A Coupled Flow Approach to Imitation Learning

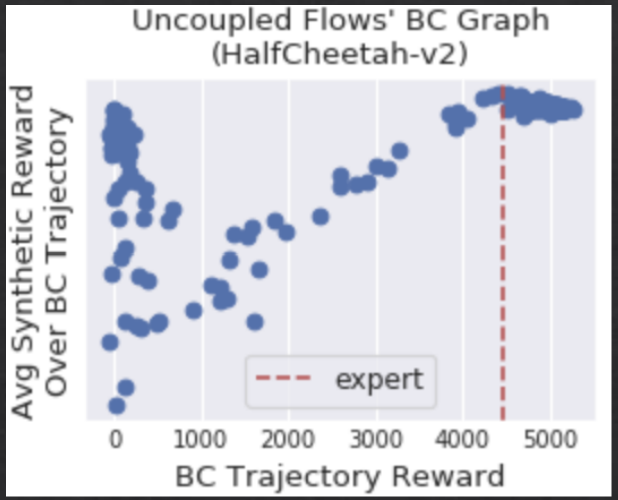
Gideon Freund, Elad Sarafian, Sarit Kraus

Bar-Ilan University

Motivation & Background

- In RL & IL, the agent's policy induces a state distribution $d_\pi(s)$ and state-action distribution $p_\pi(s,a) = \pi(a|s) \cdot d_\pi(s)$
- They are of central importance, appearing all across the literature:
  - The Policy Gradient Theorem: A fundamental theorem from which all policy-based methods are derived. [Sutton et al., 2000]
  - All distribution matching approaches in imitation learning [Ke et al. 2020]
  - Other: Curiosity based exploration [Pathak et al. 2017]; Constrained RL [Qin et al. 2021]; Batch "offline" RL [Fujimoto et al. 2019]; Convex RL [Mutti et al. 2022]
- Despite their importance, $d_\pi(s)$ and $p_\pi(s,a)$ are mostly discussed indirectly and theoretically, rather than being modeled explicitly.
  - This work concentrates on modeling them explicitly with normalizing flows, focusing on imitation learning.
- Imitation learning
  - Simple approach: Behavioral cloning (BC)
  - Distribution matching: $\min D_f(p_\pi || p_{exp})$
    - Hinges on the one-to-one relationship between a and p_π. Has shown significant improvement over BC, particularly when few expert trajectories are available or expert trajectories are subsampled.

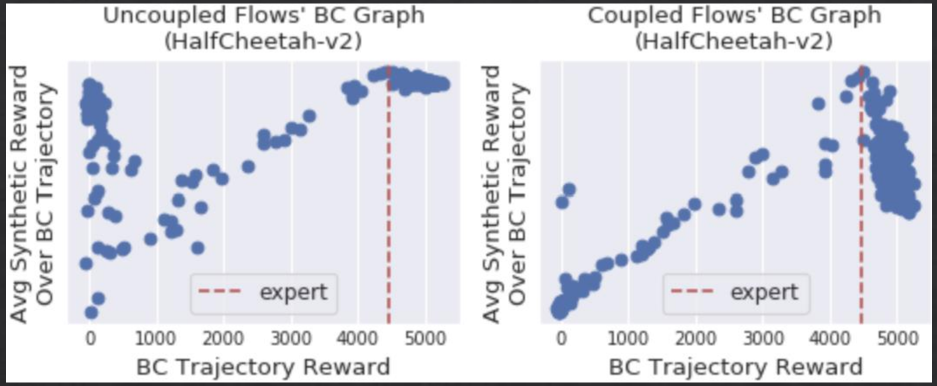
Our Approach



- Reverse KL: $argmin_{\pi} D_{KL}(p_{\pi}||p_e)$
- This IL objective is: $argmax_{\pi} J\left(\pi, r = \log\frac{p_e}{p_{\pi}}\right)$
- Thus, given an estimate of the ratio, any RL algorithm can be used for solving the IL objective.
- Naïve approach: Train two flows independently
  - Practically, this means alternating between learning them and using their logratio as reward
  - Fails: Overall normalized score of 0.158
  - BC graph
    - Intuitively, no upward trend implies failure in RL
    - Formalized in the paper
  - Problem of OOD: Flows values are meaningless when evaluated on each others data
- They must be coupled!

Our Approach

- To couple them, we employ the Donsker-Varadhan form of the KL:
  - $D_{KL}(p_\pi || p_{exp}) = \sup_{x:S\times A\rightarrow R} E_{p_\pi(s,a)}[x(s,a)] - \log E_{p_e(s,a)}[e^{x(s,a)}]$
  - Optimality point: $x^* = \log \frac{p_\pi}{p_e} + C$
    - Precisely the negative log distribution ratio!
    - Can compute x through the DV and use –x as reward in an RL algorithm
    - Set $x_{\psi,\phi}(s,a) = log p_\psi(s,a) - log q_\phi(s,a)$
  - Guarantees more meaningful values when the flows are evaluated on each others data
    - Note the drop at the end occurs precisely beyond expert level
- Additional components:
  - Squasher
  - Flow regularization
  - Smoothing
- Coupled Flow Imitation Learning (CFIL)



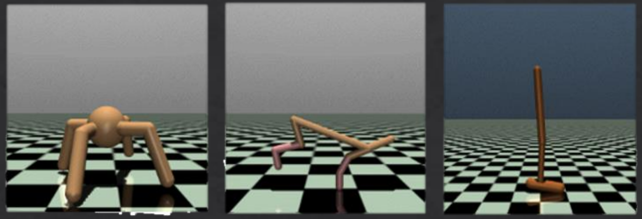
Algorithm 1 CFIL

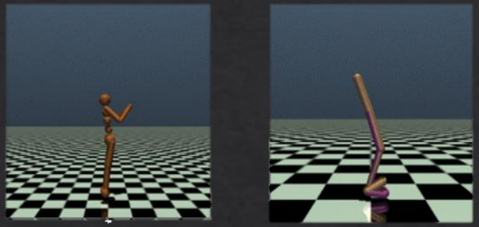
Input: Expert demos $\mathcal{R}_E = \{(s_e, a_e)\}_{t=1}^N$; parameterized flow pair p_ψ, q_ϕ; off-policy RL algorithm $\mathcal{A}$; density update rate k; squashing function σ; regularization and smoothing coefficients α, β.

Define: $x_{\psi,\phi} = \sigma(\log p_\psi - \log q_\phi)$

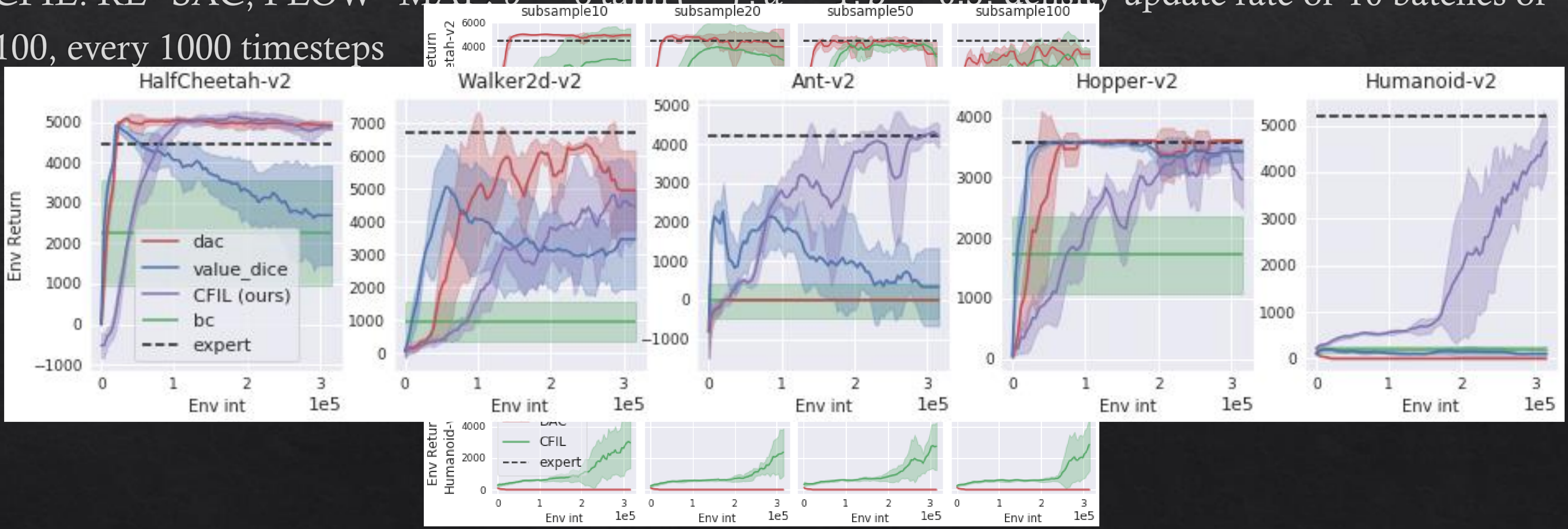
for timestep $t = 0, 1, \ldots,$ **do**
  Take a step in $\mathcal{A}$ with reward $r = -x_{\psi,\phi}$, while filling agent buffer $\mathcal{R}_A$ and potentially updating the policy and value networks according to $\mathcal{A}$'s settings.
  if $t \mod k = 0$ **then**
    Sample expert and agent batches:
    $\{(s_e^t, a_e^t)\}_{t=1}^M \sim \mathcal{R}_E$ and $\{(s^t, a^t)\}_{t=1}^M \sim \mathcal{R}_A$
    if smooth **then**
      $(s,a) \mathrel{+}= \beta \cdot (s,a) \odot u, \;\; u \sim U(-\frac{1}{2}, \frac{1}{2})^{dim((s,a))}$
    end if
    Compute loss:
    $\mathcal{J} = \log \frac{1}{M} \sum_{i=1}^M e^{x(s_e^i, a_e^i)} - \frac{1}{M} \sum_{i=1}^M x(s^i, a^i)$
    if flow reg **then**
      Compute regularization loss:
      $\mathcal{L} = -\frac{1}{M} \sum_{i=1}^M \log q_\phi(s_e^i, a_e^i) + \log p_\psi(s^i, a^i)$
      $\mathcal{J} = \mathcal{J} + \alpha\mathcal{L}$
    end if
    Update $\psi \leftarrow \psi - \eta \nabla_\psi \mathcal{J}$
    Update $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{J}$
  end if
end for

Experiments





1. Standard Mujoco benchmarks, comparing with SOTA on a single expert trajectory.
2. Plot means and standard deviations across 5 random seeds.
3. CFIL: RL=SAC; FLOW=MAF; $\sigma = 6\tanh\left(\frac{x}{\;}\right)$; $\alpha = 1, \beta = 0.5$; density update rate of 10 batches of 100, every 1000 timesteps



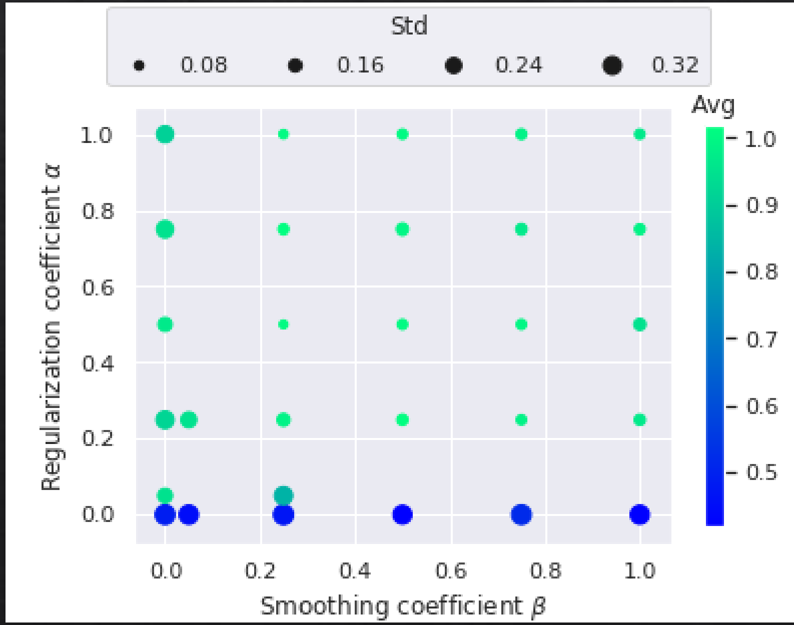
"We now turn to the state-only and subsampled regimes. Settings in which ValueDICE finds no dice:"

Ablation

- Left: We put into question the need for our squasher, our coupling and our inductive bias
- Right: We vary CFIL's smoothing and regularization coefficients to test its sensitivity

Each point and value summarizes 25 seeds (5 per environment).

	SCORE
EXPERT	1
CFIL	**1.012**
NOSQUASH	−0.091
REGULARNET	0.196 | 0.190
INDFLOW	0.158 | 0.127
INDFLOWNS	0.090 | 0.072
NUMERATOR	−0.051 | −0.001



All the alternatives fail,
demonstrating the necessity of CFIL's components

Shows both the utility of the smoothing and regularization as well as CFIL's robustness to them

Conclusion

- Presented CFIL: A unique approach to imitation learning.
  - Outperforms SOTA in a variety of settings.
  - Many novelties: estimator; smoothing & regularization; employment of flows; BC graphs.
  - Future work could include coupled flows for general ratio estimation.
- Overall, we pointed out the great potential—then demonstrated the utility—of explicitly modeling the state and state-action distributions and aim to inspire more research incorporating such models all across the reinforcement learning literature.

Mahalo