Distributed Online Stochastic Optimization with Myopic Agents

Ankur Mani
joint work with Ilan Lobel [2], Josh Reed [2],
Dhananjay Tiwari [1], and Srinivasa Salapaka [1]

[1]University of Illinois Urbana-Champaign

[2]NYU Stern School of Business

Centre for Networked Intelligence
Indian Institute of Science
August 13, 2026

Online Stochastic Optimization

- D : Set of available actions
- $r : D \to \mathcal{R}$, $r(x)$ is the uncertain reward from action $x \in D$ with unknown distribution.

Online Stochastic Optimization

- D : Set of available actions
- $r : D \to \mathcal{R}$, $r(x)$ is the uncertain reward from action $x \in D$ with unknown distribution.
- Order of events at each time until the end of time horizon
  - agent takes action based upon history
  - agent observes reward and updates history

Online Stochastic Optimization

- D : Set of available actions
- $r : D \to \mathcal{R}$, $r(x)$ is the uncertain reward from action $x \in D$ with unknown distribution.
- Order of events at each time until the end of time horizon
  - agent takes action based upon history
  - agent observes reward and updates history
- Objective : maximize the expected cumulative reward until the end of time horizon.

Online Stochastic Optimization

- D : Set of available actions
- $r : D \to \mathcal{R}$, $r(x)$ is the uncertain reward from action $x \in D$ with unknown distribution.
- Order of events at each time until the end of time horizon
  - agent takes action based upon history
  - agent observes reward and updates history
- Objective : maximize the expected cumulative reward until the end of time horizon.
- Learn the reward distribution through experimentation early and then optimize.

Online Stochastic Optimization

- D : Set of available actions
- $r : D \to \mathcal{R}$, $r(x)$ is the uncertain reward from action $x \in D$ with unknown distribution.
- Order of events at each time until the end of time horizon
  - agent takes action based upon history
  - agent observes reward and updates history
- Objective : maximize the expected cumulative reward until the end of time horizon.
- Learn the reward distribution through experimentation early and then optimize.
- Two formulations

Multiarmed Bandits	Online Linear Optimization
Finite D, independent rewards	Infinite D, correlated rewards
[Lobel, Mani, Reed, 2025]	[Tiwari, Mani, Salapaka, 2026]

Distributed Experimentation

- Central planners often rely on experimentation by independent and myopic agents.

Distributed Experimentation

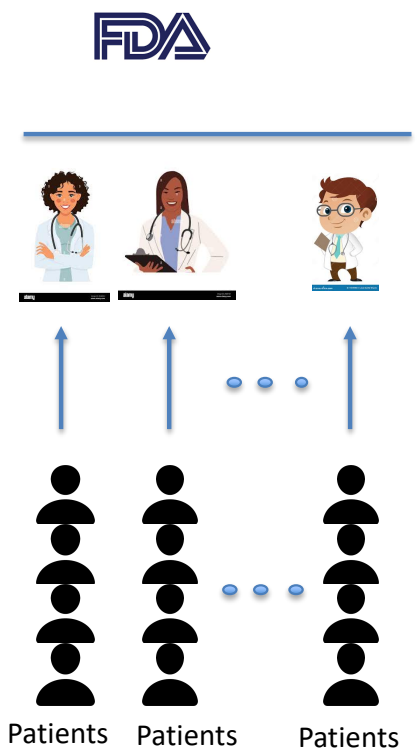
- Central planners often rely on experimentation by independent and myopic agents.
  - platforms

Distributed Experimentation

- Central planners often rely on experimentation by independent and myopic agents.
  - platforms, agentic AI

Distributed Experimentation

- Central planners often rely on experimentation by independent and myopic agents.
  - platforms, agentic AI, right-to-try,selling through independent agents



[Ayer, Mani, 2026]

Case: Moderna Covid-19 Vaccine

Google

How long did Moderna take to create the vaccine after the gene sequence was released by China?

AI Mode | All | News | Shopping | Images | Videos | Web | More | Tools

AI Overview

Moderna took **two days** to design the gene sequence for its vaccine candidate (mRNA-1273) after Chinese scientists published the novel coronavirus's genetic sequence online on January 11, 2020.

The key dates in Moderna's initial timeline are:

- **January 11, 2020:** Chinese scientists released the genetic sequence of the coronavirus.
- **January 13, 2020:** Moderna's team and scientists at the National Institutes of Health (NIH) finalized the sequence for the vaccine.
- **February 24, 2020:** The first batch of the experimental vaccine was shipped to the NIH for testing. This was 42 days after the sequence selection was finalized.
- **March 16, 2020:** The first human dose was administered in a Phase 1 clinical trial, 63 days from sequence selection to first human dosing.

The rapid design time was possible due to the speed of mRNA technology, which only required the digital genetic sequence of the virus, not the cultivation of the live virus itself.

Moderna Designed Its Coronavirus Vaccine in Just 2 Days

Dec 18, 2020 — Follow Aria Bendix * The FDA granted emergency authorization to Moderna'...

Business Insider

How the Moderna Covid-19 mRNA vaccine was made so quickly

Jul 2, 2021 — The timeline: A vaccine ...

CNBC

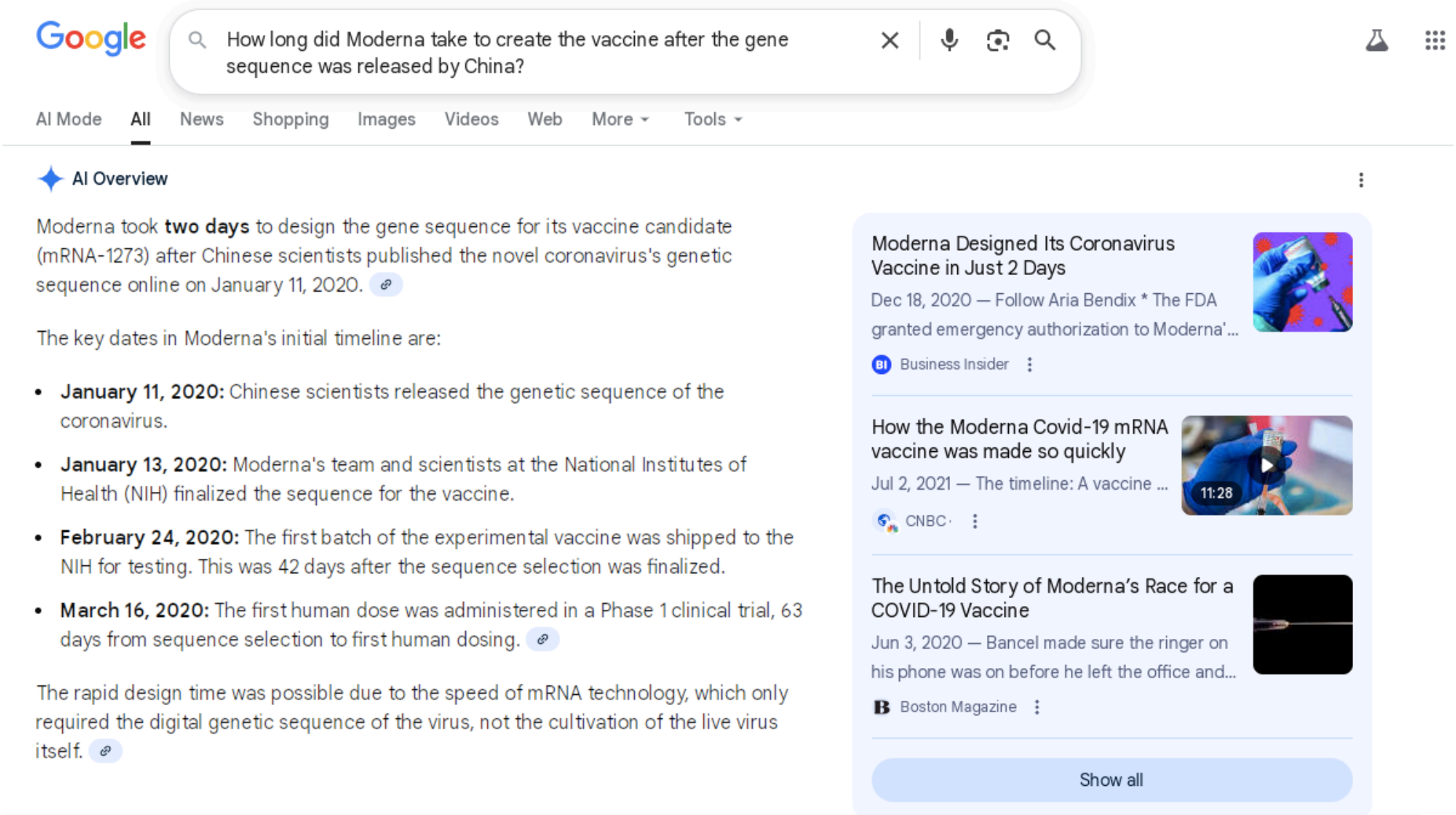
The Untold Story of Moderna's Race for a COVID-19 Vaccine

Jun 3, 2020 — Bancel made sure the ringer on his phone was on before he left the office and...

Boston Magazine

Show all

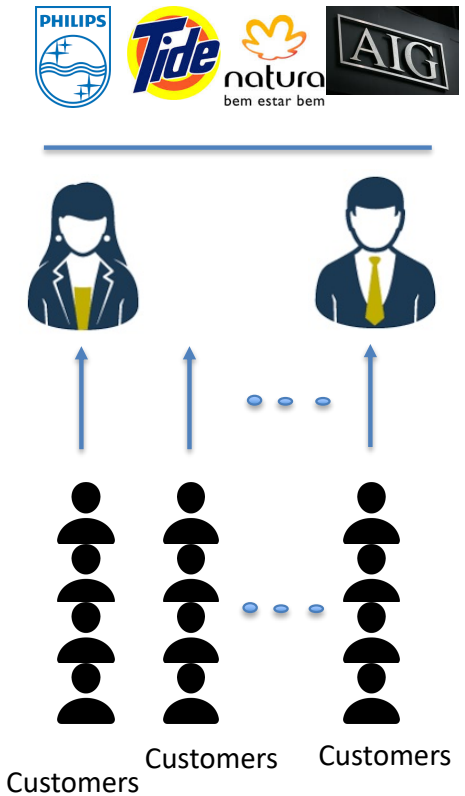
Case: Moderna Covid-19 Vaccine



After testing the Moderna vaccine on 30,000 volunteers, on Dec. 18, 2020, the FDA authorized it for emergency public use

Distributed Experimentation

- Firms often rely on experimentation by independent and myopic agents.
  - platforms, agentic AI, right-to-try, selling through independent agents



Distributed Experimentation

- Central planners often rely on experimentation by independent and myopic agents.
  - platforms, agentic AI, right-to-try, selling through independent agents
- What can the central planner do?

Distributed Experimentation

- Central planners often rely on experimentation by independent and myopic agents.
  - platforms, agentic AI, right-to-try, selling through independent agents
- What can the central planner do?
  - Incentivizing exploration [Frazier et. al. 2014, Halac et. al. 2016, Chen et. al. 2018]
  - Information design [Kremer et. al. 2014, Che and Horner, 2018, Papanastasiou et. al. 2018, Mansour et. al. 2020, Immorlica et. al. 2020, Acemoglu et. al. 2022]

Distributed Experimentation

- Central planners often rely on experimentation by independent and myopic agents.
  - platforms, agentic AI, right-to-try, selling through independent agents
- What can the central planner do?
  - Incentivizing exploration [Frazier et. al. 2014, Halac et. al. 2016, Chen et. al. 2018]
  - Information design [Kremer et. al. 2014, Che and Horner, 2018, Papanastasiou et. al. 2018, Mansour et. al. 2020, Immorlica et. al. 2020, Acemoglu et. al. 2022]
- Can a much simpler control work?
  - What if the central planner irrevocably discards arms it deems suboptimal?

Distributed Experimentation

- Central planners often rely on experimentation by independent and myopic agents.
  - platforms, agentic AI, right-to-try, selling through independent agents
- What can the central planner do?
  - Incentivizing exploration [Frazier et. al. 2014, Halac et. al. 2016, Chen et. al. 2018]
  - Information design [Kremer et. al. 2014, Che and Horner, 2018, Papanastasiou et. al. 2018, Mansour et. al. 2020, Immorlica et. al. 2020, Acemoglu et. al. 2022]
- Can a much simpler control work?
  - What if the central planner irrevocably discards arms it deems suboptimal?
  - Such control can be construed as a very simple information design, i.e.- broadcast a signal whenever an arm is deemed suboptimal.

Model: Overview

Firm offers an assortment of products

Model: Overview

Firm offers an assortment of products

Agents make selling decisions

Model: Overview

Firm can discontinue products

Agents make selling decisions

Model: Central Planner and Agents

- Central planner has a set $\mathbb{J}$ of arms and makes a subset $\mathbb{J}_t \subseteq \mathbb{J}$ available at time t to I agents.

Model: Central Planner and Agents

- Central planner has a set $\mathbb{J}$ of arms and makes a subset $\mathbb{J}_t \subseteq \mathbb{J}$ available at time t to I agents.
- Each agent makes its own effort allocation.

Model: Central Planner and Agents

- Central planner has a set $\mathbb{J}$ of arms and makes a subset $\mathbb{J}_t \subseteq \mathbb{J}$ available at time t to I agents.
- Each agent makes its own effort allocation.
- The cumulative effort spent by agent i on arms until time t is $\boldsymbol{\tau}^i(t) = (\tau_1^i(t), \tau_2^i(t), ..., \tau_J^i(t))$ [Mandelbaum, 1987] ,

$$\tau_j^i(t) = \int_0^t 1\{j \in \mathbb{J}_s\} \partial \tau_j^i(s), \text{ for all } t > 0$$

$$\sum_{j \in \mathbb{J}} \tau_j^i(t) = t, \text{ for all } t > 0$$

Model: Rewards

- The uncertain reward from arm j for agent i for a cumulative effort of s is $R_j^i(s)$.

Model: Rewards

- The uncertain reward from arm j for agent i for a cumulative effort of s is $R_j^i(s)$.
- The rewards from the same arm for different agents are correlated.

Model: Rewards

- The uncertain reward from arm j for agent i for a cumulative effort of s is $R_j^i(s)$.
- The rewards from the same arm for different agents are correlated.
- In particular, μ_j is the drift of arm j.
- The reward

$$R_j^i(s) = \mu_j s + B_j^i(s)$$

  where $B_j^i(s)$ is a standard Brownian motion.
- The Brownian motion is independent for each arm-agent pair.

Model: Information Processing

- Agents observe only their own rewards and efforts.
- Agent i's estimate of the drift of arm j at time t is

$$\hat{\mu}_j^i\left(\tau_j^i(t)\right) = \frac{1}{1+\tau_j^i(t)} R_j^i(\tau_j^i(t)).$$

- The central planner observes the rewards and efforts of all agents.

Model: Misalignment between Planner and Agents

- The central planner is forward looking but the agents are myopic.

Model: Misalignment between Planner and Agents

- The central planner is forward looking but the agents are myopic.
- Agent puts all effort on the arm with the highest drift estimate a.e.

Model: Misalignment between Planner and Agents

- The central planner is forward looking but the agents are myopic.
- Agent puts all effort on the arm with the highest drift estimate a.e.
- The central planner's utility is the reward over the whole period T,

$$U^p(T) = \sum_{j \in \mathbb{J}, 1 \leq i \leq I} R_j^i(\tau_j^i(T)).$$

Main Questions

- Expected regret over a finite horizon under a central planner's policy π is

$$Z_\pi(T, \boldsymbol{\mu}) = I\mu_{\{1\}} T - \sum_{j \in \mathbb{J}, 1 \le i \le I} \mu_j E_\pi \left[\tau_j^i(T) \right].$$

If the central planner can fully control the decision of what arms to put effort on, then

$$\max_{\mu \in \mathrm{R}^J} Z_\pi(T, \boldsymbol{\mu}) = \Omega(J \log IT).$$

[Lai, Robbins 1985; Auer, Cesa-Bianchi, Fischer 2002]

Main Questions

- Expected regret over a finite horizon under a central planner's policy π is

$$Z_\pi(T, \boldsymbol{\mu}) = I\mu_{\{1\}} T - \sum_{j \in \mathbb{J}, 1 \le i \le I} \mu_j E_\pi \left[\tau_j^i(T)\right].$$

If the central planner can fully control the decision of what arms to put effort on, then

$$\max_{\mu \in \mathrm{R}^J} Z_\pi(T, \boldsymbol{\mu}) = \Omega(J \log IT).$$

[Lai, Robbins 1985; Auer, Cesa-Bianchi, Fischer 2002]

- Can a central planner's strategy of dropping arms achieve the regret bound?
  Consider policies such that $\mathbb{J}_t \subseteq \mathbb{J}_s$, for all $0 \le s < t$.
- How many agents are needed to achieve the regret bound?

Notes on Continuous Time Model

- Equivalence of effort allocation and escape time of the drift estimate process
- Allows for writing appropriate Fokker-Planck equation and solve for effort distributions, thus providing finer results.

Notes on Continuous Time Model

- Equivalence of effort allocation and escape time of the drift estimate process
- Allows for writing appropriate Fokker-Planck equation and solve for effort distributions, thus providing finer results.
- Regret under different settings

Setting	Asymptotic Regret
Single agent without planner	$T\sum_{j\in\mathbb{J}}(\mu_{max}-\mu_j)\mathbb{P}(k^\star=j\|\boldsymbol{\mu})+\Theta(J)$
Single agent with planner	$T\sum_{j\in\mathbb{J}}(\mu_{max}-\mu_j)\mathbb{P}(k^\star=j\|\boldsymbol{\mu})+\Theta(J)$
Multiple agents with planner	Let's find out

if $\mu > 0$ then

$$\mathbb{P}\left(k^\star = j|\boldsymbol{\mu}\right) = \frac{1}{J} + \frac{\sqrt{2\pi}\left(\mu_j - \bar{\mu}\right)}{\sqrt{J}} e^{\frac{J\bar{\mu}^2}{2}} \Phi\left(-\sqrt{J}\bar{\mu}\right),$$

where $\bar{\mu} = \frac{1}{J}\sum_{k\in\mathbb{J}}\mu_k$

Central planner's Confidence Bounds

Definition

Upper confidence bound for the drift of j is

$$C_j^+\left(\boldsymbol{\tau}_j^f(t)\right) = \frac{R_j^f\left(\tau_j^f(t)\right)}{\tau_j^f(t)} + \frac{3}{2}\sqrt{\log\left(\frac{JIT}{\sqrt{2\pi}}\right)}\left(\frac{1}{\sqrt{\tau_j^f(t)}} + \frac{1}{\tau_j^f(t)}\right)$$

Definition

Lower confidence bound for the drift of j is

$$C_j^-\left(\boldsymbol{\tau}_j^f(t)\right) = \frac{R_j^f\left(\tau_j^f(t)\right)}{\tau_j^f(t)} - \frac{3}{2}\sqrt{\log\left(\frac{JIT}{\sqrt{2\pi}}\right)}\left(\frac{1}{\sqrt{\tau_j^f(t)}} + \frac{1}{\tau_j^f(t)}\right)$$

where $\boldsymbol{\tau}_j^f(t) = \left(\tau_j^1(t), \ldots, \tau_j^I(t)\right)$, $\tau_j^f(t) = \sum_{i=1}^I \tau_j^i(t)$ and $R_j^f\left(\tau_j^f(t)\right) = \sum_{i=1}^I R_j^i\left(\tau_j^i(t)\right)$.

Central planner's Confidence Bounds

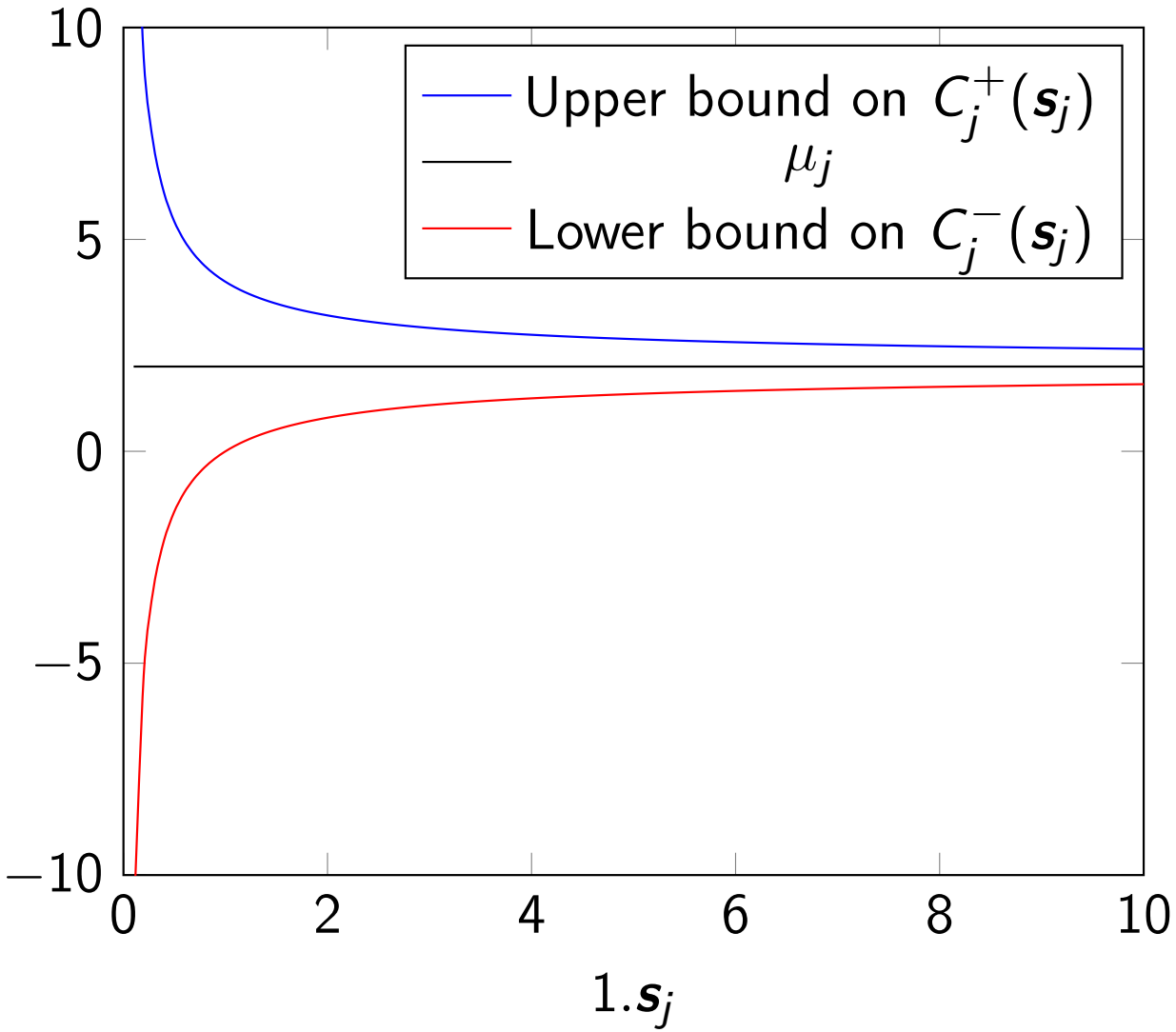


Figure: Concentration of the confidence bounds: With high probability the upper(lower) confidence bound lies between the blue(red) curve and true drift.

Anchor Rate

Definition

Anchor rate: At any time t, we define the anchor rate

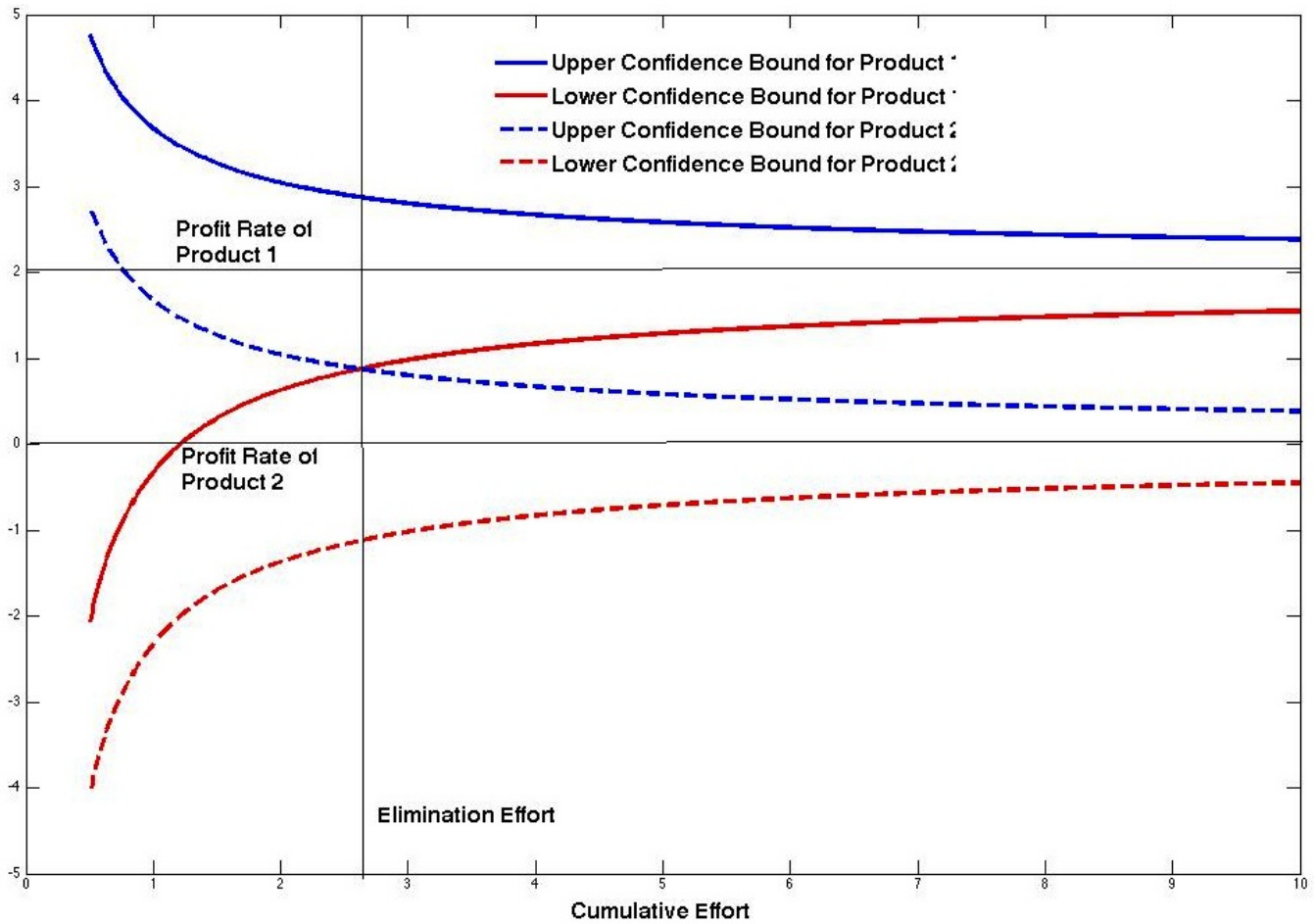
$$l^*(t) = \max_{j \in \mathbb{J}} \sup_{0<s<t} C_j^- \left(\tau_j^f(s) \right).$$

The anchor rate is the highest lower confidence bound over all arms until the current time.

Central planner's Policy

Definition

Central planner's policy π^*: Discard arm j at time $t > 0$ if $C_j^+\left(\tau_j^f(t)\right) \leq l^*(t)$.



Sources of Error

- The arm with the highest expected drift is discarded.
- The anchor rate stays low for a long time.

Erroneously Discarding the Best arm

The probability of discarding the arm with the highest expected drift is very low.

Theorem

With probability at least $1 - \frac{2\sqrt{2\pi}(\log_4 IT+2)}{JIT}$, *the true drift of any arm* j *lies between* $\max_{0<1^T s_j \leq IT} C_j^-(s_j)$ *and* $\min_{0<1^T s_j \leq IT} C_j^+(s_j)$, *i.e.-*

$$P\left(\max_{0<1^T s_j \leq IT} C_j^-(s_j) < \mu_j < \min_{0<1^T s_j \leq IT} C_j^+(s_j)\right) > 1 - \frac{2\sqrt{2\pi}\,(\log_4 IT + 2)}{JIT}.$$

Erroneously Discarding the Best arm

The probability of discarding the arm with the highest expected drift is very low.

Theorem

With probability at least $1 - \frac{2\sqrt{2\pi}(\log_4 IT+2)}{JIT}$, *the true drift of any arm* j *lies between* $\max_{0<1^T \boldsymbol{s}_j \leq IT} C_j^- (\boldsymbol{s}_j)$ *and* $\min_{0<1^T \boldsymbol{s}_j \leq IT} C_j^+ (\boldsymbol{s}_j)$, *i.e.-*

$$P\left(\max_{0<1^T \boldsymbol{s}_j \leq IT} C_j^- (\boldsymbol{s}_j) < \mu_j < \min_{0<1^T \boldsymbol{s}_j \leq IT} C_j^+ (\boldsymbol{s}_j)\right) > 1 - \frac{2\sqrt{2\pi}\,(\log_4 IT + 2)}{JIT}.$$

Corollary

Under the proposed discarding policy π^*, *the probability that the arm with the highest drift is discarded is at most* $2\sqrt{2\pi}\,(\log_4 IT + 2)\,/IT$.

Erroneously Keeping the Anchor Rate Low

The probability that anchor rate stays low for a long time is very low.

Theorem

Under the proposed policy π^*, *if* $I > 12 \log T$ *then by time* $t = \frac{112\alpha^2}{\Delta^2 I}(J-1)$ *the anchor rate satisfies* $I^*(t) \geq \mu_{\{1\}} - \Delta$ *with probability at least* $1 - \frac{2\sqrt{2\pi}(\log_4 IT+2)}{IT} - \log_2 Je^{-\frac{I}{12}}$.

$\Delta > 0$ is the minimum difference between the drifts of any two arms.

$\alpha = \frac{3}{2}\sqrt{\log \frac{JIT}{\sqrt{2\pi}}}$

Lower bound on Anchor Rate

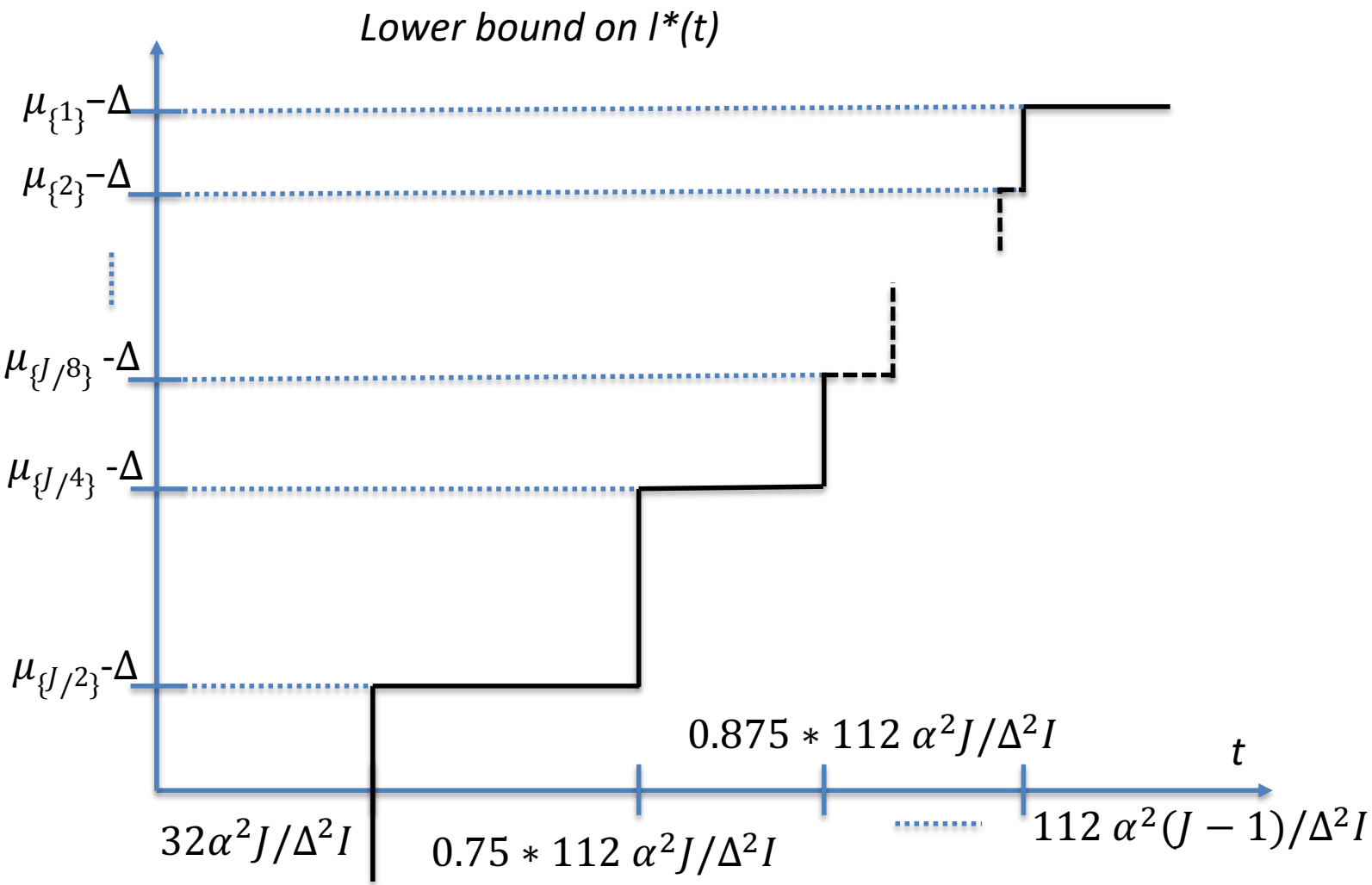


Figure: High probability lower bound on the anchor rate.

Effort on Suboptimal arms

Effort on suboptimal arms grows logarithmically in time.

Theorem

If there are at least $l = 12 \log T$ *agents, then under the discarding policy* π^*, *for any* $\boldsymbol{\mu} \in \mathbb{R}^J$ *with* $\Delta > 0$, *the expected total cumulative effort on suboptimal arms is*

$$\mathbb{E}\left[\sum_{j \neq \{1\}} \tau^f_{\pi^*,j}(T)\right] < \left(\frac{576J}{\Delta^2} + \frac{\sqrt{2\pi}}{\log 2}\right) \log lT + \frac{l \log_2 J}{T^{\frac{l}{12 \log T} - 1}}.$$

$\Delta > 0$ is the minimum difference between the drifts of any two arms.

Optimal Regret

Rate optimal regret under π^*.

Corollary

If there are at least $12 \log T$ *agents then under policy* π^**, for any* $\boldsymbol{\mu} \in \mathbb{R}^J$ *with* $\Delta > 0$*, the regret* $\tilde{Z}_{\pi^*}(T, \boldsymbol{\mu})$ *is less than*

$$(\mu_{\{1\}} - \mu_{\{J\}}) \left(\left(\frac{576J}{\Delta^2} + \frac{\sqrt{2\pi}}{\log 2} \right) \log IT + \frac{I \log_2 J}{T^{\frac{I}{12 \log T} - 1}} \right).$$

Minimum Number of Agents

$\log(T)$ is the critical rate at which the number of agents should be added as the horizon increases.

Theorem

As both the horizon T and number of agents I_T increase, if $\lim_{T\to\infty} \frac{I_T}{\log T} = 0$, *then for each* $\boldsymbol{\mu} \in \mathrm{R}^J$ *(not all equal),*

$\lim_{T\to\infty} \frac{Z_\pi(T,\boldsymbol{\mu})}{\log T} = \infty$ *under any policy* π.

Summary of Contributions

- Continuous time multi-armed bandit framework with numerous myopic agents making independent decisions.

Summary of Contributions

- Continuous time multi-armed bandit framework with numerous myopic agents making independent decisions.
- If the number of agents grows $o(\log T)$ with T then a sublinear regret cannot be achieved by discarding arms.

Summary of Contributions

- Continuous time multi-armed bandit framework with numerous myopic agents making independent decisions.
- If the number of agents grows $o(\log T)$ with T then a sublinear regret cannot be achieved by discarding arms.
- If the number of agents grow at least $12 \log T$ with T then $O(\log T)$ regret can be achieved.

Summary of Contributions

- Continuous time multi-armed bandit framework with numerous myopic agents making independent decisions.
- If the number of agents grows $o(\log T)$ with T then a sublinear regret cannot be achieved by discarding arms.
- If the number of agents grow at least $12 \log T$ with T then $O(\log T)$ regret can be achieved.
- Learning occurs in constant time and the learning speed grows linearly in the number of agents.

Thank you!

Online Stochastic LP: Single Agent with Central Planner

- Agent chooses actions $\boldsymbol{x}(t) \in D \subset \mathbb{R}^d$
- Reward:
$$dR(t) = \boldsymbol{\mu}^\top \boldsymbol{x}(t)\, dt + \boldsymbol{x}^\top(t)\, d\boldsymbol{B}(t)$$
- Myopic agent:
$$\hat{\boldsymbol{x}}(t) = \arg\max_{\boldsymbol{x} \in D} \hat{\boldsymbol{\mu}}(t)^\top \boldsymbol{x}$$
- Planner restricts actions: $D_t^\pi \subset D$

Bayesian Inference & Confidence Ellipsoids

- Observed history:

$$\mathcal{H}_t = \{\{R^i(s), x^i(s)\}_{s \le t} : i = 1, \ldots, N\}$$

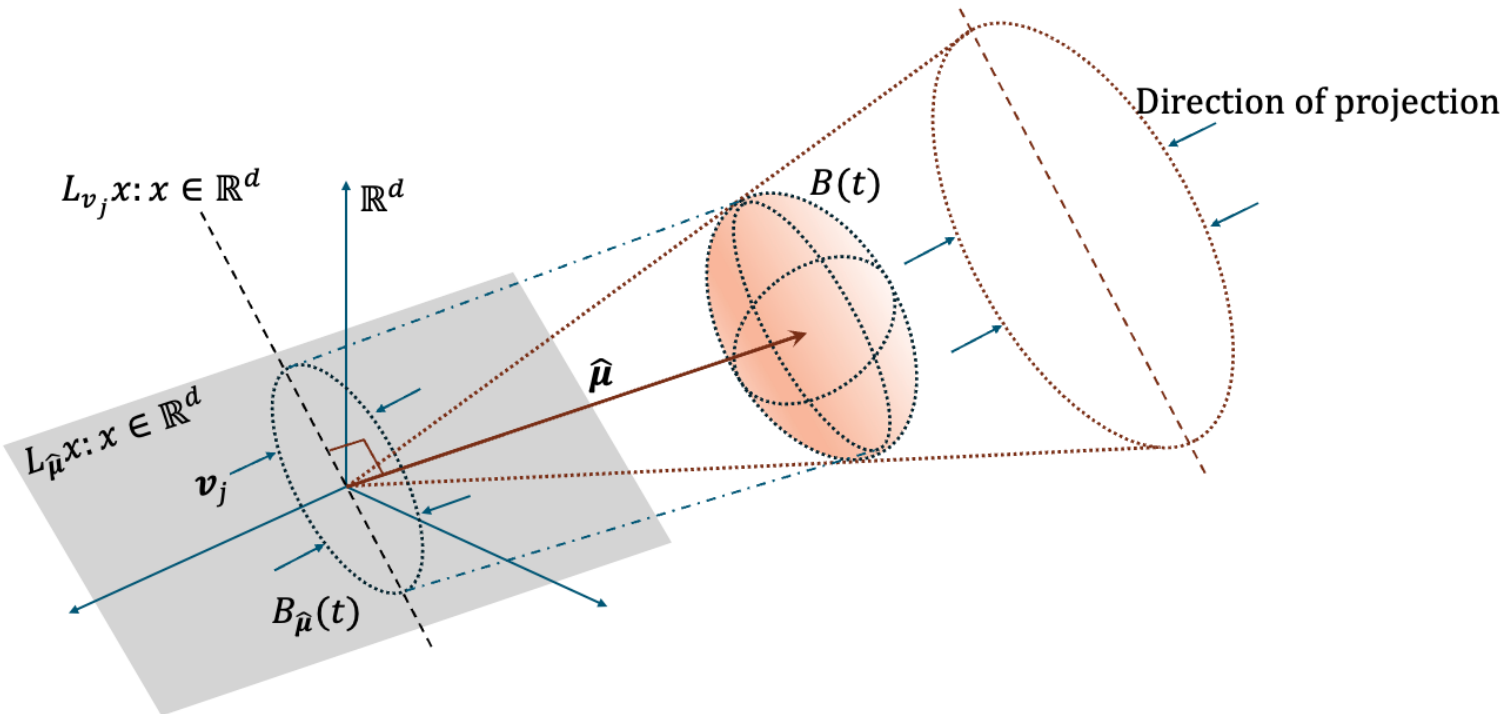
- Posterior estimate and covariance: $\hat{\boldsymbol{\mu}}(t),\ \tilde{\Sigma}(t)$
- Confidence ellipsoid (α):

$$B(t) = \left\{\boldsymbol{\nu} : (\boldsymbol{\nu} - \hat{\boldsymbol{\mu}}(t))^\top \tilde{\Sigma}(t)^{-1} (\boldsymbol{\nu} - \hat{\boldsymbol{\mu}}(t)) \le \alpha^2\right\}$$

- Diagonal $\tilde{\Sigma}(t) \Rightarrow$ axis-aligned ellipsoid; widths reflect exploration effort

Central Planner's Policy

- $B(t)$ not axis-aligned $\Rightarrow$ operate along principal axes
- Project $B(t)$ onto plane orthogonal to $\hat{\boldsymbol{\mu}}(t)$ to obtain $B_{\hat{\mu}}(t)$
- Reduce uncertainty by collapsing along the direction of $B_{\hat{\mu}}(t)$



Future Directions

- Distributed experimentation with information sharing over sparse networks.

Future Directions

- Distributed experimentation with information sharing over sparse networks.
- Applications in drug testing, revenue management, distributed sensing, path planning.

Thank you!