Safe RL

研究者

Philip S. Thomas safe RL的talk视频

Safe RL的一些概念

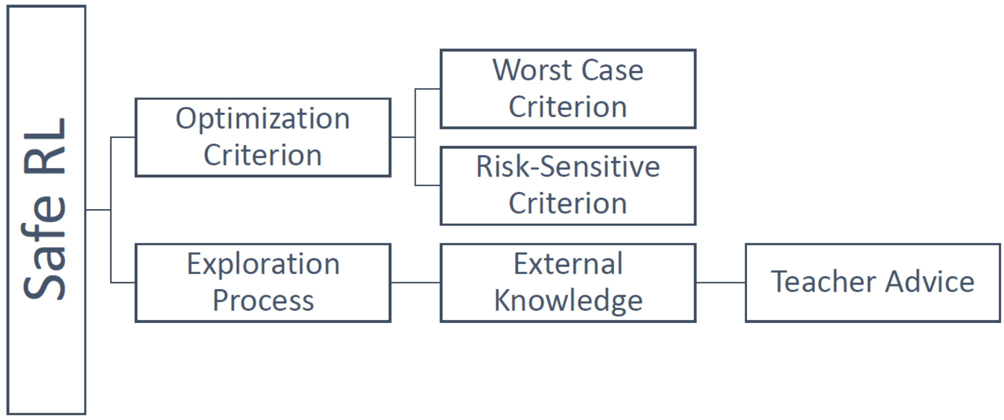
Safe RL的Safe是有不同的定义方法的。但是无论如何定义都是要保证其算法性能的提高。

“I guarantee that with probability at least $1-\delta$, I will not change your policy to one that is worse than the current policy.”

一些限制条件

- 假定初始policy是可以获得的
- 假定初始policy已知
- 假定初始policy是stochastic的

分类

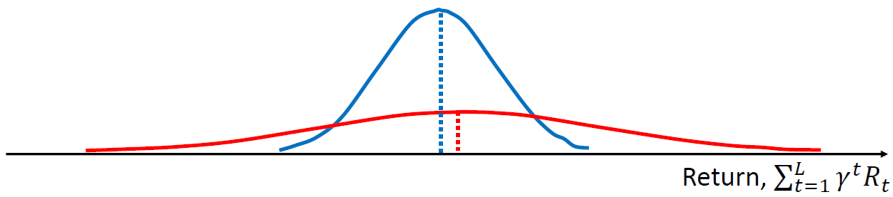


- Worst Case Criterion包括the parameter uncertainty
- Risk-Sensitive Criterion则是针对于Risk进行评判

Risk-Sensitive Criterion

- Expected return:

$$J(\pi) = \mathbf{E}[\sum_{t=1}^{L} \gamma^t R_t \, |\pi]$$



- Which policy is better if I am a casino?
- Which policy is better if I am a doctor?

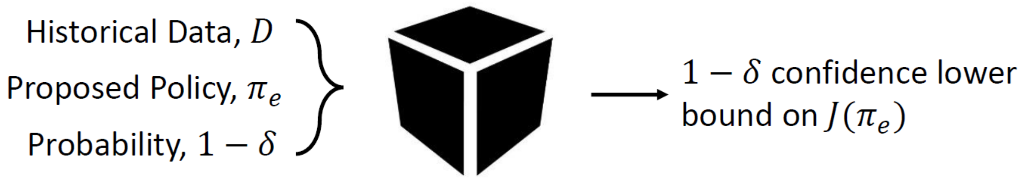
Penalize variance:

$$J(\pi) = \mathbf{E}\left[\sum_{t=1}^{L} \gamma^t R_t | \pi\right] - \lambda \operatorname{Var}\left(\sum_{t=1}^{L} \gamma^t R_t | \pi\right)$$

- External Knowledge: (i) providing initial knowledge, (ii) deriving a policy from a nite set of demonstrations and, (iii) providing teach advice.

例子

High confidence off-policy policy evaluation (HCOPE)



通过hoeffding不等式得到

$$\mathrm{E}\left[X_i\right] \geq \frac{1}{n} \sum_{i=1}^{n} X_i - b\sqrt{\frac{\ln(1/\delta)}{2n}}$$

$$\frac{1}{n} \sum_{i=1}^{n} \left(w_i \sum_{t=1}^{L} \gamma^t R_t^i\right)$$

推荐阅读

- Importance sampling for RL (IS, PDIS, WIS, CWPDIS)
    - D. Precup, R. S. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In Proceedings of the 17th International Conference on Machine Learning, pages 759–766, 2000. [NOTE: WPDIS estimator has a typo]
    - P. S. Thomas. Safe reinforcement learning. PhD Thesis, UMass Amherst, 2015.
- Doubly robust importance sampling and MAGIC for RL
    - N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. ICML 2016
    - P. S. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. ICML 2016.
- Other importance sampling estimators for RL (more for bandits)
    - P. S. Thomas and E. Brunskill. Importance Sampling with Unequal Support. AAAI 2017
    - P. S. Thomas., G. Theocharous, M. Ghavamzadeh, I. Durugkar, and E. Brunskill. Predictive Off-Policy Policy Evaluation for Nonstationary Decision Problems, with Applications to Digital Marketing. IAAI 2017.
    - S. Daroudi, P. S. Thomas, and E. Brunskill. Importance Sampling for Fair Policy Selection. UAI 2017.
    - Z. Guo, P. S. Thomas, and E. Brunskill. Using Options for Long-Horizon Off-Policy Evaluation. RLDM 2017.
    - Y. Liu, P. S. Thomas, and E. Brunskill. Model Selection for Off-Policy Policy Evaluation. RLDM 2017.
    - P. S. Thomas, S. Niekum, G. Theocharous, and G.D. Konidaris. Policy Evaluation Using the Omega-Return. NIPS 2015.
- HCOPE
    - L. Bottou, J. Peters, J. Quinonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. JMLR 2013.
    - J.P. Hanna, P. Stone, and S. Niekum. Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation. AAMAS 2017.
    - P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High Confidence Off-Policy Evaluation. AAAI 2015.
    - P. S. Thomas . Safe reinforcement learning. PhD Thesis, UMass Amherst, 2015.
- Safe Policy Improvement
    - P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High Confidence Policy Improvement. ICML 2015
    - P. S. Thomas. Safe reinforcement learning. PhD Thesis, UMass Amherst, 2015.

Meta RL

研究者

Flood Sung [知乎首页] [github]

Meta RL的研究基本上是Sergey Levine团队，而Meta Learning在Few Shot Learning上则比较百花齐放。

Chelsea Finn

Meta Learning的一些概念

Meta learning 也称为 Learning to learn，即学会如何学习。

深度学习技术视角的Meta

包含了以下这些类别：

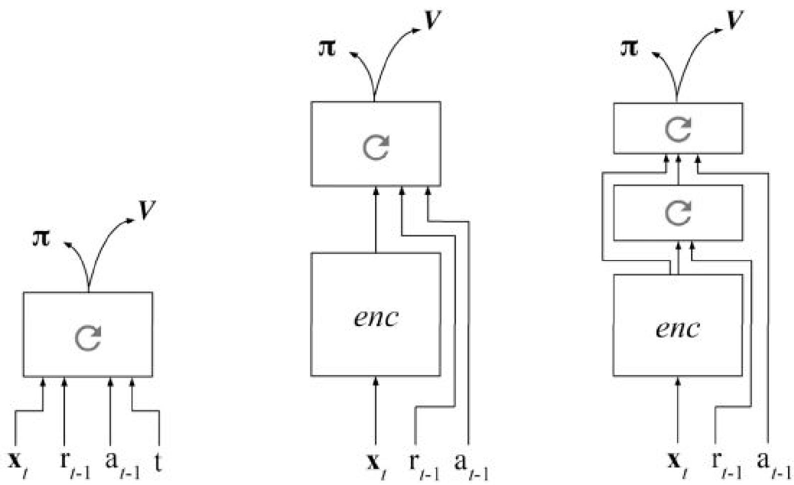
1. 训练超参数Hyper Parameters：包括Learning rate，Batch Size，input size等等目前要人为设定的参数
2. 神经网络的结构
3. 神经网络的初始化
4. 优化器Optimizer的选择。比如SGD，Adam，RMSProp
5. 神经网络参数
6. 损失函数的定义。
7. 反向传播Back-propagation。

Meta RL概念

meta RL的基本思想非常简单，就是在输入增加上一次的reward，或者用之前的（state,action,reward）来推断Meta知识。

Meta RL和hierarchical RL很相似，我们可以通过多个类似的任务来学习一个meta knowledge，这个meta knowledge就是hierarchy，就是高层的知识。

简单例子



(a) LSTM A2C (b) LSTM A3C (c) Stacked-LSTM A3C

Meta RL中目前为止最有名的算法是MAML，MAML的做法是先用之前的trajectory对神经网络做一次更新，然后再使用更新后的网络进一步训练，通过二次梯度更新整个网络参数。这样本质上也是充分利用历史信息来学习一个好的prior （在MAML中就是一个好的初始化）。

推荐阅读

[1] Wang, Jane X., et al. "**Learning to reinforcement learn**." *arXiv preprint arXiv:1611.05763*(2016).

[2] Wang, Jane X., et al. "**Prefrontal cortex as a meta-reinforcement learning system**." *Nature neuroscience*21.6 (2018): 860.

[3] Duan, Yan, et al. "**RL2: Fast Reinforcement Learning via Slow Reinforcement Learning**." *arXiv preprint arXiv:1611.02779*(2016).

[4] Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "**Model-agnostic meta-learning for fast adaptation of deep networks**." *arXiv preprint arXiv:1703.03400*(2017).

[5] Mishra, Nikhil, et al. "**A simple neural attentive meta-learner**." (2018).

[6] Houthooft, Rein, et al. "**Evolved policy gradients**." *arXiv preprint arXiv:1802.04821*(2018).

[7] Gupta, Abhishek, et al. "**Meta-Reinforcement Learning of Structured Exploration Strategies**." *arXiv preprint arXiv:1802.07245*(2018).

[8] Stadie, Bradly C., et al. "**Some considerations on learning to explore via meta-reinforcement learning**." *arXiv preprint arXiv:1803.01118*(2018).

[9] Xu, Tianbing, et al. "**Learning to Explore with Meta-Policy Gradient**." *arXiv preprint arXiv:1803.05044*(2018).

[10] Clavera, Ignasi, et al. "**Learning to Adapt: Meta-Learning for Model-Based Control**." *arXiv preprint arXiv:1803.11347*(2018).

[11] Xu, Zhongwen, Hado van Hasselt, and David Silver. "**Meta-Gradient Reinforcement Learning**." *arXiv preprint arXiv:1805.09801*(2018).

[12] Xu, Kelvin, et al. "**Learning a Prior over Intent via Meta-Inverse Reinforcement Learning**." *arXiv preprint arXiv:1805.12573*(2018).

[13] Gupta, Abhishek, et al. "**Unsupervised Meta-Learning for Reinforcement Learning**." *arXiv preprint arXiv:1806.04640*(2018).

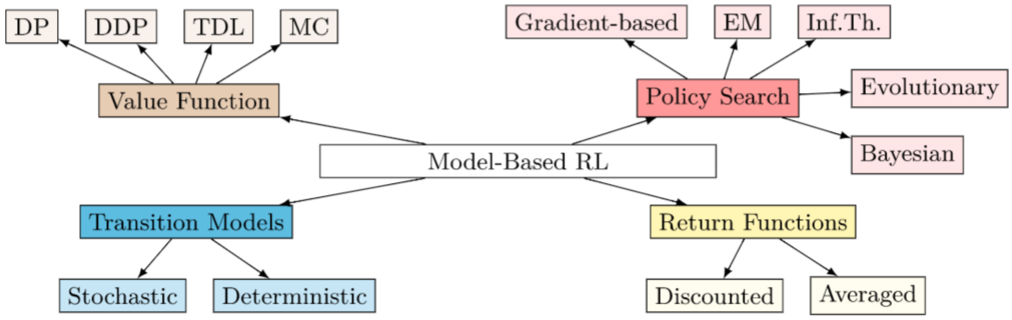
Model-based RL

研究者

Sergey Levine团队

Chelsea Finn

分类



Model-based和Model-free的比较

RL Methods	Advantages	Disadvantages
Model-based RL	– Small number of interactions between robot & environment – Faster convergence to optimal solution	–Depend on transition models – Model accuracy has a big impact on learning tasks
Model-free RL	– No need for prior knowledge of transitions – Easily implementable	– Slow learning convergence – High wear & tear of the robot – High risk of damage

Model-Based vs. Model-Free Algorithms

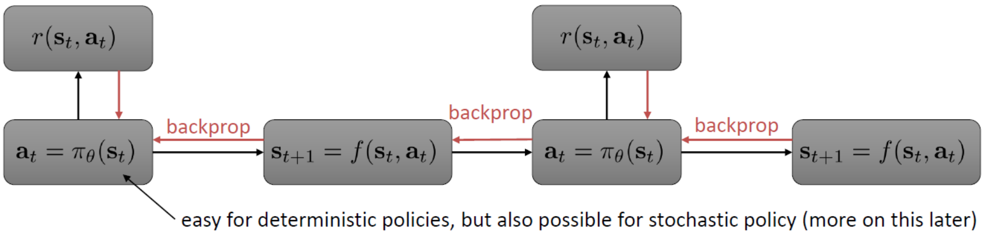
Models:

\+ Easy to collect data in a scalable way (self-supervised)
\+ Possibility to transfer across tasks
\+ Typically require a smaller quantity of supervised data
\- Models don't optimize for task performance
\- Sometimes harder to learn than a policy
\- Often need assumptions to learn complex skills (continuity, resets)

Model-Free:

\+ Makes little assumptions beyond a reward function
\+ Effective for learning complex policies
\- Require a lot of experience (slower)
\- Not transferable across tasks

基本流程



model-based reinforcement learning version 2.0:

1. run base policy $\pi_0(\mathbf{a}_t|\mathbf{s}_t)$ (e.g., random policy) to collect $\mathcal{D} = \{(\mathbf{s}, \mathbf{a}, \mathbf{s}')_i\}$
2. learn dynamics model $f(\mathbf{s}, \mathbf{a})$ to minimize $\sum_i \|f(\mathbf{s}_i, \mathbf{a}_i) - \mathbf{s}'_i\|^2$
3. backpropagate through $f(\mathbf{s}, \mathbf{a})$ into the policy to optimize $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$
4. run $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$, appending the visited tuples $(\mathbf{s}, \mathbf{a}, \mathbf{s}')$ to $\mathcal{D}$

推荐阅读

Further Reading on Model-based RL

Use known model: Tassa et al. IROS '12, Tan et al. TOG '14, Mordatch et al. TOG '14
Guided policy search: Levine*, Finn* et al. JMLR '16, Mordatch et al. RSS '14, NIPS '15
Backprop through model: Deisenroth et al. ICML '11, Heess et al. NIPS '15, Mishra et al. ICML '17, Degrave et al. '17, Henaff et al. '17
Inverse models: Agrawal et al. NIPS '16
MBRL in latent space: Watter et al. NIPS '15, Finn et al. ICRA '16
MPC with deep models: Lenz et al. RSS '15, Finn & Levine ICRA '17
Combining Model-Based & Model-Free:
- use roll-outs from model as experience: Sutton '90, Gu et al. ICML '16
- use model as baseline: Chebotar et al. ICML '17
- use model for exploration: Stadie et al. arXiv '15, Oh et al. NIPS '16
- model-free policy with planning capabilities: Tamar et al. NIPS '16, Pascanu et al. '17
- model-based look-ahead: Guo et al. NIPS '14, Silver et al. Nature '16