表现出色，与强化学习、Sub-TB 和 Q-Much 相比，能够获得更高的平均奖励、更快收敛，并保持有竞争力的多样性。

6 讨论与结论

在这项工作中，我们建立了状态流函数 $F(s)$ 与评估函数 $V(s)$ 之间的联系。基于此，我们提出了一种名为 Sub-EB 的新目标，用于学习评估函数 $V(s)$。通过四组实验，我们提供了实证证据，表明与 λ-TD 目标相比，新的 Sub-EB 目标能够更稳定、灵活地学习 V，从而提升基于策略的方法在 GFlowNet 训练中的表现。原则上，Sub-EB 目标允许灵活选择权重系数，并且适用范围超越了策略梯度方法。关于如何设计最优权重系数，以及如何将 Sub-EB 目标整合到更先进的基于策略的方法中，将留待未来研究。

致谢

本研究部分得到了美国国家科学基金会 (NSF) 项目 SHF-2215573 和 IIS-2212419 的资助。本研究的部分工作使用了德州农工大学高性能研究计算中心提供的先进计算资源。

参考文献

Alekh Agarwal、Sham M Kakade、Jason D Lee 和 Gaurav Mahajan。策略梯度方法理论：最优性、近似与分布偏移。机器学习研究杂志，22(98):1–76，2021年。

Lazar Atanackovic 和 Emmanuel Bengio。探究生成流网络的泛化行为。发表于*ICML 2024* 结构化概率推理与生成建模研讨会，2024年。

Mohammad Gheshlaghi Azar、Ian Osband 和 R´emi Munos。强化学习的极小极大遗憾界。发表于国际机器学习会议，第263–272页。PMLR，2017年。

Emmanuel Bengio、Moksh Jain、Maksym Korablyov、Doina Precup 和 Yoshua Bengio。用于非迭代多样候选生成的基于流网络的生成模型。神经信息处理系统进展，34：27381–27394，2021年。

Yoshua Bengio、Salem Lahlou、Tristan Deleu、Edward J Hu、Mo Tiwari 和 Emmanuel Bengio。GFlowNet 基础。机器学习研究期刊，24(210)：1–55，2023年。

Tristan Deleu、Ant´onio G´ois、Chris Emezue、Mansi Rankawat、Simon Lacoste-Julien、Stefan Bauer 和 Yoshua Bengio。利用生成流网络进行贝叶斯结构学习。载于人工智能中的不确定性，第518–528页。PMLR，2022。

Tristan Deleu、Padideh Nouri、Nikolay Malkin、Doina Precup 和 Yoshua Bengio。在多路径环境中将离散概率推断作为控制。载于人工智能中的不确定性，第997–1021页。PMLR，2024年。

Timofei Gritsaev、Morozov Nikita、Samsonov Sergey 与 Daniil Tiapkin。通过轨迹似然最大化优化 GFlowNets 中的反向策略。见第十三届国际学习表征会议。

Tuomas Haarnoja、Haoran Tang、Pieter Abbeel 与 Sergey Levine。基于深度能量型策略的强化学习。见国际机器学习会议，第1352–1361页。PMLR，2017年。

Tuomas Haarnoja、Aurick Zhou、Pieter Abbeel 与 Sergey Levine。软演员-评论家：带随机演员的离策略最大熵深度强化学习。见国际机器学习会议，第1861–1870页。PMLR，2018年。

Haoran He、Emmanuel Bengio、Qingpeng Cai 与 Ling Pan。随机策略评估揭示生成流网络的策略。见第四十二届国际机器学习大会，第22351–22367页。PMLR，2025年。

Hyosoon Jang、Minsu Kim 和 Sungsoo Ahn。《学习 GFlowNets 部分推断的能量分解》。载于第十二届国际学习表征会议，2023年。

Hyosoon Jang、Yunhui Jang、Minsu Kim、Jinkyoo Park 和 Sungsoo Ahn。《面向 GFlowNets 的悲观后向策略》。神经信息处理系统进展，2024年。

Chi Jin、Zeyuan Allen-Zhu、Sebastien Bubeck 和 Michael I Jordan。Q学习是否可证明高效？神经信息处理系统进展，31，2018年。

Minsu Kim、Joohwan Ko、Dinghuai Zhang、Ling Pan、Taeyoung Yun、Woo Chang Kim、Jinkyoo Park 和 Yoshua Bengio。《学习缩放对数几率以适应温度条件生成流网络》。发表于*NeurIPS 2023* 人工智能促进科学研讨会，2023a。

Minsu Kim、Taeyoung Yun、Emmanuel Bengio、Dinghuai Zhang、Yoshua Bengio、Sungsoo Ahn 和 Jinkyoo Park。《GFlowNet 局部搜索》。发表于第十二届国际学习表征会议，2023b。

Jack Kuipers、Giusi Moffa 与 David Heckerman。关于高斯有向无环图模型评分的补遗。2014。

Salem Lahlou、Joseph D Viviano 与 Victor Schmidt。torchgfn：PyTorch GFlowNet 库。*arXiv* 预印本 *arXiv:2305.14594*，2023年。

Kanika Madan、Jarrid Rector-Brooks、Maksym Korablyov、Emmanuel Bengio、Moksh Jain、Andrei Cristian Nica、Tom Bosc、Yoshua Bengio 与 Nikolay Malkin。从部分回合中学习 GFlowNet 以改善收敛与稳定性。收录于国际机器学习会议，第 23467–23483 页。PMLR，2023 年。

Nikolay Malkin、Moksh Jain、Emmanuel Bengio、Chen Sun 与 Yoshua Bengio。轨迹平衡：改进 GFlowNets 中的信用分配。神经信息处理系统进展，35:5955–5967，2022a。

Nikolay Malkin、Salem Lahlou、Tristan Deleu、Xu Ji、Edward J Hu、Katie E Everett、Dinghuai Zhang 和 Yoshua Bengio。GFlowNet 与变分推断。发表于第十一届国际学习表征会议，2022b。

Pierre M´enard、Omar Darwiche Domingues、Anders Jonsson、Emilie Kaufmann、Edouard Leurent 和 Michal Valko。用于强化学习纯探索的快速主动学习。发表于国际机器学习会议，第 7599–7608 页。PMLR，2021。

Puhua Niu、Shili Wu、Mingzhou Fan 和 Xiaoning Qian。通过策略梯度训练 GFlowNet。发表于第四十一届国际机器学习会议，2024年。

Ian Osband、Benjamin Van Roy、Daniel J Russo 和 Zheng Wen。通过随机化价值函数进行深度探索。《机器学习研究杂志》，20(124):1–62，2019年。

Ling Pan、Nikolay Malkin、Dinghuai Zhang 和 Yoshua Bengio。通过局部信用和不完整轨迹改进 GFlowNets 训练。载于国际机器学习会议，第 26878–26890 页。PMLR，2023年。

Jarrid Rector-Brooks、Kanika Madan、Moksh Jain、Maksym Korablyov、Cheng-Hao Liu、Sarath Chandar、Nikolay Malkin 和 Yoshua Bengio。用于提升 GFlowNets 探索能力的汤普森采样。载于*ICML 2023*年结构化概率推断与生成模型研讨会，2023年。

Robert W Robinson。计数无标号无环有向图。载于组合数学 *V*：第五届澳大利亚会议论文集，会议于*1976*年*8*月*24*日至*26*日在皇家墨尔本理工大学举行，第 28–43 页。Springer，2006年。John Schulman、Philipp Moritz、Sergey Levine、Michael Jordan 和 Pieter Abbeel。使用广义优势估计进行高维连续控制。载于学习表征国际会议论文集（*ICLR*），2016年。John Schulman、Xi Chen 和 Pieter Abbeel。策略梯度与软 Q 学习之间的等价性。*arXiv* 预印本 *arXiv:1704.06440*，2017年。Max W Shen、Emmanuel Bengio、Ehsan Hajiramezanali、Andreas Loukas、Kyunghyun Cho 和 Tommaso Biancalani。迈向理解并改进 GFlowNet 训练。载于国际机器学习会议，第 30956–30975 页。PMLR，2023年。Tiago Silva、de Souza da Silva Eliezer 和 Mesquita Diego。关于训练 GFlowNets 的散度度量。神经信息处理系统进展，2024年。Tiago Silva、Alves Rodrigo Barreto、da Silva Eliezer de Souza、Souza Amauri H、Garg Vikas、Kaski Samuel 和 Mesquita Diego。GFlowNets 何时能够学到正确的分布？载于第十三届国际学习表征会议，2025年。

Richard S Sutton 和 Andrew G Barto。强化学习：导论。麻省理工学院出版社，2018年。

Daniil Tiapkin、Morozov Nikita、Naumov Alexey 和 Vetrov P Dmitry。生成流网络作为熵正则化强化学习。载于国际人工智能与统计会议。PMLR，2024年。

Heiko Zimmermann、Fredrik Lindsten、Jan-Willem van de Meent 和 Christian A Naesseth。《生成流网络的变分视角》。机器学习研究汇刊，2022。

大语言模型使用情况披露

大语言模型仅用于文本润色（语法和文风）。所有科学内容均由作者撰写并核实。

A 理论分析

A.1 定理 3.1 的证明

证明。 我们首先证明 Sub-EB 条件 (5) 的**充分性**。假设 V 是定义在 $\mathcal{S}$ 上的评估函数，并且对于给定的前向策略 π_F 满足 Sub-EB 条件。
那么，对于任意 $h \in [H]$:

$$\mathbb{E}_{P_F(s_h)\pi_F(s_{h+1}|s_h)}\left[\log \pi_B(s_h|s_{h+1}) + V(s_{h+1}) - \log \pi_F(s_{h+1}|s_h) - V(s_h)\right] = 0 \tag{14}$$

$$\Downarrow$$

$$\mathbb{E}_{\pi_F(s_{h+1}|s_h)}\left[\log \pi_B(s_h|s_{h+1}) + V(s_{h+1}) - \log \pi_F(s_{h+1}|s_h) - V(s_h)\right] = 0 \tag{15}$$

$$\Downarrow$$

$$V(s_h) = \mathbb{E}_{\pi_F(s_{h+1}|s_h)}\left[\log \frac{\pi_B(s_h|s_{h+1})}{\pi_F(s_{h+1}|s_h)} + V(s_{h+1})\right]. \tag{16}$$

这里，第二个等式成立，是因为我们假设任何有效的 π_F都会引入一个轨迹分布 $P_F(\tau)$，且对于任意 $\tau \in \mathcal{T}$，该分布都严格为正。因此，对于任意 $s \in \mathcal{S}$，相应的状态概率$P_F(s)$都严格为正。

根据 (16)，前述定义 $\pi_B(x|s_f)$ exp $V(s_f) := R(x)$，我们有：

$$\begin{aligned}
V(x) &= \mathbb{E}_{\pi_F(s_f|x)}\left[\log \frac{R(x)}{\pi_F(s_f|x)}\right] \\
&= \mathbb{E}_{\pi_F(s_f|x)}\left[\log \frac{F^*(x \to s_f)}{\pi_F(s_f|x)}\right] \\
&= \mathbb{E}_{P_F(x\to s_f|x)}\left[\log \frac{P_B(x \to s_f)Z^*}{P_F(x \to s_f|x)}\right] \\
&= \mathbb{E}_{P_F(x\to s_f|x)}\left[\log \frac{P_B(x \to s_f|x)}{P_F(x \to s_f|x)}\right] + \log P_B(x)Z^* \\
&= \log F^*(x) - D_{\mathrm{KL}}(P_F(\tau_{H:}|s_H)\|P_B(\tau_{H:}|s_H)), \\
V(s_{H-1}) &= \mathbb{E}_{\pi_F(x|s_{H-1})}\left[\log \frac{\pi_B(s_{H-1}|x)}{\pi_F(x|s_{H-1})} + \mathbb{E}_{P_F(x\to s_f|x)}\left[\log \frac{P_B(x \to s_f)Z^*}{P_F(x \to s_f|x)}\right]\right] \\
&= \mathbb{E}_{P_F(\tau_{H-1:}|s_{H-1})}\left[\log \frac{P_B(\tau_{H-1:})Z^*}{P_F(\tau_{H-1:}|s_{H-1})}\right] \\
&= \mathbb{E}_{P_F(\tau_{H-1:}|s_{H-1})}\left[\log \frac{P_B(\tau_{H-1:}|s_{H-1})}{P_F(\tau_{H-1:}|s_{H-1})}\right] + \log P_B(s_{H-1})Z^* \\
&= \log F^*(s_{H-1}) - D_{\mathrm{KL}}(P_F(\tau_{H-1:}|s_{H-1})\|P_B(\tau_{H-1:}|s_{H-1})), \qquad (17) \\
&\vdots \\
V(s_h) &= \mathbb{E}_{\pi_F(s_{h+1}|s_h)}\left[\log \frac{\pi_B(s_h|s_{h+1})}{\pi_F(s_{h+1}|s_h)} + \mathbb{E}_{P_F(\tau_{h+1:}|s_{h+1})}\left[\log \frac{P_B(\tau_{h+1:})Z^*}{P_F(\tau_{h+1:}|s_{h+1})}\right]\right] \\
&= \mathbb{E}_{P_F(\tau_{h:}|s_h)}\left[\log \frac{P_B(\tau_{h:})Z^*}{P_F(\tau_{h:}|s_h)}\right] \qquad (18) \\
&= \mathbb{E}_{P_F(\tau_{h:}|s_h)}\left[\log \frac{P_B(\tau_{h:}|s_h)}{P_F(\tau_{h:}|s_h)}\right] + \log P_B(s_h)Z^* \\
&= \log F^*(s_h) - D_{\mathrm{KL}}(P_F(\tau_{h:}|s_h)\|P_B(\tau_{h:}|s_h)), \qquad (19)
\end{aligned}$$

$$
\begin{aligned}
&\vdots \\
V(s_0) &= \mathbb{E}_{P_F(\tau|s_h)}\left[\log \frac{P_B(\tau)Z^*}{P_F(\tau)}\right] \\
&= \mathbb{E}_{P_F(\tau)}\left[\log \frac{P_B(\tau)}{P_F(\tau)}\right] + \log Z^* \\
&= \log F^*(s_0) - D_{\mathrm{KL}}(P_F(\tau)\|P_B(\tau)).
\end{aligned} \tag{20}
$$

需要注意的是，评估函数 (3) 的定义与方程 (19) 一致，因为

$$
\begin{aligned}
V(s_h) &= \mathbb{E}_{P_F(\tau_{h:}|s_h)}\left[\sum_{i=h}^{H} \log \frac{\widetilde{\pi}_B(s_i|s_{i+1})}{\pi_F(s_{i+1}|s_i)}\right] = \mathbb{E}_{P_F(\tau_{h:}|s_h)}\left[\sum_{i=h}^{H} \log \frac{\pi_B(s_i|s_{i+1})}{\pi_F(s_{i+1}|s_i)}\right] + \log Z^* \\
&= \mathbb{E}_{P_F(\tau_{h:}|s_h)}\left[\log \frac{P_B(\tau_{h:})Z^*}{P_F(\tau_{h:}|s_h)}\right].
\end{aligned} \tag{21}
$$

现在，我们证明 Sub-EB 条件 (5) 的**必要性**。假设 V 是与前向策略 π_F 相关联的评估函数。那么，$\forall h \in [H]$:

$$
\begin{aligned}
V(s_h) &= \mathbb{E}_{P_F(\tau_{h:}|s_h)}\left[\sum_{i=h}^{H} \log \frac{\pi_B(s_i|s_{i+1})}{\pi_F(s_{i+1}|s_i)}\right] + \log Z^* \\
&= \mathbb{E}_{\pi_F(s_{h+1}|s_h)}\left[\log \frac{\pi_B(s_h|s_{h+1})}{\pi_F(s_{h+1}|s_h)} + \mathbb{E}_{P_F(\tau_{h+1:}|s_{h+1})}\left[\log \frac{P_B(\tau_{h+1:})Z^*}{P_F(\tau_{h+1:}|s_{h+1})}\right]\right] \\
&= \mathbb{E}_{\pi_F(s_{h+1}|s_h)}\left[\log \frac{\pi_B(s_h|s_{h+1})}{\pi_F(s_{h+1}|s_h)} + V(s_{h+1})\right].
\end{aligned} \tag{22}
$$

因此，

$$
\mathbb{E}_{P_F(s_h)\pi_F(s_{h+1}|s_h)}\left[\log \pi_B(s_h|s_{h+1}) + V(s_{h+1}) - \log \pi_F(s_{h+1}|s_h) - V(s_h)\right] = 0 \tag{23}
$$

$$
\Downarrow
$$

$$
\sum_{l=i}^{j-1} \mathbb{E}_{P_F(s_l \to s_{l+1})}\left[\log \pi_B(s_l|s_{l+1}) + V(s_{l+1}) - V(s_l) - \log \pi_F(s_{l+1}|s_l)\right] = 0 \tag{24}
$$

$$
\Downarrow
$$

$$
\mathbb{E}_{P_F(\tau_{i:j})}\left[\sum_{l=i}^{j-1} \log \pi_B(s_l|s_{l+1}) + V(s_j) - V(s_i) - \sum_{l=i}^{j-1} \log \pi_F(s_{l+1}|s_l)\right] = 0 \tag{25}
$$

$$
\Downarrow
$$

$$
\mathbb{E}_{P_F(\tau_{i:j})}\left[\log P_B(\tau_{i:j}|s_j) + V(s_j) - \log P_F(\tau_{i:j}|s_i) - V(s_i)\right] = 0 \tag{26}
$$

对于任意 $i < j \in [H+1]$。 □

A.2 定理 3.2 的证明

证明。 我们首先证明 Sub-TB 条件 (1) 的**充分性** 。假设 F 是 $\mathcal{S}$上的一个状态流函数，且 π_F 是一个前向策略，并且它们满足 Sub-TB 条件。那么，对于任意 $h \in [H]$:

$$
\mathbb{E}_{P_{\mathcal{D}}(s_h \to s_{h+1})}\left[\log \pi_B(s_h|s_{h+1})F(s_{h+1}) - \log \pi_F(s_{h+1}|s_h)F(s_h)\right] = 0 \tag{27}
$$

$$
\Downarrow
$$

$$
\log \pi_B(s_h|s_{h+1})F(s_{h+1}) - \log \pi_F(s_{h+1}|s_h)F(s_h) = 0 \tag{28}
$$

$$
\Downarrow
$$

$$
\log F(s_h) = \log \frac{\pi_B(s_h|s_{h+1})}{\pi_F(s_{h+1}|s_h)} + \log F(s_{h+1}). \tag{29}
$$

这里，第二个等式成立有以下两个原因。首先，根据我们的假设，由 $\pi_{\mathcal{D}}$诱导的轨迹分布 $P_{\mathcal{D}}$对$\mathcal{T}$中的所有轨迹都赋予正概率。因此，对于任意 $(s \to s') \in \mathcal{E}$，边缘概率 $P_{\mathcal{D}}(s \to s')$ 都严格为正。其次，$\pi_{\mathcal{D}}$是任意构造的，可能与 π_F不同。令

$$\pi^\dagger(s'|s) := \frac{\pi_B(s|s')F(s')}{\sum_{s'} \pi_B(s|s')F(s')}. \tag{30}$$

然后，对 π_F 取式（29）的期望，对于任意 $h \in [H]$，有：

$$\begin{aligned}
\log F(s_h) &= \mathbb{E}_{\pi_F(s_{h+1}|s_h)}\left[\log \frac{\pi_B(s_h|s_{h+1})}{\pi_F(s_{h+1}|s_h)} + \log F(s_{h+1})\right] &(31)\\
&= \mathbb{E}_{\pi_F(s_{h+1}|s_h)}\left[\log \frac{\pi^\dagger(s_{h+1}|s_h)}{\pi_F(s_{h+1}|s_h)}\right] + \log \sum_{s_{h+1}} \pi_B(s_h|s_{h+1})F(s_{h+1}) \\
&= -D_{\mathrm{KL}}(\pi_F(\cdot|s_h)\|\pi^\dagger(\cdot|s_h)) + \log \sum_{s_{h+1}} \pi_B(s_h|s_{h+1})F(s_{h+1}). &(32)
\end{aligned}$$

注意，最后一个等式中的第二项与 $\pi_F(\cdot|s_h)$ 无关。因此，给定 $\pi_B(s_h|\cdot)$ 和 $F(s_{h+1})$，当 $\pi_F(\cdot|s_h) = \pi^\dagger(\cdot|s_h)$ 时，$F(s_h)$ 取得最大值。相应地，给定 $\pi_B(s_{h+1}|\cdot)$ 和 $F(s_{h+2})$，当 $\pi_F(\cdot|s_{h+1}) = \pi^\dagger(\cdot|s_{h+1})$ 时，$F(s_{h+1})$ 取得最大值。从 $h = 0$ 到 $h = H$ 持续进行这一递归，我们得出结论：对于任意 $(s \to s') \in \mathcal{E}$，当 $\pi_F = \pi^\dagger$ 时，F 取得最大值。

对式（28）两边取指数并对 s_{h+1} 求和，得到$\sum_{s_{h+1}} \log \pi_B(s_h|s_{h+1})F(s_{h+1}) = \sum_{s_{h+1}} \log F(s_h)\pi_F(s_{h+1}|s_h) = F(s_h)$。将其代回式（28），得到 $\pi_F = \pi^\dagger$。达到最大值时，有 $\log F(x) = \log \sum_{s_f} \pi(x|s_f)F(s_f) := \log R(x) = \log F^*(x), \log F(s_{H-1}) = \log \sum_x \pi_B(s_{H-1}|x)F^*(x) = \log F^*(s_{H-1}) \ldots, \log F(s_0) = \log \sum_{s_1} \pi_B(s_0|s_1)F^*(s_1) = F^*(s_0)$。因此，对于任意 $s \in \mathcal{S}$，有 $F(s) = F^*(s)$。结合式（28），得到 $\pi_F = \pi_F^*$，其中 $\pi_F^*(s'|s) = \pi_B(s|s')F^*(s')\, F^*(s)$ 是由 π_B 诱导的最优前向策略。~~此外，~~根据式（31）以及与 V 类似的推导，对于马尔可夫的 P_B 和任意 $h \in [H]$，有：

$$\log F(s_h) := \log F^*(s_h) - D_{\mathrm{KL}}(P_F(\tau_{h:}|s_h)\|P_B(\tau_{h:}|s_h)). \tag{33}$$

因此，当 $\pi_F = \pi_F^*$ 时，我们有：

$$\log F(s_h) = \log F^*(s_h) - D_{\mathrm{KL}}(P_F^*(\tau_{h:}|s_h)\|P_B(\tau_{h:}|s_h)). \tag{34}$$

现在我们证明 Sub-TB 条件 (1) 的**必要性**。假设 F 和 π_F 满足上述方程 (34)。那么，方程 (33) 对 F 和 π_F也必然成立，并且对于任意 $h \in [H]$，我们有：

$$\begin{aligned}
\log F(s_h) &= \mathbb{E}_{P_F(\tau_{h:}|s_h)}\left[\log \frac{P_B(\tau_{h:})Z^*}{P_F(\tau_{h:}|s_h)}\right] \\
&= \mathbb{E}_{P_F(\tau_{h:}|s_h)}\left[\sum_{i=h}^{H} \log \frac{\pi_B(s_i|s_{i+1})}{\pi_F(s_{i+1}|s_i)}\right] + \log Z^* \\
&= \mathbb{E}_{\pi_F(s_{h+1}|s_h)}\left[\log \frac{\pi_B(s_h|s_{h+1})}{\pi_F(s_{h+1}|s_h)} + \mathbb{E}_{P_F(\tau_{h+1:}|s_{h+1})}\left[\log \frac{P_B(\tau_{h+1:})Z^*}{P_F(\tau_{h+1:}|s_{h+1})}\right]\right] \\
&= \mathbb{E}_{\pi_F(s_{h+1}|s_h)}\left[\log \frac{\pi_B(s_h|s_{h+1})}{\pi_F(s_{h+1}|s_h)} + \log F(s_{h+1})\right]. &(35)
\end{aligned}$$

我们可以将上述方程改写如下：

$$\log F(s_h) := -D_{\mathrm{KL}}(\pi_F(\cdot|s_h)\|\pi^\dagger(\cdot|s_h)) + \log \sum_{s_{h+1}} \pi_B(s_h|s_{h+1})F(s_{h+1}), \quad \forall h \in [H]. \tag{36}$$

由于 $\pi_F(\cdot|s_h)$ 与 $\pi_B(s_h|\cdot)$ 和 $F(s_{h+1})$ 相互独立，假设 $\log F(s_h)$ 在 s_h 处取得最大值，表明 $\pi_F(\cdot|s_h) = \pi^\dagger(\cdot|s_h)$。因此，我们得到 $\forall h \in [H]$:

$$\log \pi_F(s_{h+1}|s_h)F(s_h) = \log \pi_B(s_h|s_{h+1})F(s_{h+1}), \tag{37}$$

而且 Sub-TB 条件可以很容易地由此推导出来。 □

A.3 定理 3.3 的证明

证明。 我们首先证明反向 Sub-EB 条件 (11) 的**充分性**。假设 W 是定义在 $\mathcal{S}\backslash\{s_f\}$ 上、满足给定后向策略 π_B 对应的反向 Sub-EB 条件的评估函数。那么，对于任意 $h \in [H-1]$:

$$\mathbb{E}_{\pi_B(s_h|s_{h+1})P_B(s_{h+1})}\left[\log \pi_B(s_h|s_{h+1}) + W(s_{h+1}) - \log \pi_F(s_{h+1}|s_h) - W(s_h)\right] = 0 \tag{38}$$

$$\Downarrow$$

$$\mathbb{E}_{\pi_B(s_h|s_{h+1})}\left[\log \pi_B(s_h|s_{h+1}) + W(s_{h+1}) - \log \pi_F(s_{h+1}|s_h) - W(s_h)\right] = 0 \tag{39}$$

$$\Downarrow$$

$$W(s_{h+1}) = \mathbb{E}_{\pi_B(s_h|s_{h+1})}\left[\log \frac{\pi_F(s_{h+1}|s_h)}{\pi_B(s_h|s_{h+1})} + W(s_h)\right]. \tag{40}$$

此处，第二个等式成立，是因为根据我们的假设，任何有效的 π_B 和 R 都会引入一个轨迹分布 $P_B(\tau)$，且对于任意 $\tau \in \mathcal{T}$，该分布都严格为正。因此，对于任意 $s \in \mathcal{S}$，相应的状态概率 $P_B(s)$也严格为正。

根据 (40)，我们有:

$$\begin{aligned}
W(s_1) &= \mathbb{E}_{\pi_B(s_0|s_1)}\left[\log \frac{\pi_F(s_1|s_0)}{\pi_B(s_0|s_1)}\right] + W(s_0) \\
&= \mathbb{E}_{P_B(s_0\to s_1|s_1)}\left[\log \frac{P_F(s_0 \to s_1)}{P_B(s_0 \to s_1|s_1)}\right] + \log F(s_0) \\
&= \mathbb{E}_{P_B(s_0\to s_1|s_1)}\left[\log \frac{P_F(s_0 \to s_1|s_1)}{P_B(s_0 \to s_1|s_1)}\right] + \log F(s_0) P_F(s_1) \\
&= \log F(s_1) - D_{\mathrm{KL}}(P_B(\tau_{:1}|s_1)\|P_F(\tau_{:1}|s_1)),
\end{aligned} \tag{41}$$

$$\vdots$$

$$\begin{aligned}
W(s_{h+1}) &= \mathbb{E}_{\pi_B(s_h|s_{h+1})}\left[\text{对数}\ \frac{\pi_F(s_{h+1}|s_h)}{\pi_B(s_h|s_{h+1})} + \mathbb{E}_{P_B(\tau_{:h}|s_h)}\left[\text{对数}\ \frac{P_F(\tau_{:h})\ P_B(\tau_{:h}|s_h)}{\tau_{:h+1}|s_{h+1})}\right]\right] + W(s_0) \\
&= \mathbb{E}_{P_B(\tau_{:h+1}|s_{h+1})}\left[\text{日志}\ \frac{P_F(\tau_{:h+1})}{P_B(\tau_{:h+1}|s_{h+1})} + \text{日志}\ F(s_0)\right] \\
&= \mathbb{E}_{P_B(\tau_{:h+1}|s_{h+1})}\left[\log \frac{P_F(\tau_{:h+1}|s_{h+1})}{P_B(\tau_{:h+1}|s_{h+1})}\right] + \log F(s_0) P_F(s_{h+1}) \\
&= \log F(s_{h+1}) - D_{\mathrm{KL}}(P_F(\tau_{:h+1}|s_{h+1})\|P_B(\tau_{:h}|s_{h+1})),
\end{aligned} \tag{42, 43}$$

$$\vdots$$

$$\begin{aligned}
W(x) &= \mathbb{E}_{\pi_B(s_{H-1}|x)}\left[\log \frac{\pi_F(x|s_{H-1})}{\pi_B(s_{H-1}|x)} + \mathbb{E}_{P_B(\tau_{:H-1}|s_{H-1})}\left[\log \frac{P_F(\tau_{:H-1})}{P_B(\tau_{:H-1}|s_{H-1})}\right]\right] + W(s_0) \\
&= \mathbb{E}_{P_B(\tau|x)}\left[\log \frac{P_F(\tau)}{P_B(\tau|x)}\right] + \log F(s_0) \\
&= \mathbb{E}_{P_B(\tau|x)}\left[\log \frac{P_F(\tau|x)}{P_B(\tau|x)}\right] + \log F(s_0) P_F(x) \\
&= \log F(x) - D_{\mathrm{KL}}(P_B(\tau|x)\|P_F(\tau|x)).
\end{aligned} \tag{44}$$

需要注意的是，评估函数 (9) 的定义与其中的方程 (43) 一致

$$\begin{aligned}
W(s_{h+1}) &= \mathbb{E}_{P_B(\tau|x)}\left[\sum_{i=0}^{h} \log \frac{\widetilde{\pi}_F(s_{i+1}|s_i)}{\pi_B(s_i|s_{i+1})}\right] = \mathbb{E}_{P_B(\tau|x)}\left[\sum_{i=0}^{h} \log \frac{\pi_F(s_{i+1}|s_i)}{\pi_B(s_i|s_{i+1})}\right] + \log F(s_0) \\
&= \mathbb{E}_{P_F(\tau_{h:}|s_h)}\left[\log \frac{P_F(\tau_{:h+1})}{P_B(\tau_{:h+1}|s_{h+1})}\right] + \log F(s_0).
\end{aligned} \tag{45}$$

特别是，当 $h = H$ 满足反向 Sub-EB 条件时，我们有

$$\mathbb{E}_{P_{\mathcal{D}}(x)}[\log \pi_F(s_f|x) + W(x) - \log \pi_B(x|s_f) - W(s_f)] = 0 \tag{46}$$

$$\Downarrow$$

$$\mathbb{E}_{P_{\mathcal{D}}(x)}[W(x) - \log R(x)] = 0. \tag{47}$$

由于 π_F 可以任意选取，只要对任意 $\tau \in \mathcal{T}$， $P_{\mathcal{D}}(\tau) > 0$，因此可得，对任意 $x \in \mathcal{X}$， $P_{\mathcal{D}}(x) > 0$，且 $W(x) = \log R(x)$。

现在，我们证明向后方向的 Sub-EB 条件 (11) 是**必要的**。假设 W 是给定前向策略 π_B 的评估函数。那么， $\forall h \in [H-1]$:

$$\begin{aligned} W(s_{h+1}) &= \mathbb{E}_{P_B(\tau_{:h+1}|s_{h+1})}\left[\sum_{i=0}^{h} \log \frac{\pi_F(s_{i+1}|s_i)}{\pi_B(s_i|s_{i+1})}\right] + W(s_0) \\ &= \mathbb{E}_{\pi_B(s_h|s_{h+1})}\left[\log \frac{\pi_F(s_{h+1}|s_h)}{\pi_B(s_h|s_{h+1})} + \mathbb{E}_{P_B(\tau_{:h}|s_h)}\left[\log \frac{P_F(\tau_{:h})F(s_0)}{P_B(\tau_{:h}|s_h)}\right]\right] \\ &= \mathbb{E}_{\pi_B(s_h|s_{h+1})}\left[\log \frac{\pi_F(s_{h+1}|s_h)}{\pi_B(s_h|s_{h+1})} + W(s_h)\right]. \end{aligned} \tag{48}$$

且 $W(x) = \log R(x)$。因此，

$$\mathbb{E}_{P_B^{\mathcal{D}}(s_{h+1})\pi_B(s_h|s_{h+1})}\left[\log \pi_B(s_h|s_{h+1}) + W(s_{h+1}) - \log \pi_F(s_{h+1}|s_h) - W(s_h)\right] = 0 \tag{49}$$

$$\Downarrow$$

$$\sum_{l=i}^{j-1} \mathbb{E}_{P_B^{\mathcal{D}}(s_l \to s_{l+1})}\left[\log \pi_B(s_l|s_{l+1}) + W(s_{l+1}) - W(s_l) - \log \pi_F(s_{l+1}|s_l)\right] = 0 \tag{50}$$

$$\Downarrow$$

$$\mathbb{E}_{P_B^{\mathcal{D}}(\tau_{i:j})}\left[\sum_{l=i}^{j-1} \log \pi_B(s_l|s_{l+1}) + W(s_j) - W(s_i) - \sum_{l=i}^{j-1} \log \pi_F(s_{l+1}|s_l)\right] = 0 \tag{51}$$

$$\Downarrow$$

$$\mathbb{E}_{P_B^{\mathcal{D}}(\tau_{i:j})}\left[\log P_B(\tau_{i:j}|s_j) + W(s_j) - \log P_F(\tau_{i:j}|s_i) - W(s_i)\right] = 0 \tag{52}$$

对于任意 $i < j \in [H+1]$ □

A.4 使用总流量估计器的评估函数

采用总流量估计器 Z 时，真实评估函数 $V^\dagger$ 仍具有 (3) 的形式，唯一不同之处在于 $\widetilde{\pi}_B(x|s_f)$，其重新定义为 $R(x)/Z$。在这种情况下，可以验证 $V^\dagger(s_0;\theta) = \log \frac{Z^*}{Z_\theta} - D_{\mathrm{KL}}(P_F(\tau)\|P_B(\tau))$。于是，$-\nabla_\theta \mathbb{E}_{\mu(s_0;\theta)}[V^\dagger(s_0;\theta)]$ 的梯度等于 $D_{\mathrm{KL}}(P_F(\tau;\theta)\|P_B(\tau)) + \frac{1}{2}\nabla_\theta(\log Z(\theta) - \log Z^*)^2$, 的梯度，其中 $\mu(s_0;\theta) := Z(\theta)/Z$，因为 s_0 是唯一的（Niu 等，2024）。

Sub-EB 条件和目标保持不变，唯有将 $P_B(x|s_f)\exp V(s_f)$ 细化为 $R(x)/Z$。

推论 A.1 (定理 3.1 的推论)。 假设 V 是定义在 $\mathcal{S}$ 上的任意评估函数， F^* 是由后向策略 π_B 诱导的最优流量。给定一个前向策略 π_F，

$$\forall h \in [H] : V(s_h) = \log \frac{F^*(s_h)}{Z} - D_{\mathrm{KL}}(P_F(\tau_{h:}|s_h)\|P_B(\tau_{h:}|s_h)), \tag{53}$$

当且仅当 V 满足 *Sub-EB* 条件 *(5)* 时。

证明。 证明方法是将定理 3.1 的证明中的 $R(x)$ 替换为 $R(x)/Z$ □

最后，在算法 1 中， $\nabla_\theta \mathbb{E}_{\mu(s_0;\theta)}[V^\dagger(s_0;\theta)] = \mathbb{E}_{\mu(s_0)}[V^\dagger(s_0)\nabla_\theta \log Z(\theta) + \nabla_\theta V^\dagger(s_0;\theta)]$可近似为 $\widehat{\nabla}_\theta \mathbb{E}_{\mu(s_0;\theta)}[V^\dagger(s_0;\theta)] := \frac{1}{K}\sum_{\tau\in\mathcal{D}}[\sum_{h=0}^{H} R(s_h \to s_{h+1})\nabla_\theta \log Z(\theta) + \sum_{h=0}^{H} A^\gamma(s_h \to s_{h+1})\nabla_\theta \log \pi_F(s_{h+1}|s_h;\theta)]$。

A.5 λ-TD 与 SUB-EB 目标的比较

传统的 λ-TD目标对于 $V(\cdot;\phi)$ 可表示如下:

$$\mathbb{E}_{P_F(\tau)}\left[\sum_{h=0}^{H}\left(V^{\lambda}(s_h)-V(s_h;\phi)\right)^2\right],$$
$$V^{\lambda}(s_h):=V(s_h)+\sum_{i=h}^{H}\lambda^{i-h}\delta_V(s_i\to s_{i+1}), \tag{54}$$

其中，在计算相对于 ϕ 的梯度时， V^{λ} 被视为常数；并且， $\delta_V(s_i\to s_{i+1})$ 等于式 (8) 中定义的 $\delta_V(\tau_{i,\ i+1})$。不考虑梯度计算时，目标值的表达式可简化为:

$$\mathbb{E}_{P_F(\tau)}\left[\sum_{h=0}^{H}\sum_{i=h}^{H}\lambda^{i-h}(\delta_V(s_i\to s_{i+1}))^2\right]. \tag{55}$$

相比之下，Sub-EB 目标值的表达式为:

$$\mathbb{E}_{P_F(\tau)}\left[\sum_{\tau_{i,j}\,:\,i<j\in[H+1]} w_{j-i}(\delta_V(\tau_{i:j}))^2\right]. \tag{56}$$

可以看出， λ-TD 目标只考虑从步骤 h 开始的事件，以学习 $V(s_h)$ 和逐边失配 $\delta_V(s_i\to s_{i+1})$。相比之下，Sub-EB 目标通过考虑从 $i(\geq h)$ 开始的子轨迹失配 $\delta_V(\tau_{i:(\cdot)})$，以及在 $j(\leq h)$ 结束的子轨迹失配 $\delta_V(\tau_{(\cdot):j})$，纳入了从 h 开始和在 h 结束的事件信息。这使得 $V(s_h)$ 的学习更加均衡可靠。此外， w 的形式可以自由选择，而在 λ-TD 目标中必须为 λ^{i-h} （Schulman 等人，2016年）。

A.6 GFlowNet 与强化学习

强化学习方法大致可分为两种主要框架（Sutton 和 Barto，2018）：第一种是 Q 学习框架，第二种是演员-评论家框架。

Q学习 在 Q 学习框架下，核心思想是学习一个参数化函数 Q ，使其针对具有边奖励 $\log \pi_B(s|s')$ 和 $\log\ \pi_B(x|s_f)\ +V(s_f)$ 的转移环境 $\mathcal{G}$，最小化**离线** 贝尔曼目标：$=\log R(x)$ （Haarnoja 等，2017）。该目标函数可写为

$$\mathbb{E}_{P_{\mathcal{D}}(s\to s')}[(\delta_Q(s\to s'))^2],\quad \delta_Q(s\to s'):=\log\pi_B(s|s')+V(s')-Q(s,s'), \tag{57}$$

其中 $V(s):=\log\sum_{s'}\exp\ Q(s,s')$ 且 $\pi_F(s'|s):=\frac{\exp Q(s,s')}{\exp V(s)}$。任何使不匹配为零的函数 Q 都保证等于最优软Q函数 Q^*，以及相应的 $V=V^*$ 和$\pi_F=\pi_F^*$。Tiapkin等人（2024年）指出，如果将 $Q(s,s')$ 视为 $\log\ F(s\to s')$，使得 $V(s)\ =\log\sum_{s'}F(s\to s')\ =\log\ F(s)$，那么贝尔曼目标就会转化为 DB目标。两者的区别仅在于参数化方式：不直接对 $Q(s,s')$ 参数化并由此推导出 V和 π_F，而是可以联合对 (π_F,V) 参数化，并定义 $Q(s,s'):=\pi_F(s'|s)\ +\log\ V(s)$[3]。作者进一步指出，从软Q学习的角度看，贝尔曼目标的最优解与DB目标的最优解一致。然而，正如作者所承认的，这一结果仅适用于固定 π_B 的 DB目标。为解决这一局限，Deleu 等人（2024年）从基于梯度的角度证明了路径一致性学习（将软Q学习作为特例）与 Sub-TB 目标（将 DB 作为特例）之间的等价性。与 Deleu 等人（2024年）相比，我们的定理 3.2 沿着这一强化学习方向给出了更直接、明确的联系。

[3]具体而言，我们有 $\delta_Q(s'\to s')\ =\log\ \pi_B(s|s')\exp\ V(s')\ -\log\ \pi_F(s'|s)\exp\ V(s)\ =\log\ \pi_B(s|s')F(s')\ -\log\ \pi_F(s'|s)F(s)$，这与 DB 目标中的表述一致。

类似于软Q学习，我们也可以改为对 V 参数化，定义 $Q(s,\ s') := \log \pi_F(s'|s) + V(s)$ 和 $\pi_F(s'|s) := \frac{\pi_B(s|s') \exp V(s')}{\sum_{s'} \pi_B(s|s') \exp V(s')}$，并通过状态级贝尔曼目标学习 V：

$$\mathbb{E}_{P_{\mathcal{D}}(s)}[(\delta_E(s))^2], \quad \delta_E(s) := \log \sum_{s'} \pi_B(s|s') \exp V(s') - \log \sum_{s'} \exp Q(s, s'), \tag{58}$$

其中 $\log \sum_{s'} \exp Q(s, s') = \log \sum_{s'} \pi_F(s'|s) \exp V(s) = V(s)$。值得注意的是，RFI（He 等，2025）遵循这一公式，不同之处在于，它将 $F(s)$ 参数化为 $\exp V(s)$，并将 $F(s)$ 乘以一个经整流的标量 $g(s)$。然而，在采样轨迹时，这种对 π_F 的定义可能会导致计算可扩展性问题。令 $Ch(s) : \{s' \in \mathcal{S} | (s \to s') \in \mathcal{E}\}$ 表示状态 s 的子状态集合。由于 $\pi_{\mathcal{D}}$ 等于 π_F 或基于 π_F 设计，从状态 s 处的 $\pi_{\mathcal{D}}(\cdot \mid s)$ 进行状态转移，需要通过 V（一个输出标量的神经网络）进行 $|Ch(s)|$ 次独立前向计算，以获得每个 $s' \in Ch(s)$ 对应的 $V(s')$ 值。相比之下，借助 Q（一个输出向量的神经网络），可以轻松获得 $\pi_F(\cdot \mid s)$；在 s 处进行一次前向计算，即可得到每个 $s' \in Ch(s)$ 对应的 $Q(s, s')$。在贝叶斯网络结构学习等任务中，$|Ch(s)|$ 可能大于 10^2,，因此，只对 V 进行参数化的方法，可能远慢于联合参数化 (V, π_F) 或直接参数化 Q 的方法。

演员-评论家本文以及 Niu 等人（2024）提出的工作基于策略梯度理论（Agarwal 等人，2021），该理论采用演员-评论家框架（Sutton 和 Barto，2018；Haarnoja 等，2018）。在此框架中，软演员-评论器算法的核心思想是学习一个参数化评估函数 V，使其最小化（但不一定）**在线**贝尔曼目标：

$$\mathbb{E}_{P_F(s \to s')}[(\delta_S(s \to s'))^2], \quad \delta_S(s \to s') := \log \pi_F(s'|s) + V(s) - Q(s, s'), \tag{59}$$

其中 $Q(s,\ s') := \log \pi_B(s|s') + V(s')$，且 $Q(x,\ s_f) := \log R(x)$，其中上文的 $\mathbb{E}_{P_F(s \to s')}[\ldots]$ 可以推广为 $\mathbb{E}_{P_{\mathcal{D}}(s)}[\mathbb{E}_{\pi_F(s'|s)}[\ldots]]$，但仍必须保留内层在线期望。值得注意的是，在线贝尔曼目标可以视为所提出的 Sub-EB 目标的一个特例。**此时，强化学习的两个方向开始出现分歧**（Schulman 等人，2017）。当 V达到该目标的最优解（记为 $V^\dagger$）时，我们有：

$$\begin{aligned}
V(s) = V^\dagger(s) &= \mathbb{E}_{\pi_F(s'|s)}[Q(s, s') - \log \pi_F(s'|s)] \\
&= \mathbb{E}_{\pi_F(s'|s)} \left[\frac{\exp Q(s, s')}{\log \sum_{s'} \exp Q(s, s')} - \log \pi_F(s'|s) + \log \sum_{s'} \exp Q(s, s') \right] \\
&= -D_{KL}\left(\pi_F(\cdot|s) \| \pi_Q(\cdot|s)\right) + \log \sum_{s'} \exp Q(s, s'), \quad \pi_Q(s'|s) := \frac{\exp Q(s, s')}{\sum_{s'} \exp Q(s, s')}.
\end{aligned}$$

定义 $F(s \to s') := \exp Q(s,\ s')$ 可得到与正文中所呈现结果一致的表达式。这解释了为什么 $V^\dagger$（及其近似形式 V）可作为评论器：它评估行动者 π_F 距离局部最优策略 π_Q还有多远。这里，π_Q 仅是局部最优的，因为 $Q(s,\ s')$ 仍可能偏离全局最优点 $Q^*(s,\ s')$。随后，使用 V相对于 π_F最小化散度，使得 π_F的 $V^\dagger$ 更加接近最优的那个 V^*。实现这一点有两种方式：对于所有采样状态，直接将 $\pi_F(\cdot \mid s)$ 设为 $\pi_Q(\cdot \mid s)$；或者沿采样轨迹应用策略梯度更新。

A.7 GFlowNet 的软演员-评论器

为 GFlowNet 训练量身定制的软演员-评论器算法如算法 3 所示。当通过算法 3 中的采样访问所有状态，且 V 的在线贝尔曼目标达到零时，即 $\forall s \in \mathcal{S}$：$D_{KL}(\pi_F(\cdot \mid s)\ \pi_Q(\cdot \mid s)) = 0$，并且 $V = V^\dagger$，我们将在下文证明，策略 π_F 和 V 将收敛到最优量 π_F^* 和 V^*。从终止状态开始，我们有：

$$\begin{aligned}
V(x) &= -D_{KL}(\pi_F(\cdot|x) \| \pi_Q(\cdot|x)) + Q(x, s_f) \\
&= Q(x, s_f) := \log R(x) \qquad (60) \\
Q(s_{H-1}, x) &:= \log \pi_B(s_{H-1}|x) + V(x) \\
&= \log \frac{F^*(s_{H-1} \to x)}{F^*(x)} + \log R(x) = \log F^*(s_{H-1} \to x), \qquad (61)
\end{aligned}$$

其中我们使用了 π_B 的定义。接下来，我们有：

$$V(s_{H-1}) := -D_{KL}\left(\pi_F(\cdot|s_{H-1})\|\pi_Q(\cdot|s_{H-1})\right) + \log\sum_x \exp Q(s_{H-1},x)$$
$$= \log\sum_x F^*(s_{H-1},x) = \log F^*(s_{H-1}), \tag{62}$$

$$\pi_F(x|s_{H-1}) = \pi_Q(x|s_{H-1}) := \frac{\exp Q(s_{H-1},x)}{\sum_x \exp Q(s_{H-1},x)}$$
$$= \frac{F^*(s_{H-1}\to x)}{F^*(s_{H-1})} := \pi_F^*(x|s_{H-1}). \tag{63}$$

继续递归下去，我们将得到对于所有状态和边的 $V(s) = \log F^*(s)$、$Q(s\to s') = \log F^*(s\to s')$ 以及 $\pi_F(s'|s) = \pi_F^*(s'|s)$。

对 π_F 的在线期望会使算法 3 的计算开销很大，并阻碍对逐边形式 $\delta_S(s\to s')$的改进。为解决这一问题，可以将其修改为算法 4，其中的关键区别在于从 $P_F(\tau)$ 中采样在线轨迹。$\delta_S(\tau_{i,\ j})$ 的定义与 $\delta_V(\tau_{i,\ j})$ 相同。

算法 3 软演员-评论家工作流

要求：$\pi_F(\cdot|\cdot;\theta)$、$\pi_B(\cdot|\cdot)$、$V(\cdot;\phi)$、批量大小 K、总迭代次数 $N$for $n = 1,\ldots,N$ do$\mathcal{D}\leftarrow\{\tau^k|\tau^k\sim P_{\mathcal{D}}(\tau)\}_{k=1}^K$基于 $\mathcal{D}$，依据关于 $\frac{1}{K}\sum_{\tau\in\mathcal{D}}[\sum_{s\in\tau}\mathbb{E}_{\pi_F}(s'|s)$ 的梯度更新 $\phi\,[\delta_S(s\to s';\phi)^2]]$。基于 $\mathcal{D}$ 和 V，将 $\pi_F(_|s)$ 设为 $\pi_Q(\cdot|s)$，对任意 $s(\neq s_f)$ 均如此 $\in\mathcal{D}$。结束循环 返回 $\pi_F(\cdot|\cdot;\theta)$、$V(\cdot;\phi)$

算法 4改进的软演员-评论家工作流

需要：$\pi_F(\cdot|\cdot;\theta)$、$\pi_B(\cdot|\cdot)$、$V(\cdot;\phi)$、批量大小 K、总迭代次数 N对于 $n = 1,\ldots,N$ 执行$\mathcal{D}\leftarrow\{\tau^k|\tau^k\sim P_F(\tau)\}_{k=1}^K$根据 $\mathcal{D}$，通过相对于 $\frac{1}{K}\sum_{\tau\in\mathcal{D}}[\sum(s\to s')$ 的梯度更新 $\phi\in\tau\,\delta_S(s\to s';\phi)^2]$ (或$\frac{1}{K}\sum_{\tau\in\mathcal{D}}[\sum_{\tau_{i,j}\in\tau}w_{j-i}\delta_S(\tau_{i,j};\phi)^2]$)。根据 $\mathcal{D}$ 和 V，对于任意 $s(\neq s_f)\in\mathcal{D}$，将 $\pi_E(\cdot|s)$ 设为 $\pi_Q(\cdot|s)$。结束循环 返回 $\pi_F(\cdot|\cdot;\theta)$、$V(\cdot;\phi)$

与算法 4 相比，算法 1 运用策略梯度理论改进了基础策略优化算子，使其适用于实际实现。由于原始策略梯度方法严格遵循同策略方式运行，因此引入算法 2，以支持使用离线采样器 $P_{\mathcal{D}}$。

B 实验设置与结果

超参数 对于原始的基于策略的方法，以及采用 Sub-TB 目标的所提方法（RL 和 Sub-EB），根据 Niu 等人（2024）报告的消融研究结果，我们将超参数 γ 设为 0.99。对于 Sub-TB 的数据收集策略 $\pi_{\mathcal{D}}$，超参数 α 从 1.0 开始，并以 0.99 的速率指数衰减；该衰减率同样根据 Niu 等人（2024）的消融研究结果确定。在 Sub-TB 目标中，按照 Madan 等人（2023）的消融研究，将超参数 λ 设为 0.9。对于 Sub-EB 目标，λ 设为 0.9，该值是根据图 6 所示的消融研究结果从 $\{0.1, 0.2, \ldots, 0.9, 0.99\}$ 中选出的。所有实验均在 Grace 集群上进行，该集群是德克萨斯农工大学高性能计算系统的主要组成部分之一。每个节点均配备 Intel Xeon Cascade Lake CPU。实验中，我们申请了一个节点和 20 GB 内存，未使用 GPU。附录 B.5 报告了各实验中所考察方法的运行时间。

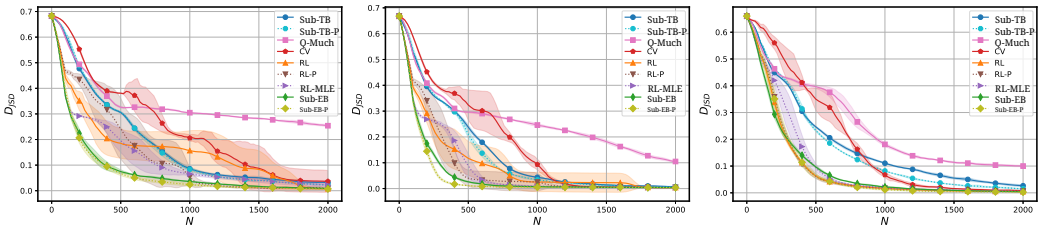


图 4：根据每种方法的五次随机初始化运行结果，绘制了不同训练方法在 256×256（左）、128×128（中）和 $64 \times 64 \times 64$（右）网格上的 D_{JSD} 均值和标准差（以阴影区域表示）。

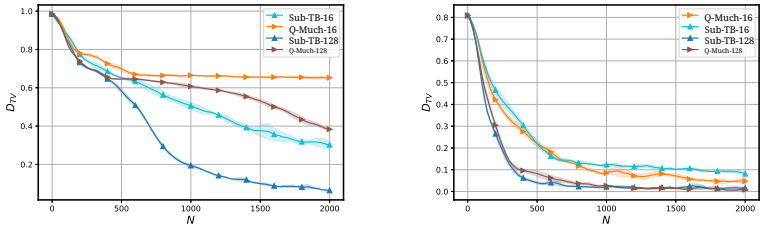


图 5：根据 Sub-TB-16、Sub-TB-128、Q-much-16 和 Q-much128 的五次随机初始化运行结果，绘制了不同训练方法在 128×128（左）和 20×20（右）网格上的 D_{TV}（右）均值和标准差（以阴影区域表示）。这里，“16”和“128”表示训练批量大小。

优化 优化采用 Adam优化器。按照 Niu 等人（2024）的做法，每次优化迭代的样本批量大小设为 128。 $\pi_F(\cdot|\cdot;\theta)$ 和 $\log F(\cdot;\theta)$ 的学习率均为 1×10^{-3}，该值根据 Sub-TB 在 256×256 网格上的表现从 $\{5 \times 10^{-3}, 1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$中选出。 $V(\cdot;\phi)$ 的学习率设为 5×10^{-3}，该值根据 RL 在 256×256 网格上的表现从 $\{10^{-2}, 5 \times 10^{-3}, 10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$ 中选出。在所有实验中，每种训练方法均运行五次，分别使用五个不同的随机种子进行初始化。

模型架构 前向策略 $\pi_F(\cdot\,|\,\cdot;\theta)$ 和评估函数 $V(\cdot;\phi)$ 均由具有四个隐藏层的神经网络参数化，每层的隐藏维度为 256。后向策略 $\pi_B(\cdot\,|\,\cdot)$ 是在有效转移（边）上的均匀分布。在超网格、序列设计和分子图设计实验中，坐标元组、整数序列和整数矩阵在输入神经网络之前均经过 K-hot 编码。在贝叶斯网络结构学习中，邻接矩阵直接作为输入传入神经网络，无需编码。

性能指标 第一个指标是 D_{TV} $P_F(x)$ 与 $P^*(x)$ 之间的总变差，其定义为： $D_{\mathrm{TV}}(P_F(x)\ P^*(x)) = \frac{1}{2}\sum_{x\in\mathcal{X}}|P_F(x) - P^*(x)|$。文献中采用的另一种性能指标是平均 l_1- 距离，其定义为 $|\mathcal{X}|^{-1}\sum_x |P^*(x) - P_F(x)|$。我们改用 D_{TV} 的原因如下。设计空间 $|\mathcal{X}|$ 通常很大（$> 10^4$），且$\sum_x |P^*(x) - P_F(x)| \le 2$，导致平均 l_1-距离被 $|\mathcal{X}|$大幅缩小。我们还使用詹森-香农散度 D_{JSD}作为第二项指标来评估不同方法，其表达式为： $D_{\mathrm{JSD}}(P_F(x)\ P^*(x)) = \frac{1}{2}D_{\mathrm{KL}}(P_F(x)\ P_M(x)) + \frac{1}{2}D_{\mathrm{KL}}(P^*(x)\ P_M(x))$，其中 $P_M := \frac{1}{2}(P_F + P^*)$。

B.1 超网格

超网格实验的生成过程定义如下。对于高度为 H、宽度为 D 的网格，除最终状态 $\mathcal{S}\backslash\{s_f\}$ 之外的状态空间由所有形如 $\{s = ([s]_1、\ldots\ldots、[s]_d、\ldots\ldots、[s]_D) \mid [s]_d \in \{0, \ldots, H-1\}\}$ 的 D 维坐标向量构成。生成

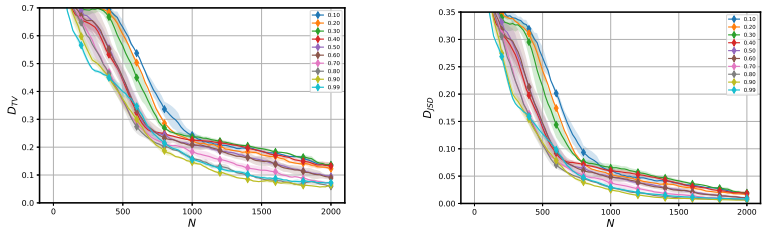


图 6： 128×128 网格上 Sub-EB 各次运行中， D_{TV} （左）和 D_{JSD} （右）的均值及标准差（以阴影区域表示）。Sub-EB 的超参数 λ 范围为 0.1 至 0.99。结果基于每种设置下五次 Sub-EB 运行。

过程从初始状态 $s_0 =$(0、.....、0) 开始，并在最终状态 s_f 结束，我们将其记为 (−1、.....、 −1)。从任意状态 $s \in \mathcal{S}\backslash\{s_f\}$ 出发，都有 $D+1$ 种有效转移（边）：(1) 对于每个 $d \in \{0,\ldots,D\}$，第 d 个坐标可以加一，从而到达一个新状态 $s' =([s]_1$、 ..... 、 $[s]_d+1$、 ..... 、 $[s]_D)$；(2) 如果 $s \in \mathcal{X}$，则可以通过执行转移 $(s \to s_f)$ 停止该过程，并将 s 作为终止坐标元组 x 返回。根据上述定义，GFlowNet 有向无环图 $\mathcal{G}$ **并非** 分层的，这意味着每个状态（除 s_f 外）都可以作为终止状态返回。与终止状态相关的奖励函数定义如下：

$$R(x) = R_0 + R_1 \prod_{d=1}^{D} \mathbb{I}\left[\left|\frac{[s]_d}{N-1} - 0.5\right| \in (0.25, 0.5]\right] + R_2 \prod_{d=1}^{D} \mathbb{I}\left[\left|\frac{[s]_d}{N-1} - 0.5\right| \in (0.3, 0.4]\right],$$

其中， $R_0 = 10^{-2}$、 $R_1 = 0.5$ 和 $R_2 = 2$ 是我们实验中的参数。

其他实验结果 正如附录 A.6 中所讨论的，与（Sub-）轨迹级方法相比，Q-Much 可能更倾向于利用，并且在具有明显分离众数的高难度任务上表现较差。为验证这一点，我们在 20×20 网格和 128×128 网格上进行了超网格任务实验，批量大小分别设为 16 和 128（默认设置）。理由如下：对于高度为 N、维度为 D 的网格，奖励模式之间的最小距离仅取决于高度 N。当 D 适中且 N较小时，即使是均匀策略也仍有相当大的概率到达这些模式。因此， N 对任务难度的影响更为显著。实验结果见图 5。在批量大小为 16 的 20×20 网格上，Q-Much 的表现优于 Sub-TB。然而，批量大小为 128 时，这一优势便不复存在。在批量大小为 128 的 128×128 网格上，Q-Much 的表现明显差于 Sub-TB。将批量大小降至 16 后，两种方法都未能收敛到令人满意的解；尽管如此，Q-Much 的表现仍逊于 Sub-TB。

关于.....的消融研究 λ 对于 128×128 网格，我们对 Sub-EB 权重 w_{j-i} 的超参数λ 进行消融研究，以考察其对基于策略的 GFlowNet 训练的影响。我们使用 λ 分别为 0.1，0.2、.....、0.90 和 0.99 的 Sub-EB 方法进行实验。实验结果如图 6 所示。可以观察到，采用 $\lambda = 0.9$的 Sub-EB 方法取得了最佳性能。尽管收敛速率和最终性能有所不同，但所有配置下的 Sub-EB 方法均表现出良好的稳定性，这表明所提出的 Sub-EB 目标能够实现评估函数 V 的可靠学习。

B.2 序列设计

这组实验的生成过程定义如下。除最终状态 $\mathcal{S}\backslash\{s_0, s_f\}$ 外，状态空间等于 $\bigcup_{t=1}^{H}\{0,\ldots,N-1\}^t$，其中每个状态都是由介于 0 和 $N-1$之间的整数构成的序列。集合 $\{0,\ldots,N-1\}$ 对应于 N种构建块。该过程从初始状态 $s_0 = \emptyset$开始，即空序列。对于任意中间状态 $s_h \in \mathcal{S}_h$，其中包含从 $\{0,\ldots,N-1\}$中选取的 h 个元素。对于满足 $t < D$的 s_t，有 $N \times 2$种可能的状态转移，分别对应于从 $\{0,\ldots,N-1\}$ 中取一个元素并将其添加到当前序列的末尾或开头。该过程持续进行，直到序列长度达到 D，从而得到 $\mathcal{X} = \{0,\ldots,N-1\}^D$中的终止状态，并转移至 s_f:=(N，.....，N)。根据生成过程的定义，GFlowNet 有向无环图 $\mathcal{G}$ 是分层的，且 $\mathcal{S}_D = \mathcal{X}$。

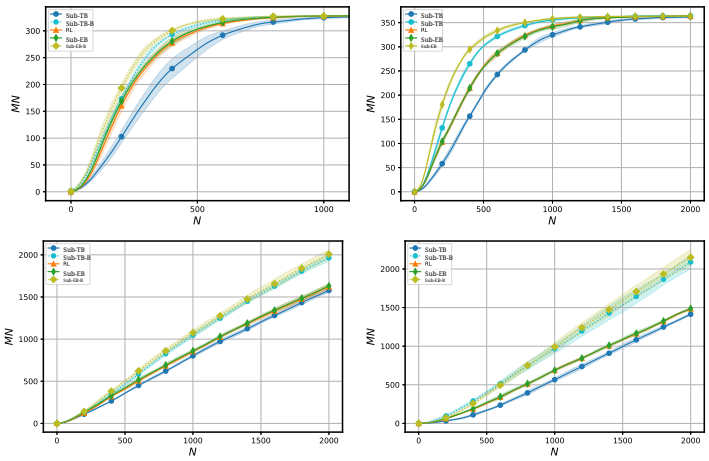


图 7：基于每种方法的 5 次随机初始化运行，绘制了 SIX6（左上）、QM9（右上）、PHO4（左下）和 sEH（右下）数据集的 MN 均值和标准差（以阴影区域表示）。

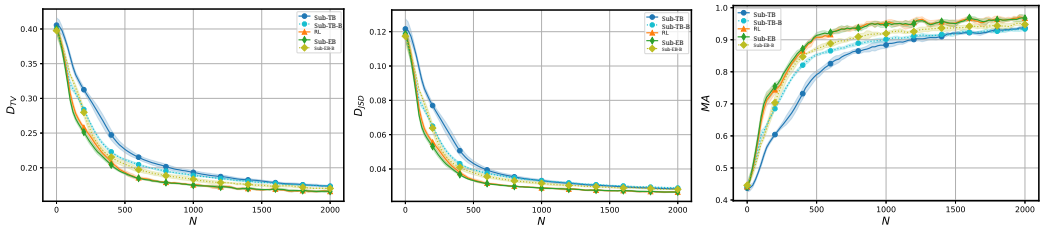


图 8：基于每种方法的五次随机初始化运行，展示 SIX6 数据集上 D_{TV}（左）、 D_{JSD}（中）和模式准确率（右）的均值和标准差（以阴影区域表示）。

在实践中，每个状态 s_t 都表示为长度固定为 D的序列，其中前 t 个元素是来自 $\{0,\ldots,N-1\}$的整数，其余元素均等于 -1。具体而言，初始状态 s_0 表示为（-1，.....，-1）。我们使用 Shen 等人（2023）提供的核苷酸序列数据集（SIX6 和 PHO4）以及分子序列数据集（QM9 和 sEH）。奖励定义为数据集原始分数的指数形式，即 $R(x) = \text{score}^{\beta}(x)$。对于 SIX6、QM9、PHO4 和 sEH，超参数 β分别设为 3、5、3 和 6，奖励则分别归一化为 $[10^{-3},10]$、$[10^{-3},10]$、$[0,10]$ $[10^{-3},10]$ 和 。在此实验设置中，我们考虑 Shen 等人（2023）引入的另一项指标，即 $P_F(x)$相对于 $P^*(x)$的模式准确率（MA），定义如下：

$$\mathrm{MA}(P_F(x),P^*(x)) = \min\left(\frac{\mathbb{E}_{P_F(x)}[R(x)]}{\mathbb{E}_{P^*(x)}[R(x)]},1\right). \tag{64}$$

我们使用动态规划（Malkin等人，2022a）计算 $P_F(x)$ 和精确的模式准确率，以及 $P_F(x)$ 与 $P^*(x)$ 之间的 D_{TV}和 D_{JSD}。

在这组实验中，我们不仅关注分布建模，还关注模式发现，其目标是找出高奖励的终止状态。除了强化学习、Sub-EB 和 Sub-TB 方法外，我们还考虑了 Sub-TB-B 和 Sub-EB-B，它们在轨迹采样过程中明确促进对高奖励状态的探索。

分布建模结果 我们通过由模式准确率、 D_{TV} 和 D_{JSD}衡量的性能来比较不同的训练方法，如图 8、9、10和11所示，分别对应SIX、QM9、PHO4和sEH数据集。可以看出，Sub-EB的表现略优于强化学习。这可以归因于强化学习在这些实验中具有足够的稳定性，使得Sub-TB目标所带来的优势不那么明显。尽管如此，Sub-EB和强化学习在以下两方面均优于Sub-TB

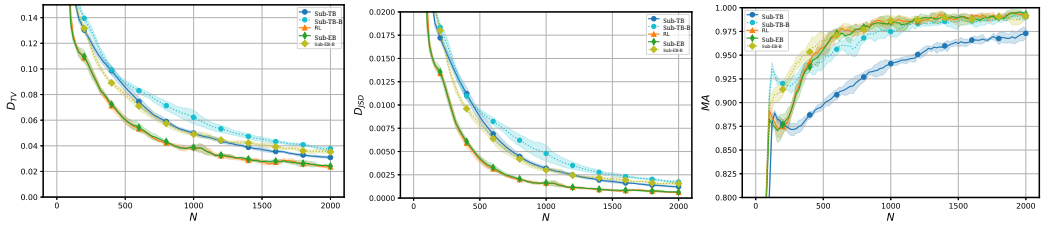


图 9：基于每种方法的五次随机初始化运行，绘制 QM9 数据集上 D_{TV}（左）、 D_{JSD}（中）和模式准确率（右）的均值和标准差（以阴影区域表示）。

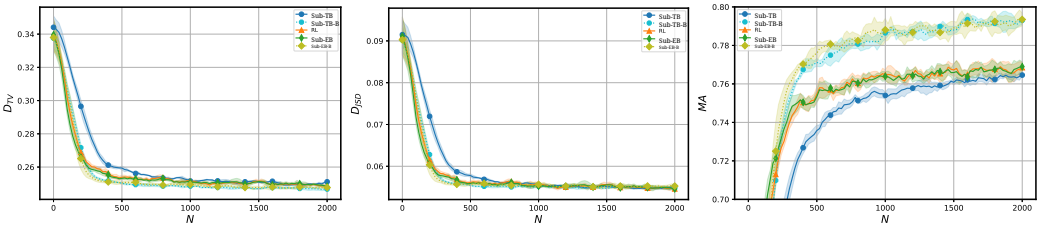


图 10：基于每种方法的五次随机初始化运行，绘制 PHO4 数据集上 D_{TV}（左）、 D_{JSD}（中）和模式准确率（右）的均值和标准差（以阴影区域表示）。

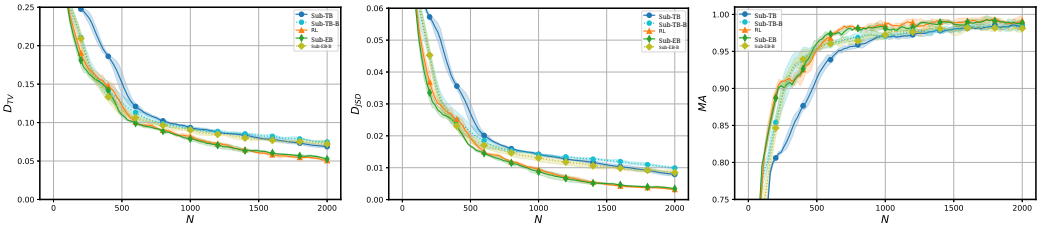


图 11：基于每种方法的五次随机初始化运行，绘制 sEH 数据集上 D_{TV}（左）、 D_{JSD}（中）和模式准确率（右）的均值和标准差（以阴影区域表示）。

收敛率和最终性能。对于离线变体，可以观察到Sub-EB-B的表现略优于Sub-TB-B，但除PHO4外，在所有数据集上的表现均不如Sub-EB。实验结果验证了所提出的Sub-EB目标能够有效应用于基于策略的训练流程，以学习估值函数 V。

模式发现结果 我们使用模式数量（MN）衡量性能，其定义为训练过程中发现的高奖励终止状态的不同数量。如果某个终止状态的奖励位于 SIX6、QM9 和 PHO4 数据集奖励的前 0.5%，或 sEH 数据集奖励的前 0.01%，则将其视为高奖励状态。如图 7 所示，尽管除 PHO4 外，所有数据集上的分布建模性能都有所下降，Sub-EB-B 和 Sub-TB-B 仍能更快地找到固定数量的不同众数，并在固定数量的优化迭代内发现更多不同众数。这些结果验证了我们的离线基于策略的训练流程，并支持了我们的主张：所提出的 Sub-EB 目标能够整合离线采样技术。

结合图 7 中的模式发现结果，这表明鼓励生成高奖励终止状态的离线技术有助于模式发现，但可能会妨碍准确的分布建模。

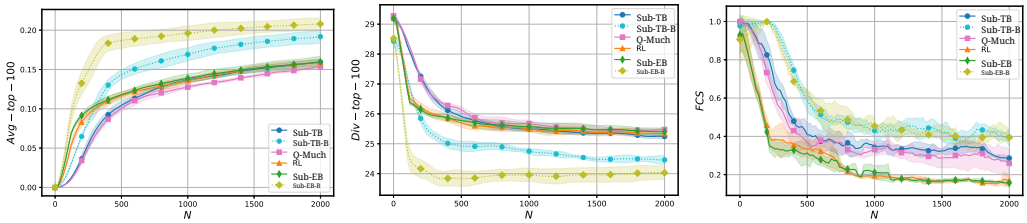


图 12：基于每种方法的五次随机初始化运行，展示了 15 个节点的前 100 个互不相同候选图的平均奖励（左）、多样性（中）和 FCS（右）的均值及标准差（以阴影区域表示）。

B.3 贝叶斯网络结构学习

贝叶斯网络是一种概率模型，用于表示 N 随机变量的联合分布，其因子分解由网络结构 x决定，而网络结构是一个有向无环图。因此，该分布可写为 $P(y_1,\ldots,y_N)=\prod_{n=1}^{N}P(y_n|Pa_x(y_n))$，其中 $Pa_x(y_n)$ 表示图 x 中 y_n 的父节点。由于任何图结构都可编码为邻接矩阵，因此除最终状态外的状态空间定义为 $\mathcal{S}\backslash\{s_f\}:=\{s|\mathcal{C}(s)=0,s\in\{0,1\}^{N\times N}\}$ 其中 $\mathcal{C}$对应于 Deleu 等人（2022）提出的无环图约束。需要注意的是，每个贝叶斯网络有向无环图 $x\in\mathcal{X}$都对应于 GFlowNet 有向无环图 $\mathcal{G}$中的一个状态 $s\in\mathcal{S}$。生成过程从初始状态 $s_0=0^{N\times N}$开始，该状态表示不含边的图，并在 $s_f:=-1^{N\times N}$处结束。对于任意 $s\in\mathcal{S}\backslash s_f$，一次状态转移（$s\to s'$）对应于将邻接矩阵中的一个零值翻转为一，从而添加一条边，前提是所得状态 s' 仍然无环。或者，生成过程也可以通过状态转移到s_f 而停止，并返回 s 作为终止图结构。根据生成过程的定义，在这组实验中不对 $\mathcal{G}$ 进行评分。给定由 $y_{1:N}$构成的观测样本集 $\mathcal{D}_y$，结构学习的目标是近似后验分布 $P(x|\mathcal{D}_y)\propto P(\mathcal{D}_y|x)P(x)$。在缺乏关于 x的先验知识时，通常假设先验分布 $P(x)$是均匀分布，从而将任务简化为最大化似然 $P(\mathcal{D}_y|x)(\propto P(x|\mathcal{D}_y))$。

真实图结构 x^* 及其对应的数据集 $\mathcal{D}_y$（大小为 $|\mathcal{D}_y|=10^3$）由 Erd˝os–R´enyi 模型（Deleu 等人，2022）模拟生成。我们使用 BGe 分数（Kuipers 等人，2014）评估生成的图结构，并将奖励定义为 $R(x)=\left(\begin{smallmatrix}\Delta(x;\mathcal{D}_y)\\ C\end{smallmatrix}\right)^{\beta}$，其中 $\Delta(x;\mathcal{D}_y)=$~~BGe(x~~; $\mathcal{D}_y$) –BGe(s_0; $\mathcal{D}_y$)，$\beta=10$ 使奖励函数更偏向高分结构，而$C:=\Delta(x^*;\mathcal{D}_y)$ 对奖励进行归一化，使得 $R(x^*)=1$。具有 n 个节点的有向无环图的确切数量记为 $a(n)$，满足（Robinson，2006年）：

$$a(n)=\sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}2^{k(n-k)}a(n-k), \tag{65}$$

其中 $a(0)=1$。为便于评估分布建模性能，Malkin 等人（2022b）将节点数设为 5，由此产生约 2.92×10^4 个可能的有向无环图。在这一设置中，真实有向无环图包含 5 条边。除了这个小规模案例外，我们还考虑了两个规模大得多的案例，节点数分别为 10 和 15，对应约 4.18×10^{18} 和 2.38×10^{35} 个可能的有向无环图。在这两个案例中，我们分别将真实有向无环图设为包含 10 条边和 15 条边。

其他实验结果 对于小规模情况，我们使用动态规划（Malkin等人，2022a）显式计算已学习 π_F 的 $P_F(x)$，并计算 $P_F(x)$与 $P^*(x)$之间的精确 D_{TV} 和 D_{JSD}。图 13 分别绘制了 Sub-TB、强化学习、Sub-EB 和 Q-Much 五次运行的 D_{TV} 和 D_{JSD}的均值与标准差。可以看出，Sub-EB 的表现优于强化学习、Sub-TB 和 Q-Much。这些结果证实了我们的结论：与 λ-TD目标相比，Sub-EB 目标能够更可靠地学习评估函数 V，从而提升基于策略的方法的性能。我们还给出了 Sub-EB-B 和 Sub-TB-B 的结果。可以看出，它们的表现都不如各自的原始版本。这些结果进一步支持了我们的结论：显式鼓励对高奖励区域进行探索，可能无益于分布建模性能。

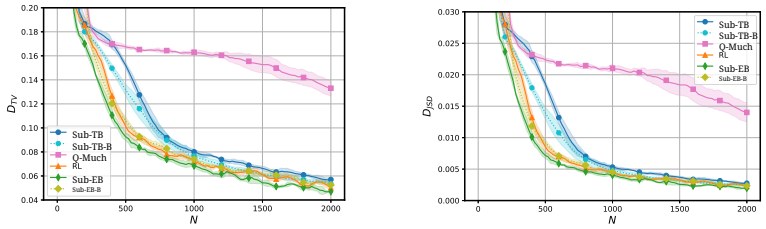


图 13：对于包含 5 个节点的贝叶斯网络结构学习任务，基于每种方法的五次随机初始化运行，绘制了D_{TV}（左）和 D_{JSD}（右）的均值和标准差（以阴影区域表示）。

B.4 分子图设计

每个分子（图）由至多 8 个构建块组成，这些构建块从 Bengio 等人（2021）提供的 $N = 105$个预定义分子子结构中选取。每个构建块（图节点）都有一个列表，其中包含至多 M 个可用的原子索引，用于与其他构建块形成化学键（图边）。每个构建块的列表恰好包含一个目标原子和 0 至 $M-1$ 个源原子。通过将一个构建块中的源原子与另一个构建块中的目标原子连接，可以形成化学键。图 14 展示了 4 个示例构建块的图示。遵循 Bengio 等人（2021）的做法，我们将分子图中的最大构建块数量限制为 $H = 8$。每个状态 $s \in \mathcal{S}$ 表示为一个 $H \times 2$矩阵。生成过程从初始状态 $s_0 = -1^{H\times 2}$开始，该状态表示空图，并在 $s_f = 105^{H\times 2}$处结束。对于任意状态 $s_h \in \mathcal{S}_h$:

1. 第一列包含从 $\{0, \ldots, N-1\}$ 中抽取的 h 个整数，表示分子中所含构建块的索引。
2. 第二列包含从 $\{0, \ldots, H \times M - 1\}$ 中抽取的 $h-1$ 个整数，表示编码连通结构的成键位点索引。具体而言，如果一个整数等于 $(i-1) \times H + j$，则表示新添加构建块的目标原子将与第 j 个构建块的第 i 个源原子相连。

对于任意 $s \in \mathcal{S} \setminus (s_f, s_0)$，状态转移 $(s \to s')$ 对应一个两级动作:

1. 选择要添加的构建块，由 $\{0, \ldots, N-1\}$ 中的一个整数表示。2. 选择一个成键位点，由整数 $k \in \{0, \ldots, H \times M - 1\}$ 表示。

当执行状态转移 $(x \to s_f)$、没有可用于形成化学键的源原子，或状态达到 H 个块的上限时，生成过程就会停止。根据生成过程的定义，GFlowNet 有向无环图 $\mathcal{G}$ **并非** 分层，这意味着除 s_f 以外的任何状态都可以是终止状态 $x \in \mathcal{X}$。由于给定一对块之间可能存在多种类型的化学键，终止状态空间的大小大于 $\frac{105!}{(105-8)!} \approx 10^{16}$。*pyTDC* 软件包提供的辛醇-水分配系数（LogP）和 c-Jun N端激酶3（JNK3）评分被直接用作奖励函数 $R(x)$。由于该生成过程中涉及两个层级的动作，我们使用独立的神经网络分别表示每个动作组成部分对应的策略。动作的对数概率通过将两个动作层级对应网络输出的对数概率相加来计算。最后，遵循 Bengio 等人（2021）的做法，我们选择参数化的 π_B，并使用批量大小为 4，以模拟查询现实世界分子预言机的计算成本高昂的情形。

实验结果如图15和图16所示。我们报告了训练过程中发现的前100个唯一图的平均奖励和多样性。分子集合的多样性通过*pyTDC* 软件包，基于 Morgan 指纹的 Tanimoto 距离计算两两不相似度的平均值。对于 LogP 任务，强化学习、Sub-EB 和 Q-Much 获得了相近的平均奖励，且均优于 Sub-TB。此外，强化学习和 Sub-EB 的收敛速度远快于 Sub-TB 和 Q-Much。四种方法的多样性相近。对于 JNK3 任务，Sub-EB 的平均奖励最高，且在所有方法中收敛最快。其多样性与 Sub-TB 相近，高于强化学习。尽管 Q-Much 的多样性最高，但其平均奖励几乎最低。

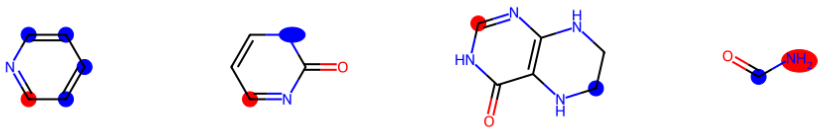


图14：四个示例构建块。红点和蓝点表示可用于成键的原子，其中每条成键边均从红点出发，以蓝点为终点。

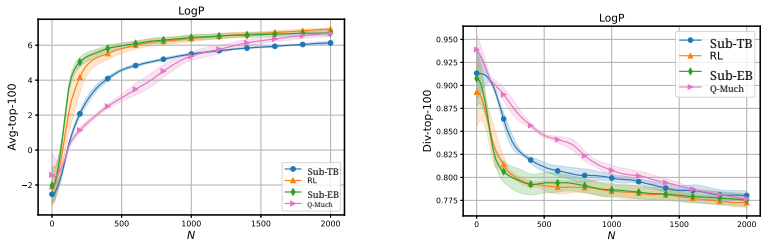


图15：LogP 任务中排名前 100 的唯一分子的平均奖励（左）和多样性（右，分子 Morgan 指纹的 Tanimoto 多样性）均值和标准差图。结果基于每种方法的五次随机初始化运行。

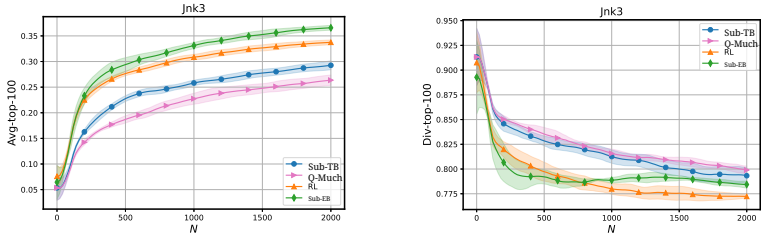


图16：JNK3 任务中排名前 100 的唯一分子的平均奖励（左）和多样性（右，分子 Morgan 指纹的 Tanimoto 多样性）均值和标准差图。结果基于每种方法的五次随机初始化运行。

总体而言，这些结果表明，Sub-EB 在平均奖励和收敛速度方面表现最佳，同时保持了合理的多样性，证实了其在大规模任务中的有效性。

B.5 运行时间报告

方法	Mean	Std
Sub-TB	1 小时 08 分钟	1.09
Sub-TB-P	1 小时 14 分钟	10.36
RL	**1小时03分**	0.92
RL-P	1小时27分	0.57
RL-MLE	1小时09分	1.78
Sub-EB	1小时17分	2.31
Sub-EB-P	1小时17分钟	1.49

方法	Mean	Std
Sub-TB	0小时52分钟	0.27
Sub-TB-P	0小时58分钟	7.92
RL	**0小时44分钟**	1.43
RL-P	1小时06分钟	2.21
RL-MLE	0小时58分钟	3.27
Sub-EB	0小时54分钟	2.66
Sub-EB-P	1小时11分钟	0.82

方法	Mean	Std
Sub-TB	1小时59分钟	1.53
Sub-TB-P	2小时19分钟	4.40
RL	**1小时27分钟**	7.37
RL-P	2小时03分钟	10.68
RL-MLE	2小时28分钟	9.58
Sub-EB	2 小时 04 分钟	4.11
Sub-EB-P	2小时41分钟	15.46

表1：各方法在 $64 \times 64 \times 64$（左）、$128 \times 128 \times 128$（中）和 $256 \times 256 \times 256$（右）网格上的总运行时间均值和标准差（标准差值）。这里，“h”表示小时，“m”表示分钟，所有标准差值均以分钟为单位。

方法	Mean	Std
Sub-TB	0小时05分	0.21
Sub-TB-B	0小时16分	0.67
RL	**0小时04分**	0.06
Sub-EB	**0小时04分**	0.34
Sub-EB-B	0小时17分	1.63

方法	Mean	Std
Sub-TB	**0小时04分**	1.10
Sub-TB-B	0小时12分钟	0.57
RL	**0小时04分钟**	0.12
Sub-EB	**0小时04分钟**	0.06
Sub-EB-B	0小时13分钟	0.71

表 2：各方法在 SIX6 和 QM9 数据集上的总运行时间均值和标准差（std）。其中，“h”表示小时，“m”表示分钟，所有标准差值均以分钟为单位。

方法	平均时间	标准差（分钟）
Sub-TB	0小时19分钟	2.05
Sub-TB-B	0小时27分钟	0.42
RL	**0小时18分钟**	0.10
Sub-EB	**0小时18分钟**	0.15
Sub-EB-B	0小时30分钟	1.09

方法	平均时间	标准差（分钟）
Sub-TB	**3小时30分钟**	10.23
Sub-TB-B	3小时46分钟	11.59
RL	3小时42分钟	13.78
Sub-EB	3小时41分钟	13.97
Sub-EB-B	3小时58分钟	7.57

表3：各方法在 PHO4 和 sEH 数据集上的总运行时间均值和标准差（std）。其中，“h”表示小时，“m”表示分钟，所有标准差值均以分钟为单位。

方法	Mean	Std
Sub-TB	0小时57分钟	1.11
Sub-TB-B	1小时23分钟	4.27
RL	0小时56分钟	0.46
Sub-EB	**0小时49分钟**	0.87
Sub-EB-B	1小时07分钟	1.43
Q-Much	0小时59分钟	3.57

方法	Mean	Std
Sub-TB	**0小时10分钟**	0.80
Sub-TB-B	0小时36分钟	1.82
RL	0小时11分钟	3.58
Sub-EB	**0小时10分钟**	0.91
Sub-EB-B	0小时31分钟	2.27
Q-Much	0小时12分钟	1.34

方法	Mean	Std
Sub-TB	0小时16分钟	1.28
Sub-TB-B	0小时42分钟	3.82
RL	**0小时12分钟**	1.79
Sub-EB	0小时14分钟	0.28
Sub-EB-B	0小时39分钟	1.13
Q-Much	0小时19分钟	0.62

表4：各方法在5节点、10节点和15节点贝叶斯网络任务中的总运行时间均值和标准差（std）。其中，“h”表示小时，“m”表示分钟，所有标准差值均以分钟为单位。由于在训练过程中需要显式计算 D_{TV} 以进行性能比较，5节点案例的运行时间最长。

方法	平均时间	标准差（分钟）
RL	**0小时 9分钟**	0.36
RLEval	0小时 13分钟	0.17
Sub-TB	0小时 13分钟	0.93
Q-Much	0小时19分钟	3.50

方法	平均时间	标准差（分钟）
RL	0小时12分钟	1.33
RLEval	**0小时9分钟**	0.08
Sub-TB	**0小时9分钟**	0.29
Q-Much	0小时10分钟	0.26

表5：各方法在 LogP 和 JNK3 任务上的总运行时间均值与标准差（std）。其中，“h”表示小时，“m”表示分钟，所有标准差均以分钟为单位报告。