Learning Fair Policies in Multiobjective (Deep) Reinforcement Learning with Average and Discounted Rewards

Umer Siddique, Paul Weng, and Matthieu Zimmer

University of Michigan-Shanghai Jiao Tong University Joint Institute

ICML 2020

Overview

1. Motivation and Problem
2. Theoretical Discussions & Algorithms
3. Experimental Results
4. Conclusion

Motivation: Why should we care about fair systems?

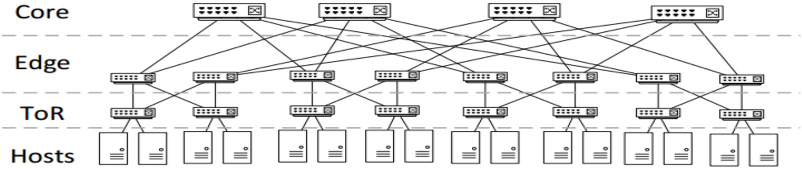


Figure: Network with a fat-tree topology from Ruffy et al. (2019).

Motivation: Why should we care about fair systems?

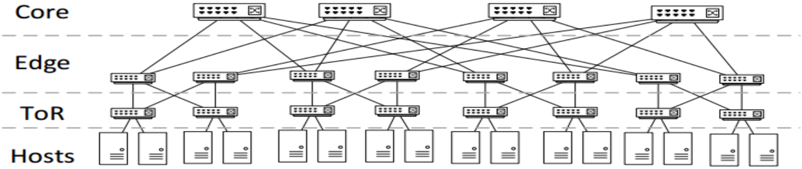


Figure: Network with a fat-tree topology from Ruffy et al. (2019).

- Fairness consideration to users is crucial

Motivation: Why should we care about fair systems?

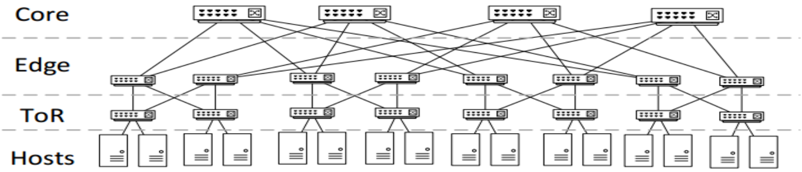


Figure: Network with a fat-tree topology from Ruffy et al. (2019).

- Fairness consideration to users is crucial
- Existing approaches to tackle this issue includes:
  - Utilitarian approach
  - Egalitarian approach

Fairness

- Fairness includes:
  - Efficiency
  - Impartiality
  - Equity

Fairness

- Fairness includes:
  - Efficiency
  - Impartiality
  - Equity
- Fairness encoded in a Social Welfare Function (SWF)

Fairness

- Fairness includes:
  - Efficiency
  - Impartiality
  - Equity
- Fairness encoded in a Social Welfare Function (SWF)
- We focus on *generalized Gini social welfare function* (GGF)

Problem Statement

- GGF can be defined as:

$$\mathrm{GGF}_{\boldsymbol{w}}(\boldsymbol{v}) = \sum_{i=1}^{D} \boldsymbol{w}_i \boldsymbol{v}_i^{\uparrow}$$

Problem Statement

- GGF can be defined as:

$$\mathrm{GGF}_{\boldsymbol{w}}(\boldsymbol{v}) = \sum_{i=1}^{D} \boldsymbol{w}_i \boldsymbol{v}_i^{\uparrow} = \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \dots & \boldsymbol{w}_D \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^{\uparrow} \\ \boldsymbol{v}_2^{\uparrow} \\ \dots \\ \boldsymbol{v}_D^{\uparrow} \end{bmatrix}$$

Problem Statement

- GGF can be defined as:

$$\mathrm{GGF}_{\boldsymbol{w}}(\boldsymbol{v}) = \sum_{i=1}^{D} \boldsymbol{w}_i \boldsymbol{v}_i^{\uparrow} = [\boldsymbol{w}_1 > \boldsymbol{w}_2 > \ldots > \boldsymbol{w}_D] \begin{bmatrix} \boldsymbol{v}_1^{\uparrow} \\ \leq \\ \boldsymbol{v}_2^{\uparrow} \\ \leq \\ \ldots \\ \leq \\ \boldsymbol{v}_D^{\uparrow} \end{bmatrix}$$

Problem Statement

- GGF can be defined as:

$$\text{GGF}_{\boldsymbol{w}}(\boldsymbol{v}) = \sum_{i=1}^{D} \boldsymbol{w}_i \boldsymbol{v}_i^{\uparrow} = [\boldsymbol{w}_1 > \boldsymbol{w}_2 > \ldots > \boldsymbol{w}_D] \begin{bmatrix} \boldsymbol{v}_1^{\uparrow} \\ \leq \\ \boldsymbol{v}_2^{\uparrow} \\ \leq \\ \ldots \\ \leq \\ \boldsymbol{v}_D^{\uparrow} \end{bmatrix}$$

- Fair optimization problem in RL:

$$\arg\max_{\pi} \text{GGF}_{\boldsymbol{w}}(\boldsymbol{J}(\pi)) \quad (1)$$

Problem Statement

- GGF can be defined as:

$$\text{GGF}_{\boldsymbol{w}}(\boldsymbol{v}) = \sum_{i=1}^{D} \boldsymbol{w}_i \boldsymbol{v}_i^{\uparrow} = [\boldsymbol{w}_1 > \boldsymbol{w}_2 > \ldots > \boldsymbol{w}_D] \begin{bmatrix} \boldsymbol{v}_1^{\uparrow} \\ \leq \\ \boldsymbol{v}_2^{\uparrow} \\ \leq \\ \ldots \\ \leq \\ \boldsymbol{v}_D^{\uparrow} \end{bmatrix}$$

- Fair optimization problem in RL:

$$\arg\max_{\pi} \text{GGF}_{\boldsymbol{w}}(\boldsymbol{J}(\pi)) \tag{1}$$

$$\text{where } \boldsymbol{J}(\pi) = \mathbb{E}_{\boldsymbol{P}_\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \boldsymbol{R}_t\right] \quad \text{or} \quad \boldsymbol{J}(\pi) = \lim_{h\to\infty} \frac{1}{h} \mathbb{E}_{\boldsymbol{P}_\pi}\left[\sum_{t=1}^{h} \boldsymbol{R}_t\right].$$

γ-discounted rewards — average rewards

Theoretical Discussion

Assumption: MDPs are weakly-communicating

Theoretical Discussion

Assumption: MDPs are weakly-communicating

- Sufficiency of Stationary Markov Policies
  - Existence of stationary Markov fair optimal policy.

Theoretical Discussion

Assumption: MDPs are weakly-communicating

- Sufficiency of Stationary Markov Policies
  - Existence of stationary Markov fair optimal policy.
- Possibly State-Dependent Optimality
  - With average reward, fair optimality stays state-independent.

Theoretical Discussion

Assumption: MDPs are weakly-communicating

- Sufficiency of Stationary Markov Policies
  - Existence of stationary Markov fair optimal policy.
- Possibly State-Dependent Optimality
  - With average reward, fair optimality stays state-independent.

Contribution on Approximation Error

- Approximate *average-optimal* policy (π_1^*) with γ-*optimal* policy (π_γ^*).

Theoretical Discussion

Assumption: MDPs are weakly-communicating

- Sufficiency of Stationary Markov Policies
  - Existence of stationary Markov fair optimal policy.
- Possibly State-Dependent Optimality
  - With average reward, fair optimality stays state-independent.

Contribution on Approximation Error

- Approximate *average-optimal* policy (π_1^*) with γ-*optimal* policy (π_γ^*).

Theorem:

$$\mathrm{GGF}_{\boldsymbol{w}}(\boldsymbol{\mu}(\pi_\gamma^*)) \geq \mathrm{GGF}_{\boldsymbol{w}}(\boldsymbol{\mu}(\pi_1^*)) - \overline{\boldsymbol{R}}(1-\gamma)\Big(\rho(\gamma, \sigma(\boldsymbol{H}_{\boldsymbol{P}_{\pi_1^*}})) + \rho(\gamma, \sigma(\boldsymbol{H}_{\boldsymbol{P}_{\pi_\gamma^*}}))\Big)$$

where $\overline{\boldsymbol{R}} = \max_\pi \|\boldsymbol{R}_\pi\|_1$ and $\rho(\gamma, \sigma) = \frac{\sigma}{\gamma - (1-\gamma)\sigma}$.

- DQN: Q network takes values in $\mathbb{R}^{|\mathcal{A}| \times D}$, instead of $\mathbb{R}^{|\mathcal{A}|}$, trained with target:

$$\hat{\boldsymbol{Q}}_\theta(s, a) = \boldsymbol{r} + \gamma \hat{\boldsymbol{Q}}_{\theta'}(s', a^*),$$

where $a^* = \text{argmax}_{a' \in \mathcal{A}} \; \text{GGF}_{\boldsymbol{w}}\big(\boldsymbol{r} + \gamma \hat{\boldsymbol{Q}}_{\theta'}(s', a')\big)$.

Value Based and Policy Gradient Algorithms

- DQN: Q network takes values in $\mathbb{R}^{|\mathcal{A}|\times D}$, instead of $\mathbb{R}^{|\mathcal{A}|}$, trained with target:

$$\hat{\boldsymbol{Q}}_\theta(s,a) = \boldsymbol{r} + \gamma \hat{\boldsymbol{Q}}_{\theta'}(s', a^*),$$

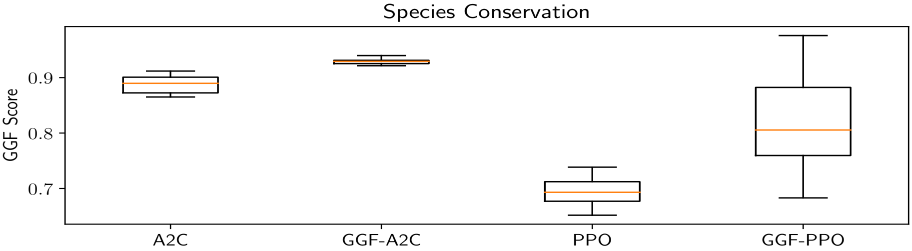
where $a^* = \operatorname{argmax}_{a' \in \mathcal{A}}\ \mathrm{GGF}_{\boldsymbol{w}}\big(\boldsymbol{r} + \gamma \hat{\boldsymbol{Q}}_{\theta'}(s', a')\big)$.

- To optimize the GGF with policy gradient:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \mathrm{GGF}_{\boldsymbol{w}}(\boldsymbol{J}(\pi_{\boldsymbol{\theta}})) &= \nabla_{\boldsymbol{J}(\pi_{\boldsymbol{\theta}})} \mathrm{GGF}_{\boldsymbol{w}}(\boldsymbol{J}(\pi_{\boldsymbol{\theta}})) \cdot \nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\pi_{\boldsymbol{\theta}}) \\ &= \boldsymbol{w}_\sigma^\intercal \cdot \nabla_{\boldsymbol{\theta}} \boldsymbol{J}(\pi_{\boldsymbol{\theta}}). \end{aligned}$$

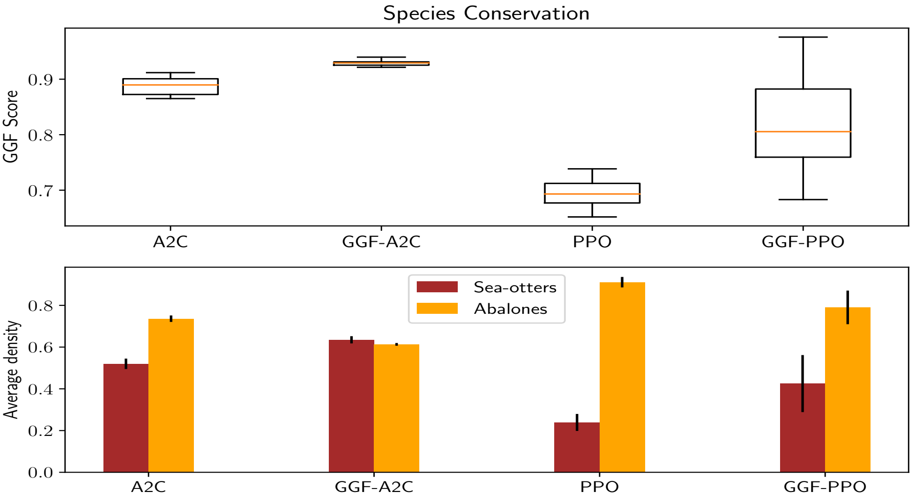
Experimental Results

What is the impact of optimizing GGF instead of the average of the objectives?



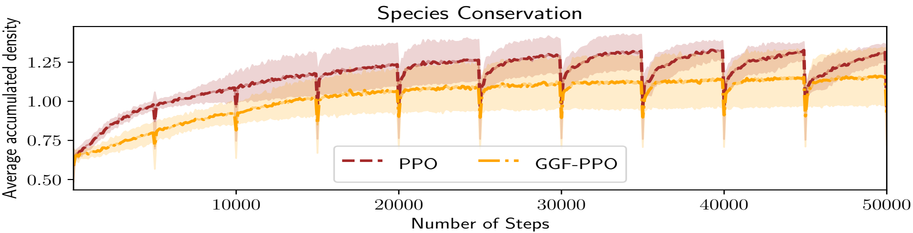
Experimental Results

What is the impact of optimizing GGF instead of the average of the objectives?



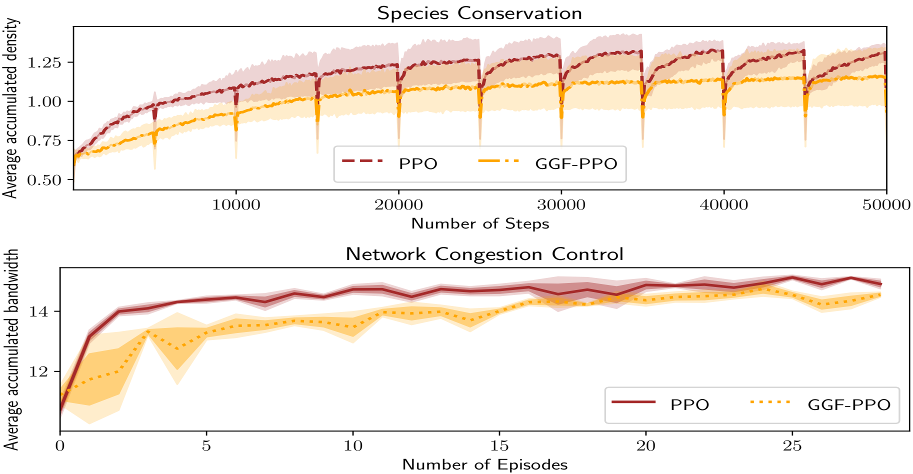
Experimental Results

What is the price of fairness?
How those algorithms performs in continuous domains?



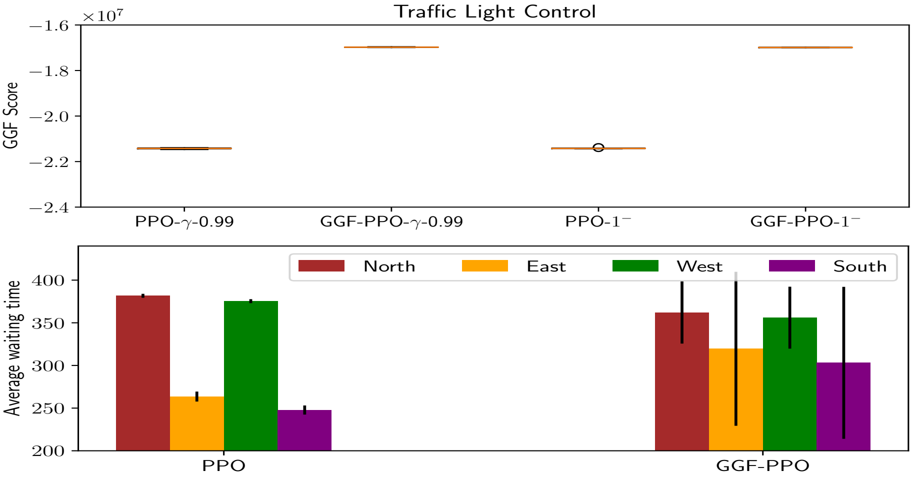
Experimental Results

What is the price of fairness?
How those algorithms performs in continuous domains?



Experimental Results (Traffic Light Control)

What is the effect of γ with respect to GGF-average optimality?



Conclusion

- Fair optimization in RL setting
- Theoretical discussion with a new bound
- Adaptations of DQN, A2C and PPO to solve this problem.
- Experimental validation in 3 domains

Conclusion

- Fair optimization in RL setting
- Theoretical discussion with a new bound
- Adaptations of DQN, A2C and PPO to solve this problem.
- Experimental validation in 3 domains

Future Works:

- Extend to distributed control
- Consider other fair social welfare functions
- Directly solve average reward problems

Ruffy, F., Przystupa, M., and Beschastnikh, I. (2019). Iroko: A framework to prototype reinforcement learning for data center traffic control. In *Workshop on ML for Systems at NeurIPS*.