Aprendizaje por recompensas
Lecturas seleccionadas

Juan Bazerque

5 de junio de 2021

Extensiones de Reinforcement Learning

- Safe Exploration: agrega restricciones de seguridad
  - Provably efficient safe exploration via primal-dual policy optimization
- Meta Learning: busca hacer el aprendizaje más rápido
  - Efficient Off-Policy Meta-RL via Probabilistic Context Variables
- Aplicación: robótica con *human in the loop*
  - Socially Aware Motion Planning with Deep Reinforcement Learning

Safe RL[1]

Objetivo: agregar restricciones al problema
Idea: se combina RL con primal-dual methods

- Se tienen rewards r e indicadores de restricciones g

$$V_{r,h}^{\pi}(x) = \mathbb{E}_{\pi}\left[\sum_{i=h}^{H} r_i\left(x_i, a_i\right) \mid x_h = x\right]$$

$$V_{g,h}^{\pi}(x) = \mathbb{E}_{\pi}\left[\sum_{i=h}^{H} g_i\left(x_i, a_i\right) \mid x_h = x\right]$$

- Se busca resolver

$$\underset{\pi \in \Delta(\mathcal{A}|\mathcal{S},H)}{\text{maximize}}\ V_{r,1}^{\pi}\left(x_1\right) \text{ subject to } V_{g,1}^{\pi}\left(x_1\right) \geq b \tag{1}$$

- Se resuelve por dualidad con multiplicador Y

$$\underset{\pi \in \Delta(\mathcal{A}|\mathcal{S},H)}{\text{maximize}}\ \underset{Y \geq 0}{\text{minimize}}\left(V_{r,1}^{\pi}\left(x_1\right) + Y\left(V_{g,1}^{\pi}\left(x_1\right) - b\right)\right)$$

[1] Ding D, Wei X, Yang Z, Wang Z, Jovanovic M. Provably efficient safe exploration via primal-dual policy optimization. International Conference on Artificial Intelligence and Statistics, 2021.

Meta-learning[3]

Objetivo: Aprender con pocas muestras luego de pre-entrenar con varias tareas
Idea: nueva tarea desconocida $\Rightarrow$ identificar la preentreanada Z más similar

- Critic evaluates the advantage function

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_{(\mathrm{s},\mathrm{a},r,\mathrm{s}')\sim\mathcal{B},\quad \mathrm{z}\sim q_\phi(\mathrm{z}|\mathrm{c})}\left[Q_\theta(\mathrm{s},\mathrm{a},\mathrm{z}) - \left(r + \bar{V}\left(\mathrm{s}',\bar{\mathrm{z}}\right)\right)\right]^2$$

- The observed trajectory is denoted by c
- A neural network $q(Z|c)$ classifies the most likely task
- Soft actor improves policy by moving it towards an exponential of Q [2]

$$\mathcal{L}_{\text{actor}} = \mathbb{E}_{\mathrm{s}\sim\mathcal{B},\mathrm{a}\sim\pi_\theta}\mathrm{z}\sim\mathrm{q}_\phi(\mathrm{z}|\mathrm{c})\left[D_{\mathrm{KL}}\left(\pi_\theta(\mathrm{a}\mid\mathrm{s},\bar{\mathrm{z}})\|\frac{\exp\left(Q_\theta(\mathrm{s},\mathrm{a},\bar{\mathrm{z}})\right)}{\mathcal{Z}_\theta(\mathrm{s})}\right)\right]$$

---

[2] Haarnoja, T., Zhou, A., Abbeel, P. and Levine, S., Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, (ICML), 2018

[3] Rakelly, K., Zhou, A., Finn, C., Levine, S. and Quillen, D., Efficient off-policy meta-reinforcement learning via probabilistic context variables. International conference on machine learning (ICML), 2019.

RL for robotics[4]

Objetivo: Hacer que un robot se mueva por un aeropuerto
Idea: Humanos como agentes y recompensa que favorece *normass sociales*

$$
\begin{aligned}
R_{\text{norm}}\left(\mathrm{s}^{jn}, \mathrm{u}\right) =& q_n I\left(\mathrm{s}^{jn} \in \mathcal{S}_{\text{norm}}\right) \\
\text{s.t.} \quad \mathcal{S}_{\text{norm}} =& \mathcal{S}_{\text{pass}} \cup \mathcal{S}_{\text{ovtk}} \cup \mathcal{S}_{\text{cross}} \\
\mathcal{S}_{\text{pass}} =& \Big\{\mathrm{s}^{jn} \mid d_g > 3, \quad 1 < \tilde{p}_x < 4, \\
& -2 < \tilde{p}_y < 0, \quad |\tilde{\phi} - \psi| > 3\pi/4\Big\} \quad (10) \\
\mathcal{S}_{\text{ovtk}} =& \Big\{\mathrm{s}^{jn} \,|d_g > 3, \quad 0 < \tilde{p}_x < 3, \quad |\,\mathrm{v}| > |\tilde{\mathrm{v}}\,| \\
& 0 < \tilde{p}_y < 1, \quad |\tilde{\phi} - \psi| < \pi/4\Big\} \\
\mathcal{S}_{\text{cross}} =& \Big\{\mathrm{s}^{jn} \mid d_g > 3, \quad \tilde{d}_a < 2, \quad \tilde{\phi}_{\text{rot}} > 0, \\
& -3\pi/4 < \tilde{\phi} - \psi < -\pi/4\},
\end{aligned}
\tag{2}
$$

[4] Chen, Y.F., Everett, M., Liu, M. and How, J.P., Socially aware motion planning with deep reinforcement learning. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2017.

Anuncios sobre la finalización del curso

- Llenar las encuestas
  - Las críticas son bienvenidas
- Entregar el obligatorio 4 para el 13 de junio
  - Solo se requiere uno de los ejercicios
- Tomar el oral en fecha a definir
  - Defensa de lo realizado en los obligatorios