Exploration and Regularization in Reinforcement Learning

Gergely Neu

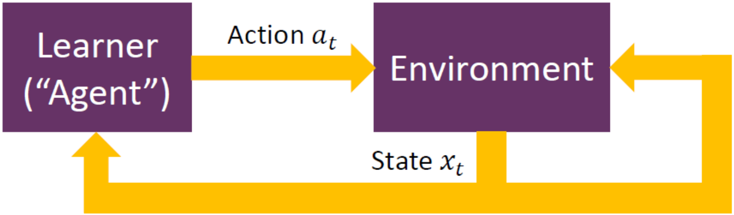
Universitat Pompeu Fabra
Barcelona, Spain

Based on joint work with Anders Jonsson and Vicenç Gómez

Outline

1. MDP basics in 5 minutes
2. Exploration and regularization in RL
3. Entropy-regularized RL
   - Recent trends
   - A unifying theory
   - An algorithmic framework
   - Some results

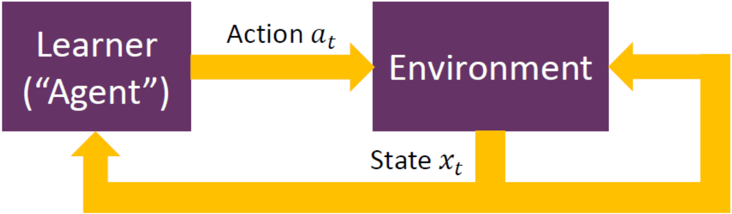
Markov decision processes



Repeat for $t = 1, 2, \ldots$:

- Learner
  - observes state x_t and plays action a_t
  - obtains reward $r(x_t, a_t)$,
- Environment generates next state $x_{t+1} \sim P(\cdot|x_t, a_t)$.

Markov decision processes



Repeat for $t = 1, 2, \ldots$:

- LEARNER
  - observes state x_t and plays action a_t
  - obtains reward $r(x_t, a_t)$,
- ENVIRONMENT generates next state $x_{t+1} \sim P(\cdot | x_t, a_t)$.

GOAL: gather as much reward as possible

Optimal control in MDPs

A 5-minute summary

- Average-reward criterion:

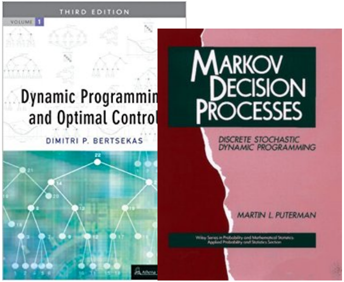
$$\liminf_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r(x_t, a_t)\right].$$

- Basic fact: enough to consider *stationary policies*

$$\pi(a|x) = \mathbb{P}\left[a_t = a \middle| x_t = x\right].$$

- Under mild assumptions, every π induces stationary distribution μ_π:

$$\mu_\pi(x, a) = \lim_{t\to\infty} \mathbb{P}\left[x_t = x, a_t = a\right].$$



Optimal control in MDPs

A 5-minute summary

- Average-reward criterion:

$$\liminf_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r(x_t, a_t)\right].$$

- Basic fact: enough to consider *stationary policies*

$$\pi(a|x) = \mathbb{P}\left[a_t = a \middle| x_t = x\right].$$

- Under mild assumptions, every π induces stationary distribution μ_π:

$$\mu_\pi(x, a) = \lim_{t\to\infty} \mathbb{P}\left[x_t = x, a_t = a\right].$$

Notice: average reward of π is linear in μ_π:

$$\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r(x_t, a_t)\right]$$

$$= \sum_{x,a} \mu_\pi(x, a) r(x, a)$$

$$= \langle \mu_\pi, r\rangle$$

Optimal control in MDPs

The LP formulation

Primal LP

$$\rho^* = \max_{\mu \in \Delta} \langle \mu, r \rangle$$

$$\Delta = \left\{ \text{distribution } \mu : \sum_b \mu(y, b) = \sum_{x,a} P(y|x, a)\mu(x, a) \quad (\forall y) \right\}$$

Optimal control in MDPs

The LP formulation

Primal LP

$$\rho^* = \max_{\mu \in \Delta} \langle \mu, r \rangle$$

$$\Delta = \left\{ \text{distribution } \mu : \sum_b \mu(y, b) = \sum_{x,a} P(y|x, a)\mu(x, a) \quad (\forall y) \right\}$$

Dual LP

$$\rho^* = \min_{\rho \in \mathbb{R}} \rho$$

$$\text{s.t.} \quad V(x) \geq r(x, a) - \rho + \sum_y P(y|x, a) V(y) \quad (\forall x, a)$$

Optimal control in MDPs

The LP formulation

Primal LP

$$\rho^* = \max_{\mu \in \Delta} \langle \mu, r \rangle$$

$$\Delta = \left\{ \text{distribution } \mu : \sum_b \mu(y,b) = \sum_{x,a} P(y|x,a)\mu(x,a) \quad (\forall y) \right\}$$

Dual "LP" $\equiv$ The Bellman equations

$$V^*(x) = \max_a \left(r(x,a) - \rho^* + \sum_y P(y|x,a)\, V^*(y) \right) \quad (\forall x, a)$$

Reinforcement Learning in MDPs

Reinforcement Learning

$$\approx$$

learning optimal policies in unknown MDPs

Reinforcement Learning in MDPs

Reinforcement Learning

$\approx$

learning optimal policies in unknown MDPs

Exactly solving imperfectly known MDPs is foolish!

Reinforcement Learning in MDPs

Reinforcement Learning

$\approx$

learning optimal policies in unknown MDPs

Exactly solving imperfectly known MDPs is foolish!

- Overfitting: too little data $\Rightarrow$ bad policy

Reinforcement Learning in MDPs

Reinforcement Learning

$\approx$

learning optimal policies in unknown MDPs

Exactly solving imperfectly known MDPs is foolish!

- Overfitting: too little data $\Rightarrow$ bad policy
- Under-exploration: tons of bad data $\Rightarrow$ bad policy

Reinforcement Learning in MDPs

Reinforcement Learning

$\approx$

learning optimal policies in unknown MDPs

Exactly solving imperfectly known MDPs is foolish!

- Overfitting: too little data $\Rightarrow$ bad policy
- Under-exploration: tons of bad data $\Rightarrow$ bad policy

SOLUTION:
Regularization!

A recent trend: (Entropy-)Regularized RL

Two popular approaches

Idea 1: Soften the max in the Bellman optimality equations!

$$V^*(x) = \max_a \left(r(x,a) - \rho^* + \sum_y P(y|x,a) V^*(y) \right)$$

A recent trend: (Entropy-)Regularized RL

Two popular approaches

Idea 1: Soften the max in the Bellman optimality equations!

$$V_\eta^*(x) = \frac{1}{\eta} \log \sum_a \exp\left(\eta \left(r(x,a) - \rho_\eta^* + \sum_y P(y|x,a)\, V_\eta^*(y)\right)\right)$$

[Marcus et al., 1997, Ruszczyński, 2010, Ziebart et al., 2010, Ziebart, 2010, Braun et al., 2011, Azar et al., 2012, Rawlik et al., 2012, Fox et al., 2016, Asadi and Littman, 2017, Haarnoja et al., 2017, Schulman et al., 2017, Nachum et al., 2017] . . . and who knows how many more NIPS'17 submissions

A recent trend: (Entropy-)Regularized RL

Two popular approaches

Idea 1: Soften the max in the Bellman optimality equations!

$$V_\eta^*(x) = \frac{1}{\eta} \log \sum_a \exp\left(\eta\left(r(x,a) - \rho_\eta^* + \sum_y P(y|x,a)\, V_\eta^*(y)\right)\right)$$

[Marcus et al., 1997, Ruszczyński, 2010, Ziebart et al., 2010, Ziebart, 2010, Braun et al., 2011, Azar et al., 2012, Rawlik et al., 2012, Fox et al., 2016, Asadi and Littman, 2017, Haarnoja et al., 2017, Schulman et al., 2017, Nachum et al., 2017] ...and who knows how many more NIPS'17 submissions

Idea 2: Maximize a regularized objective!

$$\rho(\mu) = \langle \mu, r \rangle$$

A recent trend: (Entropy-)Regularized RL

Two popular approaches

Idea 1: Soften the max in the Bellman optimality equations!

$$V_\eta^*(x) = \frac{1}{\eta} \log \sum_a \exp\left(\eta \left(r(x,a) - \rho_\eta^* + \sum_y P(y|x,a) V_\eta^*(y) \right)\right)$$

[Marcus et al., 1997, Ruszczyński, 2010, Ziebart et al., 2010, Ziebart, 2010, Braun et al., 2011, Azar et al., 2012, Rawlik et al., 2012, Fox et al., 2016, Asadi and Littman, 2017, Haarnoja et al., 2017, Schulman et al., 2017, Nachum et al., 2017] ...and who knows how many more NIPS'17 submissions

Idea 2: Maximize a regularized objective!

$$\rho_\eta(\mu) = \langle \mu, r \rangle - \frac{1}{\eta} R(\mu)$$

[Peters et al., 2010, Montgomery and Levine, 2016, Schulman et al., 2015, Mnih et al., 2016, O'Donoghue et al., 2017]

A recent trend: (Entropy-)Regularized RL

Two popular approaches

Idea 1: Soften the max in the Bellman optimality equations!

$$V^*_\eta(x) = \frac{1}{\eta}\log\sum \exp\left(\eta\left(r(x,a) - \rho^* + \sum P(y|x,a)V^*(y)\right)\right)$$

[Marcus ... , Azar
et al., 20 ... 017,
Schulma ... issions

Idea ...

$$\rho_\eta(\mu) = \langle\mu, r\rangle - \frac{1}{\eta}R(\mu)$$

[Peters et al., 2010, Montgomery and Levine, 2016, Schulman et al., 2015, Mnih et al., 2016, O'Donoghue et al., 2017]

Numerous open questions:

- are these approaches connected?
- do the derived algorithms converge anywhere?
- does a solution even exist?

A unified framework for entropy-regularized MDPs

Neu, Jonsson and Gómez (2017)

Primal LP

$$\rho^* = \max_{\mu \in \Delta} \langle \mu, r \rangle$$

Dual "LP"

$$V^*(x) = \max_a \left(r(x,a) - \rho^* + \sum_y P(y|x,a) V^*(y) \right) \quad (\forall x, a)$$

A unified framework for entropy-regularized MDPs

Neu, Jonsson and Gómez (2017)

Primal convex program

$$\rho^*_\eta = \max_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta} R(\mu) \right)$$

Dual "convex program"

$$V^*_\eta(x) = \frac{1}{\eta} \log \sum_a \exp \left(\eta \left(r(x,a) - \rho^*_\eta + \sum_y P(y|x,a) V^*_\eta(y) \right) \right)$$

A unified framework for entropy-regularized MDPs

Neu, Jonsson and Gómez (2017)

Primal convex program

$$\rho^*_\eta = \max_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta} R(\mu) \right)$$

$R(\mu) = ???$

Dual "convex program"

$$V^*_\eta(x) = \frac{1}{\eta} \log \sum_a \exp \left(\eta \left(r(x,a) - \rho^*_\eta + \sum_y P(y|x,a) V^*_\eta(y) \right) \right)$$

Conditional entropy regularization

Neu, Jonsson and Gómez (2017)

Theorem

The two convex programs are connected by Lagrangian duality with the choice

$$
\begin{aligned}
R(\mu) &= \sum_{x,a} \mu(x,a) \log \frac{\mu(x,a)}{\sum_b \mu(x,b)} \\
&= \sum_{x,a} \mu(x,a) \log \pi_\mu(a|x)
\end{aligned}
$$

Conditional entropy regularization

Neu, Jonsson and Gómez (2017)

Theorem

The two convex programs are connected by Lagrangian duality with the choice

$$R(\mu) = \sum_{x,a} \mu(x,a) \log \frac{\mu(x,a)}{\sum_b \mu(x,b)} = \sum_{x,a} \mu(x,a) \log \pi_\mu(a|x)$$

Lemma

The conditional entropy $R(\mu)$ is convex in μ and the associated Bregman divergence is

$$D\left(\mu \| \mu'\right) = \sum_{x,a} \mu(x,a) \log \frac{\pi_\mu(a|x)}{\pi_{\mu'}(a|x)} \geq 0.$$

A unified framework for entropy-regularized MDPs

Neu, Jonsson and Gómez (2017)

Primal convex program

$$\rho_\eta^* = \max_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta} R(\mu) \right)$$

Dual "convex program"

$$V_\eta^*(x) = \frac{1}{\eta} \log \sum_a \exp \left(\eta \left(r(x,a) - \rho_\eta^* + \sum_y P(y|x,a) V_\eta^*(y) \right) \right)$$

A unified framework for entropy-regularized MDPs

Neu, Jonsson and Gómez (2017)

Primal convex program

$$\rho_\eta^* = \max_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta} \sum_{x,a} \mu(x,a) \log \pi_\mu(a|x) \right)$$

Dual "convex program"

$$V_\eta^*(x) = \frac{1}{\eta} \log \sum_a \exp\left(\eta \left(r(x,a) - \rho_\eta^* + \sum_y P(y|x,a) V_\eta^*(y) \right) \right)$$

A unified framework for entropy-regularized MDPs

Neu, Jonsson and Gómez (2017)

Primal convex program

Immediate consequences:

- existence & uniqueness results
- well-defined contractive DP operators
- policy gradient theorems...

$$V_\eta^*(x) = \frac{1}{\eta} \log \sum_a \exp\left(\eta\left(r(x,a) - \rho_\eta^* + \sum_y P(y|x,a)\,V_\eta^*(y)\right)\right)$$

A unified algorithmic framework

Neu, Jonsson and Gómez (2017)

A unified algorithmic framework

Neu, Jonsson and Gómez (2017)

Every algorithm is either Mirror Descent or Dual Averaging!

A unified algorithmic framework

Neu, Jonsson and Gómez (2017)

Every algorithm is either Mirror Descent or Dual Averaging!

- provides a common analytic framework
- ensures convergence
- explains numerous recent algorithms

Example 1:

Trust-region policy optimization $\approx$ Mirror Descent

Neu, Jonsson and Gómez (2017)

Mirror descent

$$\mu_{t+1} = \arg\max_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta} D\left(\mu \| \mu_t\right) \right)$$

Example 1:

Trust-region policy optimization $\approx$ Mirror Descent

Neu, Jonsson and Gómez (2017)

Mirror descent

$$\mu_{t+1} = \arg\max_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta} D\left(\mu \| \mu_t\right) \right)$$

Trust-Region Policy Optimization [Schulman et al., 2015]:

$$D_{\mathrm{TRPO}}\left(\mu \| \mu_{\mathrm{old}}\right) = \sum_{x,a} \nu_{\mathrm{old}}(x) \pi_\mu(a|x) \log \frac{\pi_\mu(a|x)}{\pi_{\mathrm{old}}(a|x)}$$

Example 1:

Trust-region policy optimization $\approx$ Mirror Descent

Neu, Jonsson and Gómez (2017)

Mirror descent

$$\mu_{t+1} = \arg\max_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta} D\left(\mu \| \mu_t\right) \right)$$

Trust-Region Policy Optimization [Schulman et al., 2015]:

$$\begin{aligned} D_{\mathrm{TRPO}}\left(\mu \| \mu_{\mathrm{old}}\right) &= \sum_{x,a} \nu_{\mathrm{old}}(x) \pi_\mu(a|x) \log \frac{\pi_\mu(a|x)}{\pi_{\mathrm{old}}(a|x)} \\ &\approx \sum_{x,a} \nu_\mu(x) \pi_\mu(a|x) \log \frac{\pi_\mu(a|x)}{\pi_{\mathrm{old}}(a|x)} = D\left(\mu \| \mu_{\mathrm{old}}\right) \end{aligned}$$

Example 1:
Trust-region policy optimization $\approx$ Mirror Descent

Neu, Jonsson and Gómez (2017)

Mirror descent

$$\mu_{t+1} = \arg\max_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta} D\left(\mu \| \mu_t\right) \right)$$

Trust-Region Policy Optimization [Schulman et al., 2015]:

$$D_{\mathrm{TRPO}}\left(\mu \| \mu_{\mathrm{old}}\right) = \sum_{x,a} \nu_{\mathrm{old}}(x) \pi_\mu(a|x) \log \frac{\pi_\mu(a|x)}{\pi_{\mathrm{old}}(a|x)}$$

Corollary

TRPO converges to the optimal policy!

Example 2:

A3C $\approx$ Dual Averaging

Neu, Jonsson and Gómez (2017)

Dual Averaging

$$\mu_{t+1} = \operatorname*{arg\,max}_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta_t} R(\mu) \right)$$

Example 2:

A3C $\approx$ Dual Averaging

Neu, Jonsson and Gómez (2017)

Dual Averaging

$$\mu_{t+1} = \arg\max_{\mu\in\Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta_t} R(\mu) \right)$$

"A3C" [Mnih et al., 2016, O'Donoghue et al., 2017]:

$$R_{\mathrm{A3C}}(\mu) = \sum_{x,a} \nu_{\mathrm{old}}(x) \pi_\mu(a|x) \log \pi_\mu(a|x)$$

Example 2:
A3C $\approx$ Dual Averaging

Neu, Jonsson and Gómez (2017)

Dual Averaging

$$\mu_{t+1} = \operatorname*{arg\,max}_{\mu\in\Delta}\left(\langle \mu, r\rangle - \frac{1}{\eta_t} R(\mu)\right)$$

"A3C" [Mnih et al., 2016, O'Donoghue et al., 2017]:

$$R_{\mathrm{A3C}}(\mu) = \sum_{x,a} \nu_{\mathrm{old}}(x)\pi_\mu(a|x)\log \pi_\mu(a|x)$$

$$\approx \sum_{x,a} \nu_\mu(x)\pi_\mu(a|x)\log \pi_\mu(a|x) = R(\mu)$$

Example 2:
A3C $\approx$ Dual Averaging

Neu, Jonsson and Gómez (2017)

Dual Averaging

$$\mu_{t+1} = \arg\max_{\mu \in \Delta} \left(\langle \mu, r \rangle - \frac{1}{\eta_t} R(\mu) \right)$$

"A3C" [Mnih et al., 2016, O'Donoghue et al., 2017]:

$$R_{\mathrm{A3C}}(\mu) = \sum_{x,a} \nu_{\mathrm{old}}(x) \pi_{\mu}(a|x) \log \pi_{\mu}(a|x)$$

Divergence alert!!!

A3C optimizes a non-stationary and non-convex objective!

Example 2:
A3C $\approx$ Dual Averaging

Neu, Jonsson and Gómez (2017)

Patching A3C:

- O'Donoghue et al. [2017] characterize the stationary points of A3C, but do not show its existence or that A3C would converge to this fixed point

Example 2:
A3C $\approx$ Dual Averaging
Neu, Jonsson and Gómez (2017)

Patching A3C:

- O'Donoghue et al. [2017] characterize the stationary points of A3C, but do not show its existence or that A3C would converge to this fixed point
- Our theory provides a closed-form expression for the regularized policy gradient: just replace the advantage function $A^{\pi}(x, a)$ by

$$A_{\eta}^{\pi}(x, a) = r(x, a) - \frac{1}{\eta} \log \pi(a|x) + \sum_{y} P(y|x, a) V_{\eta}^{\pi}(y) - V_{\eta}^{\pi}(x)$$

Other algorithms in our framework

Neu, Jonsson and Gómez (2017)

Mirror Descent:

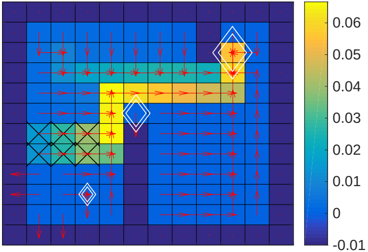
- Dynamic Policy Programming [Azar et al., 2012], Ψ-learning [Rawlik et al., 2012]
- Relative Entropy Policy Search [Peters et al., 2010, Zimin and Neu, 2013, Montgomery and Levine, 2016]

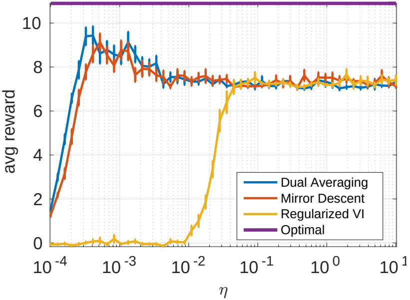
Dual Averaging:

- "MellowMax" RL algorithms of [Asadi and Littman, 2017], G-learning [Fox et al., 2016]
- "Energy-based policy search" [Haarnoja et al., 2017]
- "Path consistency learning" [Nachum et al., 2017]

Experiments

Neu, Jonsson and Gómez (2017)



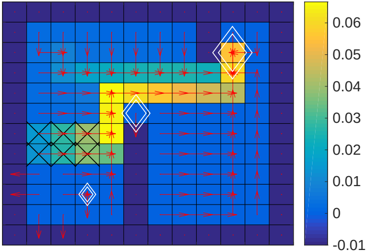


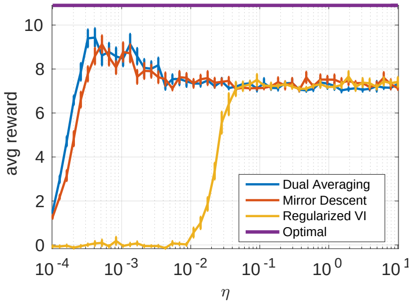
"Regularization curve":

- η too large: convergence to suboptimal goal $\leftrightarrow$ overfitting
- η too small: policy too close to uniform $\leftrightarrow$ underfitting

Experiments

Neu, Jonsson and Gómez (2017)





"Regularization curve":

- η too large: convergence to suboptimal goal $\leftrightarrow$ overfitting
- η too small: policy too close to uniform $\leftrightarrow$ underfitting

Dual Averaging perspective seems essential!

- DA theory suggests $\eta_t = t \cdot \eta_0$
- Regularized Value Iteration with constant η is bad

Outlook

Can regularization provide a useful perspective on exploration?

- "Exploration" integrated in the foundations: regularized Bellman equations
- convex optimization framework provides analysis tools and algorithmic templates

Outlook

Can regularization provide a useful perspective on exploration?

- "Exploration" integrated in the foundations: regularized Bellman equations
- convex optimization framework provides analysis tools and algorithmic templates
- BUT: no clear understanding about the statistical benefits of regularization

Outlook

Can regularization provide a useful perspective on exploration?

- "Exploration" integrated in the foundations: regularized Bellman equations
- convex optimization framework provides analysis tools and algorithmic templates
- BUT: no clear understanding about the statistical benefits of regularization

The way towards more effective algorithms?

References I

K. Asadi and M. L. Littman. A new softmax operator for reinforcement learning. *ICML*, 2017.

M. G. Azar, V. Gómez, and H. J. Kappen. Dynamic policy programming. *Journal of Machine Learning Research*, 13(Nov):3207–3245, 2012.

D. A. Braun, P. A. Ortega, E. Theodorou, and S. Schaal. Path integral control and bounded rationality. In *Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2011 IEEE Symposium on*, pages 202–209. IEEE, 2011.

R. Fox, A. Pakman, and N. Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016.

T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. *CoRR*, abs/1702.08165, 2017.

S. I. Marcus, E. Fernández-Gaucherand, D. Hernández-Hernandez, S. Coraluppi, and P. Fard. Risk sensitive markov decision processes. In *Systems and control in the twenty-first century*, pages 263–279. Springer, 1997.

References II

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

W. H. Montgomery and S. Levine. Guided policy search via approximate mirror descent. In *NIPS-29*, pages 4008–4016, 2016.

O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans. Bridging the gap between value and policy based reinforcement learning. *CoRR*, abs/1702.08892, 2017.

B. O'Donoghue, R. Munos, K. Kavukcuoglu, and V. Mnih. PGQ: Combining policy gradient and Q-learning. In *5th International Conference on Learning Representations*, 2017.

J. Peters, K. Mülling, and Y. Altun. Relative entropy policy search. In *AAAI 2010*, pages 1607–1612, 2010. ISBN 978-1-57735-463-5.

K. Rawlik, M. Toussaint, and S. Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Proceedings of Robotics: Science and Systems VIII*, 2012.

A. Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

References III

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1889–1897, 2015.

J. Schulman, P. Abbeel, and X. Chen. Equivalence between policy gradients and soft Q-learning. *CoRR*, abs/1704.06440, 2017.

B. D. Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, 2010.

B. D. Ziebart, J. A. Bagnell, and A. K. Dey. Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning (ICML)*, pages 1247–1254, 2010.

A. Zimin and G. Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In *NIPS-26*, pages 1583–1591, 2013.