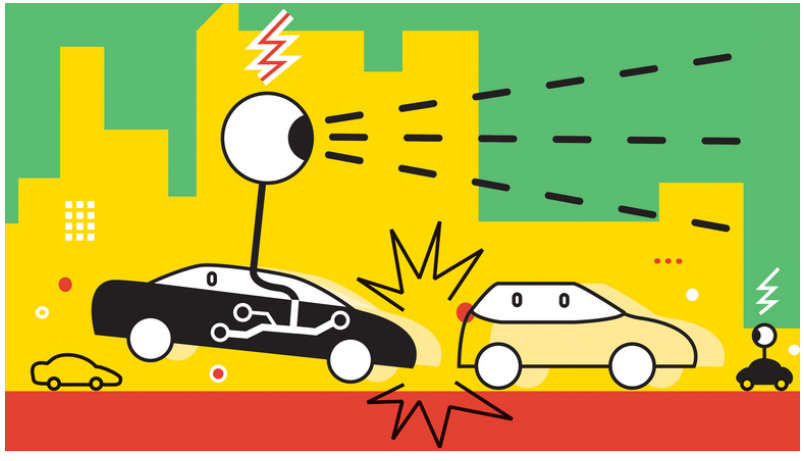
Settling the Variance of Multi-Agent Policy Gradients

Jakub Grudzien Kuba*, Muning Wen*, Yaodong Yang,
Linghui Meng, Shangding Gu, Haifeng Zhang,
David Henry Mguni, Jun Wang

Background: MARL

Multi-Agent Reinforcement Learning
—to make everyone happy

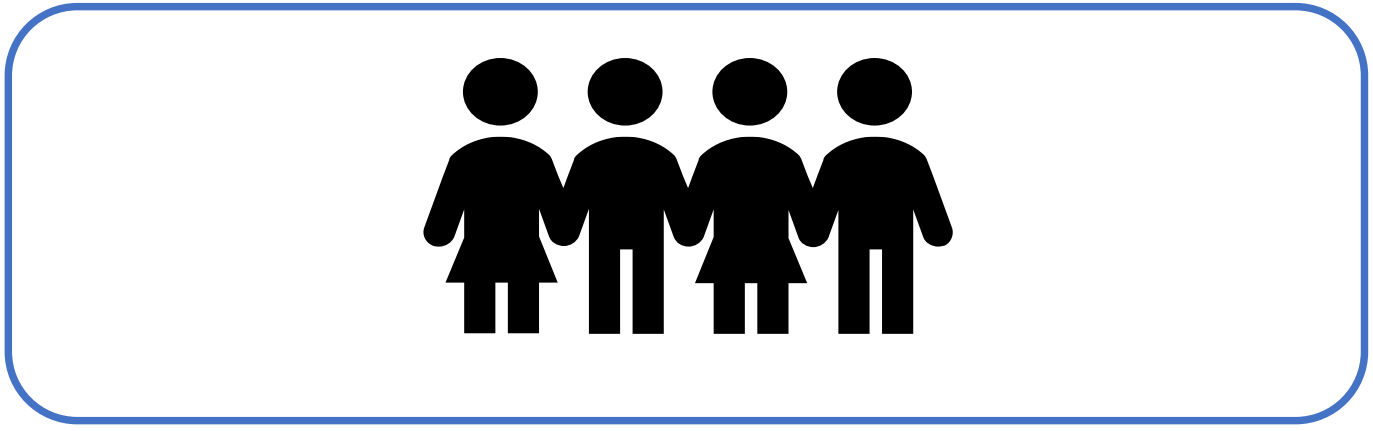


(a) A self-driving with single-agent RL

(b) A system of MARL cars

Problem Formulation

At time step t, n agents are at state s_t



state s_t

Problem Formulation

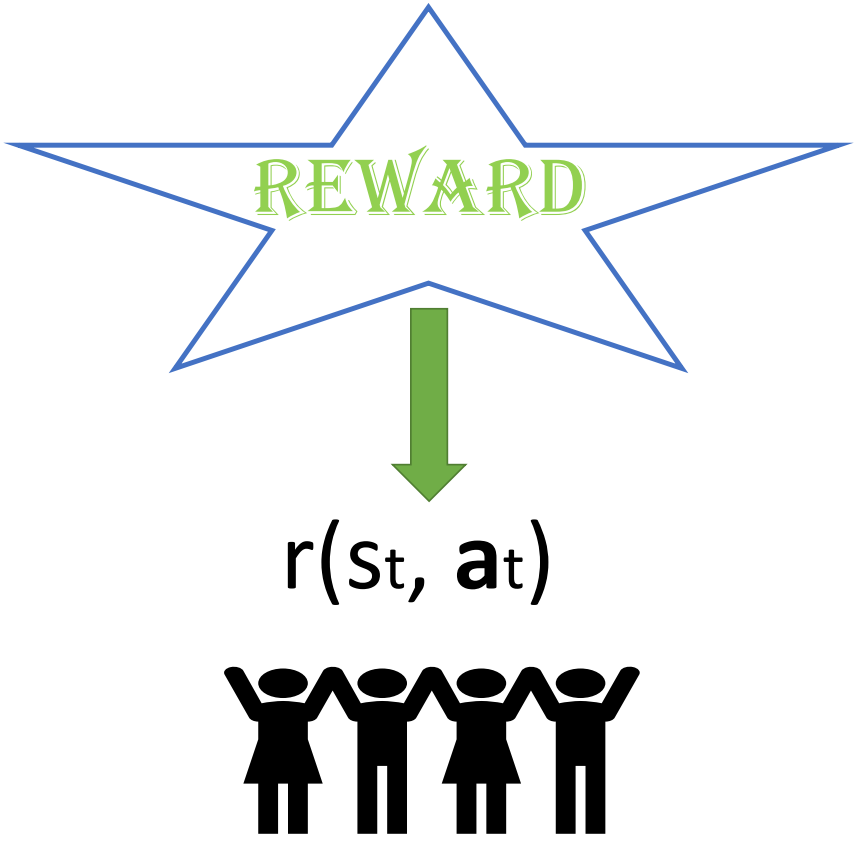
The agents take actions $a_t^1 \sim \pi_{\theta}^1(\cdot^1|s_t), \ldots, a_t^n \sim \pi_{\theta}^n(\cdot^n|s_t)$

state s_t

Equivalently $a_t \sim \pi_{\theta}(\cdot|s_t)$

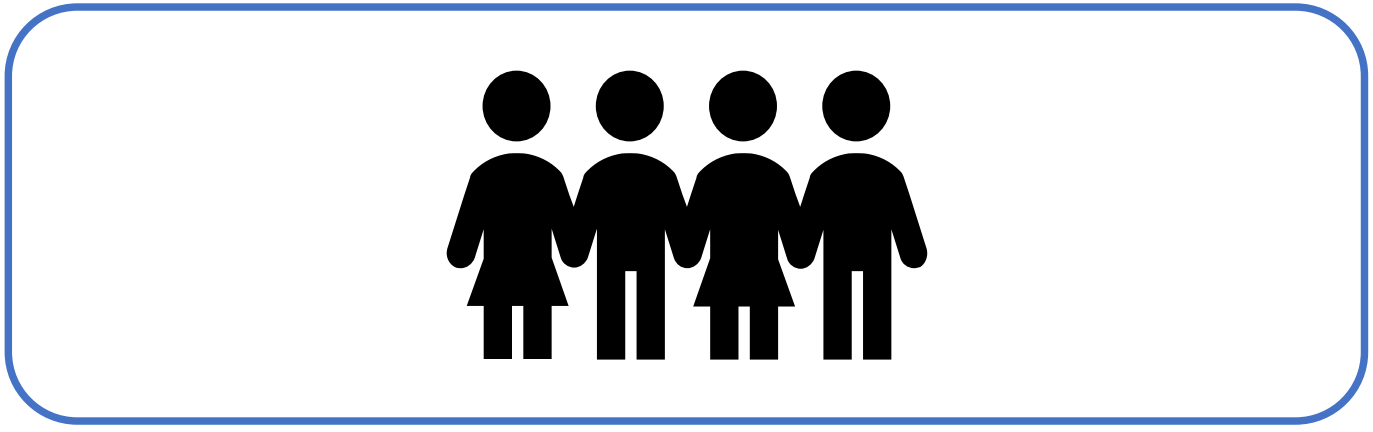
Problem Formulation

The environment emits the joint reward $r(s_t, a_t)$



Problem Formulation

The agents move to the next state $s_{t+1} \sim P(\cdot|s_t, a_t)$



state s_{t+1}

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Problem Formulation

The agents want to maximise the joint return

$$\mathcal{J}(\boldsymbol{\theta}) = \mathbb{E}_{\mathsf{s}_0 \sim d^0, a_{0:\infty} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}, \mathsf{s}_{1:\infty} \sim P} \Big[\sum_{t=0}^{\infty} \gamma^t r(\mathsf{s}_t, \mathsf{a}_t) \Big]$$

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Multi-Agent Policy Gradient (MAPG) Estimators

Decentralised Training (DT)

$$g_{\mathrm{D}}^{i} = \sum_{t=0}^{\infty} \gamma^{t} \hat{Q}^{i}(s_t, a_t^{i}) \nabla_{\theta^{i}} \log \pi_{\boldsymbol{\theta}}^{i}(a_t^{i} | s_t)$$

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Multi-Agent Policy Gradient (MAPG) Estimators

Decentralised Training (DT)

$$g_{\mathrm{D}}^{i} = \sum_{t=0}^{\infty} \gamma^{t} \hat{Q}^{i}(s_t, a_t^{i}) \nabla_{\theta^{i}} \log \pi_{\boldsymbol{\theta}}^{i}(a_t^{i}|s_t)$$

Centralised Training Decentralised Execution (CTDE)

$$g_{\mathrm{C}}^{i} = \sum_{t=0}^{\infty} \gamma^{t} \hat{Q}(s_t, a_t^{-i}, a_t^{i}) \nabla_{\theta^{i}} \log \pi_{\boldsymbol{\theta}}^{i}(a_t^{i}|s_t)$$

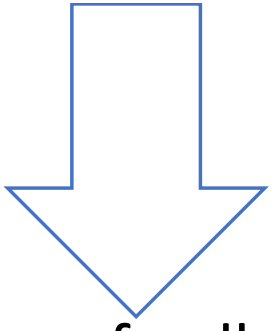
Multi-Agent Policy Gradient (MAPG) Estimators

Decentralised Training (DT)

$$g_D^i = \sum_{t=0}^{\infty} \gamma^t \hat{Q}^i(s_t, a_t^i) \nabla_{\theta^i} \log \pi_{\boldsymbol{\theta}}^i(a_t^i | s_t)$$

Centralised Training Decentralised Execution (CTDE)

$$g_C^i = \sum_{t=0}^{\infty} \gamma^t \hat{Q}(s_t, a_t^{-i}, a_t^i) \nabla_{\theta^i} \log \pi_{\boldsymbol{\theta}}^i(a_t^i | s_t)$$



same feedback

Imperial College London

HUAWEI

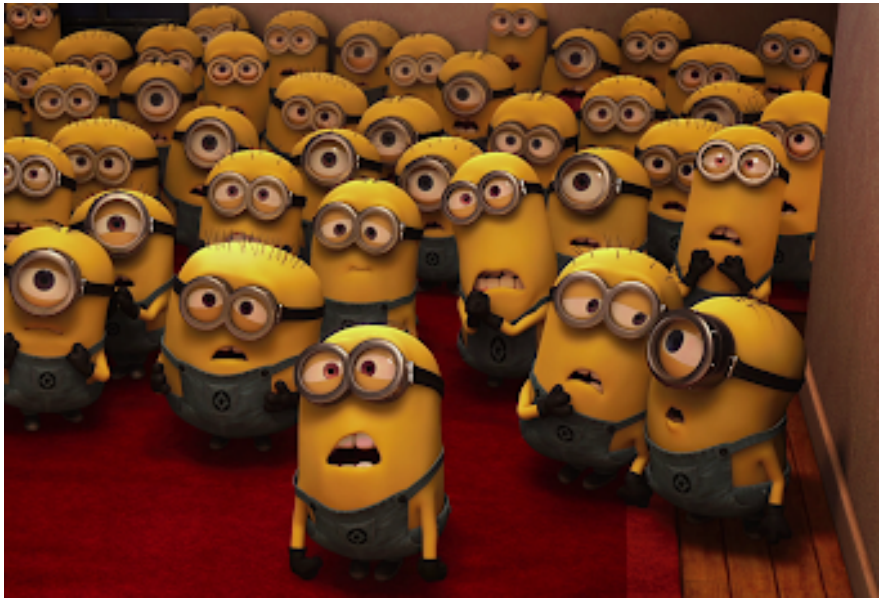
KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

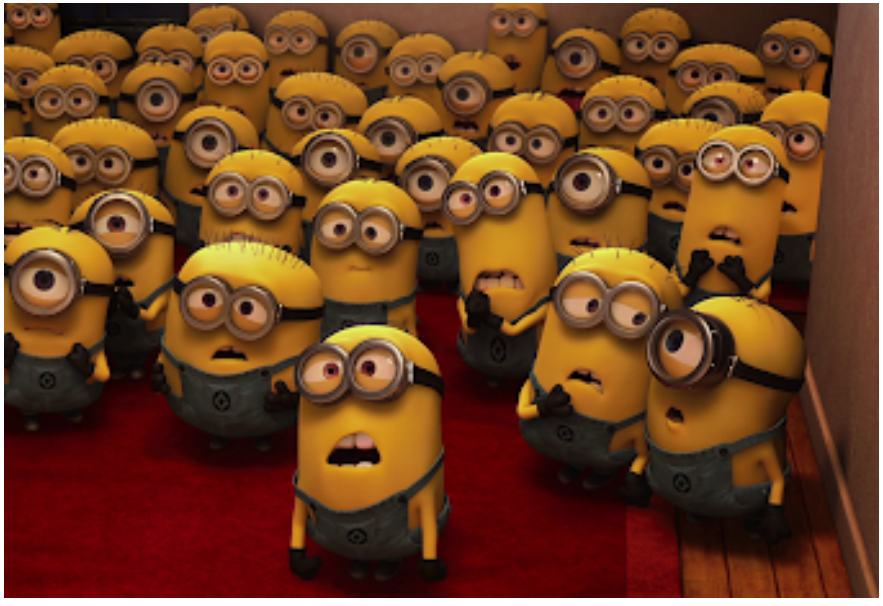
Credit Assignment Problem

$$\hat{Q}(s_t, a_t^{-i}, a_t^i)$$



Credit Assignment Problem

$$\hat{Q}(\mathrm{s}_t, \mathrm{a}_t^{-i}, \mathrm{a}_t^{i})$$



If the bias is not the problem, then what is it?

Credit Assignment Problem

$$\hat{Q}(\mathrm{s}_t, \mathrm{a}_t^{-i}, \mathrm{a}_t^{i})$$

If the bias is not the problem, then what is it?

Perhaps variance

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Settling the Variance of Multi-Agent Policy Gradients

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Settling the Variance of Multi-Agent Policy Gradients

begins with understanding how a subet of agents contributes to the return (the agents affect MAPG estimators through actions).

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Settling the Variance of Multi-Agent Policy Gradients

begins with understanding how a subet of agents contributes to the return (the agents affect MAPG estimators through actions).

Multi-agent state-action value function

$$Q_{\boldsymbol{\theta}}^{i_{1:k}}\left(s, \boldsymbol{a}^{i_{1:k}}\right) = \mathbb{E}_{\mathrm{a}^{-i_{1:k}} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i_{1:k}}}\left[Q_{\boldsymbol{\theta}}\left(s, \boldsymbol{a}^{i_{1:k}}, \mathrm{a}^{-i_{1:k}}\right)\right]$$

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Settling the Variance of Multi-Agent Policy Gradients

begins with understanding how a subet of agents contributes to the return (the agents affect MAPG estimators through actions).

Multi-agent state-action value function

$$Q_{\boldsymbol{\theta}}^{i_{1:k}}\left(s, \boldsymbol{a}^{i_{1:k}}\right) = \mathbb{E}_{\mathrm{a}^{-i_{1:k}} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i_{1:k}}}\left[Q_{\boldsymbol{\theta}}\left(s, \boldsymbol{a}^{i_{1:k}}, \mathrm{a}^{-i_{1:k}}\right)\right]$$

Multi-agent advantage function

$$A_{\boldsymbol{\theta}}^{i_{1:k}}\left(s, \boldsymbol{a}^{j_{1:m}}, \boldsymbol{a}^{i_{1:k}}\right) = Q_{\boldsymbol{\theta}}^{j_{1:m}, i_{1:k}}\left(s, \boldsymbol{a}^{j_{1:m}}, \boldsymbol{a}^{i_{1:k}}\right) - Q_{\boldsymbol{\theta}}^{j_{1:m}}\left(s, \boldsymbol{a}^{j_{1:m}}\right)$$

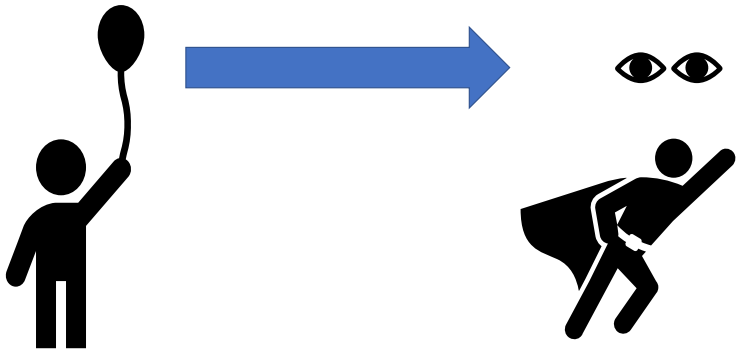
Multi-Agent Advantage Decomposition

$$A_{\theta}^{i_{1:m}}(s, \boldsymbol{a}^{i_{1:m}}) = \sum_{j=1}^{m} A_{\theta}^{i_j}(s, \boldsymbol{a}^{i_{1:j-1}}, a^{i_j})$$

$A^{1,2}(s,$  $)$

Multi-Agent Advantage Decomposition

$$A_{\theta}^{i_{1:m}}(s, \boldsymbol{a}^{i_{1:m}}) = \sum_{j=1}^{m} A_{\theta}^{i_j}(s, \boldsymbol{a}^{i_{1:j-1}}, a^{i_j})$$



$$A^{1,2}(s, \text{  }, \text{  })$$

$$= A^{1}(s, \text{  }) + A^{2}(s, \text{  }, \text{  })$$

MAPG Variance

$$\mathrm{Var}_{\mathrm{s}_{0:\infty}\sim d_{\boldsymbol{\theta}}^{0:\infty},\mathrm{a}_{0:\infty}\sim\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\mathrm{g}_{\mathrm{C}}^{i}\right]-\mathrm{Var}_{\mathrm{s}_{0:\infty}\sim d_{\boldsymbol{\theta}}^{0:\infty},\mathrm{a}_{0:\infty}\sim\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\mathrm{g}_{\mathrm{D}}^{i}\right]=\mathcal{O}\Big(\sum_{j\neq i}\epsilon_{j}^{2}\Big)$$

$$\textit{where } \epsilon_{j}=\max_{s,\boldsymbol{a}^{-j},a^{j}}\left|A_{\boldsymbol{\theta}}^{j}(s,\boldsymbol{a}^{-j},a^{j})\right|$$

MAPG Variance

$$\mathrm{Var}_{\mathsf{s}_{0:\infty}\sim d_{\boldsymbol{\theta}}^{0:\infty},\mathsf{a}_{0:\infty}\sim\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\mathsf{g}_{\mathrm{C}}^{i}\right]-\mathrm{Var}_{\mathsf{s}_{0:\infty}\sim d_{\boldsymbol{\theta}}^{0:\infty},\mathsf{a}_{0:\infty}\sim\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\mathsf{g}_{\mathrm{D}}^{i}\right]=\mathcal{O}\Big(\sum_{j\neq i}\epsilon_{j}^{2}\Big)$$

$$\textit{where } \epsilon_{j}=\max_{s,\boldsymbol{a}^{-j},a^{j}}|A_{\boldsymbol{\theta}}^{j}(s,\boldsymbol{a}^{-j},a^{j})|$$

- The more agents, the worse.

MAPG Variance

$$\mathrm{Var}_{\mathrm{s}_{0:\infty}\sim d_{\boldsymbol{\theta}}^{0:\infty},\mathrm{a}_{0:\infty}\sim\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\mathrm{g}_{\mathrm{C}}^{i}\right]-\mathrm{Var}_{\mathrm{s}_{0:\infty}\sim d_{\boldsymbol{\theta}}^{0:\infty},\mathrm{a}_{0:\infty}\sim\boldsymbol{\pi}_{\boldsymbol{\theta}}}\left[\mathrm{g}_{\mathrm{D}}^{i}\right]=\mathcal{O}\Big(\sum_{j\neq i}\epsilon_{j}^{2}\Big)$$

$$\textit{where } \epsilon_j = \max_{s,\boldsymbol{a}^{-j},a^j} |A_{\boldsymbol{\theta}}^{j}(s,\boldsymbol{a}^{-j},a^{j})|$$

- The more agents, the worse.
- The more they explore, the worse.

MAPG Variance Reduction

Modify the estimator with the baseline trick

$$
\begin{aligned}
\mathrm{g}_{\mathrm{C}}^{i}(b) &= \sum_{t=0}^{\infty} \gamma^{t} \left[\hat{Q}(\mathrm{s}_t, \mathrm{a}_t^{-i}, \mathrm{a}_t^{i}) - b(\mathrm{s}_t, \mathrm{a}_t^{-i}) \right] \nabla_{\theta^i} \log \pi_{\boldsymbol{\theta}}^{i}(\mathrm{a}_t^{i} | \mathrm{s}_t) \\
&= \sum_{t=0}^{\infty} \gamma^{t} \mathrm{g}_{\mathrm{C},t}^{i}(b)
\end{aligned}
$$

MAPG Variance Reduction

Modify the estimator with the baseline trick

$$
\begin{aligned}
\mathrm{g}^i_{\mathrm{C}}(b) &= \sum_{t=0}^{\infty} \gamma^t \left[\hat{Q}(\mathrm{s}_t, \mathrm{a}_t^{-i}, \mathrm{a}_t^i) - b(\mathrm{s}_t, \mathrm{a}_t^{-i}) \right] \nabla_{\theta^i} \log \pi^i_\theta(\mathrm{a}_t^i | \mathrm{s}_t) \\
&= \sum_{t=0}^{\infty} \gamma^t \mathrm{g}^i_{\mathrm{C},t}(b)
\end{aligned}
$$

which brings no bias

$$
\mathbb{E}_{\mathrm{a}_t^{-i} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i}, \mathrm{a}_t^i \sim \pi^i_{\boldsymbol{\theta}}} \left[\mathrm{g}^i_{\mathrm{C},t}(b) \right] = \mathbb{E}_{\mathrm{a}_t^{-i} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i}, \mathrm{a}_t^i \sim \pi^i_{\boldsymbol{\theta}}} \left[\mathrm{g}^i_{\mathrm{C},t}(0) \right]
$$

MAPG Variance Reduction

Modify the estimator with the baseline trick

$$
\begin{aligned}
\mathrm{g}_{\mathrm{C}}^{i}(b) &= \sum_{t=0}^{\infty} \gamma^{t} \left[\hat{Q}(\mathrm{s}_t, \mathrm{a}_t^{-i}, \mathrm{a}_t^{i}) - b(\mathrm{s}_t, \mathrm{a}_t^{-i}) \right] \nabla_{\theta^i} \log \pi_{\boldsymbol{\theta}}^{i}(\mathrm{a}_t^{i} | \mathrm{s}_t) \\
&= \sum_{t=0}^{\infty} \gamma^{t} \mathrm{g}_{\mathrm{C},t}^{i}(b)
\end{aligned}
$$

which brings no bias

$$
\mathbb{E}_{\mathrm{a}_t^{-i} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i}, \mathrm{a}_t^{i} \sim \pi_{\boldsymbol{\theta}}^{i}} \left[\mathrm{g}_{\mathrm{C},t}^{i}(b) \right] = \mathbb{E}_{\mathrm{a}_t^{-i} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i}, \mathrm{a}_t^{i} \sim \pi_{\boldsymbol{\theta}}^{i}} \left[\mathrm{g}_{\mathrm{C},t}^{i} \right]
$$

MAPG Varianec Reduction: Decomposition

$$\mathrm{Var}_{s_t \sim d_{\boldsymbol{\theta}}^t, a_t \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[g_{C,t}^i(b) \right] = \underbrace{\mathrm{Var}_{s_t \sim d_{\boldsymbol{\theta}}^t} \left[\mathbb{E}_{a_t \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[g_{C,t}^i(b) \right] \right]}_{\text{Variance from state}}$$

$$+ \underbrace{\mathbb{E}_{s_t \sim d_{\boldsymbol{\theta}}^t} \left[\mathrm{Var}_{a_t^{-i} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i}} \left[\mathbb{E}_{a_t^i \sim \pi_{\boldsymbol{\theta}}^i} [g_{C,t}^i(b)] \right] \right]}_{\text{Variance from other agents' actions}}$$

$$+ \underbrace{\mathbb{E}_{s_t \sim d_{\boldsymbol{\theta}}^t, a_t^{-i} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i}} \left[\mathrm{Var}_{a_t^i \sim \pi_{\boldsymbol{\theta}}^i} \left[g_{C,t}^i(b) \right] \right]}_{\text{Variance from agent } i\text{'s action}}$$

MAPG Varianec Reduction: Decomposition

$$\mathrm{Var}_{s_t \sim d_{\boldsymbol{\theta}}^t, a_t \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[g_{C,t}^i(b) \right] = \underbrace{\mathrm{Var}_{s_t \sim d_{\boldsymbol{\theta}}^t} \left[\mathbb{E}_{a_t \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}} \left[g_{C,t}^i \right] \right]}_{\text{Variance from state}}$$

$$+ \underbrace{\mathbb{E}_{s_t \sim d_{\boldsymbol{\theta}}^t} \left[\mathrm{Var}_{a_t^{-i} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i}} \left[\mathbb{E}_{a_t^i \sim \pi_{\boldsymbol{\theta}}^i} \left[g_{C,t}^i \right] \right] \right]}_{\text{Variance from other agents' actions}}$$

$$+ \underbrace{\mathbb{E}_{s_t \sim d_{\boldsymbol{\theta}}^t, a_t^{-i} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}}^{-i}} \left[\mathrm{Var}_{a_t^i \sim \pi_{\boldsymbol{\theta}}^i} \left[g_{C,t}^i(b) \right] \right]}_{\text{Variance from agent } i\text{'s action}}$$

MAPG Varianec Reduction: Objective

$$\min_{b(s,\boldsymbol{a}^{-i})} \mathrm{Var}_{a^i \sim \pi^i_\theta} \left[\left(\hat{Q}(s, \boldsymbol{a}^{-i}, a^i) - b(s, \boldsymbol{a}^{-i}) \right) \nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s) \right]$$

MAPG Varianec Reduction: Optimal Baseline

$$\min_{b(s,\boldsymbol{a}^{-i})} \text{Var}_{a^i \sim \pi^i_\theta} \left[\left(\hat{Q}(s, \boldsymbol{a}^{-i}, a^i) - b(s, \boldsymbol{a}^{-i}) \right) \nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s) \right]$$

The optimal baseline is given by

$$b^{\text{optimal}}(s, \boldsymbol{a}^{-i}) = \frac{\mathbb{E}_{a^i \sim \pi^i_\theta} \left[\hat{Q}(s, \boldsymbol{a}^{-i}, a^i) || \nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s) ||^2 \right]}{\mathbb{E}_{a^i \sim \pi^i_{\boldsymbol{\theta}}} \left[|| \nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s) ||^2 \right]}$$

MAPG Varianec Reduction: Optimal Baseline

$$\min_{b(s,\boldsymbol{a}^{-i})} \text{Var}_{a^i \sim \pi^i_\theta} \left[\left(\hat{Q}(s, \boldsymbol{a}^{-i}, a^i) - b(s, \boldsymbol{a}^{-i}) \right) \nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s) \right]$$

The optimal baseline is given by

$$b^{\text{optimal}}(s, \boldsymbol{a}^{-i}) = \frac{\mathbb{E}_{a^i \sim \pi^i_\theta} \left[\hat{Q}(s, \boldsymbol{a}^{-i}, a^i) ||\nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s)||^2 \right]}{\mathbb{E}_{a^i \sim \pi^i_{\boldsymbol{\theta}}} \left[||\nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s)||^2 \right]}$$

MAPG Varianec Reduction: Optimal Baseline

$$\min_{b(s,\boldsymbol{a}^{-i})} \mathrm{Var}_{a^i\sim\pi^i_\theta}\left[\left(\hat{Q}(s,\boldsymbol{a}^{-i},a^i)-b(s,\boldsymbol{a}^{-i})\right)\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(a^i|s)\right]$$

The optimal baseline is given by

$$b^{\text{optimal}}(s,\boldsymbol{a}^{-i})=\frac{\mathbb{E}_{a^i\sim\pi^i_\theta}\left[\hat{Q}(s,\boldsymbol{a}^{-i},a^i)||\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(a^i|s)||^2\right]}{\mathbb{E}_{a^i\sim\pi^i_\theta}\left[||\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(a^i|s)||^2\right]}$$

- No closed-form formula

MAPG Varianec Reduction: Optimal Baseline

$$\min_{b(s,\boldsymbol{a}^{-i})} \mathrm{Var}_{a^i\sim\pi^i_\theta}\left[\left(\hat{Q}(s,\boldsymbol{a}^{-i},a^i)-b(s,\boldsymbol{a}^{-i})\right)\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(a^i|s)\right]$$

The optimal baseline is given by

$$b^{\text{optimal}}(s,\boldsymbol{a}^{-i})=\frac{\mathbb{E}_{a^i\sim\pi^i_\theta}\left[\hat{Q}(s,\boldsymbol{a}^{-i},a^i)||\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(a^i|s)||^2\right]}{\mathbb{E}_{a^i\sim\pi^i_\theta}\left[||\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(a^i|s)||^2\right]}$$

- No closed-form formula
- The large dimension introduces extra variance in estimation

MAPG Varianec Reduction: Optimal Baseline

$$\min_{b(s,\boldsymbol{a}^{-i})} \mathrm{Var}_{a^i\sim\pi^i_\theta}\left[\left(\hat{Q}(s,\boldsymbol{a}^{-i},a^i)-b(s,\boldsymbol{a}^{-i})\right)\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(a^i|s)\right]$$

The optimal baseline is given by

$$b^{\text{optimal}}(s,\boldsymbol{a}^{-i})=\frac{\mathbb{E}_{a^i\sim\pi^i_\theta}\left[\hat{Q}(s,\boldsymbol{a}^{-i},a^i)||\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(a^i|s)||^2\right]}{\mathbb{E}_{a^i\sim\pi^i_\theta}\left[||\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(a^i|s)||^2\right]}$$

- No closed-form formula
- The large dimension introduces extra variance in estimation
- Multiple backpropagations are computationally expensive

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Neural Network Policies

$$\pi_\theta^i(a^i|s) = \pi^i(a^i|\psi_\theta^i(s))$$

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Neural Network Policies

Decomposition with the last layer

$$\pi^i_{\boldsymbol{\theta}}(a^i|s) = \pi^i(a^i|\psi^i_{\boldsymbol{\theta}}(s))$$

For example, $\psi^i_{\boldsymbol{\theta}}$ can be the layer of logits

$$\pi^i_{\boldsymbol{\theta}}(a^i|s) = \mathrm{softmax}\left(\psi^i_{\boldsymbol{\theta}}(s)\right)[a^i]$$

Imperial College London

HUAWEI

KING'S College LONDON

UCL

NEURAL INFORMATION PROCESSING SYSTEMS

Neural Network Policies

Decomposition with the last layer

$$\pi^i_{\boldsymbol{\theta}}(a^i|s) = \pi^i(a^i|\psi^i_{\boldsymbol{\theta}}(s))$$

Gradient decomposition with chain rule

$$\nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s) = \nabla_{\theta^i} \psi^i_{\boldsymbol{\theta}}(a^i|s) \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(a^i|\psi^i_{\boldsymbol{\theta}}(s))$$

Imperial College London HUAWEI KING'S College LONDON UCL NEURAL INFORMATION PROCESSING SYSTEMS

Neural Network Policies

Decomposition with the last layer

$$\pi^i_{\boldsymbol{\theta}}(a^i|s) = \pi^i(a^i|\psi^i_{\boldsymbol{\theta}}(s))$$

Gradient decomposition with chain rule

$$\nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s) = \nabla_{\theta^i} \psi^i_{\boldsymbol{\theta}}(a^i|s) \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(a^i|\psi^i_{\boldsymbol{\theta}}(s))$$

Imperial College London HUAWEI KING'S College LONDON UCL NEURAL INFORMATION PROCESSING SYSTEMS

Neural Network Policies

Decomposition with the last layer

$$\pi^i_{\boldsymbol{\theta}}(a^i|s) = \pi^i(a^i|\psi^i_{\boldsymbol{\theta}}(s))$$

Gradient decomposition with chain rule

$$\nabla_{\theta^i} \log \pi^i_{\boldsymbol{\theta}}(a^i|s) = \nabla_{\theta^i} \psi^i_{\boldsymbol{\theta}}(a^i|s) \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(a^i|\psi^i_{\boldsymbol{\theta}}(s))$$

MAPG Varianec Reduction: Surrogate Objective

$$\min_{b(s,\boldsymbol{a}^{-i})} \mathrm{Var}_{\mathrm{a}^i\sim\pi^i_{\boldsymbol{\theta}}}\left[\left(\hat{Q}(s,\boldsymbol{a}^{-i},\mathrm{a}^i)-b(s,\boldsymbol{a}^{-i})\right)\nabla_{\theta^i}\log\pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i|s)\right]$$

$$\propto \min_{b(s,\boldsymbol{a}^{-i})} \mathrm{Var}_{\mathrm{a}^i\sim\pi^i_{\boldsymbol{\theta}}}\left[\left(\hat{Q}(s,\boldsymbol{a}^{-i},\mathrm{a}^i)-b(s,\boldsymbol{a}^{-i})\right)\nabla_{\psi^i_{\boldsymbol{\theta}}}\log\pi^i_{\theta}(\mathrm{a}^i|\psi^i_{\boldsymbol{\theta}}(s))\right]$$

MAPG Varianec Reduction: Optimal Baseline (OB)

The optimal baseline (OB) minimising the surrogate variance is

$$b^*(s, \boldsymbol{a}^{-i}) = \frac{\mathbb{E}_{\mathrm{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[\hat{Q}(s, \boldsymbol{a}^{-i}, \mathrm{a}^i) || \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}{\mathbb{E}_{\mathrm{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[|| \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}$$

MAPG Varianec Reduction: Optimal Baseline (OB)

The optimal baseline (OB) minimising the surrogate variance is

$$b^*(s, \boldsymbol{a}^{-i}) = \frac{\mathbb{E}_{\mathrm{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[\hat{Q}(s, \boldsymbol{a}^{-i}, \mathrm{a}^i) || \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}{\mathbb{E}_{\mathrm{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[|| \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}$$

- it has an explicit formula in the discrete-action case

MAPG Varianec Reduction: Optimal Baseline (OB)

The optimal baseline (OB) minimising the surrogate variance is

$$b^*(s, \boldsymbol{a}^{-i}) = \frac{\mathbb{E}_{\mathrm{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[\hat{Q}(s, \boldsymbol{a}^{-i}, \mathrm{a}^i) || \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}{\mathbb{E}_{\mathrm{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[|| \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}$$

- it has an explicit formula in the discrete-action case
- it is easy to approximate in the continuous-action case

MAPG Varianec Reduction: Optimal Baseline (OB)

The optimal baseline (OB) minimising the surrogate variance is

$$b^*(s, \boldsymbol{a}^{-i}) = \frac{\mathbb{E}_{\mathrm{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[\hat{Q}(s, \boldsymbol{a}^{-i}, \mathrm{a}^i) || \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}{\mathbb{E}_{\mathrm{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[|| \nabla_{\psi^i_{\boldsymbol{\theta}}} \log \pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}$$

- it has an explicit formula in the discrete-action case
- it is easy to approximate in the continuous-action case
- in each case, it is cheap to compute

MAPG Varianec Reduction: Optimal Baseline (OB)

The optimal baseline (OB) minimising the surrogate variance is

$$b^*(s, \boldsymbol{a}^{-i}) = \frac{\mathbb{E}_{\mathsf{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[\hat{Q}(s, \boldsymbol{a}^{-i}, \mathsf{a}^i) || \nabla_{\psi^i} \log \pi^i_{\boldsymbol{\theta}}(\mathsf{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}{\mathbb{E}_{\mathsf{a}^i \sim \pi^i_{\boldsymbol{\theta}}} \left[|| \nabla_{\psi^i} \log \pi^i_{\boldsymbol{\theta}}(\mathsf{a}^i | \psi^i_{\boldsymbol{\theta}}(s)) ||^2 \right]}$$

- it has explicit formula in the discrete-action case
- it is easy to approximate in the continuous-action case
- in each case, it is cheap to compute

To use it in an MAPG estimator

$$\hat{X}^i(s, \mathsf{a}^{-i}, \mathsf{a}^i) \triangleq \hat{Q}(s, \mathsf{a}^{-i}, \mathsf{a}^i) - b^*(s, \mathsf{a}^{-i})$$

Empirical Results

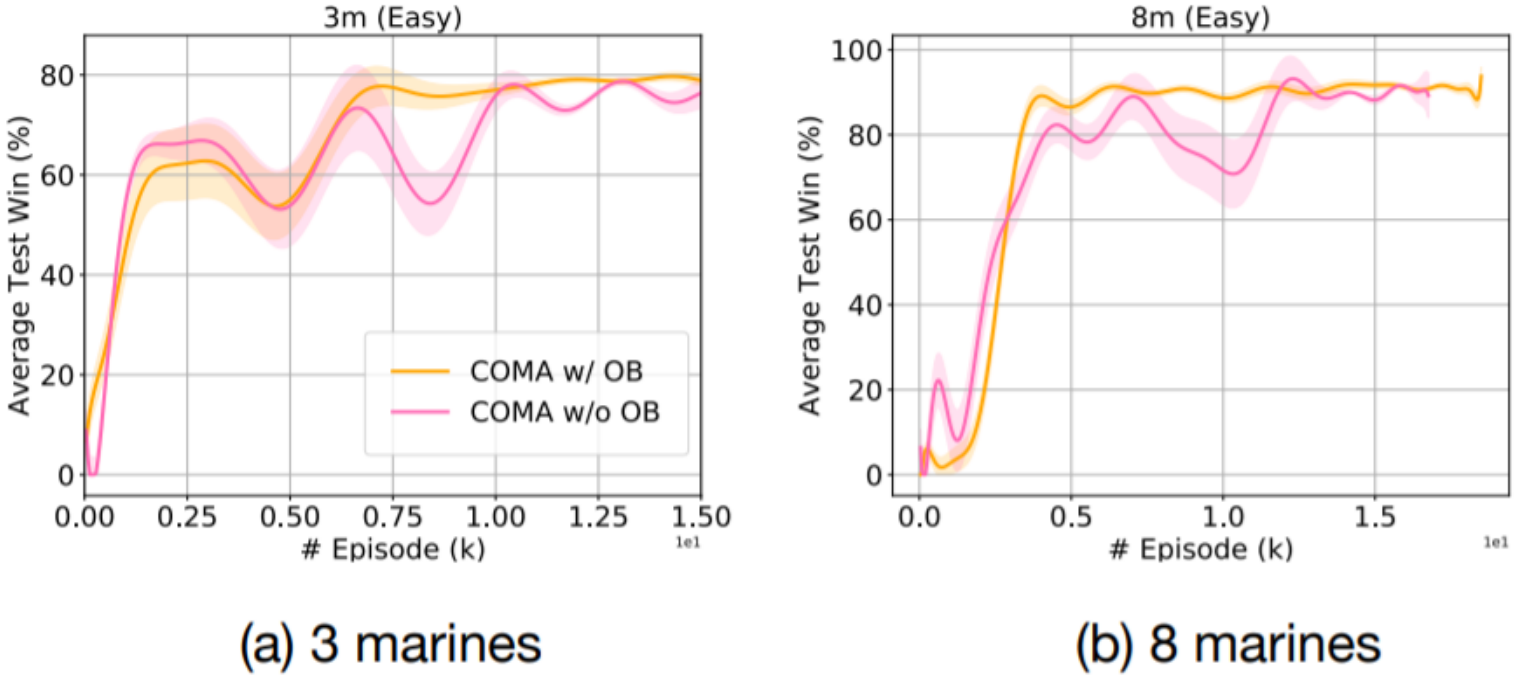
OB significantly reduces the variance of MAPG estimators

a^i	$\psi^i_{\boldsymbol{\theta}}(\mathrm{a}^i)$	$\pi^i_{\boldsymbol{\theta}}(\mathrm{a}^i)$	$x^i_{\psi^i_{\boldsymbol{\theta}}}(\mathrm{a}^i)$	$\hat{Q}(\boldsymbol{a}^{-i}, \mathrm{a}^i)$	$\hat{A}^i(\boldsymbol{a}^{-i}, \mathrm{a}^i)$	$\hat{X}^i(\boldsymbol{a}^{-i}, \mathrm{a}^i)$	Method	Variance
1	$\log 8$	0.8	0.14	2	-9.7	-41.71	MAPG	**1321**
2	0	0.1	0.43	1	-10.7	-42.71	COMA	**1015**
3	0	0.1	0.43	100	88.3	56.29	OB	**673**

Figure 2: Toy Example

Empirical Results

Lower variance leads to more stable training (example of COMA).

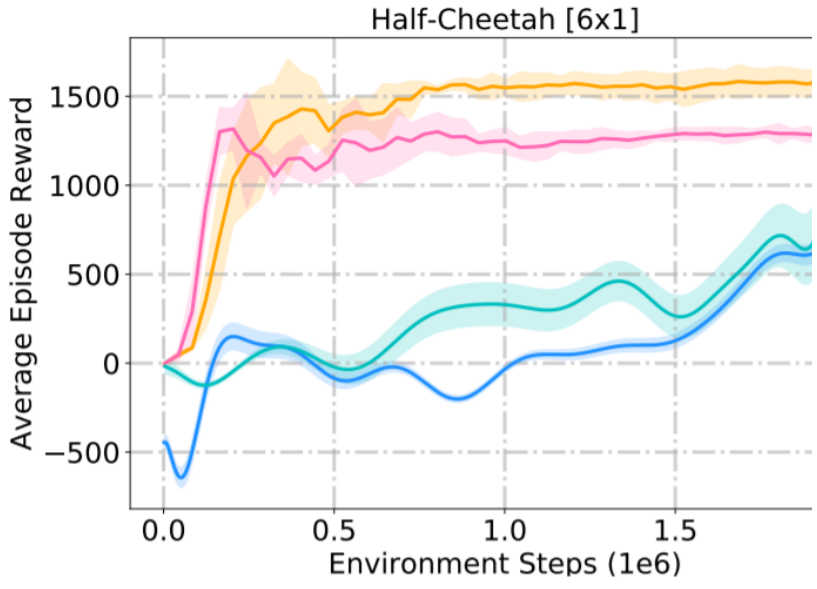


(a) 3 marines (b) 8 marines

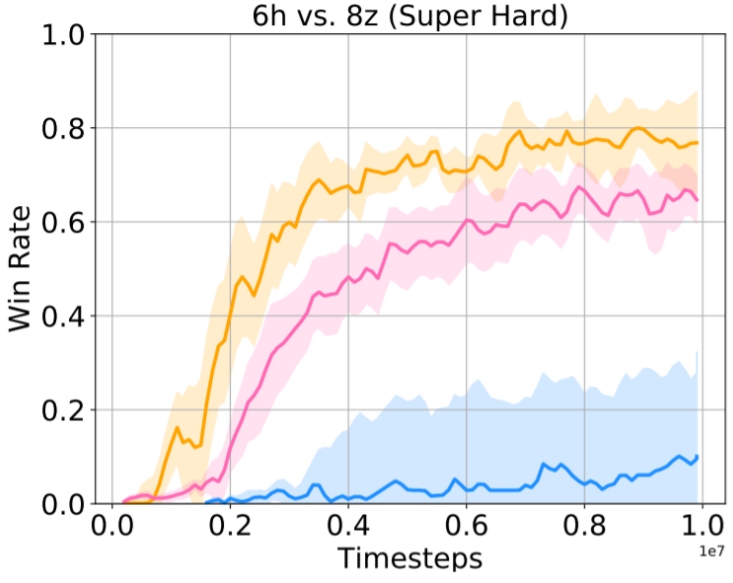
Figure 3: StarCraftII: COMA with OB vs COMA

Empirical Results

Lower variance leads to better performance (example of MAPPO).



(a) MAMuJoCo: MAPPO with OB vs MAPPO, COMIX, MADDPG



(b) StarCraftII: MAPPO with OB vs MAPPO, QMIX

Thank you for your attention!

- Jakub Grudzien Kuba (*Imperial College London, Huawei R&D UK*)
- Muning Wen (*Shanghai Jiao Tong University*)
- Yaodong Yang (*King's College London*)
- Linghui Meng (*Institute of Automation, Chinese Academy of Science*)
- Shangding Gu (*Institute of Automation, Chinese Academy of Science*)
- Haifeng Zhang (*Institute of Automation, Chinese Academy of Science*)
- David Henry Mguni (*Huawei R&D UK*)
- Jun Wang (*University College London*)