POLITECNICO
MILANO 1863

Safe Policy Optimization

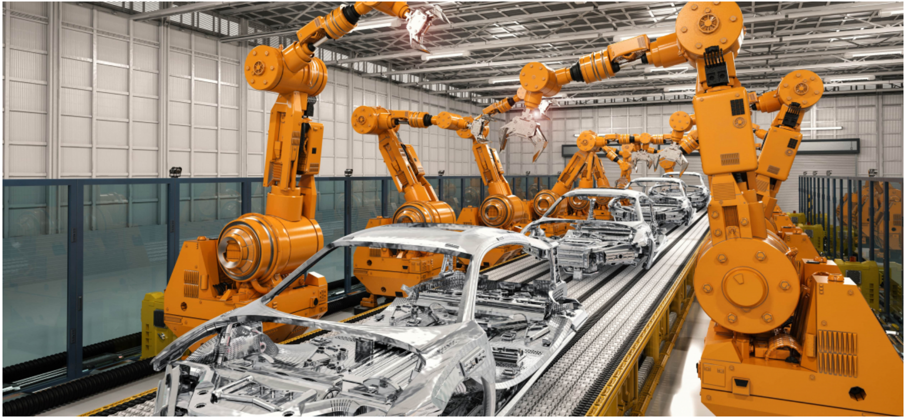
Ph.D. Course in Information Technology (Computer Science and Engineering), XXXIII cycle
Second Yearly Evaluation

Matteo Papini
Supervisor: Prof. Marcello Restelli

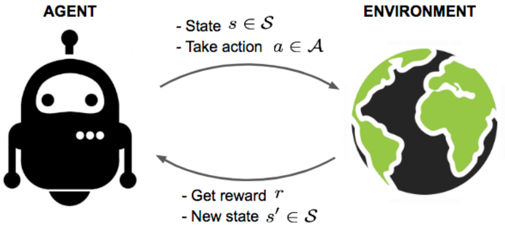
30th September 2019

Motivation

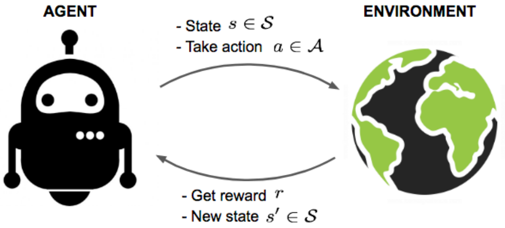
Apply **Reinforcement Learning** to **real-world** control problems



Reinforcement Learning [Sutton and Barto, 2018]

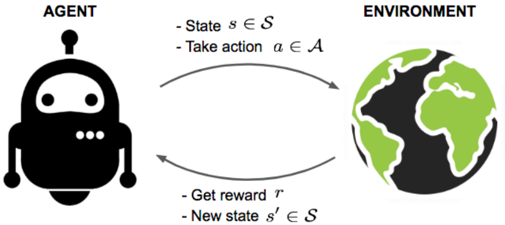


Reinforcement Learning [Sutton and Barto, 2018]



Policy π**:** agent's behavior $(s \mapsto a)$
Performance $\rho(\pi)$**:** *expected* total reward
Goal: find policy maximizing performance

Reinforcement Learning [Sutton and Barto, 2018]



Policy π: agent's behavior $(s \mapsto a)$
Performance $\rho(\pi)$: *expected* total reward
Goal: find policy maximizing performance

- Model-free
- Online
- Iterative

Safety in Reinforcement Learning

- Many notions of safety [Amodei et al., 2016, García and Fernández, 2015]
- Assume performance ρ already encodes *safety constraints*
- The optimal policy will be safe
- **The learning process itself may not be!**

Safety in Reinforcement Learning

- Many notions of safety [Amodei et al., 2016, García and Fernández, 2015]
- Assume performance ρ already encodes *safety constraints*
- The optimal policy will be safe
- **The learning process itself may not be!**

- Many notions of safety [Amodei et al., 2016, García and Fernández, 2015]
- Assume performance ρ already encodes *safety constraints*
- The optimal policy will be safe
- **The learning process itself may not be!**

- Many notions of safety [Amodei et al., 2016, García and Fernández, 2015]
- Assume performance ρ already encodes *safety constraints*
- The optimal policy will be safe
- **The learning process itself may not be!**

Continuous Reinforcement Learning

Interesting real-world control problems are **continuous**

Interesting real-world control problems are **continuous**

Policy Optimization [Deisenroth et al., 2013]:

- Scales well with state-action dimensionality
- Convergence guarantees [Sutton et al., 1999]
- Robustness to noise

Policy Optimization

- Fix a class of controllers with tunable parameters $\boldsymbol{x} \in \mathcal{X}$
- Find policy parameters maximizing performance:

$$\max_{\boldsymbol{x} \in \mathcal{X}} \rho(\boldsymbol{x})$$

- **Policy Gradient:** solve it with *Stochastic Gradient Descent*

Policy Optimization

- Fix a class of controllers with tunable parameters $\boldsymbol{x} \in \mathcal{X}$
- Find policy parameters maximizing performance:

$$\max_{\boldsymbol{x} \in \mathcal{X}} \rho(\boldsymbol{x})$$

- **Policy Gradient:** solve it with *Stochastic Gradient Descent*

Policy Optimization

- Fix a class of controllers with tunable parameters $\boldsymbol{x} \in \mathcal{X}$
- Find policy parameters maximizing performance:

$$\max_{\boldsymbol{x} \in \mathcal{X}} \rho(\boldsymbol{x})$$

- **Policy Gradient:** solve it with *Stochastic Gradient Descent*

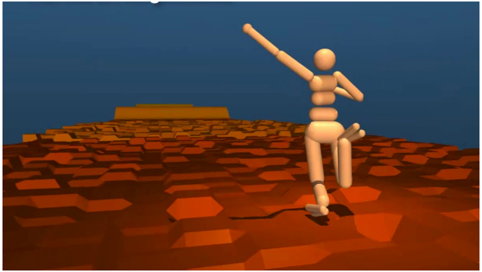
Policy Optimization

- Fix a class of controllers with tunable parameters $\boldsymbol{x} \in \mathcal{X}$
- Find policy parameters maximizing performance:

$$\max_{\boldsymbol{x} \in \mathcal{X}} \rho(\boldsymbol{x})$$

- **Policy Gradient:** solve it with *Stochastic Gradient Descent*

OpenAI [2018]



Heess et al. [2017]

Real-World Requirements

ML Engineer: ”I will improve your controller with RL!”

ML Engineer: ”I will improve your controller with RL!”

Boss:

- ”How long will it take?”
- ”Will it *actually* improve?”
- ”Will it behave safely?”
- ”How much better will it become?”

ML Engineer: "I will improve your controller with RL!"

Boss:

- "How long will it take?"
- "Will it *actually* improve?"
- "Will it behave safely?"
- "How much better will it become?"

Real-World Requirements

ML Engineer: "I will improve your controller with RL!"

Boss:

- "How long will it take?"
- "Will it *actually* improve?"
- "Will it behave safely?"
- "How much better will it become?"

Real-World Requirements

ML Engineer: "I will improve your controller with RL!"

Boss:

- "How long will it take?"
- "Will it *actually* improve?"
- "Will it behave safely?"
- "How much better will it become?"

Real-World Requirements

ML Engineer: "I will improve your controller with RL!"

Boss:

- "How long will it take?" Unkown
- "Will it *actually* improve?"
- "Will it behave safely?"
- "How much better will it become?"

Real-World Requirements

ML Engineer: "I will improve your controller with RL!"

Boss:

- "How long will it take?" Unkown
- "Will it *actually* improve?" Eventually
- "Will it behave safely?"
- "How much better will it become?"

Real-World Requirements

ML Engineer: "I will improve your controller with RL!"

Boss:

- "How long will it take?" Unkown
- "Will it *actually* improve?" Eventually
- "Will it behave safely?" Eventually
- "How much better will it become?"

Real-World Requirements

ML Engineer: "I will improve your controller with RL!"

Boss:

- "How long will it take?" Unkown
- "Will it *actually* improve?" Eventually
- "Will it behave safely?" Eventually
- "How much better will it become?" Unkown

Sample Complexity
("How long will it take?")

Sample Complexity
("How long will it take?")

First year:

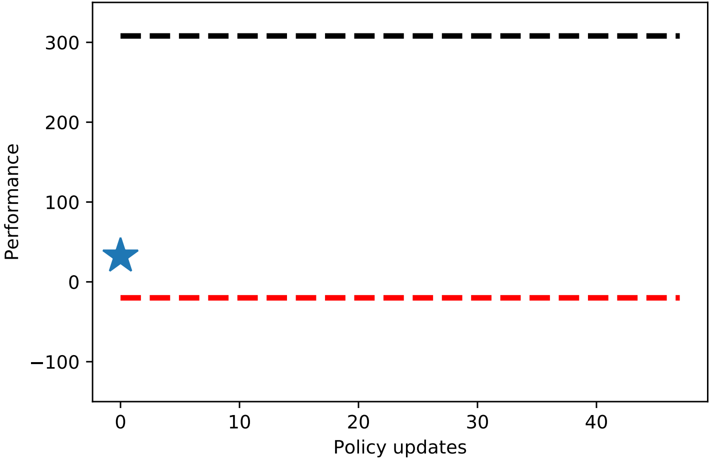
Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, Marcello Restelli: *Stochastic Variance-Reduced Policy Gradient*. **ICML 2018**: 4023-4032

Alberto Maria Metelli, **Matteo Papini**, Francesco Faccio, Marcello Restelli: *Policy Optimization via Importance Sampling*. **NeurIPS 2018**: 5447-5459

Safe Policy Updates
(”Will it actually improve?”)

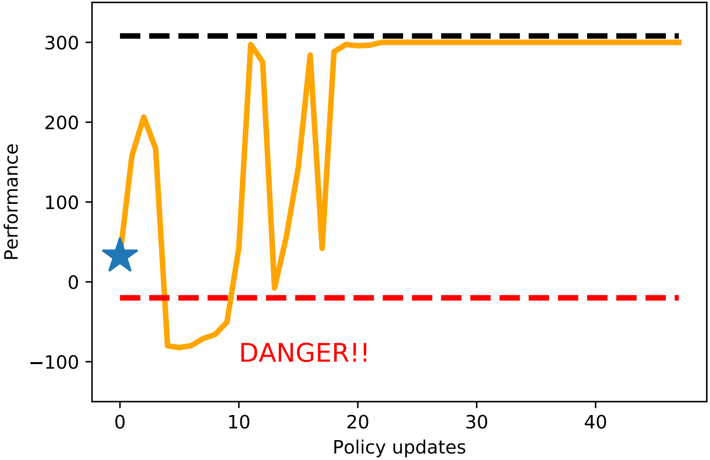
Monotonic Performance Improvement

- A concrete problem in Reinforcement Learning



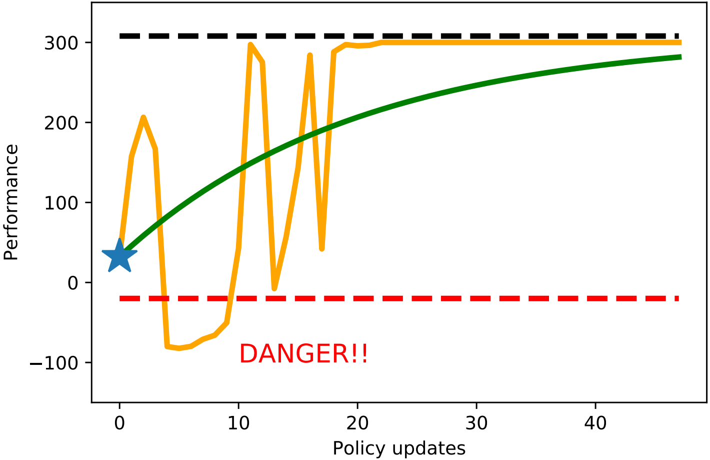
Monotonic Performance Improvement

- A concrete problem in Reinforcement Learning



Monotonic Performance Improvement

- A concrete problem in Reinforcement Learning



Smoothing Policies and Safe Policy Gradients

A policy gradient algorithm with **monotonic improvement** guarantees

A policy gradient algorithm with **monotonic improvement** guarantees

State of the art [Kakade and Langford, 2002, Pirotta et al., 2015, 2013, Schulman et al., 2015, Papini et al., 2017]:

- Restricted policy class
- Regularity assumptions on the environment

Smoothing Policies and Safe Policy Gradients

A policy gradient algorithm with **monotonic improvement** guarantees

State of the art [Kakade and Langford, 2002, Pirotta et al., 2015, 2013, Schulman et al., 2015, Papini et al., 2017]:

- Restricted policy class
- Regularity assumptions on the environment

Our method:

- General conditions on policy
- No assumptions on the environment
- Simpler formulation
- Smaller gap between theory and practice

Submitted to JMLR

Safe Exploration
("Will it behave safely?")

Safe Exploration

Exploration: perform diverse actions to gather novel information

- Necessary for improvement
- Tipically: perform *random* actions [Ahmed et al., 2019]
- Unpredictable behavior may be **unsafe**

Safe Exploration

Exploration: perform diverse actions to gather novel information

- Necessary for improvement
- Tipically: perform *random* actions [Ahmed et al., 2019]
- Unpredictable behavior may be **unsafe**

Control the amount of stochasticity:

Matteo Papini, Andrea Battistello, Marcello Restelli; *"Safely Exploring Policy Gradient"*; 14th European Workshop on Reinforcement Learning (**EWRL14**), Lille, France, 2018

*Revised version planned for **AISTATS 2020***

Direct exploration towards interesting information

OPTIMIST

Direct exploration towards interesting information

Not policy gradient, inspired by Multi-Armed-Bandit literature [Bubeck et al., 2012, Lattimore and Szepesvári, 2018]

- Global convergence
- Still requires stochasticity
- Non-monotonic performance

Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, Marcello Restelli: *Optimistic Policy Optimization via Multiple Importance Sampling*. **ICML 2019**: 4989-4999.

(Truly) Deterministic Policy Gradient

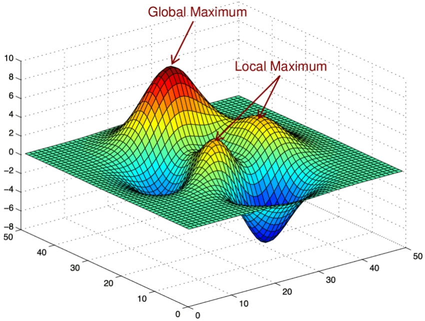
- Is (random) exploration really necessary?
- Maybe not if the world is sufficiently regular
- **Idea**: re-use similar experience instead of sampling new one

*Planned for **ICML 2020***

Quality of Solutions
("How much better will it become?")

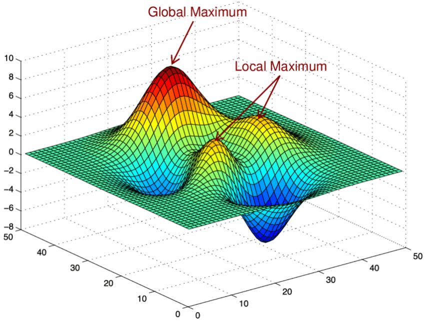
Non-Convexity

- The performance objective ρ is **nonconvex**
- Policy gradient only converges to **local optima**
- Locally optimal performance could be poor
- Locally optimal policies may be **unsafe**



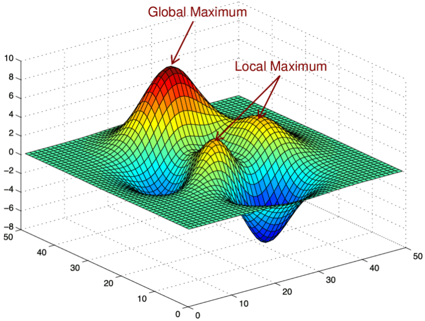
Non-Convexity

- The performance objective ρ is **nonconvex**
- Policy gradient only converges to **local optima**
- Locally optimal performance could be poor
- Locally optimal policies may be **unsafe**



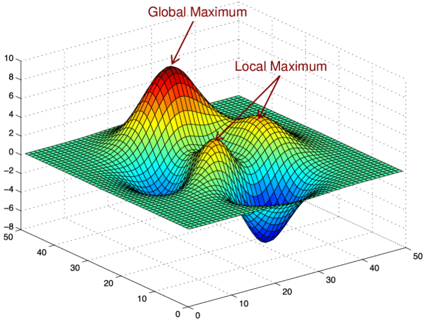
Non-Convexity

- The performance objective ρ is **nonconvex**
- Policy gradient only converges to **local optima**
- Locally optimal performance could be poor
- Locally optimal policies may be **unsafe**



Non-Convexity

- The performance objective ρ is **nonconvex**
- Policy gradient only converges to **local optima**
- Locally optimal performance could be poor
- Locally optimal policies may be **unsafe**



Global Optimality Guarantees

- The *policy* optimization problem is *special*
- Convergence to the **global optimum** is possible in some cases [Bhandari and Russo, 2019, Agarwal et al., 2019, Shani et al., 2019]
- Can we exploit tools from **convex optimization** to design *new* algorithms?

Global Optimality Guarantees

- The *policy* optimization problem is *special*
- Convergence to the **global optimum** is possible in some cases [Bhandari and Russo, 2019, Agarwal et al., 2019, Shani et al., 2019]
- Can we exploit tools from **convex optimization** to design *new* algorithms?

Global Optimality Guarantees

- The *policy* optimization problem is *special*
- Convergence to the **global optimum** is possible in some cases [Bhandari and Russo, 2019, Agarwal et al., 2019, Shani et al., 2019]
- Can we exploit tools from **convex optimization** to design *new* algorithms?

Global Optimality Guarantees

- The *policy* optimization problem is *special*
- Convergence to the **global optimum** is possible in some cases [Bhandari and Russo, 2019, Agarwal et al., 2019, Shani et al., 2019]
- Can we exploit tools from **convex optimization** to design *new* algorithms?

Possible target: **NeurIPS 2020**

Wrapping Up

- Sample Complexity [Papini et al., 2018, Xu et al., 2019a,b]
- Safe Policy Updates [Papini et al., 2019b]
- Safe Exploration (Papini et al. [2019a] + work in progress)
- Quality of Solutions (future work)

Wrapping Up

- Sample Complexity [Papini et al., 2018, Xu et al., 2019a,b]
- Safe Policy Updates [Papini et al., 2019b]
- Safe Exploration (Papini et al. [2019a] + work in progress)
- Quality of Solutions (future work)

We also want to:

- Find a trade-off between competing goals
- Apply to real problems

Publications (Overview)

	GGS	2017	2018	2109	2020
ICML	A++		[Papini et al., 2018]	[Papini et al., 2019a]	planed
NeurIPS	A++	[Papini et al., 2017]	[Metelli et al., 2018]		planned
AAAI	A++				submitted
AISTATS	A+				planned
IJCNN	B			[Beraha et al., 2019]	

Publications

First year:

- **Matteo Papini**, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, Marcello Restelli: *Stochastic Variance-Reduced Policy Gradient*. **ICML 2018**: 4023-4032
- Alberto Maria Metelli, **Matteo Papini**, Francesco Faccio, Marcello Restelli: *Policy Optimization via Importance Sampling*. **NeurIPS 2018**: 5447-5459

Second year:

- **Matteo Papini**, Alberto Maria Metelli, Lorenzo Lupo, Marcello Restelli: *Optimistic Policy Optimization via Multiple Importance Sampling*. **ICML 2019**: 4989-4999
- Mario Beraha, Alberto Maria Metelli, **Matteo Papini**, Andrea Tirinzoni, Marcello Restelli: *Feature Selection via Mutual Information: New Theoretical Insights*. IJCNN 2019

Other Publications

Before the Ph.D.:

- **Matteo Papini**, Matteo Pirotta, Marcello Restelli: *Adaptive Batch Size for Safe Policy Gradients*. **NeurIPS 2017**: 3591-3600

Workshop papers:

- **Matteo Papini**, Andrea Battistello, Marcello Restelli: *Safely Exploring Policy Gradient*. EWRL 2018

Submissions

- **Matteo Papini**, Matteo Pirotta, Marcello Restelli: *Smoothing Policies and Safe Policy Gradients*. (**JMLR**)
- Pierluca D'Oro, Alberto Maria Metelli, Andrea Tirinzoni, **Matteo Papini**, Marcello Restelli: *Gradient-Aware Model-based Policy Search*. (**AAAI 2020**)
- Lorenzo Bisi, Luca Sabbioni, Edoardo Vittori, **Matteo Papini**, Marcello Restelli: *Risk-Averse Trust Region Optimization for Reward-Volatility Reduction* (**AAAI 2020**)

Courses and Schools

Courses (25/25 CFU):

Crediti: totale 30.0, da dott. di iscrizione 10.0, da altro dott. 5.0, da scuola di dott. 10.0, da laurea triennale 0.0, da laurea magistrale 0.0, extra 5.0

Anno di frequenza	AC	Sem	Pos	Cfu	Codice	Insegnamento	Stato	Data	Voto	L	CdS	Indir	Sez	Origine
2017	1		E	5	051936	EXTERNAL COURSES WITH EVALUATION - 1	C	10/09/18	A		INFO	DOT		▲
2017	1		S	5	051845	INTRODUZIONE ALL'OTTIMIZZAZIONE CONVESSA	N				M^3I	DOT		◪
2017	1		E	5	096678	COMMUNICATING SCIENTIFIC RESEARCH	S	03/07/18	A	S	DOTT	DOT		■
2017	1		E	5	053360	SCIENTIFIC COMMUNICATION IN ENGLISH	S	08/07/18	A	S	DOTT	DOT		■
2017	1		E	5	051909	IMAGE CLASSIFICATION: MODERN APPROACHES	S	09/09/18	A	S	INFO	DOT		□
2017	1		E	5	051908	DEEP LEARNING: THEORY, TECHNIQUES AND APPLICATIONS	S	10/09/18	A	S	INFO	DOT		□

Genera file PDF in una nuova finestra

Schools:

- Deep Learning and Reinforcement Learning Summer School (DLRLSS), Toronto, Canada, 2018
- ACAI Summer School on Reinforcement Learning, Nieuwpoort, Belgium, 2017

Teaching

2017/2018

- Responsabile di laboratorio, Informatica B, Prof. Luca Cassano
- Esercitatore, Web and Internet Economics, Prof. Nicola Gatti

2018/2019

- Esercitatore, Informatica B, Prof. Luca Cassano

2019/2020

- Esercitatore, Informatica B, Prof. Luca Cassano (now)

Teaching outside Politecnico:

- Teaching Assistant for the Reinforcement Learning Summer School (RLSS), Lille, France, 2019

Other Activities

Talks and posters:

- Seminar *Temporal Credit Assignment in Off-Policy Reinforcement Learning*, DEIB, November 28th , 2017
- Poster presentation at NeurIPS 2017
- Oral and poster presentation at ICML 2018
- Poster presentation at DLRLSS 2018
- Poster presentation at EWRL14 (2018)
- Oral and poster presentation at NeurIPS 2018
- Oral and poster presentation at ICML 2019
- Invited talk at MAPLE workshop 2019

Editorial activities:

- Subreviewer for IJCAI 2018
- Reviewer for ICML 2019
- PC Member for UAI 2019
- Reviewer for NeurIPS 2019
- Reviewer for AAAI 2020 (now)
- Reviewer for AISTATS 2020 (planned)

Co-supervised master students: G. Canonaco, D. Binaghi, A. Battistello, F. Faccio, A. Mongelluzzo, L. Lupo, G. Pelosi, P. Melzi (now)

Thank you for your attention!

Questions?

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 151–160. PMLR, 2019.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.

Mario Beraha, Alberto Maria Metelli, Matteo Papini, Andrea Tirinzoni, and Marcello Restelli. Feature selection via mutual information: New theoretical insights. 2019.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Marc Peter Deisenroth, Gerhard Neumann, and Jan Peters. A survey on policy search for robotics. *Foundations and Trends in Robotics*, 2(1-2):1–142, 2013.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.*, 16:1437–1480, 2015.

Bibliography II

Nicolas Heess, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, SM Eslami, Martin Riedmiller, et al. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*, 2017.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, pages 267–274. Morgan Kaufmann, 2002.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. 2018.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In *NeurIPS*, pages 5447–5459, 2018.

OpenAI. Openai five. `https://blog.openai.com/openai-five/`, 2018.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Adaptive batch size for safe policy gradients. In *NeurIPS*, pages 3591–3600, 2017.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 4023–4032. PMLR, 2018.

Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. Optimistic policy optimization via multiple importance sampling. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 4989–4999. PMLR, 2019a.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *CoRR*, abs/1905.03231, 2019b.

Bibliography III

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. In *NIPS*, pages 1394–1402, 2013.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2-3):255–283, 2015.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897. JMLR.org, 2015.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. *arXiv preprint arXiv:1909.02769*, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, pages 1057–1063. The MIT Press, 1999.

Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *UAI*, page 191. AUAI Press, 2019a.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. *CoRR*, abs/1909.08610, 2019b.