

# Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning

Shakir Mohamed and Danilo J. Rezende  
Google DeepMind, London  
{shakir, danilor}@google.com

## 【强化学习算法 25】Empowerment



张楚珩

清华大学 交叉信息院博士在读

5 人赞同了该文章

这篇文章也没有给算法的名字了，但是其核心就是定义了empowerment并给出了其迭代求解方法，所以这里就暂且把它记为empowerment吧。

### 原文传送门

Mohamed, Shakir, and Danilo Jimenez Rezende. "Variational information maximisation for intrinsically motivated reinforcement learning." *Advances in neural information processing systems*. 2015.

### 特色

本专栏前面讲到的ICM和Curiosity基本上都需要拟合并建立一个dynamic model，并且把prediction error作为intrinsic reward。由于这样的奖励会随着智能体的经历而变化（参见概念homeostatic和heterostatic[1]），智能体可能会变得厌倦。个人感觉这虽然很符合生物的特性（类似“曾经沧海难为水，除却巫山不是云”），但是总感觉怪怪的。

考虑在一个复杂的房子里面探索，这样的智能体一旦多次探索熟悉这个环境之后，不管到哪里都不会收到任何奖励了。但是想一想，intrinsic reward的目的是什么？（结合本专栏讲的Reward Shaping Invariance）我们是为了给智能体提供一个良好的先验知识，智能体在这个房子里面的先验知识就应该是知道不管来了什么任务，如果它现在在卧室，那么大概率应该先走到卧室的门口，那么intrinsic reward最本地就应该导航这个智能体到门口去。在【强化学习思想 22】Reward Shaping Invariance的框架下就是说，要给门口附近的格子更高的势能。这篇文章就客观地定义了这样的一个势能，这里把它叫做empowerment。当然，其使用方式不仅仅可以把它作为intrinsic reward来编码环境的先验知识或者指导探索。

### 过程

#### 1. 互信息

互信息的定义如下

$$\mathcal{I}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})} \left[ \log \left( \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \right) \right], \quad (1)$$

它讲的是两个随机变量之间的关联性如何。如果它们相互独立，即  $p(\mathbf{a}, \mathbf{y}) = p(\mathbf{a})p(\mathbf{y})$ ，不难看出，互信息为零；如果  $\mathbf{a}$  和  $\mathbf{y}$  有确定性关系，即只要知道了  $\mathbf{a}$  就知道  $\mathbf{y}$ ，那么  $\mathcal{I}(\mathbf{a}, \mathbf{y}) = \mathbb{E}_{p(\mathbf{a})}[-\log p(\mathbf{a})] = H(\mathbf{a})$ ，即其中一个变量的熵。

## 2. 定义 Empowerment

Empowerment还挺不好翻译的，它是针对一个状态而言的，讲的是在某个状态上智能体最多有多少可以控制整个环境的能力。换句话说，就是从这个态出发，容不容易达到任意的其他状态。比如，机器人要是处于站着的状态，那么它既可以马上躺下，也可以马上跑步，跳跃（有较大的 empowerment）；但是如果机器人处于躺着的状态，那么要它立马跑到某个地点就会比较困难（empowerment较小）。

正式的定义如下

$$\mathcal{E}(\mathbf{s}) = \max_{\omega} \mathcal{I}^{\omega}(\mathbf{a}, \mathbf{s}' | \mathbf{s}) = \max_{\omega} \mathbb{E}_{p(\mathbf{s}' | \mathbf{a}, \mathbf{s}) \omega(\mathbf{a} | \mathbf{s})} \left[ \log \left( \frac{p(\mathbf{a}, \mathbf{s}' | \mathbf{s})}{\omega(\mathbf{a} | \mathbf{s}) p(\mathbf{s}' | \mathbf{s})} \right) \right], \quad (2)$$

其中  $\mathbf{a}$  是未来K步的开环行动序列， $\omega(\mathbf{a} | \mathbf{s})$  是基于状态的K步开环行动策略。它说的是从状态  $\mathbf{s}$  出发，智能体有多大的可能性通过自己的努力改变自己的命运。

## 3. 问题转化

首先，用信息论的公式，转化为熵

$$\mathcal{I}(\mathbf{a}, \mathbf{s}' | \mathbf{s}) = H(\mathbf{a} | \mathbf{s}) - H(\mathbf{a} | \mathbf{s}', \mathbf{s}), \quad (3)$$

where  $H(\mathbf{a} | \mathbf{s}) = -\mathbb{E}_{\omega(\mathbf{a} | \mathbf{s})} [\log \omega(\mathbf{a} | \mathbf{s})]$  and  $H(\mathbf{a} | \mathbf{s}', \mathbf{s}) = -\mathbb{E}_{p(\mathbf{s}' | \mathbf{a}, \mathbf{s}) \omega(\mathbf{a} | \mathbf{s})} [\log p(\mathbf{a} | \mathbf{s}', \mathbf{s})]$

难点在于  $p(\mathbf{a} | \mathbf{s}', \mathbf{s})$  分布也不知道，求起来也无从下手。这种情况最常用的就是Variational Lower Bound，即使用  $q(\mathbf{a} | \mathbf{s}', \mathbf{s})$  来近似  $p(\mathbf{a} | \mathbf{s}', \mathbf{s})$ 。

$$\mathcal{I}^{\omega}(\mathbf{s}) = H(\mathbf{a} | \mathbf{s}) - H(\mathbf{a} | \mathbf{s}', \mathbf{s}) \geq H(\mathbf{a} | \mathbf{s}) + \mathbb{E}_{p(\mathbf{s}' | \mathbf{a}, \mathbf{s}) \omega(\mathbf{a} | \mathbf{s})} [\log q_{\xi}(\mathbf{a} | \mathbf{s}', \mathbf{s})] = \mathcal{I}^{\omega, q}(\mathbf{s}) \quad (4)$$

于是empowerment可以写作

$$\hat{\mathcal{E}}(\mathbf{s}) = \max_{\omega, q} \mathcal{I}^{\omega, q}(\mathbf{s}) \text{ s.t. } H(\mathbf{a} | \mathbf{s}) < \epsilon$$

等效地

$$\hat{\mathcal{E}}(\mathbf{s}) = \max_{\omega, q} \mathbb{E}_{p(\mathbf{s}' | \mathbf{a}, \mathbf{s}) \omega(\mathbf{a} | \mathbf{s})} \left[ -\frac{1}{\beta} \ln \omega(\mathbf{a} | \mathbf{s}) + \ln q_{\xi}(\mathbf{a} | \mathbf{s}', \mathbf{s}) \right] \quad (5)$$

## 4. 如何求解 $q_{\xi}(\mathbf{a} | \mathbf{s}', \mathbf{s})$ ?

$q_{\xi}(\mathbf{a} | \mathbf{s}', \mathbf{s})$  的定义如下

$$q_{\xi}(\mathbf{a} | \mathbf{s}', \mathbf{s}) = q(a_1 | \mathbf{s}, \mathbf{s}') \prod_{k=2}^K q(a_k | f_{\xi}(a_{k-1}, \mathbf{s}, \mathbf{s}')), \quad (6)$$

即定义每一步行动都服从一个参数化的高斯分布

$$q(a_k) = \mathcal{N}(a_k | \mu_\xi(a_{k-1}, \mathbf{s}, \mathbf{s}'), \sigma_\xi^2(a_{k-1}, \mathbf{s}, \mathbf{s}')) \quad (19)$$

求解的方法就是通过最大化  $\mathbb{E}_{\mathbf{s}, \mathbf{s}'} [\ln q_\xi(\mathbf{a} | \mathbf{s}, \mathbf{s}')] / \beta$  来更新神经网络的参数  $\xi$ 。

#### 4. 如何求解 $w(\mathbf{a} | \mathbf{s})$ ?

对目标函数求导并且取导数为零，可以得到最优的  $w(\mathbf{a} | \mathbf{s})$ 。

$$w^*(\mathbf{a} | \mathbf{s}) = \frac{1}{Z(\mathbf{s})} \exp(\hat{u}(\mathbf{s}, \mathbf{a})) = \frac{1}{Z(\mathbf{s})} \exp(\beta \mathbb{E}_{p(\mathbf{s}' | \mathbf{s}, \mathbf{A})} [\ln q_\xi(\mathbf{a} | \mathbf{s}, \mathbf{s}')]])$$

在最优情形下

$$\mathcal{E}(\mathbf{s}) = \frac{1}{\beta} \log Z(\mathbf{s})$$

于是我们可以对这个最优的  $w(\mathbf{a} | \mathbf{s})$  做参数化

$$\omega^*(\mathbf{a} | \mathbf{s}) \approx h_\theta(\mathbf{a} | \mathbf{s}) \Rightarrow \hat{u}(\mathbf{s}, \mathbf{a}) \approx r_\theta(\mathbf{s}, \mathbf{a}); \quad r_\theta(\mathbf{s}, \mathbf{a}) = \ln h_\theta(\mathbf{a} | \mathbf{s}) + \psi_\theta(\mathbf{s}). \quad (7)$$

并且通过最小化如下均方误差来优化相应的参数

$$L(h_\theta, \psi_\theta) = \mathbb{E}_{p(\mathbf{s}' | \mathbf{s}, \mathbf{A})} [(\beta \ln q_\xi(\mathbf{a} | \mathbf{s}, \mathbf{s}') - r_\theta(\mathbf{s}, \mathbf{a}))^2]. \quad (8)$$

最后要求的empowerment也可以通过  $\mathcal{E}(\mathbf{s}) = \frac{1}{\beta} \log \psi_\theta(\mathbf{s})$  求到。

#### 5. 处理高维输入

由于这里解决的都是高维度（图像）输入的问题，所以顺带着还要学一个CNN的embedding，CNN的参数用  $\lambda$  表示。由于上面两个使用梯度下降的优化问题都需要用到这个embedding，因此就对这两个优化目标的和做梯度下降来优化这个CNN的embedding。

#### 算法

---

##### Algorithm 1: Stochastic Variational Information Maximisation for Empowerment

---

Parameters:  $\xi$  variational,  $\lambda$  convolutional,  $\theta$  source

```

while not converged do
   $\mathbf{x} \leftarrow \{\text{Read current state}\}$ 
   $\mathbf{s} = \text{ConvNet}_\lambda(\mathbf{x})$  {Compute state repr.}
   $\mathbf{A} \sim \omega(\mathbf{a} | \mathbf{s})$  {Draw action sequence.}
  Obtain data  $(\mathbf{x}, \mathbf{a}, \mathbf{x}')$  {Acting in env.}
   $\mathbf{s}' = \text{ConvNet}_\lambda(\mathbf{x}')$  {Compute state repr.}
   $\Delta \xi \propto \nabla_\xi \log q_\xi(\mathbf{a} | \mathbf{s}, \mathbf{s}') \quad (18)$ 
   $\Delta \theta \propto \nabla_\theta L(h_\theta, \psi_\theta) \quad (8)$ 
   $\Delta \lambda \propto \nabla_\lambda \log q_\xi(\mathbf{a} | \mathbf{s}, \mathbf{s}') + \nabla_\lambda L(h_\theta, \psi_\theta)$ 
end while
 $\mathcal{E}(\mathbf{s}) = \frac{1}{\beta} \log \psi_\theta(\mathbf{s})$  {Empowerment}

```

---

知乎 @张楚珩

## 实验结果

这里只截取两个实验来看看吧，学出来的结果看起来十分合理。比如，障碍物附近可以移动的空间较小，因此empowerment小；如果可移动的盒子附近，相当于手头有了工具，这样可操作空间就更大。

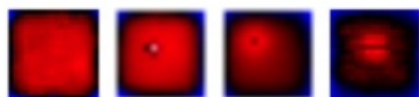


Figure 4: Empowerment for a room environment, showing a) an empty room, b) room with an obstacle c) room with a moveable box, d) room with row of moveable boxes.



Figure 5: Left: empowerment landscape for agent and key scenario. Yellow is the key and green is the door. Right: Agent in a corridor with flowing lava. The agent places a brick to stem the flow of lava.

## 本文的数学推导？

会用到以下一些数学知识

1. 信息论的公式：公式(3)，通过查wiki能够得到 Mutual Information / Conditional Entropy
2. Variational Lower Bound：公式(4)，原理很简单，对于expected log-probability，如果把分布换成任意其他的分布，数值就会更小，中间相差了一个恒大于零的KL divergence。
3. Lagrange Multiplier：  $w^*(a|s)$  的推导会用到。这里稍微写一下。

要解决的问题是

$$\max_{\mathbf{w}} \mathbb{E}_{p(s'|s,a)w(a|s)} \left[ -\frac{1}{\beta} \ln w(a|s) + \ln q_\beta(a|s',s) \right] \quad \text{s.t.} \sum_a w(a|s) = 1$$

拉格朗日函数（由于都condition on  $s$ ，这里就省略了）

$$\mathcal{L} = \left( \sum_a \sum_{s'} w(a) p(s'|a) \left[ -\frac{1}{\beta} + \ln q_\pi(a|s') \right] \right) - \lambda \left( \sum_a w(a) - 1 \right)$$

令  $\frac{\partial \mathcal{L}}{\partial w(a)} = 0$ ，有

$$\ln(Cw(a)) = \ln(w(a) \exp(\beta\lambda + 1)) = \beta \mathbb{E}_{s' \sim p(s'|s,a)} [\ln q_\pi(a'|s')] = \hat{u}(s, a)$$

再利用  $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$  能够得到归一化因子的表达式，即文中的结果。

## 参考文献

[1] Oudeyer, Pierre-Yves, and Frederic Kaplan. "What is intrinsic motivation? A typology of computational approaches." *Frontiers in neurorobotics* 1 (2009): 6.

发布于 2018-11-02

机器学习

算法

强化学习 (Reinforcement Learning)

▲ 赞同 5

▼

💬 添加评论


🔗 分享

♥ 喜欢

★ 收藏

...

## 文章被以下专栏收录

 强化学习前沿  
读呀读paper

进入专栏