Safe model-based multi-agent mean-field reinforcement learning

Conference Poster

Author(s):
Jusup, Matej; Pásztor, Barna; Janik, Tadeusz; Zhang, Kenan; Corman, Francesco; Krause, Andreas; Bogunovic, Ilija

Publication date:
2024-05

Permanent link:
https://doi.org/10.3929/ethz-b-000673391

Funding acknowledgement:
181210 - DADA - Dynamic data driven Approaches for stochastic Delay propagation Avoidance in railways (SNF)

Safe Model-Based Multi-Agent Mean-Field Reinforcement Learning

M. Jusup[1], B. Pasztor[1], T. Janik[1], K. Zhang[2], F. Corman[1], A. Krause[1], I. Bogunovic[3]
[1]ETH Zurich; [2]EPFL Lausanne; [3]UCL London

AAMAS 2024 AUCKLAND NZ

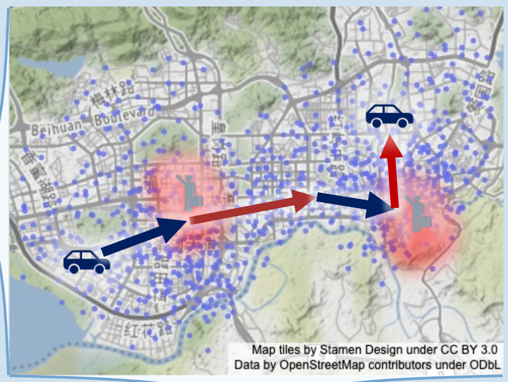
Motivation: Vehicle Repositioning

Repositioning to the high-demand areas

$ $ $

- Fair service accessibility
- Limit on number of vehicles per district

Safety constraints!



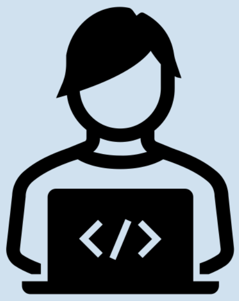
- True mean-field distribution: $\mu_{n,t}$
- True safety constraints: $h_C(\mu_{n,t}) \geq 0$

Safe Model-Based Mean-Field RL

RA collects trajectories $(z_{n,t}, s_{n,t+1})$

$$\boldsymbol{h_C(\tilde{\mu}_{n,t}) \geq 0 \Longrightarrow h_C(\mu_{n,t}) \geq 0?}$$

Controller sends updated policy $\boldsymbol{\pi}_n^*$ to RA



- Statistical model: $\tilde{f}_{n-1}(z) \approx m_{n-1}(z) + \Sigma_{n-1}(z)$
- Statistical mean-field distribution: $\tilde{\mu}_{n,t}$
- Statistical safety constraints: $h_C(\tilde{\mu}_{n,t}) \geq 0$
- Policy: $\boldsymbol{\pi}_n$

Why Safe Mean-Field Reinforcement Learning?

Objective

Learning the optimal *safe policy* $\boldsymbol{\pi}^*$ under *unknown transitions* f given *safety constraints* $h_C(\mu_{n,t}) \geq 0$

Individual interactions lead to the combinatorial state-action space

Mean-field distribution μ of cooperative identical agents

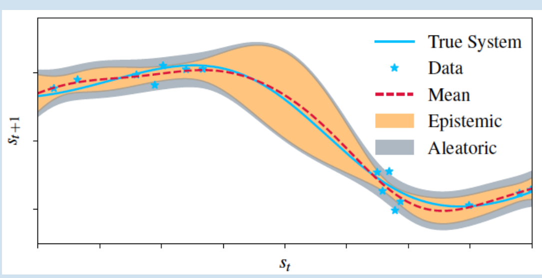
The representative agent (RA) interacts with the mean-field distribution μ

RA policy $\boldsymbol{\pi}^*$ is used to control all the agents

No individual interactions

Complex inputs $z = (s, \mu, a)$ and probabilistic transitions $U(\cdot)$

Calibrated Statistical Model of Unknown Transitions



Mean: $m_{n-1}(z)$
Covariance: $\Sigma_{n-1}(z)$
Confidence: $\sigma_{n-1}^2(z) = diag(\Sigma_{n-1}(z))$
Calibrated model:

$$|f(z) - m_{n-1}(z)| \leq const * \sigma(z)$$

Model-Based Learning Protocol in Safe-$\mathbf{M}^3$-UCRL

Input: Safety constraint $h_C(\cdot)$, initial mean-field distribution μ_0, number of episodes N, number of steps T

```
1: for n = 1, ... N do
2:     Compute C_{n,t} for t = 1, ..., T
3:     Learn a policy π*_n by optimizing the objective
4:     Execute the obtained policy π*_n
5:     Collect the trajectories from the representative agent
6:     Update the statistical model f̃_{n-1}
7: end for
Return π*_N
```

Safe-$\mathbf{M}^3$-UCRL

$$\boldsymbol{\pi}_n^* = \underset{\boldsymbol{\pi}_n \in \Pi}{\arg\max} \max_{\eta(\cdot) \in [-1,1]^p} \mathbb{E}\left[\sum_{t=0}^{T-1} r(\tilde{z}_{n,t}) \middle| \tilde{\mu}_{n,0} = \mu_0\right]$$

$$\text{subject to} \quad \tilde{a}_{n,t} = \pi_{n,t}(\tilde{s}_{n,t}, \tilde{\mu}_{n,t})$$

$$\tilde{f}_{n-1}(\tilde{z}_{n,t}) = \boldsymbol{m}_{n-1}(\tilde{z}_{n,t}) + \beta_{n-1}\boldsymbol{\Sigma}_{n-1}(\tilde{z}_{n,t})\eta(\tilde{z}_{n,t})$$

$$\tilde{s}_{n,t+1} = \tilde{f}_{n-1}(\tilde{z}_{n,t}) + \varepsilon_{n,t}$$

$$\tilde{\mu}_{n,t+1} = U(\tilde{\mu}_{n,t}, \pi_{n,t}, \tilde{f}_{n-1})$$

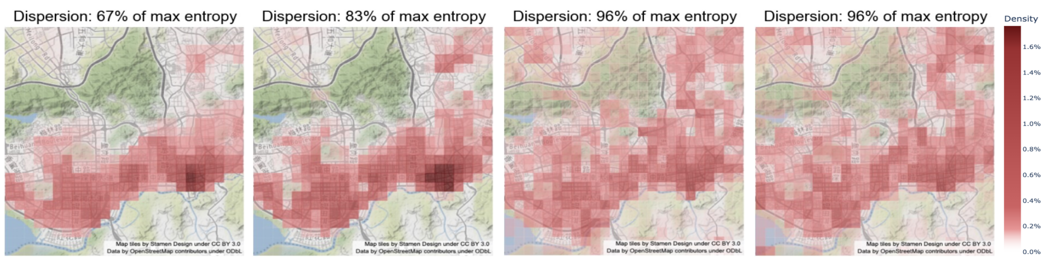
$$\boxed{h_C(\tilde{\mu}_{n,t+1}) \geq L_h C_{n,t+1}}$$

Enables safe exploration!

Contributions of Safe-$\mathbf{M}^3$-UCRL

- Safe exploration guided by budget $C_{n,t}$ induced by epistemic uncertainty $\sigma_{n-1}(z)$
- Relationship between safety constraints under statistical and true environments
  $$|h_C(\tilde{\mu}_{n,t}) - h_C(\mu_{n,t})| \leq L_h C_{n,t}$$
- Algorithm that adheres to the safety constraints throughout the entire execution (with high probability)
- **Showcasing usefulness of Mean-Field RL in real-world applications!**

Experimental Results



Observed demand ρ_0 used as a target distribution | Unconstrained M^3-UCRL under known dynamics | Safe-M^3-UCRL under known dynamics for $p = 0.85$ | Safe-M^3-UCRL under unknown dynamics for $p = 0.85$