

Stein 사후분포 수송을 이용한 Flow 기반 오프라인 강화학습

장준영^{*}, 강승완, 김기웅
한국과학기술원

{junjang99, seungwanwin, kekim}@kaist.ac.kr

Stein Posterior Transport for Flow-Based Offline Reinforcement Learning

요약

최근 Flow 기반 오프라인 강화학습 방법들은 생성된 행동에 대해 Q -gradient를 직접 역전파하는 샘플 단위 구조를 주로 사용한다. 그러나 이러한 방식은 데이터 support 보존과 value improvement가 별도의 loss term으로 경쟁하게 만들며, 학습된 정책이 단일 모드로 치우칠 위험이 있다. 본 논문은 정책 개선을 behavior prior 위의 KL-regularized posterior inference로 재정의하고, 그 해인 Boltzmann posterior를 Stein Variational Gradient Descent (SVGD)로 근사하는 Stein Posterior-guided Offline Transport (SPOT)을 제안한다. SPOT은 prior particle를 posterior 방향으로 이동시킨 뒤, one-step student로 distill한다. 이를 통해 support 제약을 base measure에 내장하고 kernel repulsion으로 다중 모드성을 보존한다. OGBench manipulation 환경에서 SPOT은 FQL을 비롯한 기존 방법 대비 일관된 성능 향상을 달성하였다.

1. 서론

Flow 기반 생성 모델은 다중 모드 분포에 대한 강한 표현력을 바탕으로 오프라인 강화학습에도 확장되고 있다. 그러나 off-policy 알고리즘으로 flow 정책을 학습할 때 multi-step ODE 적분 과정 전반에 걸쳐 Q 의 gradient를 역전파하는 것은 계산량과 gradient 안정성 측면에서 부담이 크다. 이에 사전 학습된 behavior cloning (BC) flow 정책을 teacher로 두고 one-step student 정책을 distillation하는 접근이 제안되었다 [1].

오프라인 강화학습의 critic은 데이터 support 밖에서 과대추정되기 쉬워, Q -gradient ascent가 정책을 out-of-distribution (OOD) 방향으로 끌어당기는 경향이 있다. 기존에는 이를 BC 정규화로 억제하지만, 두 항의 결합은 다음 두 구조적 한계를 남긴다. (1) **Support 보존과 개선의 경쟁**: BC 정규화와 Q -gradient ascent가 loss 단위로 더해지므로 그 균형은 하이퍼파라미터에 의존하고, 수렴 지점이 명시적 target distribution으로 해석되지 않는다. (2) **Mode collapse**: 정책을 Q -gradient ascent로 학습하면 noise별로 받는 gradient가 모두 동일한 high- Q 방향을 가리키며, particle 간 분리하는 메커니즘이 없으므로 단일 모드로 수렴하기 쉽다.

본 논문은 이러한 한계를 해결하기 위해 정책 업데이트를 KL-regularized posterior inference 문제로 재정의한다. 이때 해는 behavior prior를 base measure로 하는 Boltzmann posterior로 주어지므로 target이 명시적 분포로 존재하고, support 제약이 별도 loss가 아닌 target 자체에 내장된다. Posterior는 정규화 상수가 intractable하나, [2]에서 제시한 SVGD는 score만으로 KL-descent direction을 제공하므로 근사가 가능하다. **SPOT**은 (i) conditional flow matching [3]으로 prior를 학습하고, (ii) prior를 posterior 방향의 Stein field로 업데이트한 뒤, (iii) student로 distill한다. ¹

2. 연구 배경

2.1 Flow-based Policy

연속 행동 공간 MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, $\mathcal{A} \subseteq \mathbb{R}^d$, 오프라인 데이터셋 $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$ 와 critic $Q_\psi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ 를 가정한다. Flow 기반 정책은 noise $q_0 = \mathcal{N}(0, I_d)$ 와 결정론적 generator $g : \mathcal{S} \times \mathbb{R}^d \rightarrow \mathcal{A}$ 를 결합하여, $\xi \sim q_0$ 에 대해 $a = g(s, \xi)$ 로 상태 조건부 행동 분포를 유도한다. g 는 conditional flow matching [3]으로 학습한다: $a_t = (1-t)\xi + ta$ 에 대해

$$\mathcal{L}_{\text{FM}}(\phi) = \mathbb{E}_{(s,a) \sim \mathcal{D}, \xi \sim q_0, t \sim \text{Unif}[0,1]} [\|v_\phi(s, a_t, t) - (a - \xi)\|_2^2], \quad (1)$$

를 최소화한 뒤 v_ϕ 를 $t=0 \rightarrow 1$ 적분하여 generator $g_\phi(s, \xi)$ 를 얻는다. 본 논문에서는 g_ϕ 가 유도하는 분포를 behavior prior $\mu_\phi(\cdot | s)$ 로 표기한다. 즉 $\xi \sim q_0$ 일 때 $g_\phi(s, \xi) \approx \mu_\phi(\cdot | s)$ 이다.

2.2 Flow Q-Learning (FQL)

FQL [1]은 사전학습된 g_ϕ 로부터 one-step student 정책 f_θ 를 다음 actor loss로 학습한다.

$$\mathcal{L}_{\text{FQL}}(\theta) = -\mathbb{E}_{s \sim \mathcal{D}, \xi \sim q_0} [Q_\psi(s, f_\theta(s, \xi))] + \alpha \mathbb{E}_{s \sim \mathcal{D}, \xi \sim q_0} [\|f_\theta(s, \xi) - g_\phi(s, \xi)\|_2^2]. \quad (2)$$

이 objective는 샘플 단위이며, f_θ 가 유도하는 정책 분포가 어떤 분포로 수렴해야 하는지 직접 규정하지 않는다. 두 term은 서로 다른 방향으로 작용하므로 α 에 민감하며, Q loss로 직접 끌어당겨지는 one-step 매핑은 구조적으로 단일 모드에 수렴하기 쉽다.

¹ 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대 학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임 (IITP-2026-RS-2024-00436857)

3. 본론

3.1 Policy Improvement as Posterior Inference

고정된 상태 s 에 대해 target 정책을 KL-regularized 정책 개선 문제의 해로 정의한다.

$$\pi^*(\cdot | s) = \arg \max_{\pi(\cdot | s)} \left\{ \mathbb{E}_{a \sim \pi(\cdot | s)} [Q_\psi(s, a)] - \frac{1}{\beta} \text{KL}(\pi(\cdot | s) \parallel \mu_\phi(\cdot | s)) \right\} \quad (3)$$

여기서 $\beta > 0$ 은 inverse temperature로, $\beta \rightarrow 0$ 에서 해는 prior로, $\beta \rightarrow \infty$ 에서 Q_ψ -greedy 정책으로 수렴한다. 변분법에 의해서 식 (3)의 해는 *Boltzmann posterior* 형태를 갖는다.

$$\pi^*(a | s) = \frac{1}{Z(s)} \mu_\phi(a | s) \exp(\beta Q_\psi(s, a)), \quad (4)$$

여기서 $Z(s) = \int \mu_\phi(a | s) e^{\beta Q_\psi(s, a)} da$ 이다. 이 재정의는 target을 명시적 분포로 규정하며, μ_ϕ 를 base measure로 포함시켜 support 제약을 target 분포에 반영한다. 다만 $Z(s)$ 가 일반적으로 intractable하므로 π^* 로부터 직접 샘플링하기 어렵다.

3.2 Stein Variational Transport

식 (4)의 정규화 상수 $Z(s)$ 는 일반적으로 intractable하지만, a 에 무관하므로 score $\nabla_a \log \pi^*$ 에는 사라진다.

$$\nabla_a \log \pi^*(a | s) = \nabla_a \log \mu_\phi(a | s) + \beta \nabla_a Q_\psi(s, a). \quad (5)$$

SVGD [2]는 이 score만으로 particle 집합을 target 방향으로 이동시키는 변분 방법이며, 이동 방향은 다음과 같이 닫힌 형태로 주어진다.

$$\phi^*(\cdot) = \mathbb{E}_{a' \sim q} [k(a', \cdot) \nabla_{a'} \log \pi^*(a' | s) + \nabla_{a'} k(a', \cdot)]. \quad (6)$$

첫 항은 각 particle을 π^* 의 고밀도 영역으로 끌어당기는 attraction, 둘째 항은 particle 간 거리를 벌리는 repulsion으로 작용한다. 이 두 항이 동시에 작동하므로 SVGD는 multimodal target에 구조적으로 적합하며, 정규화 상수 없이 score만으로 transport가 가능하다.

3.3 Target Score 근사

Prior score $\nabla_a \log \mu_\phi(a | s)$ 를 정확히 계산하려면 g_ϕ 의 ODE 역적분과 Jacobian log-determinant 평가가 필요하여 강화학습 inner loop에서 사용하기에 비현실적이다. 따라서 본 논문은 Q -gradient만 사용한 근사 score를 사용한다.

$$\nabla_a \log \pi^*(a | s) \approx \beta \nabla_a Q_\psi(s, a). \quad (7)$$

이 근사의 영향은 다음 두 요소로 부분적으로 완화된다. (a) particle이 prior에서 샘플링되어 초기 위치가 support 내부에 있고, single-step에 작은 η 로 transport되므로 support 내부 근방에 머문다. (b) kernel repulsion 항은 score 근사와 무관하게 작동하여 particle 다양성을 유지한다. Prior score를 명시적으로 포함하는 확장은 향후 연구로 남긴다.

3.4 SPOT 학습 알고리즘

SPOT은 다음 세 단계로 구성된다.

Stage 1: Behavior prior. Conditional flow matching(식 (1))으로 generator g_ϕ 를 사전학습한 후, 이후 학습 전반에 걸쳐 g_ϕ 를 고정한다.

Stage 2: Stein posterior transport. 배치 내 각 상태 s 에 대해 K 개의 prior particle을 샘플링하고

$$\xi_i \sim q_0, \quad a_i^{(0)} = g_\phi(s, \xi_i), \quad i = 1, \dots, K, \quad (8)$$

근사 score (7)를 식 (6)에 대입하여 Stein field를 구성한다.

$$\hat{\phi}_s(a_i^{(0)}) = \frac{1}{K} \sum_{j=1}^K \left[\beta k(a_j^{(0)}, a_i^{(0)}) \nabla_{a_j} Q_\psi(s, a_j^{(0)}) + \nabla_{a_j} k(a_j^{(0)}, a_i^{(0)}) \right]. \quad (9)$$

Kernel은 RBF $k(a, a') = \exp(-\|a - a'\|^2 / 2h)$ 를 사용한다. Particle을 transport하여 target을 얻는다.

$$a_i^{(1)} = a_i^{(0)} + \eta \hat{\phi}_s(a_i^{(0)}). \quad (10)$$

Stage 3: Noise-matched distillation. Teacher target $\{a_i^{(1)}\}$ 을 배포 가능한 형태로 만들기 위해 one-step student $f_\theta(s, \xi)$ 를 다음 loss로 학습한다.

$$\mathcal{L}_{\text{distill}}(\theta) = \mathbb{E}_{s \sim \mathcal{D}, \xi_{1:K} \sim q_0} \left[\frac{1}{K} \sum_{i=1}^K \|f_\theta(s, \xi_i) - \text{sg}(a_i^{(1)})\|_2^2 \right]. \quad (11)$$

핵심은 student가 받는 noise ξ_i 가 prior particle $a_i^{(0)}$ 을 생성한 동일 noise라는 점이다. 이렇게 noise별 pointwise regression으로 학습되므로, student는 결과적으로 prior에서 π^* 방향으로 transport된 분포로부터 샘플을 생성하는 one-step 정책으로 수렴한다. 또한 prior에서 서로 다른 모드에 대응하던 noise들은 transport 후에도 서로 다른 모드로 매핑되므로 multimodality가 자연스럽게 유지된다.

Algorithm 1 SPOT: Stein Posterior Offline Transport**Require:** dataset $\mathcal{D}$, β , particle 수 K , step η

```

1:  $g_\phi \leftarrow$  flow matching on  $\mathcal{D}$ ;  $\phi$  고정
2: student  $f_\theta$ , critic  $Q_\psi$  초기화
3: for training step = 1, 2, ... do
4:   batch  $\{s^{(b)}\} \sim \mathcal{D}$ 
5:   for each  $s$  in batch do ▷ 배치 내 병렬
6:      $\xi_{1:K} \sim q_0$ ,  $a_{1:K}^{(0)} \leftarrow g_\phi(s, \xi_{1:K})$ 
7:      $\hat{\phi}_s(a_i^{(0)})$  계산 (식 (9))
8:      $a_i^{(1)} \leftarrow a_i^{(0)} + \eta \hat{\phi}_s(a_i^{(0)})$ 
9:   end for
10:   $\mathcal{L}_{\text{distill}}$  (식 (11))로  $\theta$  업데이트
11:  Standard critic loss로  $\psi$  업데이트
12: end for

```

전체 학습 절차. Algorithm 1이 전체 절차를 요약한다. Critic Q_ψ 는 표준 오프라인 강화학습 방식(TD 업데이트)으로 student 정책과 공동 학습된다.

4. 실험 및 결과

본 연구에서는 제안한 SPOT의 효과를 OGBench의 reward-based single-task variants에서 검증하였다. 각 환경은 서로 다른 목표를 갖는 다섯 개의 task (task1–task5)로 구성되며, 모든 실험은 offline 환경에서 수행하였다. 학습은 총 1,000,000 gradient step 동안 수행하였고, 매 100,000 step마다 50 episode로 평가하였다. 최종 성능은 3개 seed에 대해 각 seed의 마지막 세 evaluation epoch에서 측정한 성공률의 평균으로 계산하였으며, 모든 방법은 동일한 task 구성과 성공률 metric을 기준으로 비교하였다. 실험에서는 기존 알고리즘들과의 상대적인 성능 비교를 위해, 가우시안 분포 기반의 BC 모델을 비롯하여 Flow Policy 구조를 차용한 IFQL과 FQL을 비교군으로 설정하여 결과를 분석하였다.

표 1는 세 manipulation 환경에서의 평균 성공률을 나타낸다. cube-single에서는 SPOT이 96±2%의 성공률을 보여 FQL의 94±2%와 유사하거나 더 높은 성능을 달성하였다. 해당 환경은 대부분의 task에서 성공률이 높은 상한에 가깝지만, SPOT은 일부 어려운 task에서 추가적인 개선을 보였다. cube-double에서는 SPOT이 28±3%로 FQL의 17±1%를 크게 상회하였으며, 이는 다중 object 조작 환경에서 더 효과적인 정책 개선이 가능함을 시사한다. scene 환경에서도 SPOT은 62±5%로 FQL의 55±2%보다 높은 평균 성능을 달성하였다. 이러한 결과는 Stein transport 기반 posterior inference가 기존 one-step actor improvement보다 복잡한 조작 환경에서 더 안정적인 성능 향상을 제공할 수 있음을 보여준다.

표 1: OGBench 환경 내 singletask에 대한 평가 성공률 (%).

Task Category	Task	BC	IFQL	FQL	SPOT
cube-single	task1	10 ± 5	79 ± 4	96 ± 2	96 ± 1
	task2	3 ± 1	73 ± 3	97 ± 1	96 ± 1
	task3	9 ± 3	88 ± 4	97 ± 2	99 ± 0
	task4	2 ± 1	79 ± 6	93 ± 4	96 ± 2
	task5	3 ± 3	77 ± 7	86 ± 2	92 ± 5
cube-double	task1	8 ± 3	35 ± 9	61 ± 2	66 ± 7
	task2	0 ± 0	9 ± 5	15 ± 1	25 ± 1
	task3	0 ± 0	8 ± 5	7 ± 4	16 ± 2
	task4	0 ± 0	1 ± 1	2 ± 0	3 ± 1
	task5	0 ± 0	17 ± 6	1 ± 0	30 ± 7
scene	task1	19 ± 6	98 ± 3	100 ± 0	100 ± 0
	task2	1 ± 1	0 ± 0	68 ± 4	92 ± 2
	task3	1 ± 1	54 ± 19	95 ± 1	93 ± 4
	task4	2 ± 2	0 ± 0	12 ± 5	28 ± 22
	task5	0 ± 0	0 ± 0	0 ± 0	0 ± 0

5. 결론

본 논문은 flow 기반 오프라인 강화학습의 정책 개선에서 support 보존과 value improvement가 별도 loss로 경쟁하고 one-step generator가 단일 모드로 붕괴되는 한계를, KL-regularized posterior inference로 재정의하여 해결하였다. 그 해인 Boltzmann posterior를 SVGD로 근사하고 noise-matched distillation으로 one-step student에 전달한다. 이를 통해 SPOT은 support 제약을 base measure에 내장하고 multimodality를 kernel repulsion으로 구조적으로 보존한다.

OGBench 세 환경에서 SPOT은 FQL 대비 일관된 성능 향상을 보였다. cube-single에서는 상한에 가까운 성능을 유지하면서, cube-double에서는 평균 성공률을 17%에서 28%로, scene에서는 55%에서 62%로 개선하여, support 제약과 multimodality의 영향이 큰 task일수록 posterior 기반 transport의 이점이 두드러짐을 확인하였다. 향후 연구로는 prior score를 효율적으로 추정하여 SVGD에 결합하는 확장, multi-step transport의 영향 분석, 그리고 타 도메인으로의 일반화 검증을 계획한다.

참고 문헌

- [1] S. Park, Q. Li, and S. Levine, “Flow q-learning,” in *Forty-second International Conference on Machine Learning*, 2025.
- [2] Q. Liu and D. Wang, “Stein variational gradient descent: A general purpose bayesian inference algorithm,” in *Advances in Neural Information Processing Systems*, 2016.
- [3] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” 2023.