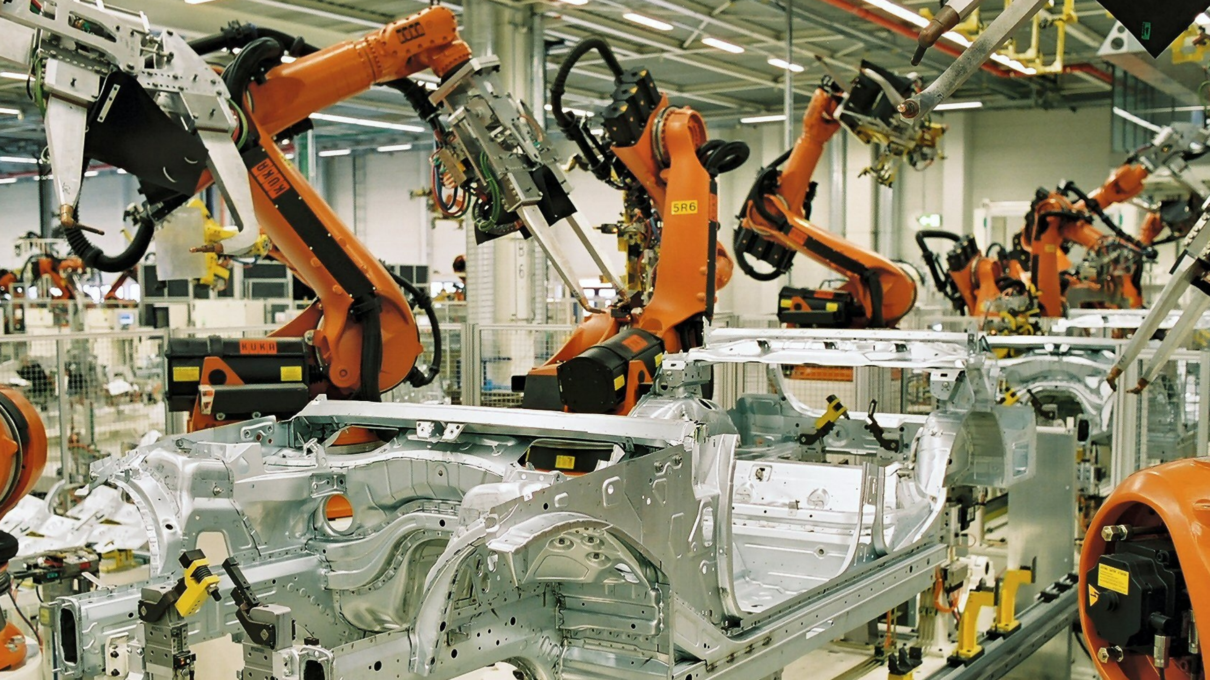
POLITECNICO
MILANO 1863

Balancing Learning Speed and Stability in Policy Gradient via Adaptive Exploration

Matteo Papini Andrea Battistello Marcello Restelli

The 23rd International Conference on Artificial Intelligence and Statistics, June 2020


KUKA
5R6
KUKA

Safe Exploration[1]

Learn safe behavior

[1]Amodei et al., "Concrete Problems in AI Safety", 2016.

Safe Exploration[1]

~~Learn safe behavior~~

[1]Amodei et al., "Concrete Problems in AI Safety", 2016.

Safe Exploration[1]

~~Learn safe behavior~~

Learn safely

[1] Amodei et al., "Concrete Problems in AI Safety", 2016.

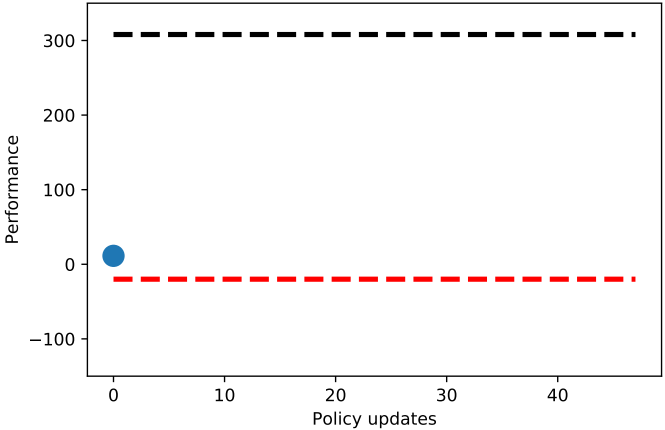
Safe Exploration[1]

~~Learn safe behavior~~

Learn safely $\implies$ Explore safely

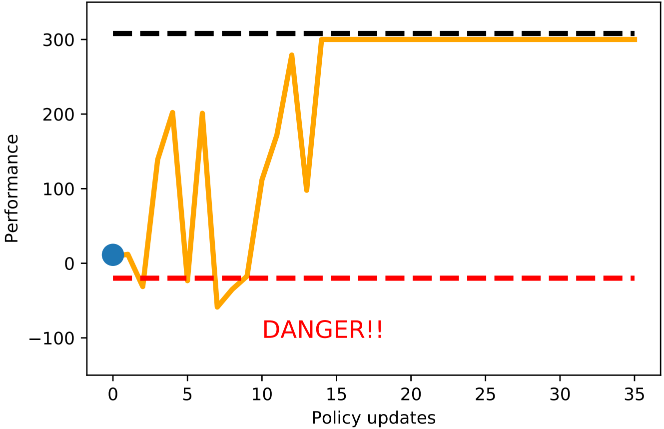
[1] Amodei et al., "Concrete Problems in AI Safety", 2016.

Monotonic Improvement



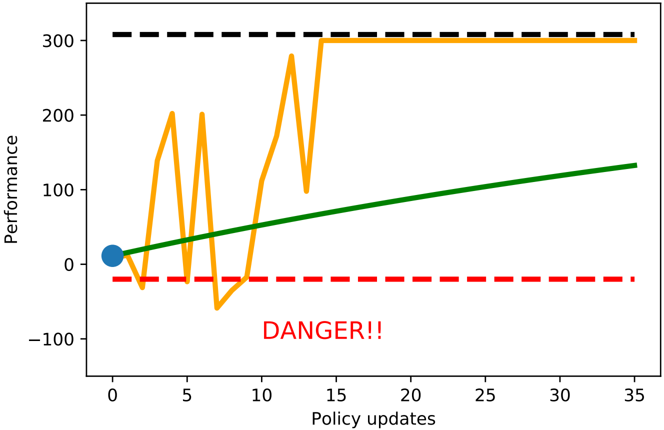
Data from Cart-Pole experiment

Monotonic Improvement



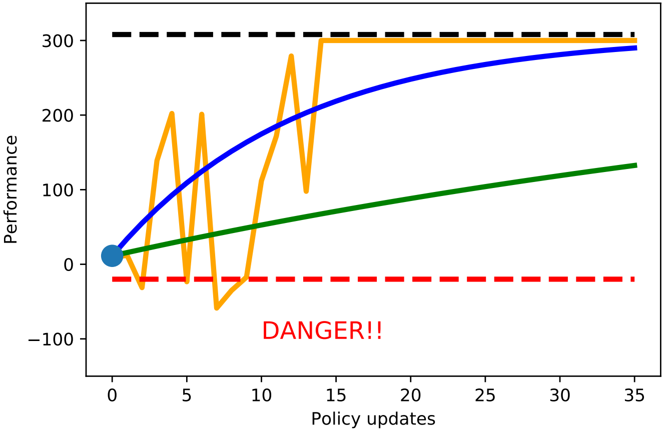
Data from Cart-Pole experiment

Monotonic Improvement



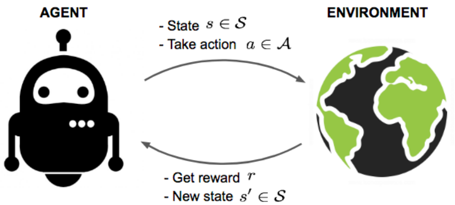
Data from Cart-Pole experiment

Monotonic Improvement



Data from Cart-Pole experiment

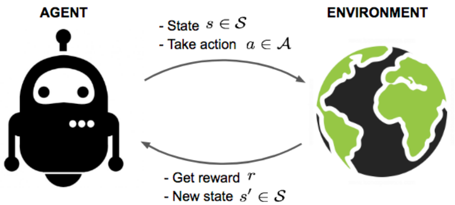
Continuous Reinforcement Learning[2]



- Policy $a \sim \pi(\cdot|s)$
- Goal: $\max_\pi \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi\right]$ (discount factor $\gamma \in (0,1)$)
- Continuous $\mathcal{S}, \mathcal{A} \subseteq \mathbb{R}^n$

[2] Sutton and Barto, *Reinforcement learning: An introduction*, 2018.

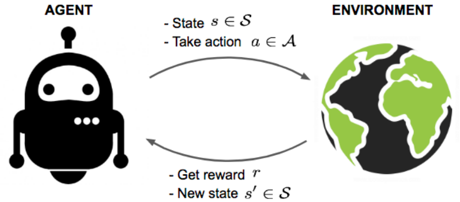
Continuous Reinforcement Learning[2]



- Policy $a \sim \pi(\cdot|s)$
- Goal: $\max_\pi \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi\right]$ (discount factor $\gamma \in (0,1)$)
- Continuous $\mathcal{S}, \mathcal{A} \subseteq \mathbb{R}^n$

[2]Sutton and Barto, *Reinforcement learning: An introduction*, 2018.

Continuous Reinforcement Learning[2]

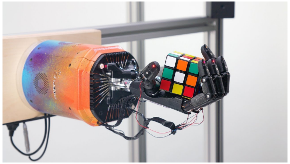


- Policy $a \sim \pi(\cdot|s)$
- Goal: $\max_\pi \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi\right]$ (discount factor $\gamma \in (0,1)$)
- Continuous $\mathcal{S}, \mathcal{A} \subseteq \mathbb{R}^n$

[2] Sutton and Barto, *Reinforcement learning: An introduction*, 2018.

Policy Gradient

- Parametric policy $\pi_{\boldsymbol{\theta}}$ $(\boldsymbol{\theta} \in \mathbb{R}^d)$
- Performance $J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi_{\boldsymbol{\theta}}\right]$
- $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \nabla J(\boldsymbol{\theta})$
- Best for continuous control [3]
  - Convergence guarantees
  - Robustness to noise
  - Freedom in policy design



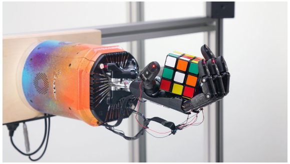
OpenAI 2019

Google/BAIR 2020

[3] Deisenroth, Neumann, and Peters, "A Survey on Policy Search for Robotics", 2013

Policy Gradient

- Parametric policy $\pi_{\boldsymbol{\theta}}$ $\quad (\boldsymbol{\theta} \in \mathbb{R}^d)$
- Performance $J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi_{\boldsymbol{\theta}}\right]$
- $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \nabla J(\boldsymbol{\theta})$
- Best for continuous control [3]
  - Convergence guarantees
  - Robustness to noise
  - Freedom in policy design

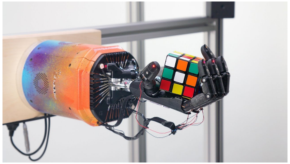


OpenAI 2019

Google/BAIR 2020

[3] Deisenroth, Neumann, and Peters, "A Survey on Policy Search for Robotics", 2013

- Parametric policy $\pi_{\boldsymbol{\theta}}$ $\quad (\boldsymbol{\theta} \in \mathbb{R}^d)$
- Performance $J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi_{\boldsymbol{\theta}}\right]$
- $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \nabla J(\boldsymbol{\theta})$
- Best for continuous control [3]
  - Convergence guarantees
  - Robustness to noise
  - Freedom in policy design



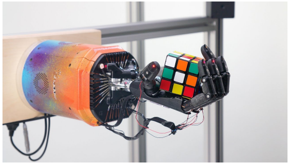
OpenAI 2019

Google/BAIR 2020

[3] Deisenroth, Neumann, and Peters, "A Survey on Policy Search for Robotics", 2013

Policy Gradient

- Parametric policy $\pi_{\boldsymbol{\theta}}$ $(\boldsymbol{\theta} \in \mathbb{R}^d)$
- Performance $J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi_{\boldsymbol{\theta}}\right]$
- $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \nabla J(\boldsymbol{\theta})$
- Best for continuous control [3]
  - Convergence guarantees
  - Robustness to noise
  - Freedom in policy design



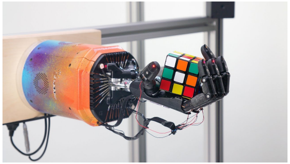
OpenAI 2019

Google/BAIR 2020

[3] ibid., 2013

Policy Gradient

- Parametric policy $\pi_{\boldsymbol{\theta}}$ $\quad (\boldsymbol{\theta} \in \mathbb{R}^d)$
- Performance $J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi_{\boldsymbol{\theta}}\right]$
- $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \nabla J(\boldsymbol{\theta})$
- Best for continuous control [3]
  - Convergence guarantees
  - Robustness to noise
  - Freedom in policy design



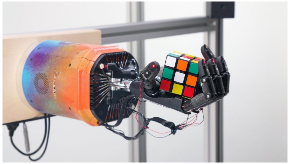
OpenAI 2019

Google/BAIR 2020

[3] ibid., 2013

Policy Gradient

- Parametric policy $\pi_{\boldsymbol{\theta}}$ $\quad (\boldsymbol{\theta} \in \mathbb{R}^d)$
- Performance $J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi_{\boldsymbol{\theta}}\right]$
- $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \nabla J(\boldsymbol{\theta})$
- Best for continuous control [3]
  - Convergence guarantees
  - Robustness to noise
  - Freedom in policy design



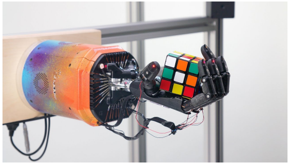
OpenAI 2019

Google/BAIR 2020

[3] ibid., 2013

Policy Gradient

- Parametric policy $\pi_{\boldsymbol{\theta}}$ $\quad (\boldsymbol{\theta} \in \mathbb{R}^d)$
- Performance $J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_t \gamma^t r_{t+1} \mid a_t \sim \pi_{\boldsymbol{\theta}}\right]$
- $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \nabla J(\boldsymbol{\theta})$
- Best for continuous control [3]
    - Convergence guarantees
    - Robustness to noise
    - Freedom in policy design



OpenAI 2019

Google/BAIR 2020

[3] ibid., 2013

Smoothing Policies

- $J(\boldsymbol{\theta})$ is nonconvex

- *Smooth* $J(\boldsymbol{\theta})$ allows *monotonic improvement*

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq 0$$

- Conditions on $\nabla \log \pi_{\boldsymbol{\theta}}, \nabla^2 \log \pi_{\boldsymbol{\theta}} \implies J(\boldsymbol{\theta})$ smooth [4]

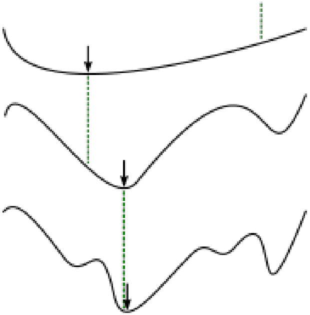
[4] Papini, Pirotta, and Restelli, "Smoothing Policies and Safe Policy Gradients", 2019

Smoothing Policies

- $J(\boldsymbol{\theta})$ is nonconvex
- *Smooth* $J(\boldsymbol{\theta})$ allows *monotonic improvement*

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq 0$$

- Conditions on $\nabla \log \pi_{\boldsymbol{\theta}}, \nabla^2 \log \pi_{\boldsymbol{\theta}} \implies J(\boldsymbol{\theta})$ smooth [4]



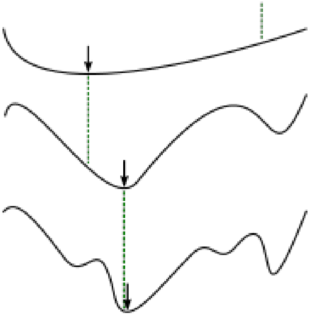
[4] ibid., 2019

Smoothing Policies

- $J(\boldsymbol{\theta})$ is nonconvex
- *Smooth* $J(\boldsymbol{\theta})$ allows *monotonic improvement*

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq 0$$

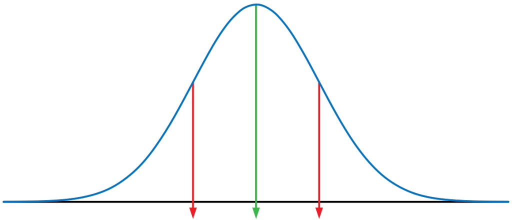
- Conditions on $\nabla \log \pi_{\boldsymbol{\theta}}, \nabla^2 \log \pi_{\boldsymbol{\theta}} \implies J(\boldsymbol{\theta})$ smooth [4]



[4] ibid., 2019

Gaussian Policy

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\mu_{\boldsymbol{\theta}}(s) - a)^2}{2\boldsymbol{\sigma^2}}\right\}$$

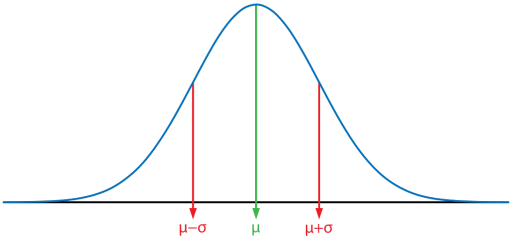


- Variance $\boldsymbol{\sigma^2}$ controls the *amount of exploration*
- Gaussian policies are smoothing[5]

[5] Papini, Pirotta, and Restelli, "Smoothing Policies and Safe Policy Gradients", 2019.

Gaussian Policy

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\mu_{\boldsymbol{\theta}}(s) - a)^2}{2\sigma^{\mathbf{2}}}\right\}$$

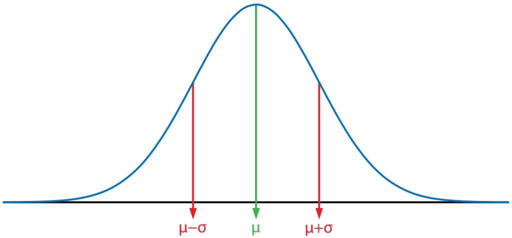


- Variance $\sigma^{\mathbf{2}}$ controls the *amount of exploration*
- Gaussian policies are smoothing[5]

[5] Ibid., 2019.

Gaussian Policy

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\mu_{\boldsymbol{\theta}}(s)-a)^2}{2\sigma^{\mathbf{2}}}\right\}$$



- Variance $\sigma^{\mathbf{2}}$ controls the *amount of exploration*
- Gaussian policies are smoothing[5]

[5] Ibid., 2019.

Speed vs Safety

- $J(\boldsymbol{\theta})$ is $C/\sigma^{\mathbf{2}}$-smooth
- Larger $\sigma \implies$ faster convergence [6]
- Large $\sigma \implies$ not safe!

[6] Ahmed et al., "Understanding the Impact of Entropy on Policy Optimization", 2019

Speed vs Safety

- $J(\boldsymbol{\theta})$ is $C/\sigma^{\mathbf{2}}$-smooth
- Larger $\sigma \implies$ faster convergence [6]
- Large $\sigma \implies$ not safe!

[6] ibid., 2019

Speed vs Safety

- $J(\boldsymbol{\theta})$ is $C/\sigma^{\mathbf{2}}$-smooth
- Larger $\sigma \implies$ faster convergence [6]
- Large $\sigma \implies$ not safe!

[6] ibid., 2019

Adaptive Exploration (State of the Art)

- ~~Hyper-parameter tuning~~
- Learn σ as any other parameter[7] $\implies$ *policy collapse*
- Entropy bonus[8]

[7] Duan et al., "Benchmarking Deep Reinforcement Learning for Continuous Control", 2016.
[8] Haarnoja et al., "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor", 2018.

Adaptive Exploration (State of the Art)

- ~~Hyper-parameter tuning~~
- Learn σ as any other parameter[7] $\implies$ *policy collapse*
- Entropy bonus[8]

[7] Duan et al., "Benchmarking Deep Reinforcement Learning for Continuous Control", 2016.

[8] Haarnoja et al., "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor", 2018.

Adaptive Exploration (State of the Art)

- ~~Hyper-parameter tuning~~
- Learn σ as any other parameter[7] $\implies$ *policy collapse*
- Entropy bonus[8]

[7] Duan et al., "Benchmarking Deep Reinforcement Learning for Continuous Control", 2016.
[8] Haarnoja et al., "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor", 2018.

Meta-Exploring Policy Gradient (MEPG)

1. Special exploration parameter ω

$$\sigma_\omega = e^\omega$$

2. Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \sigma_\omega^{\mathbf{2}} \frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3. Far-sighted update for ω:

$\mathcal{L}(\omega)$ (Exploratory objective)

$\nabla_\omega \mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta})$

$\omega' \leftarrow \omega + \eta \nabla_\omega \mathcal{L}(\omega)$ (Meta-gradient)

Meta-Exploring Policy Gradient (MEPG)

1. Special exploration parameter ω

$$\sigma_\omega = e^\omega \implies J_\omega(\boldsymbol{\theta}) = J_\omega(\boldsymbol{\theta})$$

2. Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha\sigma_\omega^{\mathbf{2}} \frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3. Far-sighted update for ω:

$\mathcal{L}(\omega)$ (Exploratory objective)

$\nabla_\omega \mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta})$

$\omega' \leftarrow \omega + \eta \nabla_\omega \mathcal{L}(\omega)$ (Meta-gradient)

Meta-Exploring Policy Gradient (MEPG)

1. Special exploration parameter ω

$$\sigma_\omega = e^{\omega} \implies J_\omega(\boldsymbol{\theta}) = J_\omega(\boldsymbol{\theta})$$

2. Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \sigma_\omega^{\mathbf{2}} \frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3. Far-sighted update for ω:

$\mathcal{L}(\omega)$ (Exploratory objective)

$\nabla_\omega \mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta})$

$\omega' \leftarrow \omega + \eta \nabla_\omega \mathcal{L}(\omega)$ (Meta-gradient)

Meta-Exploring Policy Gradient (MEPG)

1. Special exploration parameter ω

$$\sigma_\omega = e^\omega \implies J_\omega(\boldsymbol{\theta}) = J_\omega(\boldsymbol{\theta})$$

2. Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha\sigma_\omega^{\mathbf{2}} \frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3. Far-sighted update for ω:

$\mathcal{L}(\omega) = J_\omega(\boldsymbol{\theta}')$ (Exploratory objective)

$\nabla_\omega \mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta})$

$\omega' \leftarrow \omega + \eta \nabla_\omega \mathcal{L}(\omega)$ (Meta-gradient)

1. Special exploration parameter ω

$$\sigma_\omega = e^\omega \implies J_\omega(\boldsymbol{\theta}) = J_\omega(\boldsymbol{\theta})$$

2. Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha\sigma_\omega^{\mathbf{2}} \frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3. Far-sighted update for ω:

$$\mathcal{L}(\omega) = J_\omega\left(\boldsymbol{\theta} + \alpha\sigma_\omega^{\mathbf{2}} \nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\right) \qquad \text{(Exploratory objective)}$$

$$\nabla_\omega \mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta})$$

$$\omega' \leftarrow \omega + \eta \nabla_\omega \mathcal{L}(\omega) \qquad \text{(Meta-gradient)}$$

Meta-Exploring Policy Gradient (MEPG)

1 Special exploration parameter ω

$$\sigma_\omega = e^{\omega} \implies J_\omega(\boldsymbol{\theta}) = J_\omega(\boldsymbol{\theta})$$

2 Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha\sigma_\omega^{\mathbf{2}} \frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3 Far-sighted update for ω:

$$\mathcal{L}(\omega) \simeq J_\omega(\boldsymbol{\theta}) + \alpha\sigma_\omega^{\mathbf{2}} \|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\| \qquad \text{(Exploratory objective)}$$

$$\nabla_\omega \mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta})$$

$$\omega' \leftarrow \omega + \eta \nabla_\omega \mathcal{L}(\omega) \qquad \text{(Meta-gradient)}$$

Meta-Exploring Policy Gradient (MEPG)

1. Special exploration parameter ω

$$\sigma_\omega = e^\omega \implies J_\omega(\boldsymbol{\theta}) = J_\omega(\boldsymbol{\theta})$$

2. Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha\sigma_\omega^{\mathbf{2}} \frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3. Far-sighted update for ω:

$$\mathcal{L}(\omega) \simeq J_\omega(\boldsymbol{\theta}) + \alpha\sigma_\omega^{\mathbf{2}} \|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\| \qquad \text{(Exploratory objective)}$$

$$\nabla_\omega \mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta})$$

$$\omega' \leftarrow \omega + \eta \nabla_\omega \mathcal{L}(\omega) \qquad \text{(Meta-gradient)}$$

Meta-Exploring Policy Gradient (MEPG)

1. Special exploration parameter ω

$$\sigma_\omega = e^\omega \implies J_\omega(\boldsymbol{\theta}) = J_\omega(\boldsymbol{\theta})$$

2. Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha\sigma_\omega^{\mathbf{2}} \frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3. Far-sighted update for ω:

$$\mathcal{L}(\omega) \simeq J_\omega(\boldsymbol{\theta}) + \alpha\sigma_\omega^{\mathbf{2}} \|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\| \qquad \text{(Exploratory objective)}$$

$$\nabla_\omega \mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta}) + 2\alpha\sigma_\omega^{\mathbf{2}} \|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|$$

$$\omega' \leftarrow \omega + \eta\nabla_\omega\mathcal{L}(\omega) \qquad \text{(Meta-gradient)}$$

Meta-Exploring Policy Gradient (MEPG)

1. Special exploration parameter ω

$$\sigma_\omega = e^\omega \implies J_\omega(\boldsymbol{\theta}) = J_\omega(\boldsymbol{\theta})$$

2. Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha\sigma_\omega^{\mathbf{2}}\frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3. Far-sighted update for ω:

$$\mathcal{L}(\omega) \simeq J_\omega(\boldsymbol{\theta}) + \alpha\sigma_\omega^{\mathbf{2}}\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\| \qquad \text{(Exploratory objective)}$$

$$\nabla_\omega\mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta}) + 2\alpha\sigma_\omega^{\mathbf{2}}\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\| + \alpha\sigma_\omega\nabla_\omega\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|$$

$$\omega' \leftarrow \omega + \eta\nabla_\omega\mathcal{L}(\omega) \qquad \text{(Meta-gradient)}$$

Meta-Exploring Policy Gradient (MEPG)

1. Special exploration parameter ω

$$\sigma_\omega = e^\omega \implies J_\omega(\boldsymbol{\theta}) = J_\omega(\boldsymbol{\theta})$$

2. Exploration-aware policy update

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha\sigma_\omega^{\mathbf{2}} \frac{\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|}$$

3. Far-sighted update for ω:

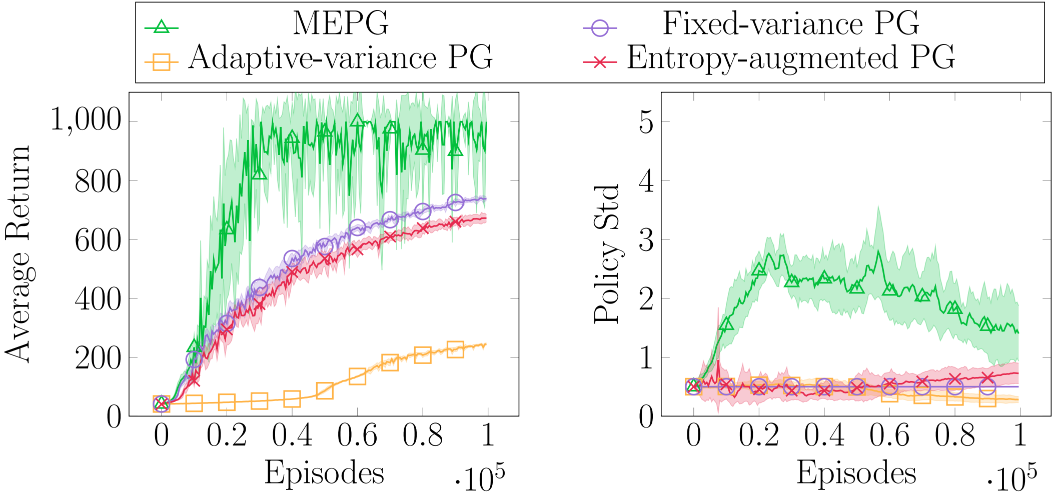
$$\mathcal{L}(\omega) \simeq J_\omega(\boldsymbol{\theta}) + \alpha\sigma_\omega^{\mathbf{2}} \|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\| \qquad \text{(Exploratory objective)}$$

$$\nabla_\omega \mathcal{L} = \nabla_\omega J_\omega(\boldsymbol{\theta}) + 2\alpha\sigma_\omega^{\mathbf{2}} \|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\| + \alpha\sigma_\omega \nabla_\omega \|\nabla_{\boldsymbol{\theta}} J_\omega(\boldsymbol{\theta})\|$$

$$\omega' \leftarrow \omega + \eta\nabla_\omega \mathcal{L}(\omega) \qquad \text{(Meta-gradient)}$$

Sutton, "Adapting Bias by Gradient Descent: An Incremental Version of Delta-Bar-Delta", 1992

Continuous Cart-Pole[9]



[9] Brockman et al., "OpenAI Gym", 2016.

MEPG:

$$\boldsymbol{\theta}' \leftarrow \alpha\sigma_{\omega}^{\mathbf{2}}\nabla_{\boldsymbol{\theta}}J_{\omega}(\boldsymbol{\theta})$$
$$\omega' \leftarrow \omega + \eta\nabla_{\omega}\mathcal{L}(\omega)$$

Safe Policy Gradient

MEPG:

$$\boldsymbol{\theta}' \leftarrow \alpha \sigma_{\omega}^{\mathbf{2}} \nabla_{\boldsymbol{\theta}} J_{\omega}(\boldsymbol{\theta})$$
$$\omega' \leftarrow \omega + \eta \nabla_{\omega} \mathcal{L}(\omega)$$

Monotonic improvement:

$$J_{\omega}(\boldsymbol{\theta}') - J_{\omega}(\boldsymbol{\theta}) \geq 0 \qquad \text{for sufficiently small } \alpha > 0$$
$$J_{\omega'}(\boldsymbol{\theta}) - J_{\omega}(\boldsymbol{\theta}) \geq 0 \qquad \text{for sufficiently small } \eta \in \mathbb{R}$$

Safe Policy Gradient

MEPG:

$$\boldsymbol{\theta}' \leftarrow \alpha\sigma_{\omega}^{\mathbf{2}}\nabla_{\boldsymbol{\theta}} J_{\omega}(\boldsymbol{\theta})$$
$$\omega' \leftarrow \omega + \eta\nabla_{\omega}\mathcal{L}(\omega)$$

Monotonic improvement:

$J_{\omega}(\boldsymbol{\theta}') - J_{\omega}(\boldsymbol{\theta}) \geq 0$ for sufficiently small $\alpha > 0$

$J_{\omega'}(\boldsymbol{\theta}) - J_{\omega}(\boldsymbol{\theta}) \geq 0$ for sufficiently small $\eta \in \mathbb{R}$

η can be negative!

Safely-Exploring Policy Gradient (SEPG)

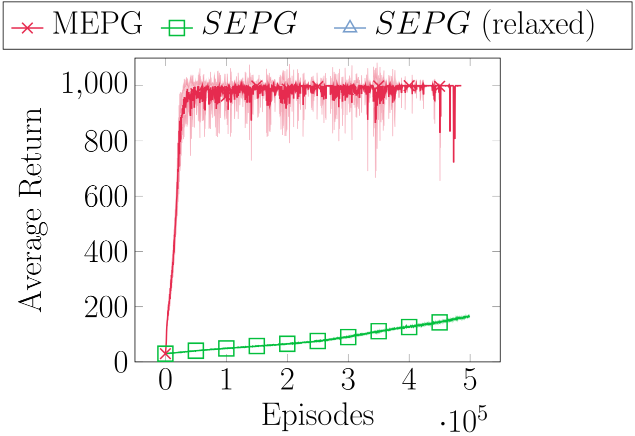
`Alternate:`

$$\boldsymbol{\theta}' \leftarrow \alpha^* \nabla_{\boldsymbol{\theta}} J_{\omega}(\boldsymbol{\theta})$$
$$\omega' \leftarrow \omega + \eta^* \nabla_{\omega} \mathcal{L}(\omega)$$

Largest safe step sizes (adaptive):

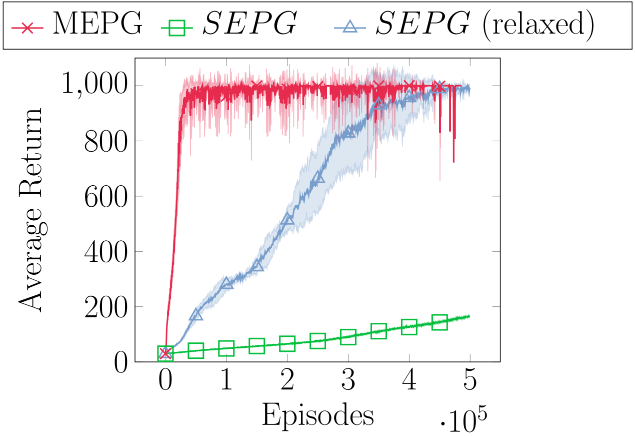
$$\alpha^* \propto \frac{\sigma_{\omega}^{\mathbf{2}}}{\|\nabla_{\boldsymbol{\theta}} J_{\omega}(\boldsymbol{\theta})\|}$$
$$\eta^* \propto \frac{1}{\|\nabla_{\omega} \mathcal{L}_{\omega}(\boldsymbol{\theta})\|}$$

Continuous Cart-Pole (cont.)



$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq \quad 0$$

Continuous Cart-Pole (cont.)



$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq 0$$

$$J(\boldsymbol{\theta}') - J(\boldsymbol{\theta}) \geq -C$$

Conclusion

- Adapting policy variance farsightedly is important
- Safe updates are possible
- Gap between theory and practice
- No epistemic uncertainty
- Remove random actions?

Conclusion

- Adapting policy variance farsightedly is important
- Safe updates are possible
- Gap between theory and practice
- No epistemic uncertainty
- Remove random actions?

Thanks for watching!

Contact: matteo.papini@polimi.it

References

- Ahmed, Zafarali et al. "Understanding the Impact of Entropy on Policy Optimization". In: *ICML*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 151–160.
- Akkaya, Ilge et al. "Solving Rubik's Cube with a Robot Hand". In: *arXiv preprint arXiv:1910.07113* (2019).
- Amodei, Dario et al. "Concrete Problems in AI Safety". In: *CoRR* abs/1606.06565 (2016).
- Brockman, Greg et al. "OpenAI Gym". In: *CoRR* abs/1606.01540 (2016).
- Deisenroth, Marc Peter, Gerhard Neumann, and Jan Peters. "A Survey on Policy Search for Robotics". In: *Foundations and Trends in Robotics* 2.1-2 (2013), pp. 1–142.
- Duan, Yan et al. "Benchmarking Deep Reinforcement Learning for Continuous Control". In: *ICML*. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 1329–1338.

References (cont.)

- Haarnoja, Tuomas et al. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *ICML*. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1856–1865.
- Papini, Matteo, Matteo Pirotta, and Marcello Restelli. "Smoothing Policies and Safe Policy Gradients". In: *CoRR* abs/1905.03231 (2019).
- Peng, Xue Bin et al. *Learning Agile Robotic Locomotion Skills by Imitating Animals*. 2020. arXiv: 2004.00784 [cs.RO].
- Sutton, Richard S. "Adapting Bias by Gradient Descent: An Incremental Version of Delta-Bar-Delta". In: *AAAI*. AAAI Press / The MIT Press, 1992, pp. 171–176.
- Sutton, Richard S and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.