# On convergence and stability of learning dynamics in continuous games

Manxi Wu
UC Berkeley, EECS
Simons Institute for Theory of Computing

C3.ai DTI Colloquium
April 2022

# Today's outline

On convergence and stability of learning dynamics in continuous games, joint with Saurabh Amin (MIT), and Asuman Ozdaglar (MIT)

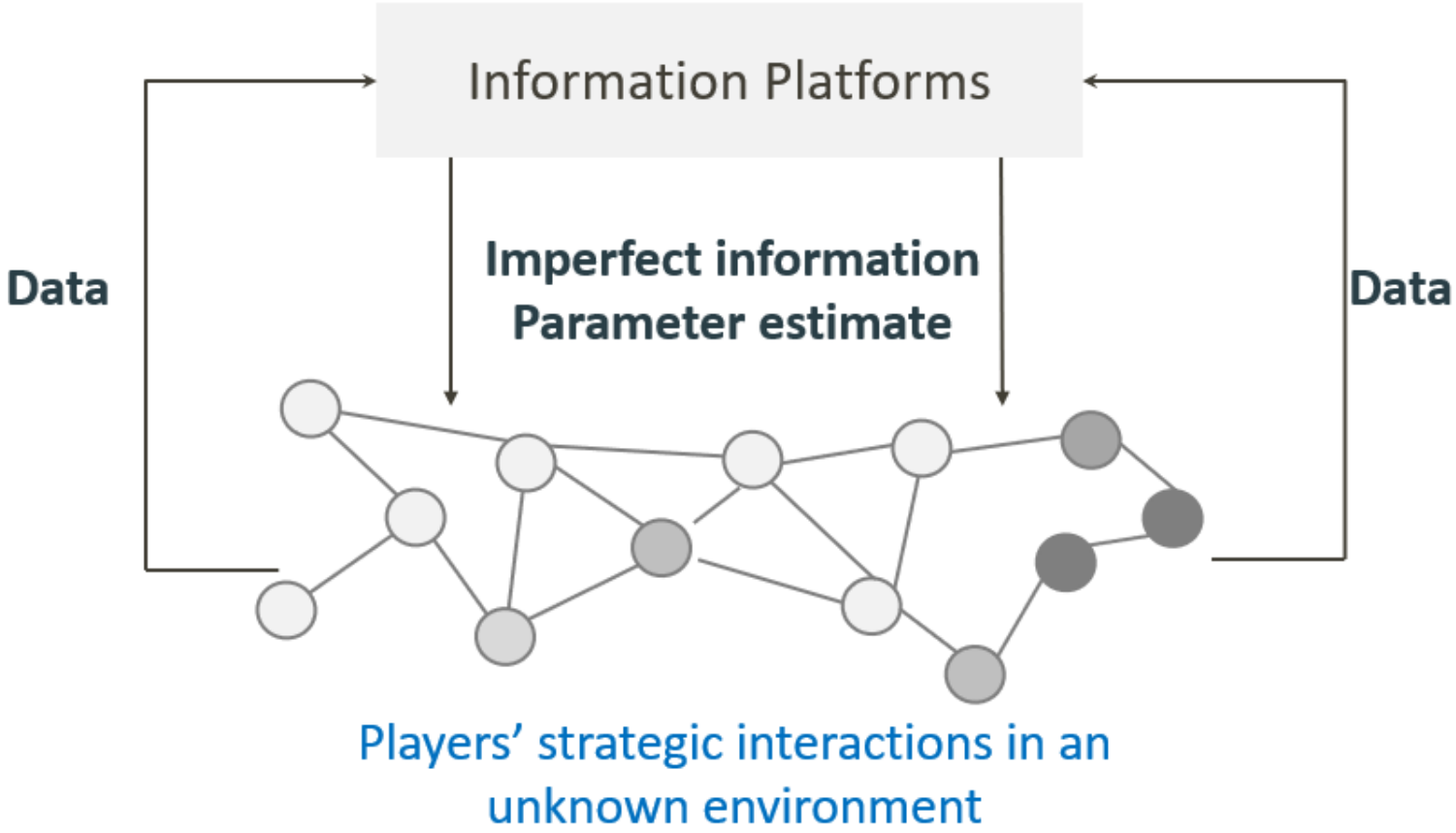
(Necsys, L4DC, preprint arXiv:2109.00719v1)

Adaptive incentive design with learning agents in engineering systems, joint with Chinmay Maheshwari, Kshitij Kulkarni, and Shankar Sastry

https://arxiv.org/pdf/2110.08879

https://arxiv.org/abs/2204.05507

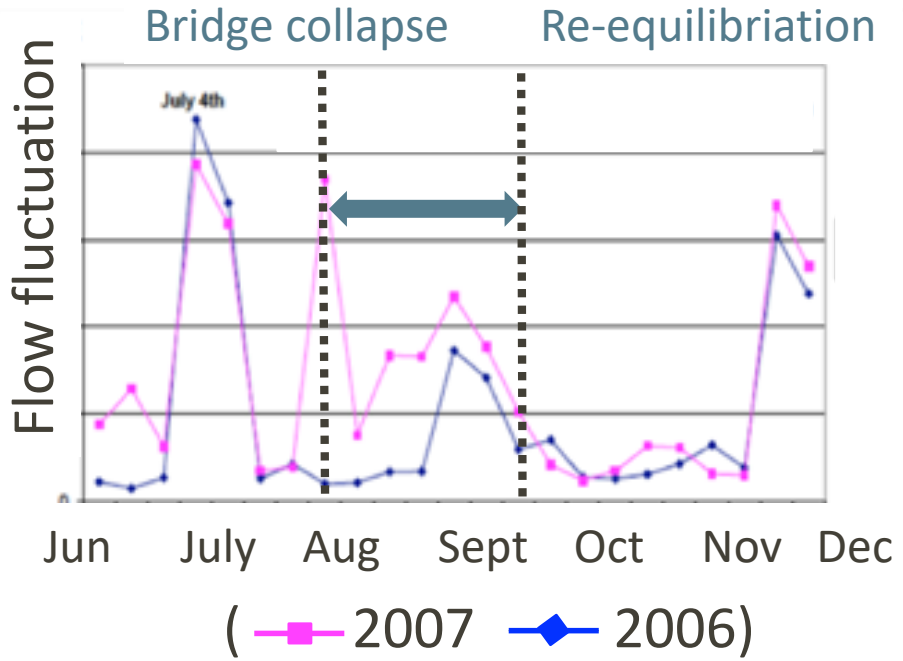
# Learning with Adaptive Strategies




How do strategic players rely on an information platform to learn and adjust their strategies in an unknown environment?

# *Re-equilibriation* after Disruptive Events

I-35W bridge collapse on 8/1/2007



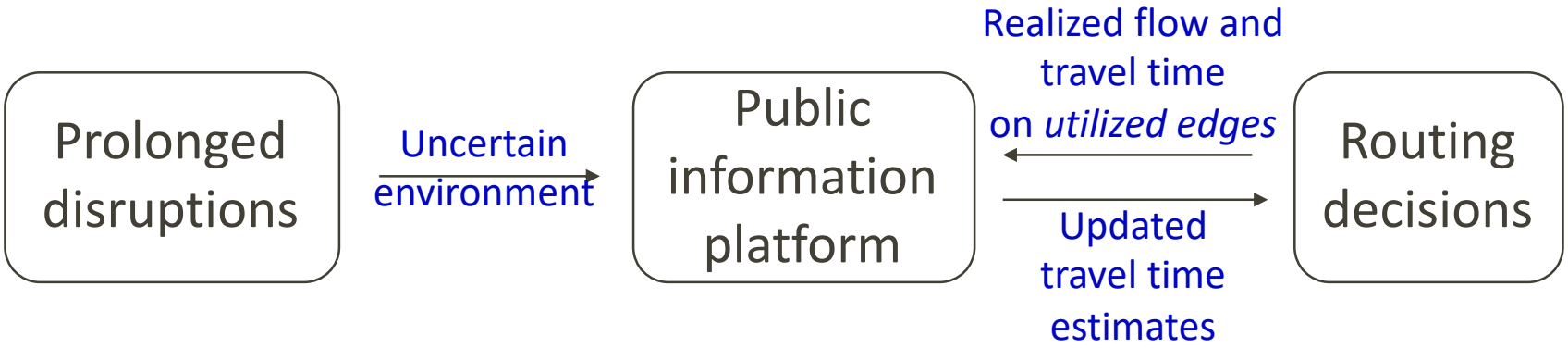

*The collapse abruptly interrupted the usual route of about 140,000 daily vehicle trips and substantially disturbed the flow pattern on the network.*

*It took several weeks for the network to re-equilibrate. Travel behavior changes after network disruption are not well-understood.*

Zhu, Levinson, Liu (2010)

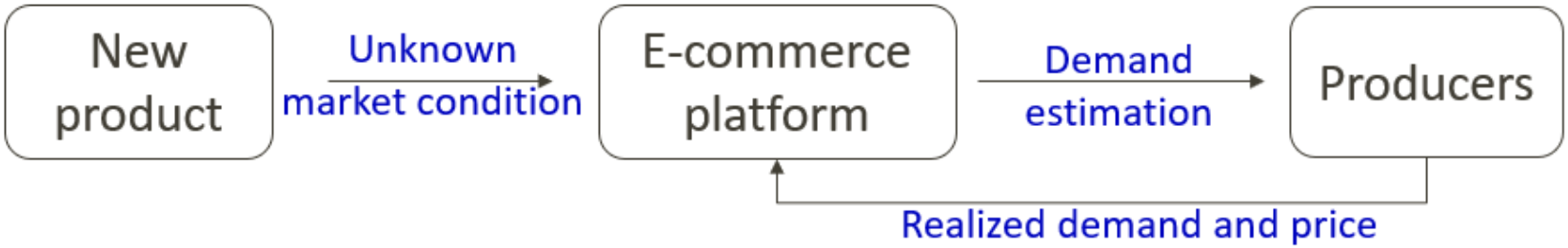
# Learning in Traffic Routing

Recall: Re-equilibriation after I-35W collapse. Travelers adapt their routing decisions while learning in the uncertain network condition




- Unknown state represents latent network condition
- Realized travel times in each stage (day-to-day routing) based on random and state-dependent costs
- Public information system (aggregator): broadcasts updated travel time estimates to all travelers

# Learning in Cournot Game




- Companies decide the production level while learning the uncertain market condition.

# Model: Bayesian Learning

Players repeatedly play a game in steps $k = 1, 2, \ldots$:

- Players $i \in I$
- Unknown parameter $s \in \mathcal{S}$
- Strategy profile in step $k$: $q^k = (q_i^k)_{i \in I} \in \mathcal{Q}$
- Players' (nonlinear) utilities depend on unknown parameter vector
- Given parameter $s$, the payoff outcomes are randomly realized according to $\phi^s(y^k | q^k)$

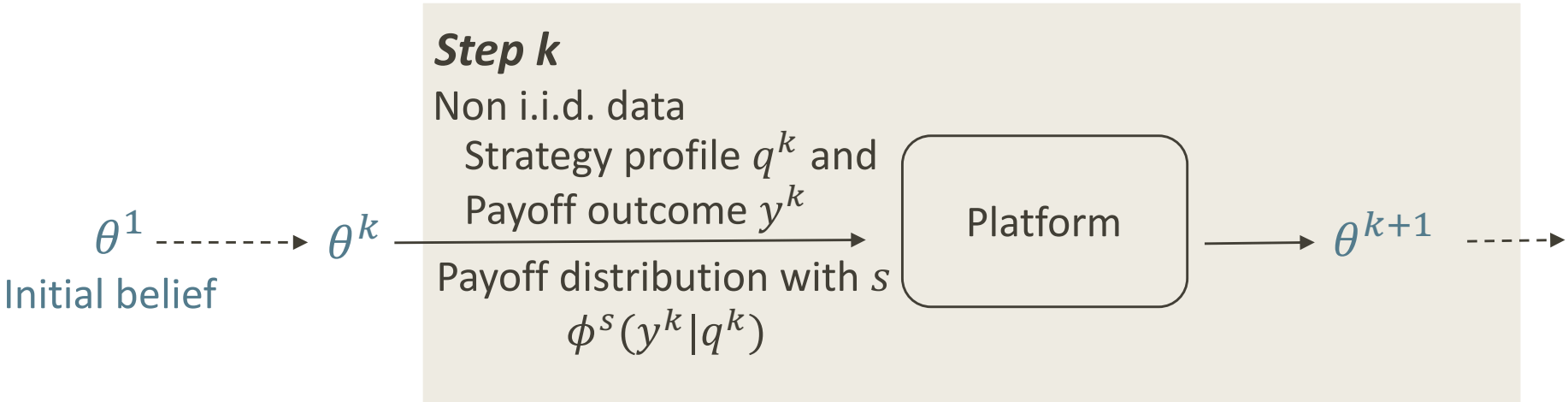
$$y_i^k = u_i^s(q^k) + \epsilon_i^s(q^k),$$

where $u_i^s(q^k)$ is average payoff function and $\epsilon_i^s(q^k)$ is the noise term with zero mean

- True parameter is $s^* \in \mathcal{S}$

# Belief Updates

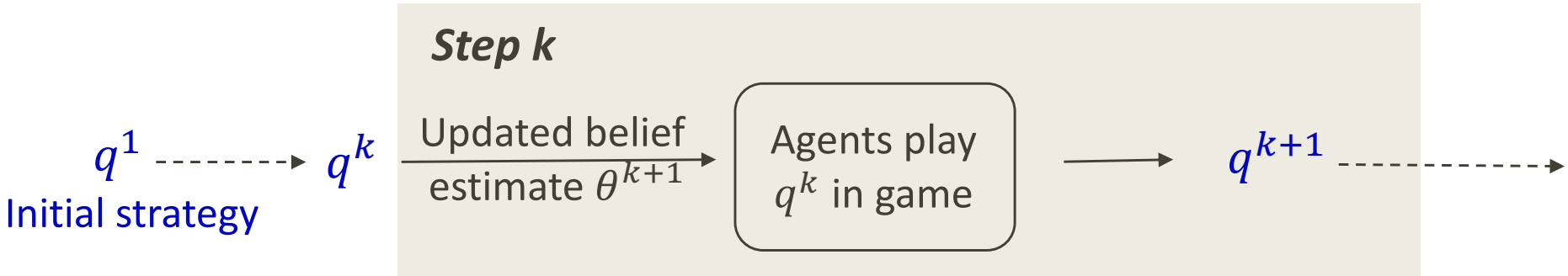
Belief estimate in step $k$ in $\theta^k = \left(\theta^k(s)\right)_{s\in\mathcal{S}}$




- Information platform observes $\left(q^k, y^k\right)$ in each step, and update the belief estimate $\theta^k$ as follows:

$$\theta^{k+1}(s) = \frac{\theta^k(s)\phi^s(y^k|q^k)}{\sum_{s'\in\mathcal{S}}\theta^k(s')\phi^{s'}(y^k|q^k)}, \quad \forall s \in \mathcal{S}$$

# Strategy Updates



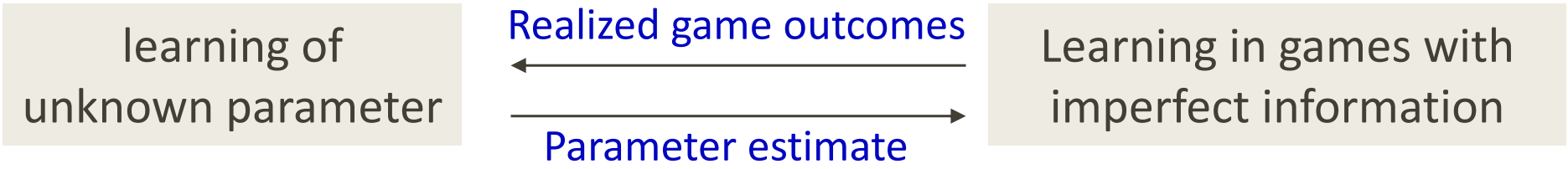

In step $k$:

- Players receive belief estimate $\theta^{k+1}$
- Player $i$'s expected utility $\mathbb{E}\left[u_i^s(q)\middle|\theta^{k+1}\right] = \sum_{s\in\mathcal{S}} \theta^{k+1}(s) u_i^s(q)$
- Strategy update: $q_i^{k+1} = g_i(\theta^{k+1}, q^k)$

# Examples of strategy updates

- **Simultaneous best response:**
  $q_i^{k+1} = g_i^{BR}(\theta^{k+1}, q_{-i}^k) = \arg\max_{q_i} \sum_{s\in S} \theta^{k+1}(s) u_i^s(q_i, q_{-i}^k)$
- **Sequential best response:**
  $$q_i^{k+1} = \begin{cases} g_i^{BR}(\theta^{k+1}, q_{-i}^k) \text{ if } \ k \mod |I| = i \\ q_i^k \quad \text{otherwise} \end{cases}$$
- **Inertial best response:**
  $q_i^{k+1} = (1-\alpha^k) q_i^k + \alpha^k g_i^{BR}(\theta^{k+1}, q_{-i}^k)$
- **Belief-based no-regret learning:**
  $$q_i^{k+1} = \arg\max_{q_i \in Q_i} \{< x_i^{k+1}, q_i > - h_i(q_i)\}, \quad \forall i \in I, \quad \forall k,$$
  where $h_i : Q_i \to \mathbb{R}$ is a continuous and strongly convex regularizer, and $(x_i^k)_{k=1}^{\infty}$ is a sequence of each player $i$'s scores such that
  $$x_i^{k+1} = x_i^k + \alpha^k \left( \sum_{s\in S} \theta^{k+1}(s) \frac{\partial u_i^s(q^k)}{\partial q_i^k} \right), \quad \forall i \in I, \quad \forall k.$$

# Features of Our Learning Model




Key feature: Dynamic interplay between statistical learning of the parameter on the platform and strategy learning in the game

- Platform aggregates information of the unknown parameter based on the non-i.i.d. outcomes generated by the players' strategies
- Players are strategic and choose strategies based on the imperfect information provided by the platform

# Related literature

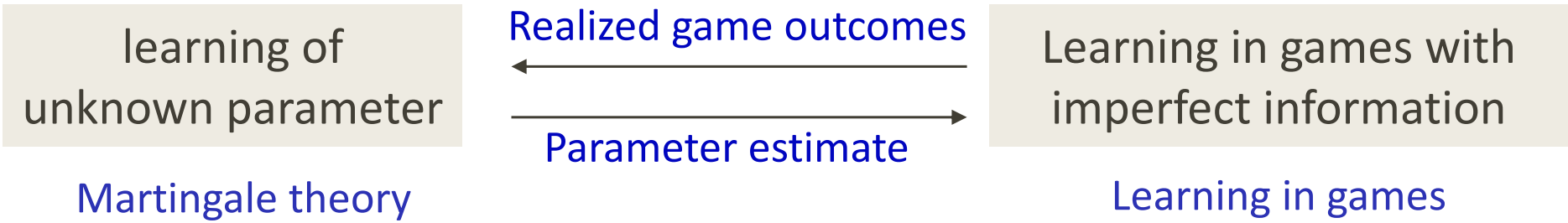
[Statistical estimates and learning] Wald (1950), Blackwell and Dubins (1962), Sorin (1999), Gale and Kariv (2003), Acemoglu, Dahleh, Lobel, and Ozdaglar (2011), Jadbabaie, Molavi and Tahbaz-Salehi (2013), ...

[Strategy learning in games]

- Best response & fictitious play Milgrom & Roberts (1990), Monderer & Shapley (1996), Hofbauer & Sorin (2006), Hofbauer & Sandholm (2002), ...
- No-regret & gradient learning Rosen (1965), Hart & Mas-Colell (2000, 2001), Papadimitriou & Roughgarden (2008), Daskalakis & Panageas (2018), Mertikopoulos & Zhou (2019), Mazumdar, Ratliff, Sastry (2020), ...
- Distributed learning in games Marden & Shamma (2012), Blume et al. (1993), Hart & Mas-Colell (2003), Daskalakis et al. (2011), Cominetti, Melo & Sorin (2010), Krichene, Drighes & Bayen (2014), ...

[Belief-based learning in extensive form game] Fudenberg & Levine (1993), Dekel, Fudenberg, & Levine (1999), ...

# Our Focus




New tools to study the dynamic interplay between statistical learning of the parameter on the platform and strategy learning in the game

- Convergence: The set of parameter estimates and strategies that arise as long-run outcomes of the stochastic dynamics
- Stability properties: local and global
- Conditions for complete learning, and how to explore and find complete information equilibrium otherwise

# Assumptions

Assumption 1: For any $\theta$ and any $q$, the strategy update $g(\theta, q)$ is upper hemicontinuous in $\theta$ and $q$.

- Satisfied by all four examples of strategy updates in continuous games

Assumption 2: Given any static belief $\theta^k = \theta$, the strategy update converges to a Nash equilibrium in the game with belief $\theta$.

- Our focus is NOT the convergence of strategy learning in static information environment, but rather to study how the interplay of belief and strategy updates affect one another
- Satisfied in potential games, zero-sum games, dominance solvable games, strictly concave games, etc.
- We will talk about how this assumption can be relaxed.

# Convergence of Beliefs and Strategies

**Theorem 1.** Under Assumption 1-2, $\lim_{k\to\infty}(\theta^k, q^k) = (\bar{\theta}, \bar{q})$ w.p.1.,

- Belief $\bar{\theta}$ consistently estimate the payoff distribution with $\bar{q}$

$$\mu(y|\bar{\theta}, \bar{q}) \triangleq \sum_{s\in\mathcal{S}} \bar{\theta}(s)\phi^s(y|\bar{q}) = \phi^{s^*}(y|\bar{q}).$$

- Players have no incentive to deviate: i.e. $\bar{q}$ is an equilibrium strategy in game with imperfect information $\bar{\theta}$

Belief converges exponentially fast: $\forall s \in \{S|D_{KL}(\phi^{s^*}(y|\bar{q})||\phi^s(y|\bar{q})) > 0\}$,

$$\lim_{k\to\infty} \frac{1}{k} \log(\theta^k(s)) = -D_{KL}(\phi^{s^*}(y|\bar{q})||\phi^s(y|\bar{q})), \quad w.p.\ 1.$$

# Proof idea

1. Belief convergence $\theta^k \to \bar{\theta}$: Belief ratio $\theta^k(s)/\theta^k(s^*)$ is martingale
2. Strategy convergence $q^k \to \bar{q}$: If strategies converge in games with a constant belief then strategies also converge with changing beliefs
   - Construct an auxiliary strategy sequence such that $\tilde{q}^k = q^k$ for all $k = 1, 2, \ldots, K$, and strategies after $K$ is updated with $\bar{\theta}$
   - Auxiliary strategy sequence converges to an equilibrium strategy associated with $\bar{\theta}$ follows from Assumption 2.
   - The distance between the auxiliary strategy sequence and the original sequence converges to zero as $K \to \infty$ follows from Assumption 1.
3. Belief $\bar{\theta}$ consistently estimate the payoff distribution with $\bar{q}$: Approximate likelihood function with i.i.d. process based on fixed point strategy

# Remark: Non-convergence

Without Assumption 2, learning may not converges.

- Belief converges, i.e. $\lim_{k\to\infty} \theta^k = \bar{\theta}, \quad w.p.1.$
- Asymptotic property (convergence, cycle, chaotic behavior) of $(q^k)_{k=1}^{\infty}$ is the same as the strategies updated with static belief $\bar{\theta}$
- Consistent payoff estimate:
  $\lim_{k\to\infty} D_{KL}\left(\phi^{s^*}(y^k|q^k)||\mu(y^k|\bar{\theta}, q^k)\right) = 0.$

# Fixed Point Properties


- Learning may not recover complete information of $s^*$ because belief $\bar{\theta}$ may form a wrong payoff estimation with strategies $q \neq \bar{q}$
  - Similar to self-confirming equilibrium (Fudenberg & Kreps 1993, 1995) Beliefs of opponents' strategies may be wrong at unreached info. sets

- Complete information fixed points: $(\theta^*, q^*)$, where $\theta^*(s^*) = 1$ and $q^* \in \mathrm{EQ}(\theta^*)$.

# Global Stability

A fixed point belief $\bar{\theta}$ and the associated equilibrium set $\mathrm{EQ}(\bar{\theta})$ are *globally stable* if

$$\forall(\theta^1, q^1), \quad \lim_{k\to\infty} \theta^k = \bar{\theta}, \text{ and } \lim_{k\to\infty} q^k \in \mathrm{EQ}(\bar{\theta}), \ w.p.1$$

**Proposition 1.** The following three statements are equivalent:

(a) The set of globally stable fixed points is non-empty.

(b) For any $\theta \in \Delta(\mathcal{S}) \setminus \{\theta^*\}$, there exists $q \in \mathrm{EQ}(\theta)$ and $s \in [\bar{\theta}]$ such that $s$ is not payoff consistent at $q$.

(c) All fixed points are complete information fixed points, i.e. $\Omega = \{(\theta^*, \mathrm{EQ}(\theta^*))\}$.

Any globally stable fixed point must be a complete information fixed point.

# Local stability

$(\bar{\theta}, \mathrm{EQ}(\bar{\theta}))$ are *locally stable* if $\forall \gamma \in (0,1)$, $\forall \bar{\epsilon}, \bar{\delta} > 0$, $\exists \epsilon^1, \delta^1 > 0$ such that for the learning dynamics starting from $\theta^1 \in N_{\epsilon^1}(\bar{\theta})$ and $q^1 \in N_{\delta^1}(\mathrm{EQ}(\bar{\theta}))$,

$$\lim_{k\to\infty} \Pr\left(\theta^k \in N_{\bar{\epsilon}}(\bar{\theta}),\ q^k \in N_{\bar{\delta}}(\mathrm{EQ}(\bar{\theta}))\right) > \gamma.$$

Local stability requires that the learning dynamics is robust to small perturbations around $\bar{\theta}$ and $\mathrm{EQ}(\bar{\theta})$.

- Local perturbations of beliefs may arise from random errors in data collection or analysis algorithms
- Local perturbations of strategies are likely to occur due to players' mistakes in choosing strategies or random local exploration

# Local Stability

**Theorem 2** A fixed point $(\bar{\theta}, \bar{q})$ is locally stable if

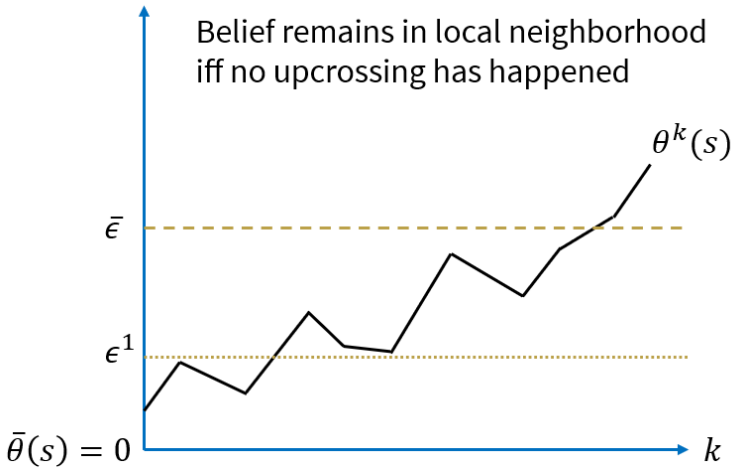
- *Local consistency:* Belief $\bar{\theta}$ consistently estimates the payoff distribution in a small neighborhood of $\bar{q}$ (instead of *just* at $\bar{q}$)
- *Local invariance:* $\exists$ a neighborhood of fixed point that is invariant to locally perturbations of both strategy and belief

Local consistency ensures that any $s$ that cannot be distinguished from $s^*$ at $\bar{q}$ is also not distinguishable when strategy is locally perturbed.

Local invariance reduces to local stability analysis of strategy learning dynamics when the belief is held constant.

# Proof idea

- Identify the initial state neighborhood to ensure that the belief does not leave $N_{\bar{\epsilon}}(\bar{\theta})$ with probability higher than $\gamma$




- Use martingale upcrossing inequality to bound the local perturbations of belief ratio for any $s \notin [\bar{\theta}]$
- Use local consistency property to show that beliefs of $s \in [\bar{\theta}]$ remain in local neighborhood of $\bar{\theta}(s)$.

- Use local invariance property to show that strategy remain in local neighborhood when beliefs remain in local neighborhood

# Local stability of complete information fixed points

**Proposition 2.** Any complete information fixed point is locally stable.

- Complete information fixed points are robust to local perturbation of beliefs and strategies.
- Other fixed point may not be locally stable unless the local consistency property and local invariance property are satisfied.

# Complete learning

Learning is complete if $\bar{q}$ is a complete information Nash equilibrium.

*Case 1:* $\bar{\theta} = \theta^*$ and $\bar{q} \in \mathrm{EQ}(\theta^*)$

- If $[\bar{\theta}]$ is a singleton set, then we know the true parameter is identified
- (Proposition 2) Case 1 w.p.1 $\Leftrightarrow$ global stability $\Leftrightarrow$ True parameter is identifiable in equilibrium

*Case 2:* $\bar{\theta} \neq \theta^*$ but $\bar{q} \in \mathrm{EQ}(\bar{\theta}) = \mathrm{EQ}(\theta^*)$. Identifying the true parameter is not necessary for playing a complete information equilibrium

- How can we ensure that $\bar{q}$ is a complete information equilibrium?
- If learning is not complete, how to find a complete information equilibrium?

# Complete learning

**Proposition 3.** For a fixed point $(\bar{\theta}, \bar{q})$, $\bar{q} = q^*$ if

(i) *Local consistency.* Belief $\bar{\theta}$ consistently estimates the payoff distribution in a small neighborhood of $\bar{q}$ (instead of *just* at $\bar{q}$)
i.e. $\bar{q}_i$ is a locally optimal strategy

(ii) *Payoff concavity.* The payoff function $u_i^s(q_i, q_{-i})$ is concave in $q_i$ for all $q_{-i} \in Q_{-i}$, all $i \in \mathcal{I}$ and all $s \in [\bar{\theta}]$.
i.e. $\bar{q}_i$ is a best response to $\bar{q}_{-i}$ with $s^*$

**Remarks:**

- Payoff concavity is assumed in most studies of continuous games
- Local consistency is equivalent to requiring that $\bar{\theta}$ provides a consistent estimate of utility gradient in local neighborhood of $\bar{q}$ – this is typically assumed for every step in no-regret learning.

# Local exploration

If these conditions are not satisfied, how can we find a complete information equilibrium?

- Concave payoff functions: Local exploration is sufficient to exclude parameters that are not locally consistent. Leads to a new fixed point which satisfies (i) and (ii), and learning is complete
- Payoff functions are not concave: Local exploration is insufficient. Need to distinguish any pair of $s, s' \in [\bar{\theta}]$ by sampling payoffs at strategy profiles outside of local neighborhood of $\bar{q}$

# Extensions

**Time scale separation:** Results hold when the belief update is at a slower time scale compared to the strategy update

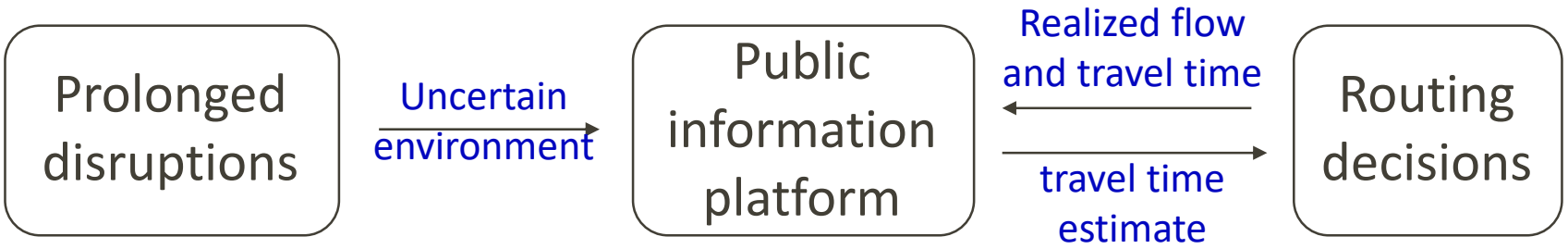
**Finite games:** Results hold in finite games. Local perturbation will always identify the complete information equilibrium

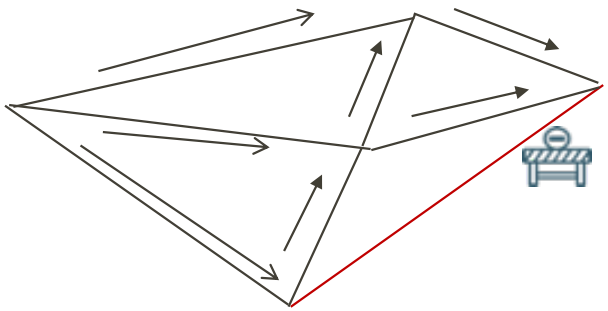
**Other parameter estimates:** Results hold with MAP, MLE, and OLS

# Summary

1. Dynamic interplay between parameter learning and strategy learning
2. Convergence of the joint evolution of beliefs and strategies
3. Fixed point: consistent payoff estimate, and no incentive to deviate
4. Local and global stability properties
5. Sufficient conditions for complete learning. Finding complete information equilibrium through local exploration.

# Learning in Traffic Routing



- Belief accurately estimates the costs of edges that are taken
- Traffic flow is eq. given the belief

- Incomplete learning: cost estimate may not be accurate on untaken edges
- Local exploration resolves the problem

# Incomplete learning

For series-parallel networks, the total cost of any fixed point routing strategy is no less than that in complete information equilibrium

$$\text{Avg. cost of all agents with any } \bar{q} \geq \text{Avg. cost of all agents with } q^*$$

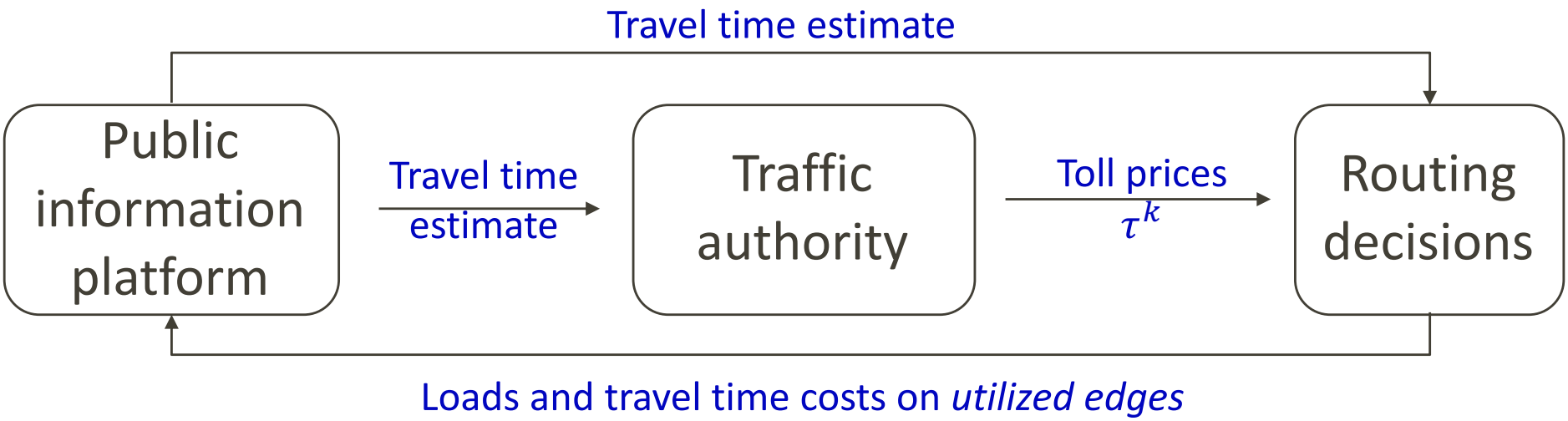
Thus, complete learning is socially desirable on series-parallel routes

Learning is guaranteed to be complete if

- The unknown network condition does not affect the free flow travel time of each edge
- The total demand of travelers is large so that all edges are utilized

# Learning for Tolling

Our results apply to adaptive tolling with unknown latency functions, where routing decisions are based on updated beliefs and tolls



# Learning for Tolling

- Bayesian update of belief $\theta^k \to \theta^{k+1}$
- Update routing strategy based on $\theta^{k+1}$ and $\tau^{k+1}$

$$q^{k+1} = (1-\alpha^k)q^k + \alpha^k g\left(\theta^{k+1}, \tau^{k+1}\right)$$

- Update toll prices $\tau^k = \left(\tau_e^k\right)_{e\in E}$: $\mathit{Toll}\left(\theta^{k+1}, q^k\right)$ is the estimated externality (i.e. marginal cost) based on $\theta^{k+1}$ and $q^k$

$$\tau^{k+1} = (1-\beta^k)\tau^k + \beta^k \mathit{Toll}\left(\theta^{k+1}, q^k\right),$$

Toll $\tau^k$ is updated at a slower time scale compared with $\theta^k$ and $q^k$

$$\lim_{k\to\infty} \beta^k/\alpha^k = 0.$$

**Proposition (Adaptive tolling leads to socially optimal outcome)** If $x_e^1 > 0$ for all $e \in E$, then learning leads to optimal tolling mechanism and socially optimal routing with complete information:

$$\text{Toll } \tau^k \ \to\ \tau^{opt} \text{ in } s^*, \text{ and Load } w^k \ \to\ w^{opt} \text{ in } s^*$$

# Extension: Adaptive incentive design with learning agents

Design of social-scale systems with learning agents

1. Update estimate of unknown pay-off relevant parameters $\theta^k \to \theta^{k+1}$
2. Learning agents update strategies $q^{k+1}$ based on information $\theta^{k+1}$ and incentive mechanism $\tau^k$
3. Externality-based incentive update incentive mechanism $\tau^k$:

$$\tau_i^{k+1} = (1-\beta^k)\tau^k + \beta^k \left( \begin{array}{c} \text{estimated externality caused by player } i \\ \text{based on } q^{k+1} \text{ and } \theta^{k+1} \end{array} \right)$$

Incentive update needs to be at a slower timescale compared to parameter learning and strategy learning

The estimated externality = the marginal cost of player $i$'s strategy for the society − the marginal cost for the player themselves.

# Summary of results

The externality-based adaptive incentive design ensures that any fixed point of the dynamics must be a socially optimal incentive mechanism

# Summary of results

The externality-based adaptive incentive design ensures that any fixed point of the dynamics must be a socially optimal incentive mechanism

This is because an optimal mechanism aligns individuals' incentives with the societal goals by asking each individual to pay for their externality

# Summary of results

The externality-based adaptive incentive design ensures that any fixed point of the dynamics must be a socially optimal incentive mechanism

This is because an optimal mechanism aligns individuals' incentives with the societal goals by asking each individual to pay for their externality

In large scale societal systems, computing the optimal incentive design on top of equilibrium outcomes is a challenging bi-level optimization problem

# Summary of results

The externality-based adaptive incentive design ensures that any fixed point of the dynamics must be a socially optimal incentive mechanism

This is because an optimal mechanism aligns individuals' incentives with the societal goals by asking each individual to pay for their externality

In large scale societal systems, computing the optimal incentive design on top of equilibrium outcomes is a challenging bi-level optimization problem

Adaptive incentive design approach allows the social planner to eventually induce optimal long-run outcomes with simple update rules

# Summary of results

The externality-based adaptive incentive design ensures that any fixed point of the dynamics must be a socially optimal incentive mechanism

This is because an optimal mechanism aligns individuals' incentives with the societal goals by asking each individual to pay for their externality

In large scale societal systems, computing the optimal incentive design on top of equilibrium outcomes is a challenging bi-level optimization problem

Adaptive incentive design approach allows the social planner to eventually induce optimal long-run outcomes with simple update rules

The adaptive incentive mechanism converges in a variety of problems, e.g. public good provision in networks, routing, and Cournot competition.

joint with Chinmay Maheshwari, Kshitij Kulkarni, and Shankar Sastry;

https://arxiv.org/pdf/2110.08879, https://arxiv.org/abs/2204.05507

**Thank you!**

manxiwu@berkeley.edu