CS4400
DEEP REINFORCEMENT LEARNING

Lecture 8: Exploration

Wendelin Böhmer
<j.w.bohmer@tudelft.nl>

TU Delft

14th of December 2023

Content of this lecture

8.1

Exploration

Exploration

Recap: exploration so far

- Uninformed ϵ-greedy exploration

$$\pi_\theta^\epsilon(a|s) \quad := \quad (1-\epsilon)\,\pi_\theta(a|s) + \epsilon\,\frac{1}{|\mathcal{A}|}$$

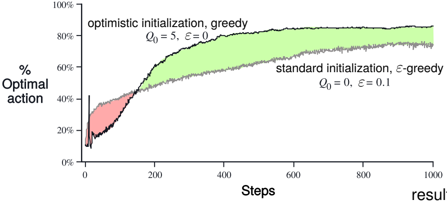
- Uninformed Boltzmann exploration

$$\pi_\theta^\beta(a|s) \quad := \quad \frac{\exp(\beta Q_\theta(s,a))}{\sum_{a'} \exp(\beta Q_\theta(s,a'))}$$

- Maximum entropy regularization/reward

$$\mathcal{L}^\alpha[\theta] \quad := \quad \mathcal{L}[\theta] \;+\; \alpha\,\mathbb{E}\Big[\ln \pi_\theta(a|s)\,\Big|\, a \sim \pi_\theta(\cdot|s)\Big]$$

$$r_t^\alpha \quad := \quad r(s_t, a_t) \;-\; \alpha\,\ln \pi(a_t|s_t)$$

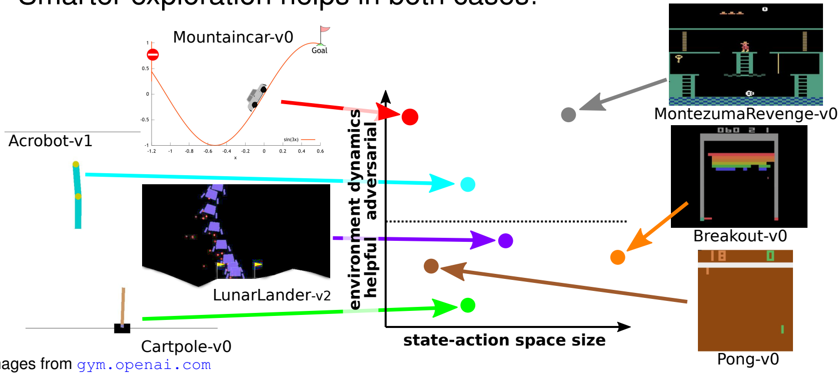


$P(r\,|\,\mathbf{a}_1) = 0.2 \qquad P(r\,|\,\mathbf{a}_2) = 0.8$

results for a 2-armed bandit task from Sutton and Barto (2018)

8.1 What makes exploration hard?

- Large state-action spaces $\rightarrow$ more episodes/generalization
  - random exploration is normal distributed, e.g. in navigation
  - many similar states easy to approximate, e.g. in Breakout
- Adversarial dynamics $\rightarrow$ more/longer episodes
- Smarter exploration helps in both cases!



images from gym.openai.com

8.1 How can we explore in a smart way?

- Choose actions with high observed reward more often
  - ϵ-greedy and Boltzman exploration
  - \+ exploration "around" exploitation policy
  - \- over-commits to easily reachable reward

this article by Lilian Weng gives a good overview over various exploration methods in deep RL

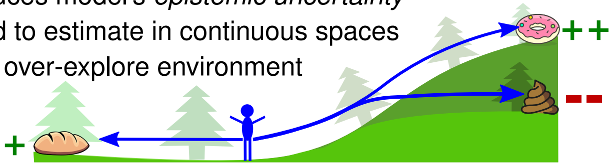
8.1 How can we explore in a smart way?

- Choose actions with high observed reward more often
  - ϵ-greedy and Boltzman exploration
  - \+ exploration “around” exploitation policy
  - \- over-commits to easily reachable reward
- Choose actions with uncertain returns
  - \+ varying returns $\rightarrow$ uncertain future decision?
  - \- might be irreducible *aleatoric uncertainty*

this article by Lilian Weng gives a good overview over various exploration methods in deep RL

8.1 How can we explore in a smart way?

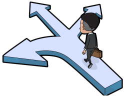
- Choose actions with high observed reward more often
  - ϵ-greedy and Boltzman exploration
  - + exploration “around” exploitation policy
  - - over-commits to easily reachable reward
- Choose actions with uncertain returns
  - + varying returns $\rightarrow$ uncertain future decision?
  - - might be irreducible *aleatoric uncertainty*
- Choose actions you have not tried yet
  - + reduces model’s *epistemic uncertainty*
  - - hard to estimate in continuous spaces
  - - can over-explore environment

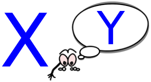


this article by Lilian Weng gives a good overview over various exploration methods in deep RL

8.1 The many faces of uncertainty

- Aleatoric uncertainty
  - stochastic environment, irreducible
- Epistemic uncertainty
  - unknown environment, reducible
- Model-bias
  - wrong model, irreducible





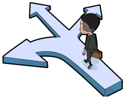
assignment sheet 3

image sources: www.wikipedia.org, www.wikipedia.org, openclipart.org

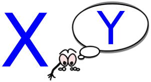
The many faces of uncertainty

- Aleatoric uncertainty
  - stochastic environment, irreducible

- Epistemic uncertainty
  - unknown environment, reducible



- Model-bias
  - wrong model, irreducible



- Example: MSE regression on $\mathcal{D}$, $x \sim \rho(\cdot), y \sim \mathcal{N}\big(\,\cdot\,|\mu(x), \sigma^2(x)\big)$
  - only one possible definition of epistemic uncertainty!

$$\mathbb{E}_{\mathcal{D}}\Big[\underbrace{\mathbb{E}\big[\big(y - f_{\mathcal{D}}(x)\big)^2 \big| \mathcal{D}\big]}_{\text{generalization error}}\Big] = \underbrace{\mathbb{E}[\sigma^2(x)]}_{\text{aleatoric}} + \underbrace{\mathbb{E}\big[\mathbb{V}_{\mathcal{D}}[f_{\mathcal{D}}(x)|x]\big]}_{\text{epistemic}} + \underbrace{\mathbb{E}\big[\big(\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)|x] - \mu(x)\big)^2\big]}_{\text{model-bias}}$$

assignment sheet 3

image sources: www.wikipedia.org, www.wikipedia.org, openclipart.org

8.1 Summary

- Some environments are hard to explore randomly
- Agents should reduce the epistemic uncertainty
- Model-bias and aleatoric uncertainty are irreducible

Learning Objectives

LO8.1: Identify which environments are harder to explore
LO8.2: Explain the different types of uncertainties

8.2

Exploration
Thompson sampling

Thompson sampling

- Bayesian perspective: learn a posterior over models
  - another possible definition of epistemic uncertainty

$$\mathbb{P}(\theta|\mathcal{D}) \quad = \quad \frac{\mathbb{P}(\mathcal{D}|\theta)\,\mathbb{P}(\theta)}{\int \mathbb{P}(\mathcal{D}|\theta)\,\mathbb{P}(\theta)\,d\theta}$$

- Choose actions proportional to their optimality under $\mathbb{P}$

$$\pi(a|s) \quad = \quad \int \mathbb{P}(\theta|\mathcal{D})\,\delta\Big(Q_\theta(s,a) = \max_{a'\in\mathcal{A}} Q_\theta(s,a')\Big)\,d\theta$$

- Efficient implementation by sampling

$$a_t = \arg\max_{a\in\mathcal{A}} Q_\theta(s_t,a)\,, \qquad \theta \sim \mathbb{P}(\cdot|\mathcal{D})$$

assignment sheet 3

see Chapelle and Li (2011) for a recent survey on Thompson sampling (original Thompson, 1933) and Wang and Yeung (2020) for a recent survey on Bayesian deep learning

Question: posteriors over neural networks

- Thompson sampling requires $\mathbb{P}(\theta|\mathcal{D})$
  - how do we represent this posterior with a neural net?
  - how do we sample Q-value functions $q_\theta(s, a)$ from it?
  - try to think out of the box!

Question: posteriors over neural networks

- Thompson sampling requires $\mathbb{P}(\theta|\mathcal{D})$
  - how do we represent this posterior with a neural net?
  - how do we sample Q-value functions $q_\theta(s, a)$ from it?
  - try to think out of the box!

- No spoilers!

- No spoilers!

- No spoilers!

Bayes by backpropagation

- How to train a neural-net posterior $\mathbb{P}(\theta|\mathcal{D})$?
  - no analytical solution for Bayes update of neural network:

$$\mathbb{P}(\theta|\mathcal{D}) \quad = \quad \frac{\mathbb{P}(\mathcal{D}|\theta)\,\mathbb{P}(\theta)}{\int \mathbb{P}(\mathcal{D}|\theta')\,\mathbb{P}(\theta')\,d\theta'}$$

for some more information see Blundell et al. (2015) and Fortunato et al. (2019)

8.2 Bayes by backpropagation

- How to train a neural-net posterior $\mathbb{P}(\theta|\mathcal{D})$?
  - no analytical solution for Bayes update of neural network:

$$\mathbb{P}(\theta|\mathcal{D}) \quad = \quad \frac{\mathbb{P}(\mathcal{D}|\theta)\,\mathbb{P}(\theta)}{\int \mathbb{P}(\mathcal{D}|\theta')\,\mathbb{P}(\theta')\,d\theta'}$$

- Approximate the posterior distribution $\mathbb{P}(\theta|\mathcal{D}) \approx p_\phi(\theta)$
  - approximation only as good as the model class of p_ϕ
  - likelihood based on loss $\mathbb{P}(\mathcal{D}|\theta) \propto \exp(-\mathcal{L}_{[\theta]})$

$$\min_\phi D_{\mathsf{KL}}\big(p_\phi(\cdot)\big\|\mathbb{P}(\cdot|\mathcal{D})\big) \quad \equiv \quad \min_\phi \mathbb{E}\Big[\mathcal{L}_{[\theta]}\,\Big|\,\theta\sim p_\phi(\cdot)\Big] + D_{\mathsf{KL}}\big(p_\phi(\cdot)\big\|\mathbb{P}(\cdot)\big)$$

see Lecture 7.2 for the reparametrization trick, for some more information see Blundell et al. (2015) and Fortunato et al. (2019)

8.2 Bayes by backpropagation

- How to train a neural-net posterior $\mathbb{P}(\theta|\mathcal{D})$?
  - no analytical solution for Bayes update of neural network:

$$\mathbb{P}(\theta|\mathcal{D}) \quad = \quad \frac{\mathbb{P}(\mathcal{D}|\theta)\,\mathbb{P}(\theta)}{\int \mathbb{P}(\mathcal{D}|\theta')\,\mathbb{P}(\theta')\,d\theta'}$$

- Approximate the posterior distribution $\mathbb{P}(\theta|\mathcal{D}) \approx p_\phi(\theta)$
  - approximation only as good as the model class of p_ϕ
  - likelihood based on loss $\mathbb{P}(\mathcal{D}|\theta) \propto \exp(-\mathcal{L}_{[\theta]})$
  - reparameterization trick $\theta =: f_\phi(\epsilon) \sim p_\phi(\cdot)\,, \quad \epsilon \sim p'(\cdot)$

$$\begin{aligned}
\min_\phi D_{\mathsf{KL}}\big(p_\phi(\cdot)\big\|\mathbb{P}(\cdot|\mathcal{D})\big) \quad &\equiv \quad \min_\phi \mathbb{E}\Big[\mathcal{L}_{[\theta]}\,\Big|\,\theta\sim p_\phi(\cdot)\Big] + D_{\mathsf{KL}}\big(p_\phi(\cdot)\big\|\mathbb{P}(\cdot)\big) \\
&= \quad \min_\phi \mathbb{E}\Big[\mathcal{L}_{[f_\phi(\epsilon)]}\,\Big|\,\epsilon\sim p'(\cdot)\Big] + D_{\mathsf{KL}}\big(p_\phi(\cdot)\big\|\mathbb{P}(\cdot)\big)
\end{aligned}$$

- Minimize average loss $\mathbb{E}\Big[\mathcal{L}_{[\theta]}\,\Big|\,\theta\sim q_\phi(\cdot)\Big]$ with gradient descend of ϕ

see Lecture 7.2 for the reparametrization trick, for some more information see Blundell et al. (2015) and Fortunato et al. (2019)

Parameterizable posteriors

- Gaussian posterior over network parameters (Noisy Nets)
  - mean $\boldsymbol{\mu} \in \mathbb{R}^{|\boldsymbol{\theta}|}$ and diagonal covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma}^2)$
  - does not model correlations between parameters
  - e.g. linear layer: $g(\boldsymbol{x})_i = \sum_j \underbrace{(\mu_{ij} + \epsilon_{ij}|\sigma_{ij}|)}_{\theta_{ij} \sim \mathbb{P}} x_j \,,\; \epsilon_{ij} \sim \mathcal{N}(\cdot|\mathbf{0}, \mathbf{I})$

Fortunato et al. (2018) use Noisy Nets, and Gal et al. (2017) use dropout (Srivastava et al., 2014) for exploration

8.2 Parameterizable posteriors

- Gaussian posterior over network parameters (Noisy Nets)
  - mean $\boldsymbol{\mu} \in \mathbb{R}^{|\boldsymbol{\theta}|}$ and diagonal covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma}^2)$
  - does not model correlations between parameters
  - e.g. linear layer: $g(\boldsymbol{x})_i = \sum_j \underbrace{(\mu_{ij} + \epsilon_{ij}|\sigma_{ij}|)}_{\theta_{ij} \sim \mathbb{P}} x_j \,, \;\; \epsilon_{ij} \sim \mathcal{N}(\cdot|\mathbf{0}, \mathbf{I})$
- Dropout learns a distribution over robust networks
  - by 'dropping' parameters with probability $p \in [0, 1)$
  - can be interpreted as posterior distribution
  - e.g. linear layer: $g(\boldsymbol{x})_i = \frac{1}{1-p} \sum_j \underbrace{\epsilon_{ij}\, \phi_{ij}}_{\theta_{ij} \sim \mathbb{P}} x_j \,, \;\; \epsilon_{ij} \sim$ Bernoulli$(\cdot|1-p)$

Fortunato et al. (2018) use Noisy Nets, and Gal et al. (2017) use dropout (Srivastava et al., 2014) for exploration

8.2 Parameterizable posteriors

- Gaussian posterior over network parameters (Noisy Nets)
  - mean $\boldsymbol{\mu} \in \mathbb{R}^{|\boldsymbol{\theta}|}$ and diagonal covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}^2)$
  - does not model correlations between parameters
  - e.g. linear layer: $g(\boldsymbol{x})_i = \sum_j \underbrace{(\mu_{ij} + \epsilon_{ij}|\sigma_{ij}|)}_{\theta_{ij} \sim \mathbb{P}} x_j \,, \; \epsilon_{ij} \sim \mathcal{N}(\cdot|\mathbf{0}, \mathbf{I})$
- Dropout learns a distribution over robust networks
  - by 'dropping' parameters with probability $p \in [0, 1)$
  - can be interpreted as posterior distribution
  - e.g. linear layer: $g(\boldsymbol{x})_i = \frac{1}{1-p} \sum_j \underbrace{\epsilon_{ij}\, \phi_{ij}}_{\theta_{ij} \sim \mathbb{P}} x_j \,, \;\; \epsilon_{ij} \sim \mathrm{Bernoulli}(\cdot|1-p)$
- Minimizing average loss drives posterior variance to zero
  - D_{KL} keeps it from collapsing completely

  $\Rightarrow$ terrible for detecting novel state-actions

Fortunato et al. (2018) use Noisy Nets, and Gal et al. (2017) use dropout (Srivastava et al., 2014) for exploration

8.2 Ensemble posterior

- Ensembles model posterior as set $\phi := \{\boldsymbol{\theta}^k\}_{k=1}^m$
  - initialize each $\boldsymbol{\theta}^k$ randomly like any neural net
  - Thompson sampling by selecting $k \sim$ Uniform$(1, \ldots, m)$
  - Bayes-by-backprop, similar to *particle filters*

$$\min_{\boldsymbol{\theta}^k} \mathbb{E}\big[\mathcal{L}_{[\boldsymbol{\theta}]} | \boldsymbol{\theta} \sim p_\phi\big] + D_{\mathsf{KL}}[p_\phi \| \mathbb{P}] \quad \equiv \quad \min_{\boldsymbol{\theta}^k} \mathcal{L}_{[\boldsymbol{\theta}^k]} + \tfrac{1}{2\sigma^2} \|\boldsymbol{\theta}^k\|_2^2 \,, \forall k$$

here we use $p_\phi(\boldsymbol{\theta}) = \frac{1}{m} \sum_{k=1}^m \delta(\boldsymbol{\theta} = \boldsymbol{\theta}^k)$ and $\mathbb{P}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} | \mathbf{0}, \sigma^2 \mathbf{I})$; en.wikipedia.org/wiki/Particle_filter

Ensemble posterior

- Ensembles model posterior as set $\phi := \{\boldsymbol{\theta}^k\}_{k=1}^m$
  - initialize each $\boldsymbol{\theta}^k$ randomly like any neural net
  - Thompson sampling by selecting $k \sim$ Uniform$(1, \ldots, m)$
  - Bayes-by-backprop, similar to *particle filters*

$$\min_{\boldsymbol{\theta}^k} \mathbb{E}\big[\mathcal{L}_{[\boldsymbol{\theta}]} | \boldsymbol{\theta} \sim p_\phi\big] + D_{\mathsf{KL}}[p_\phi \| \mathbb{P}] \quad \equiv \quad \min_{\boldsymbol{\theta}^k} \mathcal{L}_{[\boldsymbol{\theta}^k]} + \tfrac{1}{2\sigma^2} \|\boldsymbol{\theta}^k\|_2^2 \,, \forall k$$

- Ensembles work more by accident than by design
  - reasonable m are much to small to represent posterior
  - but $\boldsymbol{\theta}^k$ converge to different local minima of $\mathcal{L}$

  $\rightarrow$ predictions coincide on training set $\mathcal{D}$

  $\rightarrow$ predictions often diverge outside $\mathcal{D}$

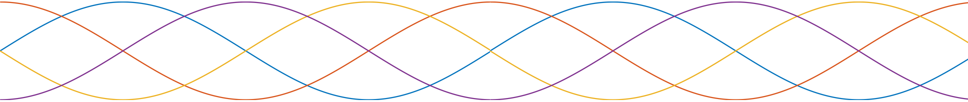
$\Rightarrow$ suitable to detect novel state-actions!

see Lu and Van Roy (2017) for ensemble sampling for exploration

8.2 Ensemble extensions

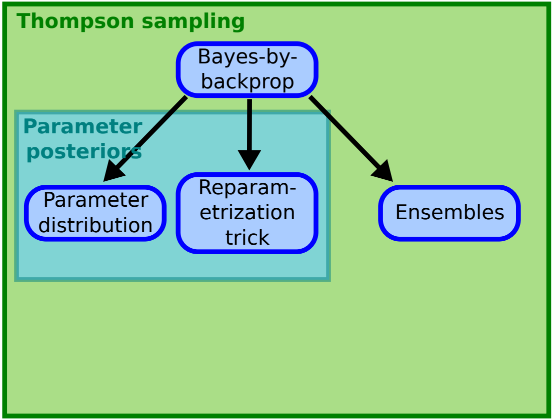
- Bootstrapped ensembles vary the data each model trains on
  - by using random masks on the mini-batches of each model
  - minor effect in comparison to different local minima

- Randomized prior functions make sure predictions diverge
  - define set of m functions g^k such that $\forall \boldsymbol{x}$:
    $\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x}')\| > \epsilon, \forall \boldsymbol{x}' \in \big\{\boldsymbol{x}' \big| \|\boldsymbol{x}' - \boldsymbol{x}\| > \epsilon' \big\}$
  - g^k is prior function of model $\boldsymbol{\theta}^k$, e.g.: $f(\boldsymbol{x})_i = \sum_j \theta^k_{ij} x_j + g^k(\boldsymbol{x})_i$
  - models learn to compensate for priors on training set $\mathcal{D}$
  - outside $\mathcal{D}$ priors guarantee divergent predictions



Osband et al. (2016) use bootstrapped ensembles and Osband et al. (2018) use randomized priors for exploration

Overview Thompson sampling



8.2 Summary

- Thompson sampling reduces epistemic uncertainty
- Posteriors are learned with Bayes-by-backpropagation
- Gaussian and dropout posteriors do not work well
- Ensemble posteriors work well, but only by accident
- Randomized prior functions improve ensembles

Learning Objectives

LO8.3: Explain Thompson sampling
LO8.4: Explain and derive Bayes-by-backpropagation
LO8.5: Explain why ensembles detect out-of-distribution samples

8.3

Exploration
Optimistic exploration

Optimistic exploration

- Upper confidence bounds (UCB) from multi-armed bandits
  - after $N(s,a)$ executions of a in s
  - variance of average $\mathbb{V}[\frac{1}{N(s,a)} \sum_{t=1}^{n} r_t \, \delta(s_t = s) \, \delta(a_t = a)] \propto \frac{1}{N(s,a)}$
  - act *optimistically* with confidence-parameter C
  - guaranteed to converge to optimum (constant regret)

$$a_t^* \quad := \quad \arg\max_{a \in \mathcal{A}} \Big(Q_\theta(s_t, a) + C \underbrace{\sqrt{\frac{\log \sum_{a'} N(s_t,a')}{N(s_t,a)}}}_{\text{bonus } \eta(s_t,a)} \Big)$$

for $\mathbb{V} \propto \frac{1}{N}$ see exercise sheet 1

see Auer et al. (2002) for UCB1

8.3 Optimistic exploration

- Upper confidence bounds (UCB) from multi-armed bandits
  - after $N(s,a)$ executions of a in s
  - variance of average $\mathbb{V}[\frac{1}{N(s,a)} \sum_{t=1}^{n} r_t \, \delta(s_t = s) \, \delta(a_t = a)] \propto \frac{1}{N(s,a)}$
  - act *optimistically* with confidence-parameter C
  - guaranteed to converge to optimum (constant regret)

$$a_t^* \quad := \quad \arg\max_{a \in \mathcal{A}} \Big(Q_\theta(s_t, a) + C \underbrace{\sqrt{\frac{\log \sum_{a'} N(s_t, a')}{N(s_t, a)}}}_{\text{bonus } \eta(s_t, a)} \Big)$$

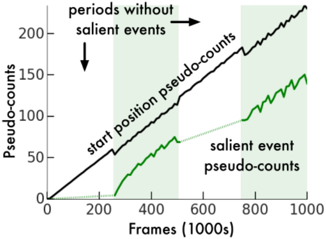
- Works decently well in tabular reinforcement learning
  - confidence bonus diminishes over time
  - special case of optimistic initialization
  - how do counts generalize to continuous spaces?

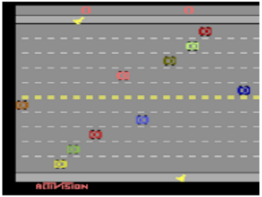
other bonuses exist, e.g. Rashid et al. (2020) add the bonus $\eta(s,a) = N(s,a)^{-m}$ to the Q-values

8.3 Visitation counts

- Counting visitations $N(s,a)$ impossible in continuous spaces
- Pseudo-counts are based on estimated density model $p(s,a)$
  - after observing n samples in s: $p(s,a) = \frac{N(s,a)}{n}$
  - after observing (s,a) again: $p'(s,a) = \frac{N(s,a)+1}{n+1}$

  $\Rightarrow \; N(s,a) = \frac{p(s,a)\,(1-p'(s,a))}{p'(s,a)-p(s,a)}$
  - perform gradient descent and compare density before and after



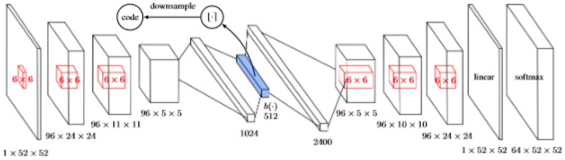


see Bellemare et al. (2016, image source) with a weak density model; and Ostrovski et al. (2017) with auto-regressive model

8.3 Visitation counts

- Counting visitations $N(s,a)$ impossible in continuous spaces
- Pseudo-counts are based on estimated density model $p(s,a)$
  - after observing n samples in s: $p(s,a) = \frac{N(s,a)}{n}$
  - after observing (s,a) again: $p'(s,a) = \frac{N(s,a)+1}{n+1}$

  $\Rightarrow \quad N(s,a) = \frac{p(s,a)\,(1-p'(s,a))}{p'(s,a)-p(s,a)}$
  - perform gradient descent and compare density before and after
- Random hash functions can divide state-action space
  - e.g. Gaussian hash $h(s,a) = \delta(\mathbf{A}^\top \boldsymbol{b}(s,a) > 0)\,, \quad A_{ij} \sim \mathcal{N}(\cdot|0,1)$
  - quality depends very much on hash function



hash counting by Tang et al. (2017)

8.3 Novelty measures

- UCB can use *any* novelty measure bonus that decays to 0
  - no theoretical motivation, but works in practice
- Random network distillation (RND)
  - two differently initialized neural nets, $\boldsymbol{f}_\phi, \boldsymbol{f}_\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$
  - keep ψ fixed and train $\min_\phi \mathbb{E}\big[\|\boldsymbol{f}_\phi(s,a) - \boldsymbol{f}_\psi(s,a)\|^2\big]$
  - distance $\eta(s,a) := \|\boldsymbol{f}_\phi(s,a) - \boldsymbol{f}_\psi(s,a)\|^2$ is novelty measure
  - ensemble without the interpretation as posterior variance

assignment sheet 4

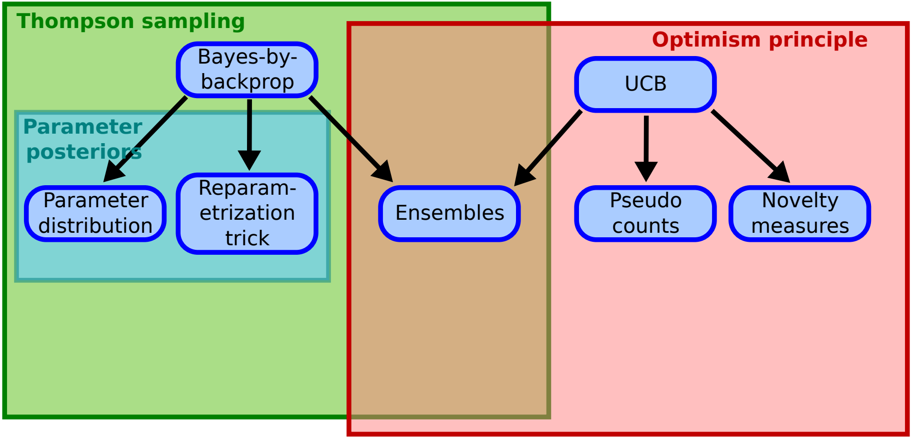
random network distillation by Burda et al. (2019)

8.3 Novelty measures

- UCB can use *any* novelty measure bonus that decays to 0
  - no theoretical motivation, but works in practice
- Random network distillation (RND)
  - two differently initialized neural nets, $\boldsymbol{f}_\phi, \boldsymbol{f}_\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$
  - keep ψ fixed and train $\min_\phi \mathbb{E}\big[\|\boldsymbol{f}_\phi(s,a) - \boldsymbol{f}_\psi(s,a)\|^2\big]$
  - distance $\eta(s,a) := \|\boldsymbol{f}_\phi(s,a) - \boldsymbol{f}_\psi(s,a)\|^2$ is novelty measure
  - ensemble without the interpretation as posterior variance
- Many other novelty measures exist
  - predicting the next state or reward
  - cosine-similarity to training samples
- Novelty measures are scale-free
  - no interpretation like variance of values
  - hyper-parameter selection even harder

novelty based on state prediction e.g. in Pathak et al. (2017), cosine-similarity in O'Donoghue et al. (2018); Böhmer et al. (2019)

8.3 Overview optimism principle



8.3 Summary

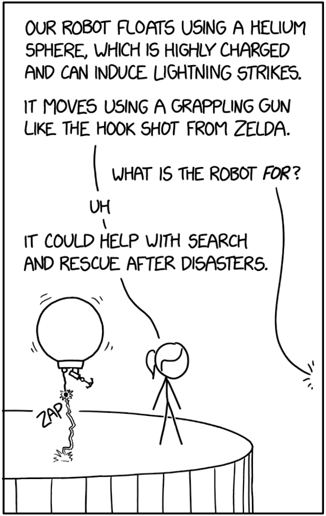
- Optimistic exploration gives bonus for uncertain actions
- Pseudo-counts and hash functions generalize visitations
- Novelty measure estimate uncertainty without interpretation

Learning Objectives

LO8.6: Explain optimistic exploration
LO8.7: Explain and compare visitation counts and novelty estimation

8.3 Next lecture

- Next lecture: **offline RL**!
- Don't forget assignment 3!
- Questions? Ask them here: `answers.ewi.tudelft.nl`



"IT COULD HELP WITH SEARCH AND RESCUE" IS ENGINEER-SPEAK FOR "WE JUST REALIZED WE NEED A JUSTIFICATION FOR OUR COOL ROBOT."

image source: xkcd.com

References I

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002. ISSN 0885-6125.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems (NIPS) 29*, pages 1471–1479, 2016. URL https://arxiv.org/abs/1606.01868.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pages 1613–1622. PMLR, 07–09 Jul 2015. URL https://arxiv.org/abs/1505.05424.

Wendelin Böhmer, Tabish Rashid, and Shimon Whiteson. Exploration with unreliable intrinsic reward in multi-agent reinforcement learning. *CoRR*, abs/1906.02138, 2019. URL http://arxiv.org/abs/1906.02138. Presented at the ICML *Exploration in Reinforcement Learning* workshop.

Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *7th International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://papers.nips.cc/paper/2011/hash/e53a0a2978c28872a4505bdb51db06dc-Abstract.html.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1706.10295.

Meire Fortunato, Charles Blundell, and Oriol Vinyals. Bayesian recurrent neural networks, 2019. URL https://arxiv.org/abs/1704.02798.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3584–3593, 2017. URL https://arxiv.org/abs/1705.07832.

References II

Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

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.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2721–2730, 2017. URL https://arxiv.org/abs/1703.01310.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. URL https://arxiv.org/abs/1705.05363.

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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

References III

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. #Exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS) 30*, pages 2753–2762. 2017. URL https://arxiv.org/abs/1611.04717.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

Hao Wang and Dit-Yan Yeung. A survey on Bayesian deep learning. *ACM Comput. Surv.*, 53(5), 2020. ISSN 0360-0300. doi: 10.1145/3409383. URL http://wanghao.in/paper/CSUR20_BDL.pdf.