Enabling Long-term Fairness in Dynamic Resource Allocation

ACM SIGMETRICS 2023

Tareq Si Salem[1], George Iosifidis[2], and Giovanni Neglia[3]

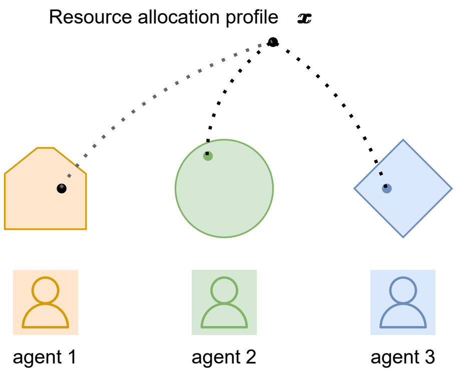
[1]Northeastern University, [2]Delft University of Technology, and [3]Inria Sophia Antipolis

June 21, 2023

Fairness in Resource Allocation — Offline Setting

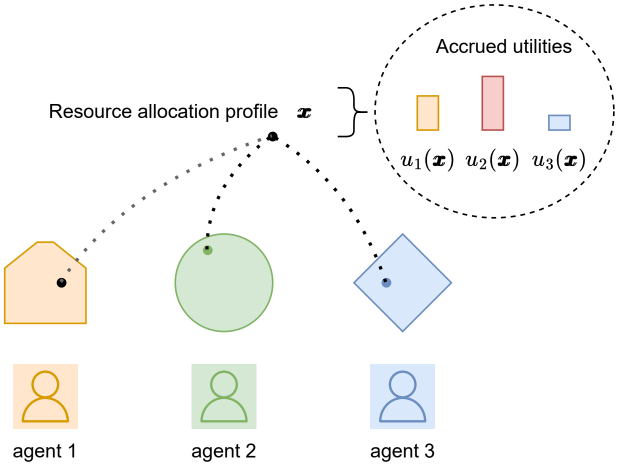
- Consider a system comprised of a set of agents $\mathcal{I}$.

Fairness in Resource Allocation — Offline Setting



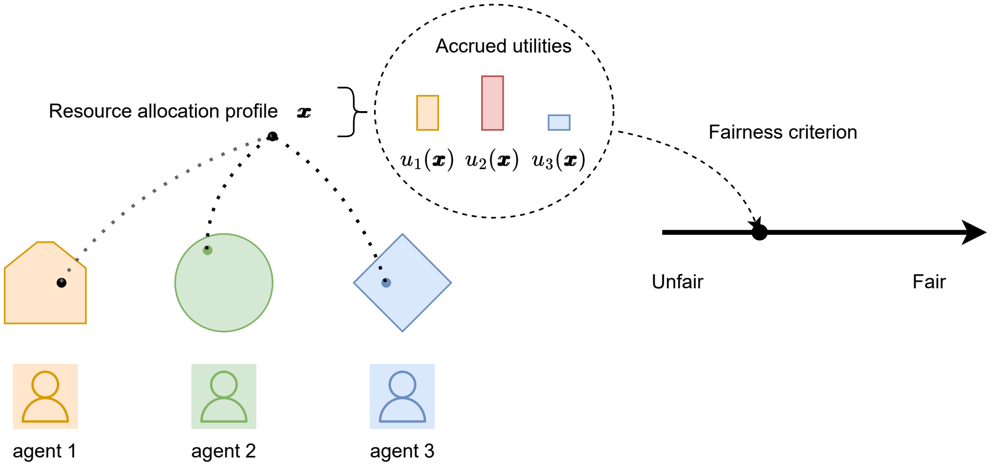
- Consider a system comprised of a set of agents $\mathcal{I}$.
- An allocation profile is selected from a convex set $\mathcal{X} = \bigtimes_{i \in \mathcal{I}} \mathcal{X}_i$.

Fairness in Resource Allocation — Offline Setting



- Consider a system comprised of a set of agents $\mathcal{I}$.
- An allocation profile is selected from a convex set $\mathcal{X} = \bigtimes_{i \in \mathcal{I}} \mathcal{X}_i$.
- The utility of each agent i is a a concave function $u_i(\boldsymbol{x})$.

Fairness in Resource Allocation — Offline Setting



- Consider a system comprised of a set of agents $\mathcal{I}$.
- An allocation profile is selected from a convex set $\mathcal{X} = \times_{i \in \mathcal{I}} \mathcal{X}_i$.
- The utility of each agent i is a a concave function $u_i(\boldsymbol{x})$.
- An α-fairness criterion is applied over the vector valued utilities $\boldsymbol{u}(\boldsymbol{x}) = (u_i(\boldsymbol{x}))_{i \in \mathcal{I}}$.

Fairness in Resource Allocation — Offline Setting

The α-fairness encompasses the utilitarian principle, proportional fairness (Nash bargaining solution), max-min fairness (Kalai–Smorodinsky bargaining).

Applications

Fairness in Resource Allocation — Offline Setting

The α-fairness encompasses the utilitarian principle, proportional fairness (Nash bargaining solution), max-min fairness (Kalai–Smorodinsky bargaining).

Applications

- Congestion control in the Internet (F.P98; MW00)

Fairness in Resource Allocation — Offline Setting

The α-fairness encompasses the utilitarian principle, proportional fairness (Nash bargaining solution), max-min fairness (Kalai–Smorodinsky bargaining).

Applications

- Congestion control in the Internet (F.P98; MW00)
- Selection transmission power in multi-user wireless networks (X. 06; Res06)

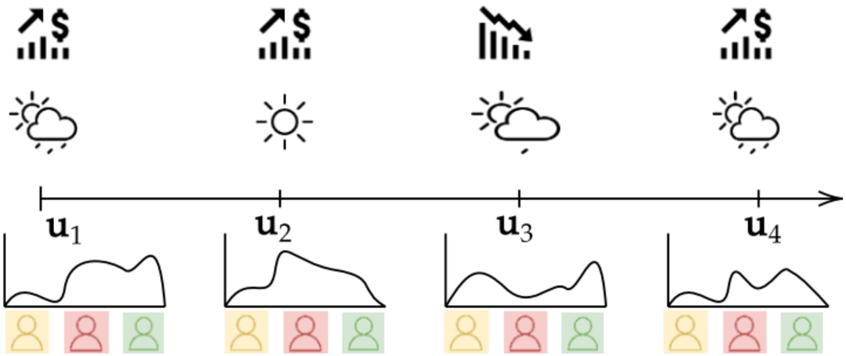
Fairness in Resource Allocation — Offline Setting

The α-fairness encompasses the utilitarian principle, proportional fairness (Nash bargaining solution), max-min fairness (Kalai–Smorodinsky bargaining).

Applications

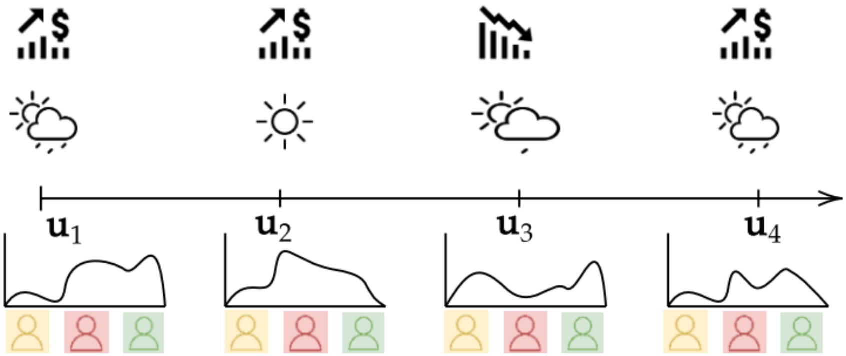
- Congestion control in the Internet (F.P98; MW00)
- Selection transmission power in multi-user wireless networks (X. 06; Res06)
- Allocate multidimensional resources in cloud computing platforms (C. 13; W. 14; T. 15)

Fairness in Resource Allocation — Offline Setting



Considering a *static* environment is very often unrealistic for today's communication and computing systems.

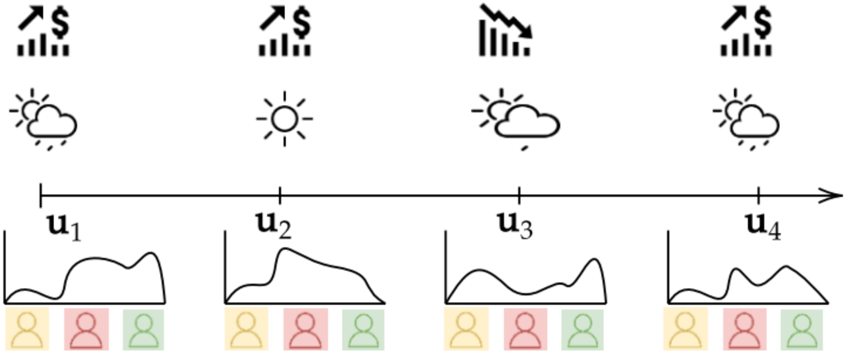
Fairness in Resource Allocation — Offline Setting



Considering a *static* environment is very often unrealistic for today's communication and computing systems.

- Small-cell mobile networks

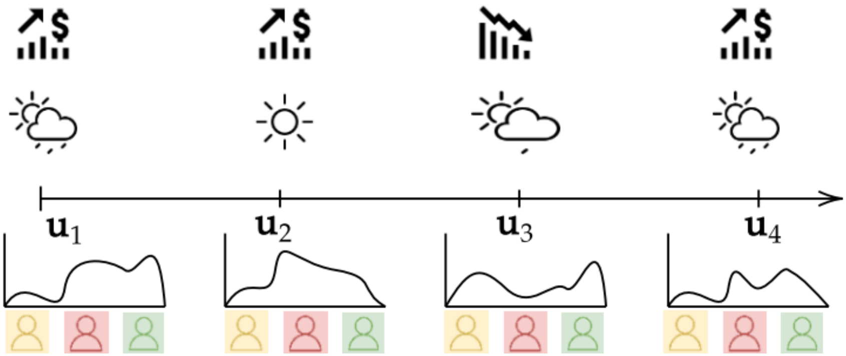
Fairness in Resource Allocation — Offline Setting



Considering a *static* environment is very often unrealistic for today's communication and computing systems.

- Small-cell mobile networks
- Placing content files at edge caches

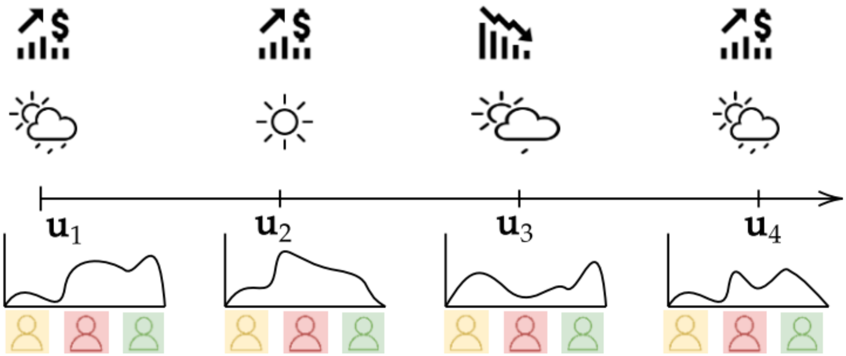
Fairness in Resource Allocation — Offline Setting



Considering a *static* environment is very often unrealistic for today's communication and computing systems.

- Small-cell mobile networks
- Placing content files at edge caches
- Increasing virtualization of communication and computing systems

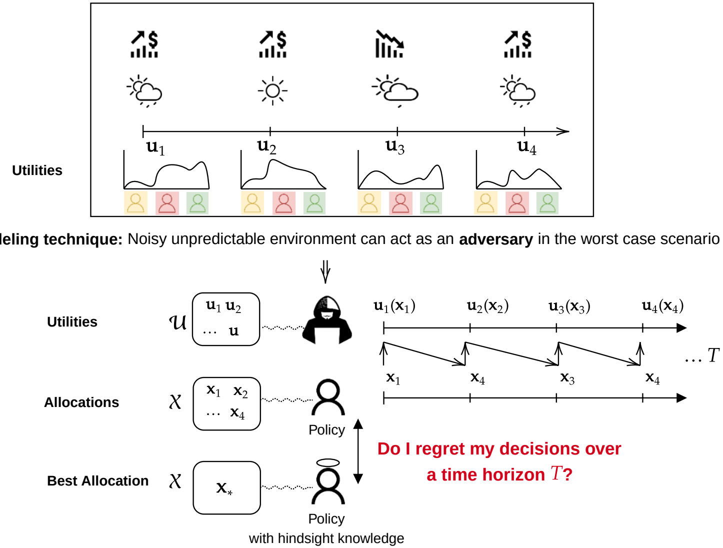
Fairness in Resource Allocation — Offline Setting



Considering a *static* environment is very often unrealistic for today's communication and computing systems.

- Small-cell mobile networks
- Placing content files at edge caches
- Increasing virtualization of communication and computing systems
- Services driven by user generated data

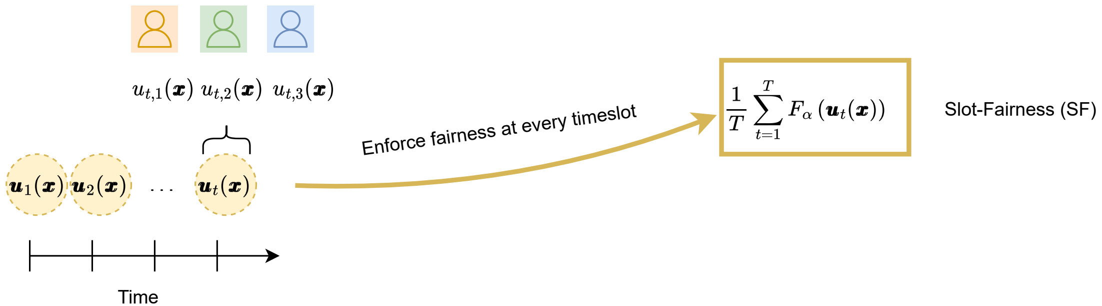
Dynamic Setting through Adversarial Analysis



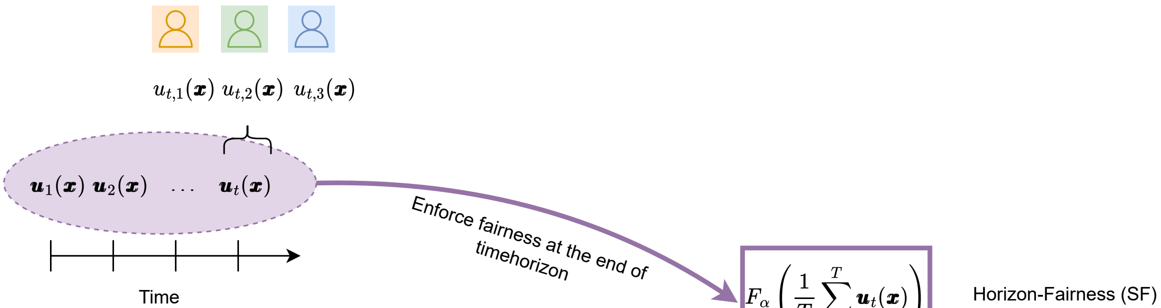
How to Measure Regret?

Describe first how the best allocation is determined.

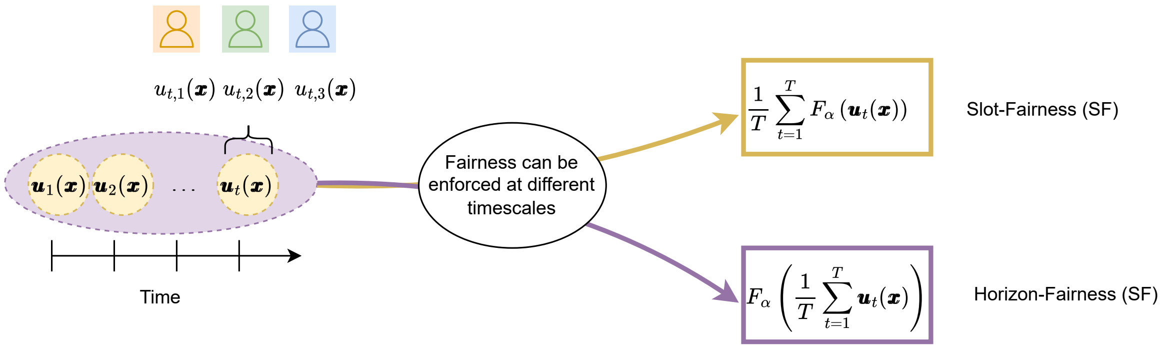
How and When to Enforce Fairness in the Dynamic Setting?



How and When to Enforce Fairness in the Dynamic Setting?



How and When to Enforce Fairness in the Dynamic Setting?



How and When to Enforce Fairness in the Dynamic Setting?

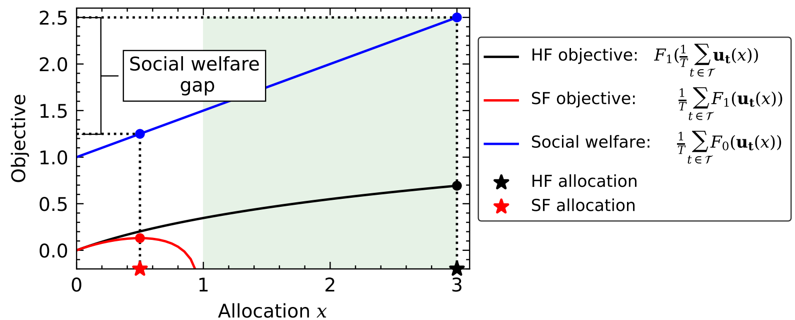
Consider a system with two agents $\mathcal{I} = \{1, 2\}$, an allocation set $\mathcal{X} = [0, x_{\max}]$ with $x_{\max} > 1$, α-fairness criterion with $\alpha = 1$, even $T \in \mathbb{N}$, and the following sequence of utilities

$$\{\boldsymbol{u}_t(x)\}_{t=1}^{T} = \{(1 + x, 1 - x), (1 + x, 1 + x), \ldots\}.$$

How and When to Enforce Fairness in the Dynamic Setting?

Consider a system with two agents $\mathcal{I} = \{1, 2\}$, an allocation set $\mathcal{X} = [0, x_{\max}]$ with $x_{\max} > 1$, α-fairness criterion with $\alpha = 1$, even $T \in \mathbb{N}$, and the following sequence of utilities

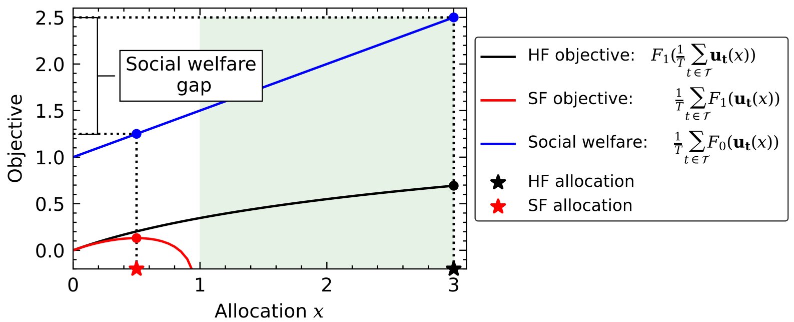
$$\{\boldsymbol{u}_t(x)\}_{t=1}^{T} = \{(1 + x, 1 - x), (1 + x, 1 + x), \dots\}.$$



How and When to Enforce Fairness in the Dynamic Setting?

Consider a system with two agents $\mathcal{I} = \{1, 2\}$, an allocation set $\mathcal{X} = [0, x_{\max}]$ with $x_{\max} > 1$, α-fairness criterion with $\alpha = 1$, even $T \in \mathbb{N}$, and the following sequence of utilities

$$\{\boldsymbol{u}_t(x)\}_{t=1}^{T} = \{(1 + x, 1 - x), (1 + x, 1 + x), \ldots\}.$$



Horizon-fairness/Slot-Fairness

Slot-Fairness is too restrictive.

Fairness Regret

We propose the *fairness regret* metric:

Definition

The long-term fairness regret of a policy $\mathcal{A}$ under α-fairness is defined as follows:

$$\mathfrak{R}_T\left(F_\alpha, \mathcal{A}\right) \triangleq \sup_{\{\boldsymbol{u}_t\}_{t=1}^T \in \mathcal{U}^T} \left\{ F_\alpha\left(\frac{1}{T}\sum_{t\in\mathcal{T}} \boldsymbol{u}_t(\boldsymbol{x}_\star)\right) - F_\alpha\left(\frac{1}{T}\sum_{t\in\mathcal{T}} \boldsymbol{u}_t(\boldsymbol{x}_t)\right) \right\}.$$

When $\lim_{T\to\infty} \mathfrak{R}_T\left(F_\alpha, \mathcal{A}\right) = 0$, policy $\mathcal{A}$ will attain the same fairness value as the static benchmark under any possible sequence of utility functions.

Is Vanishing Regret Achievable?

Impossibility Result

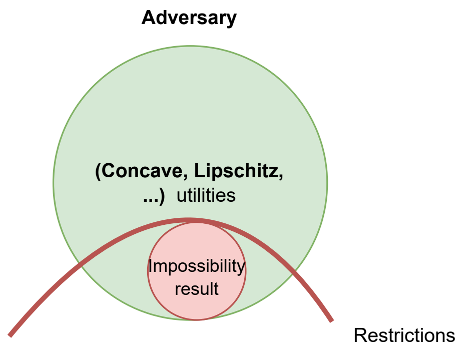
There is no online policy $\boldsymbol{\mathcal{A}}$ attaining $\mathfrak{R}_T\left(F_\alpha, \mathcal{A}\right) = o(1)$ for $|\mathcal{I}| > 1$ and $\alpha > 0$ under an unrestricted adversary (e.g., OCO's adversary).

Is Vanishing Regret Achievable?

Impossibility Result

There is no online policy $\boldsymbol{\mathcal{A}}$ attaining $\mathfrak{R}_T\left(F_\alpha, \mathcal{A}\right) = o(1)$ for $|\mathcal{I}| > 1$ and $\alpha > 0$ under an unrestricted adversary (e.g., OCO's adversary).

- (Proof sketch) We design an adversary with a choice over different sequences of utilities against two agents. No policy can have vanishing fairness regret under all sequences.



Overcoming the Impossibility Result

The quantity $\boldsymbol{\delta}_t(\boldsymbol{x})$ quantifies how much the adversary *perturbs* the average utility at time t, and $\Xi(\mathcal{T})$ is the set of all possible decompositions of $\mathcal{T}$ into sets of contiguous timeslots.

Budgeted-severity: $$\mathbb{V}_{\mathcal{T}} = \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}} \left| \delta_{t,i}(\boldsymbol{x}_\star) \right|,$$

Partitioned-severity: $$\mathbb{W}_{\mathcal{T}} = \inf_{\substack{\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K\} \\ \in \Xi(\mathcal{T})}} \left\{ \sum_{k=1}^{K} \sum_{i \in \mathcal{I}} \left| \sum_{t \in \mathcal{T}_k} \delta_{t,i}(\boldsymbol{x}_\star) \right| + \sum_{k=1}^{K} \frac{|\mathcal{T}_k|^2}{\sum_{k' < k} |\mathcal{T}_k| + 1} \right\}.$$

Overcoming the Impossibility Result

The quantity $\boldsymbol{\delta}_t(\boldsymbol{x})$ quantifies how much the adversary *perturbs* the average utility at time t, and $\Xi(\mathcal{T})$ is the set of all possible decompositions of $\mathcal{T}$ into sets of contiguous timeslots.

$$\text{Budgeted-severity:} \qquad \mathbb{V}_{\mathcal{T}} = \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{I}} \left|\delta_{t,i}(\boldsymbol{x}_\star)\right|,$$

$$\text{Partitioned-severity:} \qquad \mathbb{W}_{\mathcal{T}} = \inf_{\substack{\{\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_K\}\\ \in\,\Xi(\mathcal{T})}} \left\{ \sum_{k=1}^{K} \sum_{i\in\mathcal{I}} \left| \sum_{t\in\mathcal{T}_k} \delta_{t,i}(\boldsymbol{x}_\star) \right| + \sum_{k=1}^{K} \frac{|\mathcal{T}_k|^2}{\sum_{k'<k} |\mathcal{T}_k| + 1} \right\}.$$

- The adversary is restricted to select utilities such that: $\min\{\mathbb{V}_{\mathcal{T}}, \mathbb{W}_{\mathcal{T}}\} = o(T)$. These restrictions capture several practical utility patterns, such as non-stationary corruptions, ergodic and periodic inputs (LGK22; BLM22; ZLL[+]19; DAJJ12).

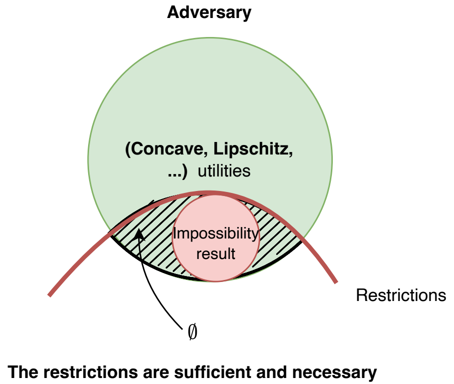
Overcoming the Impossibility Result

The quantity $\boldsymbol{\delta}_t(\boldsymbol{x})$ quantifies how much the adversary *perturbs* the average utility at time t, and $\Xi(\mathcal{T})$ is the set of all possible decompositions of $\mathcal{T}$ into sets of contiguous timeslots.

Budgeted-severity: $$\mathbb{V}_{\mathcal{T}} = \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{I}} \left|\delta_{t,i}(\boldsymbol{x}_\star)\right|,$$

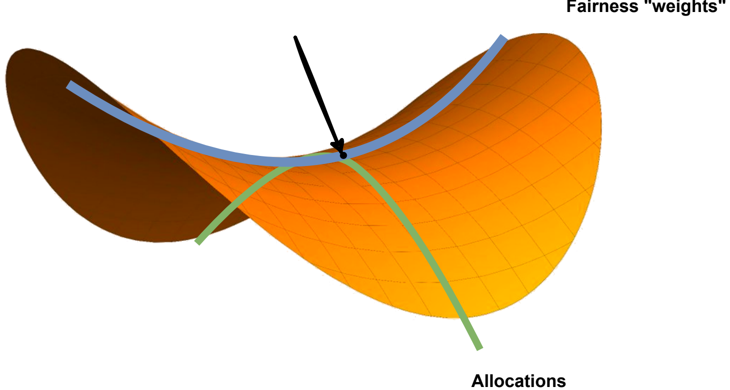
Partitioned-severity: $$\mathbb{W}_{\mathcal{T}} = \inf_{\substack{\{\mathcal{T}_1,\mathcal{T}_2,\dots,\mathcal{T}_K\}\\ \in\,\Xi(\mathcal{T})}} \left\{ \sum_{k=1}^{K}\sum_{i\in\mathcal{I}}\left|\sum_{t\in\mathcal{T}_k}\delta_{t,i}(\boldsymbol{x}_\star)\right| + \sum_{k=1}^{K}\frac{|\mathcal{T}_k|^2}{\sum_{k'<k}|\mathcal{T}_k| + 1} \right\}.$$

- The adversary is restricted to select utilities such that: $\min\{\mathbb{V}_{\mathcal{T}}, \mathbb{W}_{\mathcal{T}}\} = o(T)$. These restrictions capture several practical utility patterns, such as non-stationary corruptions, ergodic and periodic inputs (LGK22; BLM22; ZLL+19; DAJJ12).



ONLINE HORIZON FAIR (OHF) POLICY

$$\max_{\boldsymbol{x}\in\mathcal{X}} \min_{\boldsymbol{\theta}\in\Theta} \Psi(\boldsymbol{\theta}, \boldsymbol{x}) = \max_{\boldsymbol{x}\in\mathcal{X}} F_\alpha(\boldsymbol{u}(\boldsymbol{x}))$$

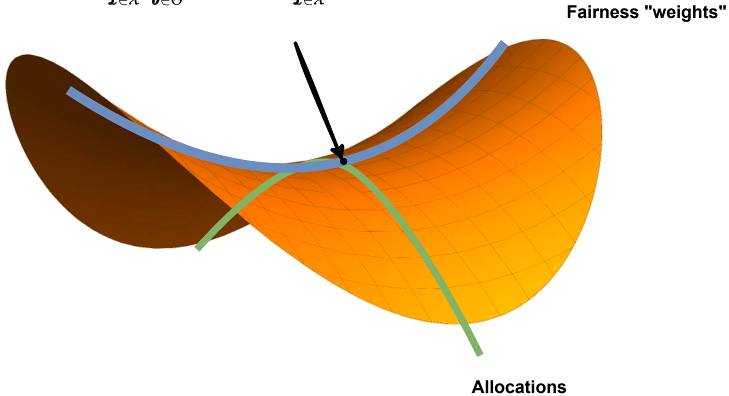


Our policy employs a *time-varying* a convex-concave function:

$$\Psi_{t,\alpha}(\boldsymbol{\theta}, \boldsymbol{x}) \triangleq \underbrace{(-F_\alpha)^\star(\boldsymbol{\theta})}_{(I)} - \underbrace{\boldsymbol{\theta} \cdot \boldsymbol{u}_t(\boldsymbol{x})}_{(II)}$$

Online Horizon Fair (OHF) Policy

$$\max_{\boldsymbol{x}\in\mathcal{X}} \min_{\boldsymbol{\theta}\in\Theta} \Psi(\boldsymbol{\theta}, \boldsymbol{x}) = \max_{\boldsymbol{x}\in\mathcal{X}} F_\alpha(\boldsymbol{u}(\boldsymbol{x}))$$

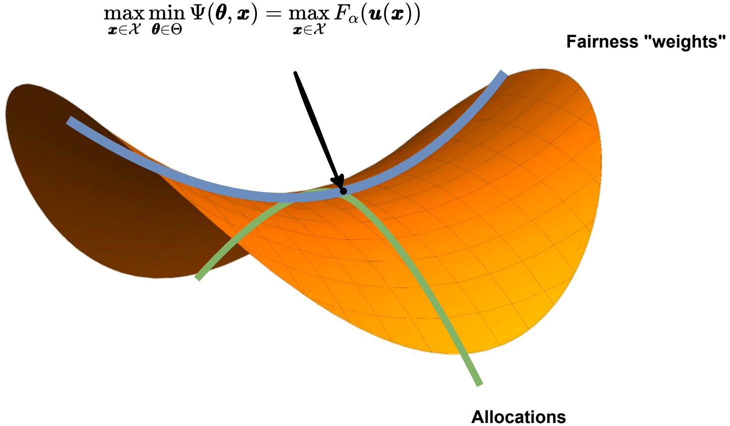


Our policy employs a *time-varying* a convex-concave function:

$$\Psi_{t,\alpha}(\boldsymbol{\theta}, \boldsymbol{x}) \triangleq \underbrace{(-F_\alpha)^\star(\boldsymbol{\theta})}_{(I)} - \underbrace{\boldsymbol{\theta} \cdot \boldsymbol{u}_t(\boldsymbol{x})}_{(II)}$$

- (I) A *convex conjugate* term that tracks the global fairness metric in a dual (conjugate) space

Online Horizon Fair (OHF) Policy

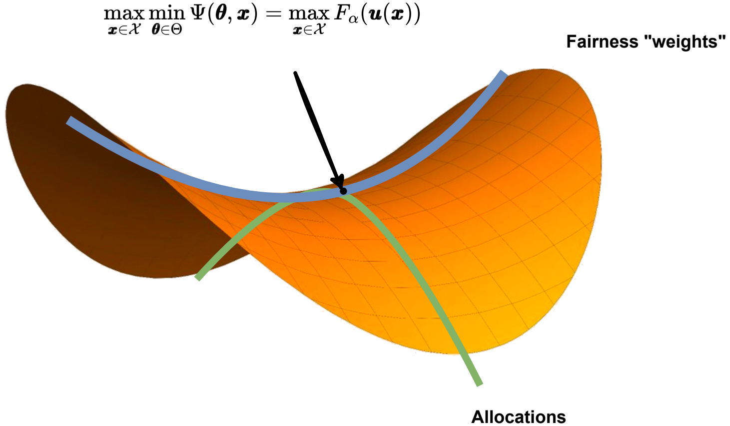


Our policy employs a *time-varying* a convex-concave function:

$$\Psi_{t,\alpha}(\boldsymbol{\theta}, \boldsymbol{x}) \triangleq \underbrace{(-F_\alpha)^\star(\boldsymbol{\theta})}_{(I)} - \underbrace{\boldsymbol{\theta}\cdot\boldsymbol{u}_t(\boldsymbol{x})}_{(II)}$$

- (I) A *convex conjugate* term that tracks the global fairness metric in a dual (conjugate) space
- (II) A weighted sum of utilities term that tracks the appropriate allocations in the primal space

Online Horizon Fair (OHF) Policy



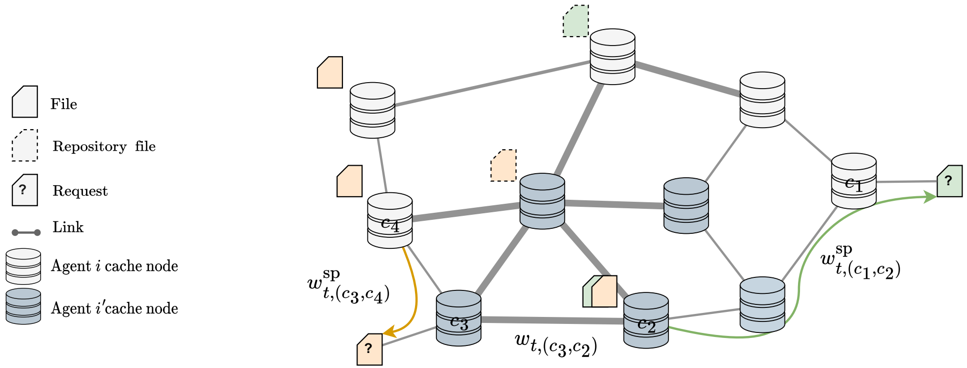
Our policy employs a *time-varying* a convex-concave function:

$$\Psi_{t,\alpha}(\boldsymbol{\theta}, \boldsymbol{x}) \triangleq \underbrace{(-F_\alpha)^\star(\boldsymbol{\theta})}_{(I)} - \underbrace{\boldsymbol{\theta} \cdot \boldsymbol{u}_t(\boldsymbol{x})}_{(II)}$$

- (I) A *convex conjugate* term that tracks the global fairness metric in a dual (conjugate) space
- (II) A weighted sum of utilities term that tracks the appropriate allocations in the primal space
- OHF policy attains the following fairness regret guarantee:

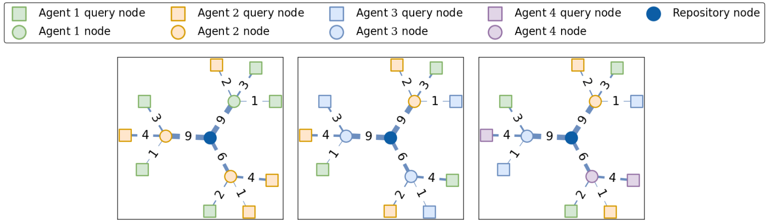
$$\mathfrak{R}_T(F_\alpha, \mathcal{A}) = \mathcal{O}\left(\frac{1}{\sqrt{T}} + \frac{\min\{\mathbb{V}_\mathcal{T}, \mathbb{W}_\mathcal{T}\}}{T}\right) = o(1).$$

An Application — Virtualized Caching Systems

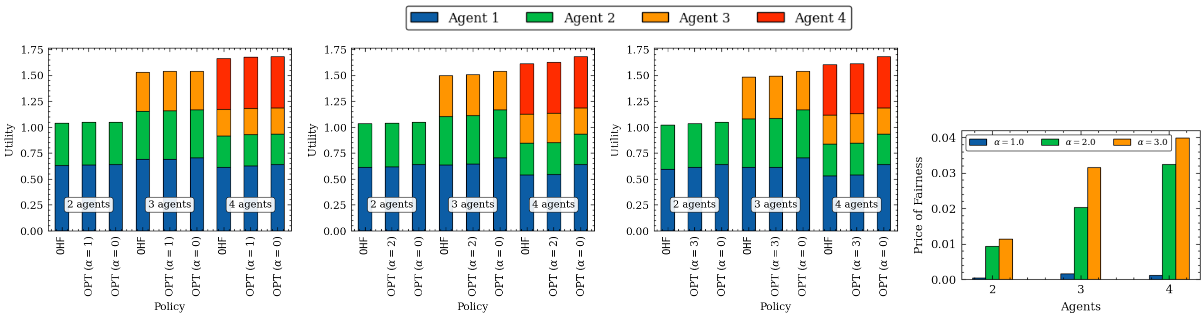


An application: a network comprised of a set of caching nodes $\mathcal{C}$. A request arrives at a cache node $c \in \mathcal{C}$, it can be partially served locally, and if needed, forwarded along the shortest retrieval path to another node to retrieve the remaining part of the file.

An Application — Virtualized Caching Systems

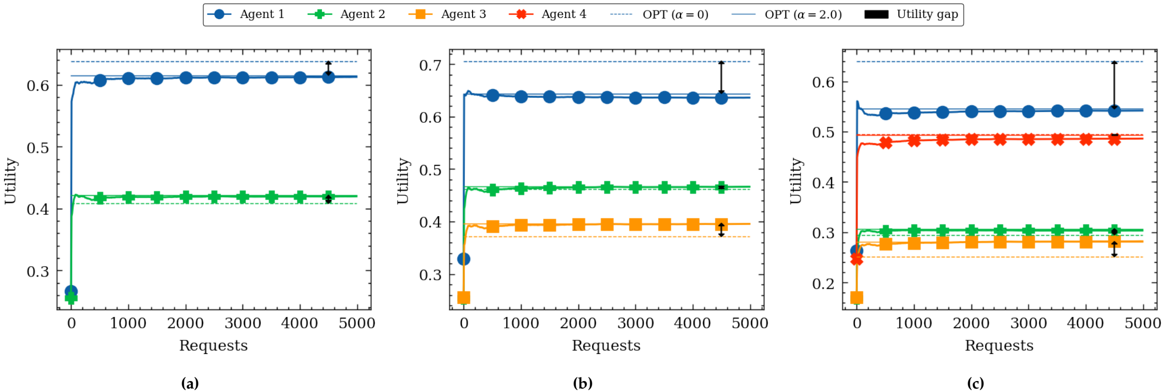


(a) Tree-1 **(b)** Tree-2 **(c)** Tree-3



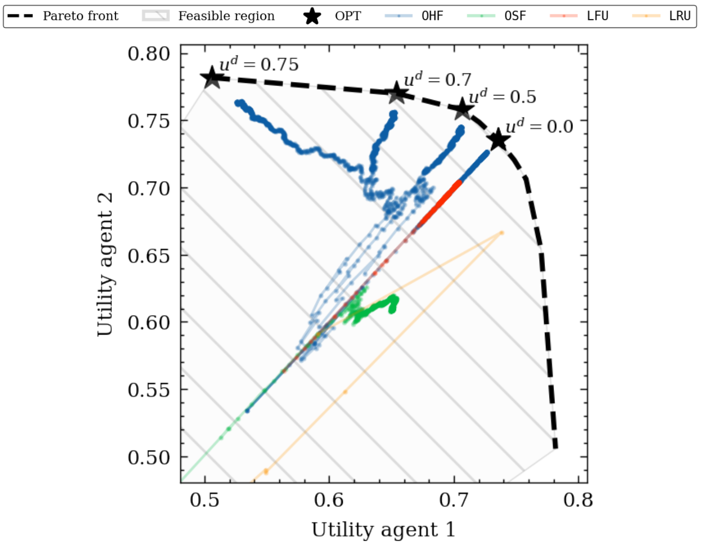
(a) $\alpha = 1$ **(b)** $\alpha = 2$ **(c)** $\alpha = 3$ **(d)** Price of Fairness

An Application — Virtualized Caching Systems



The time-averaged utility across different agents obtained by OHF policy and OPT for $\alpha = 2$ under an increasing number of agents in $\{2, 3, 4\}$ and Tree 1–3 network topology.

An Application — Virtualized Caching Systems



Future Work

- Bridge the horizon-fairness and slot-fairness criteria to target applications where the agents are interested in ensuring fairness within a target time window.
- Add support for coalition formation in our fairness framework.

Thank You

t.sisalem@northeastern.edu

References I

[BLM22] Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni, *The Best of Many Worlds: Dual Mirror Descent for Online Allocation Problems*, Operations Research (2022).

[C. 13] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang, *Multiresource Allocation: Fairness-Efficiency Tradeoffs in a Unifying Framework*, IEEE/ACM Trans. on Networking **21** (2013), no. 6, 1785–1798.

[DAJJ12] John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan, *Ergodic Mirror Descent*, SIAM Journal on Optimization **22** (2012), no. 4, 1549–1578.

[F.P98] F.P. Kelly, A. Maulloo, and D. Tan, *Rate Control for Communication Networks: Shadow Prices, Proportional Fairness, and Stability*, J. Oper. Res. Soc. **49** (1998), no. 3, 237–252.

[LGK22] Luofeng Liao, Yuan Gao, and Christian Kroer, *Nonstationary Dual Averaging and Online Fair Allocation*, ArXiv e-prints (2022).

[MW00] Jeonghoon Mo and Jean Walrand, *Fair End-to-End Window-Based Congestion Control*, IEEE/ACM Transactions on networking **8** (2000), no. 5, 556–567.

[Res06] Resource Allocation and Cross Layer Control in Wireless Networks, *L. Georgiadis, M. J. Neely, and L. Tassiulas*, Found. Trends Netw. **1** (2006), no. 1, 1–143.

[T. 15] T. Bonald, and J. W. Roberts, *Multi-Resource Fairness: Objectives, Algorithms and Performance*, ACM Sigmetrics, 2015.

[W. 14] W. Wang, B. Li, and B. Liang, *Dominant Resource Fairness in Cloud Computing Systems with Heterogeneous Servers*, IEEE INFOCOM, 2014.

[X. 06] X. Lin, N. B. Shroff, and R. Srikant, *A Tutorial on Cross-Layer Optimization in Wireless Networks*, IEEE J. Sel. Areas Commun. **24** (2006), no. 8, 1452–1463.

[ZLL+19] Yu-Hang Zhou, Chen Liang, Nan Li, Cheng Yang, Shenghuo Zhu, and Rong Jin, *Robust online matching with user arrival distribution drift*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 459–466.