Appendix: Policy Optimization with Advantage Regularization for Long-Term Fairness in Decision Systems

Eric Yang Yu UC San Diego
Zhizhen Qin UC San Diego
Min Kyung Lee UT Austin
Sicun Gao UC San Diego

A An Introduction to Lyapunov Stability

We first define an n-dimensional dynamical system Φ by:

$$\dot{x}(t) = f(x(t), u(t)), u(t) = g(x(t))$$

where $x(t) \in D$ is a state vector at time $t \in \mathbb{R}$ in the state space $D \subseteq \mathbb{R}^n$, $g : D \to \mathbb{R}^m$ is a control function, and $f : D \to \mathbb{R}^n$ is a Lipschitz-continuous vector field.

Suppose system Φ has an equilibrium point at x_e. The system is stable at x_e if for all $\epsilon \in \mathbb{R}^+$, there exists some $\delta(\epsilon) \in \mathbb{R}^+$ such that $||x(t) - x_e|| < \epsilon$ for all $t \geq 0$ if $||x(0) - x_e|| < \delta$. In other words, if an initial point $x(0)$ starts at some distance δ from equilibrium point x_e, any point along all possible solution trajectories from $x(0)$ to x_e should be smaller than some distance ϵ from x_e. To take this stability notion one step further, we say system Φ is locally asymptotically stable around x_e if $\lim_{t\to\infty} x(t) = 0$ for all $||x(0) - x_e|| < \delta$.

Next, we explore the definition of the Lyapunov function and Lie derivative. Using the existing system setup, let $V : D \to \mathbb{R}$ be a continuously differentiable function. V is a Lyapunov function if $V(0) = 0$ and $L_f V(x(t)) < 0$ and $\forall x \in D \setminus \{0\}, V(x(t)) > 0$. The Lie derivative of V over f is defined:

$$L_f V(x(t)) = \sum_{i=1}^{n} \frac{\partial V}{\partial t} = \sum_{i=1}^{n} \frac{\partial V}{\partial x_i} \frac{dx_i}{dt} = \sum_{i=1}^{n} \frac{\partial V}{\partial x_i} \dot{x}(t)$$

At a high level, the Lyapunov function V defines a field of attraction around some equilibrium point, and the Lie derivative $L_f V(x(t))$ measures the rate of convergence of V over time along the system dynamics of $x(t)$ to its equilibrium point x_e. If this Lyapunov function can be defined, system f is asymptotically stable at x_e.

One issue with formulating the Lie derivative in the context of Reinforcement Learning (RL) training is that computing it requires full access to system dynamics f, which the RL policy in training does not have access to. Thus, we must approximate the Lie derivative along sampled trajectories of the dynamical system during training:

$$L_{f,\Delta t} V(x(t)) = \frac{1}{\Delta t}(V(x(t+1)) - V(x(t)))$$

where

$$\lim_{\Delta t \to \infty} L_{f,\Delta t} V(x(t)) = L_f V(x(t))$$

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

B Hyperparameters for Case Studies

We used the following hyperparameters in the training procedures. Table 1, Table 2 and Table 3 show the hyperparameters used for PPO variations on attention allocation, lending, and infectious disease control environments, respectively. Note that for the lending and precision disease control environments, our hyperparameters are chosen with respect to a min-max normalization applied to each advantage term.

PPO Agent	ζ_0	ζ_1	ζ_2	β_0	β_1	β_2
Greedy (G-PPO)	1	0.25	0	0	0	0
Reward-Only Fairness Constrained (R-PPO)	1	0.25	10	0	0	0
Advantage Regularized (A-PPO)	1	0.25	0	0.05	0.32	0.63

Table 1: The hyperparameters used for each PPO variation during training on the base and harder attention allocation environments.

PPO Agent	ζ_0	ζ_1	β_0	β_1	β_2
Greedy (G-PPO)	1	0	0	0	0
Reward-Only Fairness Constrained (R-PPO)	1	2	0	0	0
Advantage Regularized (A-PPO)	1	0	1	0.5	0.5

Table 2: The hyperparameters used for each PPO variation during training on the lending environment.

PPO Agent	ζ_0	ζ_1	β_0	β_1	β_2
No Fairness Constraints (N PPO)	1	0	0	0	0
Only Reward Fairness Constraint (R PPO)	1	0.1	0	0	0
Only Advantage Fairness Constraint (A PPO)	1	0	0.6	0.15	0.25

Table 3: The hyperparameters used for each PPO variation during training on the precision disease control environment.

C Social network for Infectious Disease Case Study

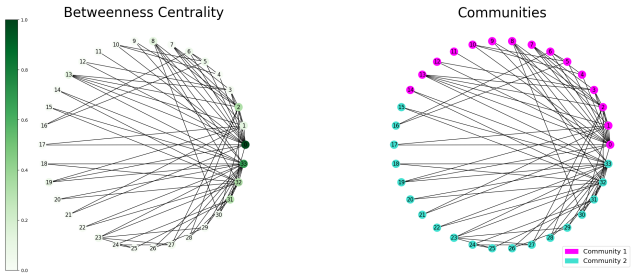

Figure 1: [Left] Node betweenness centrality visualized for the Karate Club graph in the precision disease control environment, where the intensity of the node positively correlates with its betweenness centrality value. [Right] Communities for the Karate Club graph in the precision disease control environment. These communities are obtained by applying the Girvan-Newman community detection algorithm once on the graph.

We obtain our notion of a community in the Karate Club graph using the Girvan-Newman community detection algorithm. We define edge betweenness as the number of shortest paths between two nodes that travel through an edge. In this algorithm, each edge is computed for its edge betweenness. Then, the edge with the highest betweenness is removed to reveal two communities seen on the right in Figure 1. Node betweenness centrality is defined as the combined fraction of all shortest paths between pairs that pass through a node, and can be visualized on the left in Figure 1. Although

betweenness centrality is distinct from edge betweenness and is not a part of the Girvan-Newman algorithm, we include it to provide more insight into the underlying structure of the Karate Club Graph.