CS4400
DEEP REINFORCEMENT LEARNING

Lecture 9: Offline RL

Wendelin Böhmer

<j.w.bohmer@tudelft.nl>

TU Delft

19th of December 2023

Content of this lecture

9.1

Offline RL

Deep exploration

9.1 Deep exploration

- Thompson sampling and optimism are often too "local"
  - local value predictions can be *certain* and *wrong*
  - + explores only immediate consequences
  - - ignores uncertainty of future rewards

- How can we express *long-term* future uncertainty?

9.1 Deep exploration

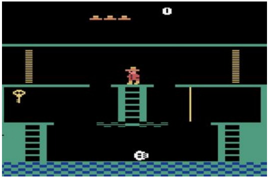
- Thompson sampling and optimism are often too "local"
  - local value predictions can be *certain* and *wrong*
  - \+ explores only immediate consequences
  - \- ignores uncertainty of future rewards

- How can we express *long-term* future uncertainty?
  - PAI/SAC *rewarded* policy entropy
  - incentives exploration of far away places

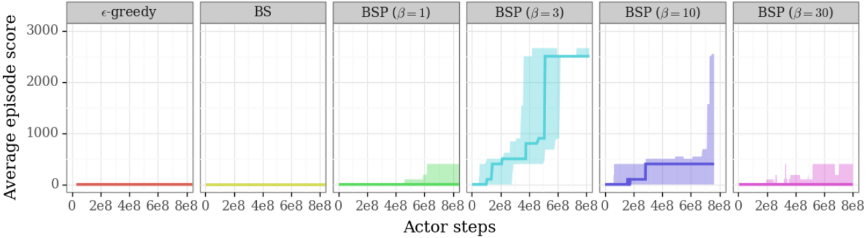
⇒ sample or learn long-term uncertainty

9.1 Episode-wise sampling

- One Thompson sample per episode
  - draw one q_θ from ensemble/posterior
  - follow q_θ until end of episode
  - different q_θ lead to different states
  - diverse q_θ explore well



- Consequent long-term exploration without propagated uncertainty



results on *Montezuma's Revenge* from Osband et al. (2018, BSP refers to 'Bootstrap with prior functions', β denotes prior scales)

9.1 Intrinsic rewards

- Add some exploration bonus $\eta(s_t, a_t)$ to reward
  - e.g., policy entropy in PAI/SAC
  - e.g., standard deviation of q_θ from noisy-net/dropout/ensemble
  - e.g., inverse square-root of pseudo or hash visitation counts
  - e.g., novelty measures like random network distillation
  - or many other models of local uncertainty or novelty

$$\bar{r}_t \;:=\; r_t + C\,\eta(s_t, a_t) \qquad \text{or} \qquad \bar{r}_t \;:=\; r_t + C\eta(s_{t+1})$$

- Deep exploration works if bonus decays to zero
  - unexplored states are only initially attractive
  - theoretical guarantees for tabular Q-learning
  - finding the right scale C is tricky
  - can be implemented with two value heads

assignment sheet 4

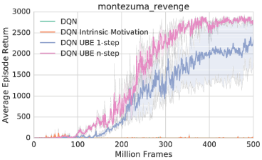
Jin et al. (2018) prove regret bounds for tabular counts, and Rashid et al. (2020) demonstrated the use of random hash counts

9.1 Propagating uncertainty

- Intrinsic reward poor substitute for “future uncertainty”
- Uncertainty Bellman equation (UBE)
  - propagates “local uncertainty” $\eta(s,a)$ through Markov chain
  - $\eta(s,a)$ depends on epistemic reward and transition variance
  - propagated uncertainty $U^\pi(s,a)$ is upper bound to variance

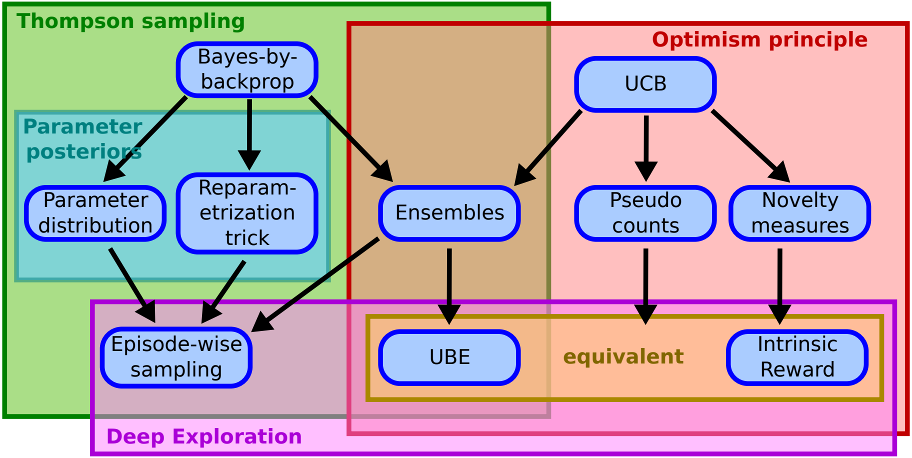
$$\mathbb{V}[Q^\pi(s,a)] \;\leq\; U^\pi(s,a) \;:=\; \eta(s,a) + \gamma^2 \mathbb{E}\Big[U^\pi(s',a') \,\Big|\, \substack{s' \sim P(\cdot|s,a) \\ a' \sim \pi(\cdot|s')}\Big]$$

- Thompson sampling $\sim \mathcal{N}\big(\cdot \,\big|\, Q^\pi(s,a), U^\pi(s,a)\big)$
  - learn $U^\pi(s,a)$ as additional heads of DQN
  - propagation speed very important
  - e.g. use n-steps targets for U^π
- Almost identical to intrinsic reward!



results and UBE definintion can be found in O’Donoghue et al. (2018)

9.1 Overview deep exploration



9.1 Summary

- Optimism/uncertainty must be propagated for deep exploration
- Intrinsic reward adds a novelty/uncertainty bonus
- UBE propagates variance/uncertainty of future rewards
- Both yield almost the exact same equations!

Learning Objectives

LO9.1: Explain how optimism/uncertainty can be propagated
LO9.2: Implement intrisic reward and RND

9.2

Offline RL

Offline learning

9.2 Offline reinforcement learning

- On-policy RL directly interacts with environment
  - stable for small policy updates, but sample inefficient
- Off-policy RL reuses old interactions but regularly samples new
  - more sample efficient, must be stabilized with tricks
  - destabilizes if #updates $\gg$ #online samples
- Offline RL cannot sample new trajectories
  - off-policy objectives quickly diverge

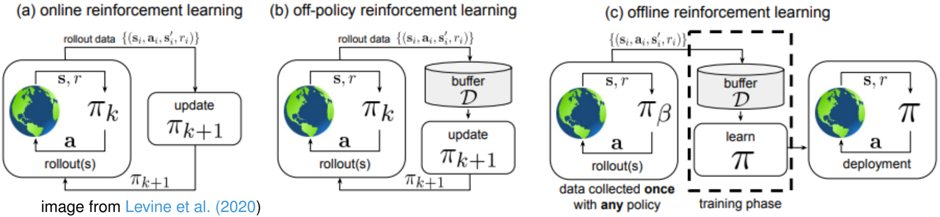


image from Levine et al. (2020)

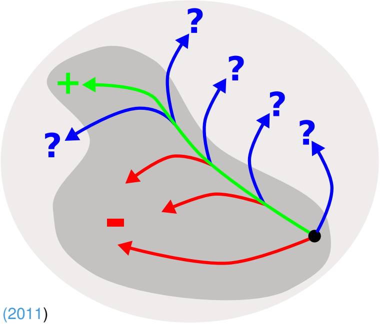
9.2 Main challenges of offline RL

- When sampling is expensive or dangerous
  - robotics, autonomous cars, healthcare, recommender systems
- Algorithms are provided with a *static* dataset
  - a.k.a. batch RL: $\mathcal{D} = \{(s_t, a_t, r_t, s'_t)\}_{t=1}^n$
  - $a_t \sim \pi_\beta(\cdot|s_t)$ sampled from *behavior policy* π_β,
  - $s_t \sim \xi_t^{\pi_\beta}(\cdot)$ sampled from induced state-distribution $\xi_t^{\pi_\beta}$
- Main challenges in offline RL:
  - *no exploration:* unknown state-actions remain unknown
  - *distribution shift:* $\pi_\theta \neq \pi_\beta$ and $\xi_t^{\pi_\theta} \neq \xi_t^{\pi_\beta}$
  - *learning stability:* errors cannot be detected/corrected

for an overview see Levine et al. (2020) or Lange et al. (2012)

9.2 Distribution shift bounds

- What is the value of actions that leave the training set?
  - one wrong decision can ruin an otherwise optimal policy
  - but how bad is it if the optimal solution is in $\mathcal{D}$?

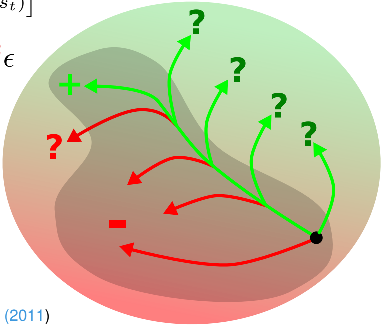


first bound from Ross and Bagnell (2010), second bound from Ross et al. (2011)

Distribution shift bounds

- What is the value of actions that leave the training set?
  - one wrong decision can ruin an otherwise optimal policy
  - but how bad is it if the optimal solution is in $\mathcal{D}$?
- Behavioral cloning (offline) error bound:
  - offline data $s_t \sim d^{\pi_\beta}(\cdot)$ with optimal actions a_t^* and horizon H
  - small error $\epsilon = \mathbb{E}_{\mathcal{D}}\Big[\delta(a_t \neq a_t^*)\Big|_{a_t \sim \pi_\theta(\cdot|s_t)}\Big]$

$$\mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{H} \delta(a_t \neq a_t^*)\Big|_{\substack{a_t \sim \pi_\theta(\cdot|s_t) \\ a_t^* \sim \pi^*(s_t)}}\Big] \leq C + H^2 \epsilon$$



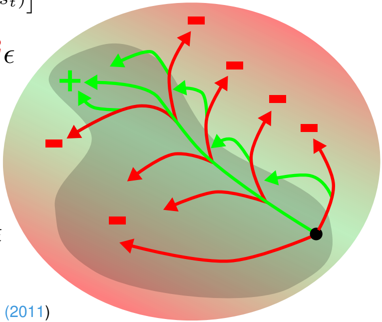
first bound from Ross and Bagnell (2010), second bound from Ross et al. (2011)

9.2 Distribution shift bounds

- What is the value of actions that leave the training set?
  - one wrong decision can ruin an otherwise optimal policy
  - but how bad is it if the optimal solution is in $\mathcal{D}$?

- Behavioral cloning (offline) error bound:
  - offline data $s_t \sim d^{\pi_\beta}(\cdot)$ with optimal actions a_t^* and horizon H
  - small error $\epsilon = \mathbb{E}_{\mathcal{D}}\Big[\delta(a_t \neq a_t^*)\Big|_{a_t \sim \pi_\theta(\cdot|s_t)}\Big]$

$$\mathbb{E}_{\pi_\theta}\Bigg[\sum_{t=0}^{H}\delta(a_t \neq a_t^*)\Bigg|_{\substack{a_t \sim \pi_\theta(\cdot|s_t)\\ a_t^* \sim \pi^*(s_t)}}\Bigg] \leq C + H^2\epsilon$$



- DAgger (online) error bound:
  - online data $s_t \sim d^{\pi_\theta}(\cdot)$

$$\mathbb{E}_{\pi_\theta}\Bigg[\sum_{t=0}^{H}\delta(a_t \neq a_t^*)\Bigg|_{\substack{a_t \sim \pi_\theta(\cdot|s_t)\\ a_t^* \sim \pi^*(s_t)}}\Bigg] \leq C + H\epsilon$$

first bound from Ross and Bagnell (2010), second bound from Ross et al. (2011)

9.2 Summary

- Offline learning differs from online learning
- No way to correct errors
- Offline learning unstable, due to distribution shift
- Distribution shift can be bounded online and offline

Learning Objectives

LO9.3: Explain how offline RL differs from online RL

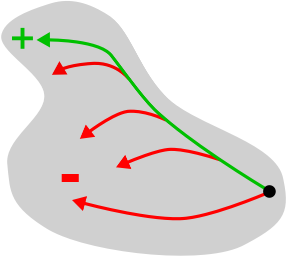
9.3

Offline RL

Offline RL approaches

Policy-Constrained Methods

- Distribution shift in $\xi_t^\pi(s)$ and $\pi_\theta(a|s) \neq \pi_\beta(a|s)$ can be jointly mitigated by restricting divergence of $\pi_\theta(a|s)$ from $\pi_\beta(a|s)$

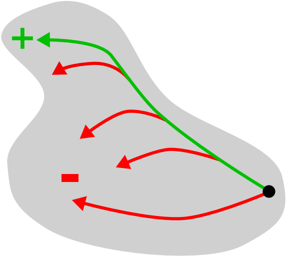


9.3 Policy-Constrained Methods

- Distribution shift in $\xi_t^\pi(s)$ and $\pi_\theta(a|s) \neq \pi_\beta(a|s)$ can be jointly mitigated by restricting divergence of $\pi_\theta(a|s)$ from $\pi_\beta(a|s)$

- TRPO with a learned reference behavior policy $\hat{\pi}_\beta$ can be a natural choice for learning π_θ in an offline setting:

$$\max_\theta \mathbb{E}\left[\frac{\pi_\theta(a|s)}{\hat{\pi}_\beta(a|s)}\big(Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)\big)\,\Big|\, s,a\sim\mathcal{D}\right]$$
$$\text{s.t. } \mathbb{E}\Big[D_{KL}[\hat{\pi}_\beta(\cdot|s)\|\pi_\theta(\cdot|s)] \leq \delta \,\Big|\, s\sim\mathcal{D}\Big]$$



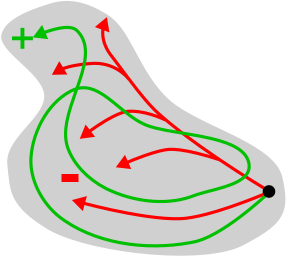
TRPO has been introduced in Lecture 6.2

9.3 Policy-Constrained Methods

- Distribution shift in $\xi_t^\pi(s)$ and $\pi_\theta(a|s) \neq \pi_\beta(a|s)$ can be jointly mitigated by restricting divergence of $\pi_\theta(a|s)$ from $\pi_\beta(a|s)$

- TRPO with a learned reference behavior policy $\hat{\pi}_\beta$ can be a natural choice for learning π_θ in an offline setting:

$$\max_\theta \mathbb{E}\left[\frac{\pi_\theta(a|s)}{\hat{\pi}_\beta(a|s)}\Big(Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)\Big)\,\Big|\, s,a\sim\mathcal{D}\right]$$
$$\text{s.t. } \mathbb{E}\left[D_{KL}[\hat{\pi}_\beta(\cdot|s)\|\pi_\theta(\cdot|s)] \leq \delta \,\Big|\, s\sim\mathcal{D}\right]$$



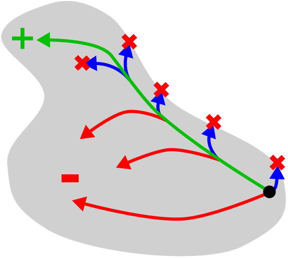
- Poor performance if π_β is highly sub-optimal
  - even if $\mathcal{A}$ is covered well (e.g. by uniform π_β)

TRPO has been introduced in Lecture 6.2

9.3 Restricting available actions

- Restrict learned policies to Π_ϵ with non-zero action probability only in the support of the empirical action distribution

$$\Pi_\epsilon = \{\pi | \pi(a|s) = 0, \forall a \text{ where } \pi_\beta(a|s) < \epsilon\}$$



9.3 Restricting available actions

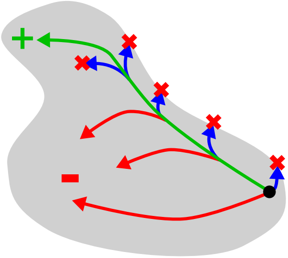
- Restrict learned policies to Π_ϵ with non-zero action probability only in the support of the empirical action distribution

$$\Pi_\epsilon = \{\pi | \pi(a|s) = 0, \forall a \text{ where } \pi_\beta(a|s) < \epsilon\}$$

- Express Π_ϵ in terms of states and actions in $\mathcal{D}$
  - distance measure $d(a, a')$ between actions
  - distance measure $d(s, s')$ between states
  - set $\mathcal{A}_\mathcal{D}(s) := \{a_t \,|\, (s_t, a_t) \in \mathcal{D},\, d(s, s_t) \leq \epsilon'\}$

$$\max_\theta \mathbb{E}\Big[Q^{\pi_\theta}(s,a)\Big|_{a\sim\pi_\theta(\cdot|s)}^{s\sim\mathcal{D}}\Big]$$

$$\text{s.t. } \mathbb{E}\Big[\min_{a'\in\mathcal{A}_\mathcal{D}(s)} d(a,a') \leq \epsilon \Big|_{a\sim\pi_\theta(\cdot|s)}^{s\sim\mathcal{D}}\Big]$$

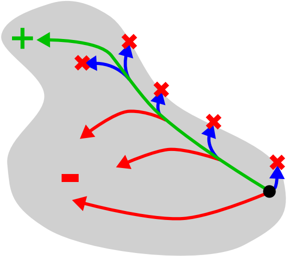


- Which distance measures d are working?

Bootstrapping error accumulation reduction

- Sample-based *maximum mean discrepancy* (MMD)
  - distance between mean embeddings in kernel Hilbert space $\mathcal{H}_\kappa$

$$\mathrm{MMD}^2\big(\{a_i\}_{i=1}^m, \{a'_i\}_{j=1}^m\big) := \frac{1}{m^2}\sum_{i,j=1}^m \Big(\kappa(a_i, a_j) - 2\,\kappa(a_i, a'_j) + \kappa(a'_i, a'_j)\Big)$$



see Kumar et al. (2019) for details

9.3 Bootstrapping error accumulation reduction

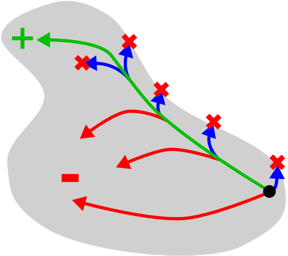
- Sample-based *maximum mean discrepancy* (MMD)
  - distance between mean embeddings in kernel Hilbert space $\mathcal{H}_\kappa$

$$\mathrm{MMD}^2\big(\{a_i\}_{i=1}^m, \{a'_i\}_{j=1}^m\big) := \frac{1}{m^2}\sum_{i,j=1}^{m}\Big(\kappa(a_i, a_j) - 2\,\kappa(a_i, a'_j) + \kappa(a'_i, a'_j)\Big)$$

- Bootstrapping error accumulation reduction (BEAR)
  - uses MMD with Gaussian kernel κ
  - π_θ can strongly diverge from π_β
  - π_θ is roughly restricted to $a \in \mathcal{A}_\mathcal{D}(s)$

$$\max_\theta \; \mathbb{E}\left[Q^{\pi_\theta}(s,a) \,\middle|\, \begin{smallmatrix} s\sim\mathcal{D} \\ a\sim\pi_\theta(\cdot|s)\end{smallmatrix}\right]$$

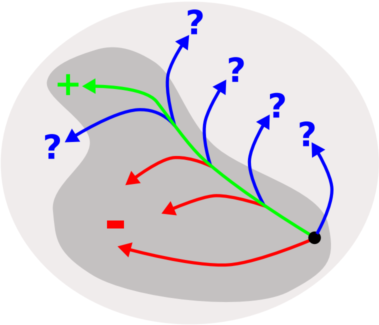
$$\text{s.t. } \mathbb{E}\left[\mathrm{MMD}^2(\{a_i\},\{a'_i\}) \leq \epsilon \,\middle|\, \begin{smallmatrix} s\sim D \\ a_i\sim\pi_\theta(\cdot|s) \\ a'_i\sim\mathcal{A}_\mathcal{D}(s)\end{smallmatrix}\right]$$



see Kumar et al. (2019) for details

9.3 Uncertainty penalized value estimation

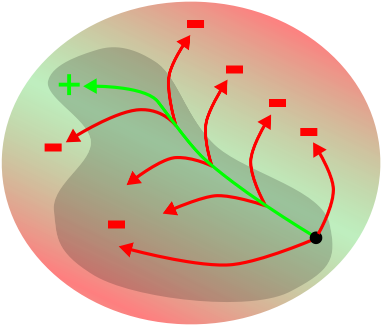
- Out-of-distribution actions should have high epistemic uncertainty



see for example (Osband et al., 2016; Kumar et al., 2019; Eysenbach et al., 2017; Agarwal et al., 2020)

9.3 Uncertainty penalized value estimation

- Out-of-distribution actions should have high epistemic uncertainty
- Penalize out-of-distribution actions $a \notin \mathcal{D}$
  - opposite of exploration: stay away of the unknown
  - can use all epistemic uncertainty estimation methods



see for example (Osband et al., 2016; Kumar et al., 2019; Eysenbach et al., 2017; Agarwal et al., 2020)

9.3 Uncertainty penalized value estimation

- Out-of-distribution actions should have high epistemic uncertainty

- Penalize out-of-distribution actions $a \notin \mathcal{D}$
  - opposite of exploration: stay away of the unknown
  - can use all epistemic uncertainty estimation methods

- Pessimism/conservatism: underestimate values that leave $\mathcal{D}$
  - e.g. pessimistic offline DQN with an ensemble of Q-values $\{Q_{\theta_i}\}_{i=1}^m$

a) select the worst possible values $\underline{Q}(s,a) := \min_i Q_{\theta_i}(s,a)$ (see TD3)

b) punish uncertain actions $\underline{Q}(s,a) := \underbrace{\mathbb{E}[Q_{\theta_i}(s,a)]}_{\text{or } Q_{\theta_i}(s,a)} - \alpha\sqrt{\mathbb{V}[Q_{\theta_i}(s,a)]}$

$$\mathcal{L}_{[\{\theta_i\}_{i=1}^m]} \quad := \quad \mathbb{E}_{\mathcal{D}}\Big[\sum_{i=1}^m \big(r_t \;+\; \gamma \max_a \underline{Q}(s_{t+1},a) \;-\; Q_{\theta_i}(s_t,a_t)\big)^2\Big]$$

see for example (Osband et al., 2016; Kumar et al., 2019; Eysenbach et al., 2017; Agarwal et al., 2020)

9.3 Summary

- Constraining the policy divergence has poor performance
- Restricting actions works, but hard for continuous actions
- Penalizing uncertain actions easy and effective

Learning Objectives

LO9.4: Explain policy-constraints, action-restriction and -penalization

Next lecture

- Next lecture: **multi-agent RL**!
- Submit **assignment sheet 3** until Thursday!
- Questions? Ask them here: `answers.ewi.tudelft.nl`

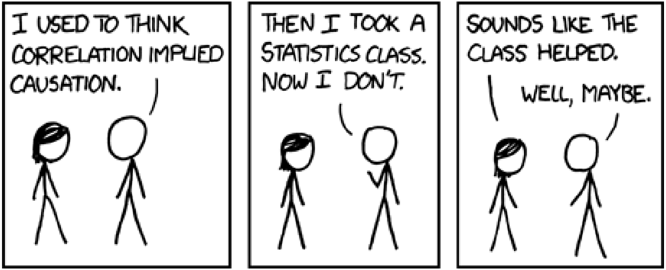


image source: `xkcd.com`

References I

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/1907.04543.

Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning. *arXiv:1711.06782 [cs]*, November 2017. URL https://arxiv.org/abs/1711.06782.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4863–4873, 2018. URL http://arxiv.org/abs/1807.03765.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.00949.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch Reinforcement Learning. In Marco Wiering and Martijn van Otterlo, editors, *Reinforcement Learning*, volume 12, pages 45–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-27644-6 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_2.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020. URL https://arxiv.org/abs/2005.01643.

Brendan O'Donoghue, Ian Osband, Rémi Munos, and Volodymyr Mnih. The uncertainty Bellman equation and exploration. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3836–3845, 2018. URL https://arxiv.org/abs/1709.05380.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, pages 2377–2386, 2016. URL https://arxiv.org/abs/1402.0635.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS) 31*, pages 8617–8629. 2018. URL https://arxiv.org/abs/1806.03335.

References II

Tabish Rashid, Bei Peng, Wendelin Böhmer, and Shimon Whiteson. Optimistic exploration even with a pessimistic initialisation. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/2002.12174.

Stephane Ross and Drew Bagnell. Efficient Reductions for Imitation Learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 661–668. JMLR Workshop and Conference Proceedings, March 2010. URL https://proceedings.mlr.press/v9/ross10a.html.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. *arXiv:1011.0686 [cs, stat]*, March 2011. URL https://arxiv.org/abs/1011.0686.