Trust Region Policy Optimization

TRPO
置信域策略优化
第一部分

东川路第一可爱猫猫虫

主要内容

- TRPO的优势
- 新旧策略期望回报的数值关系
- 折扣访问频率（状态占据度量）ρ
- 代理函数与MM算法
- KL散度与TV散度
- Pinsker不等式及其应用
- 从惩罚项到置信域
- 从最大KL散度到平均KL散度
- 基于泰勒展开的近似与海森矩阵

感谢
-泰深满嘴长牙
秋_0427
两位粉丝大佬的充电

Trust Region Policy Optimization

- 置信域

  旧策略的邻域

- 置信域策略优化

  在旧策略的邻域里优化策略

  用KL散度来衡量新旧策略的远近，并将其限制在阈值内

- 从理论上证明了单调改进

新旧策略的期望回报

- 我们的优化目标

  策略π的期望回报 $$\eta(\pi) = \mathbb{E}_{s_0, a_0, \ldots}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]$$

- Kakade & Langford, 2002

  新策略的期望回报=

  旧策略的期望回报+新策略在旧策略优势函数上的累计期望

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{s_0, a_0, \ldots \sim \tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A_{\pi}(s_t, a_t)\right]$$

Discounted visitation frequencies

- 期望具有线性性质

  若级数收敛，则求期望和求无穷级数可以交换顺序

$$\mathbb{E}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right]=\sum_{t=0}^{\infty}\gamma^t\cdot\mathbb{E}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left[A_\pi(s_t,a_t)\right]$$

- 引入折扣访问频率（状态占据度量）

$$\rho_\pi(s)=P(s_0=s)+\gamma P(s_1=s)+\gamma^2 P(s_2=s)+\ldots$$

- 表示状态s在策略π下的长期权重
- 经过t步后处于状态s的折扣概率之和

$$\rho_{\tilde{\pi}}(s)=\sum_{t=0}^{\infty}\gamma^t\cdot\mathbb{P}_{s_0\sim d,\tilde{\pi}}\left[s_t=s\right]$$

数学推导

- 将$A_\pi(s_t, a_t)$视为随机变量

  则它的期望可以分解为：

$$\mathbb{E}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left[A_\pi(s_t,a_t)\right]=\sum_s\sum_a A_\pi(s,a)\cdot\mathbb{P}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left(s_t=s,\ a_t=a\right)$$

- 求和部分的第二项：

$$\mathbb{P}\left[s_t=s,\ a_t=a\right]=\mathbb{P}\left[s_t=s\right]\cdot\mathbb{P}\left[a_t=a\mid s_t=s\right]$$

- 由于是马尔可夫链，动作选择仅依赖当前状态

$$\mathbb{P}\left[a_t=a\mid s_t=s\right]=\tilde{\pi}(a\mid s)$$

$$\mathbb{P}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left(s_t=s,\ a_t=a\right)=\mathbb{P}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left(s_t=s\right)\cdot\tilde{\pi}(a\mid s)$$

$$\mathbb{E}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left[A_\pi(s_t,a_t)\right]=\sum_s\sum_a A_\pi(s,a)\cdot\mathbb{P}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left(s_t=s\right)\cdot\tilde{\pi}(a\mid s)$$

引入代理函数

$$\mathbb{E}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left[\gamma^t A_\pi(s_t,a_t)\right] = \gamma^t \cdot \left(\sum_s \sum_a A_\pi(s,a)\cdot \mathbb{P}(s_t=s)\cdot\tilde{\pi}(a\mid s)\right)$$

$$\sum_{t=0}^{\infty}\sum_s\sum_a \gamma^t\cdot A_\pi(s,a)\cdot\mathbb{P}(s_t=s)\cdot\tilde{\pi}(a\mid s) = \sum_s\sum_a\left(\sum_{t=0}^{\infty}\gamma^t\cdot\mathbb{P}(s_t=s)\right)\cdot\tilde{\pi}(a\mid s)\cdot A_\pi(s,a)$$

$$= \sum_s \rho_{\tilde{\pi}}(s)\sum_a \tilde{\pi}(a|s)A_\pi(s,a)$$

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\tilde{\pi}}(s)\sum_a \tilde{\pi}(a|s)A_\pi(s,a)$$

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_{\pi}(s)\sum_a \tilde{\pi}(a|s)A_\pi(s,a)$$

- 在旧策略π处，L_π和η的梯度相同

  在旧策略附近优化L_π，就近似于优化η

如何衡量策略的差异

- 两个离散的概率分布p和q
- KL散度

$$D_{KL}(P||Q) = E_p\left[\log\frac{P(x)}{Q(x)}\right] = \Sigma_i p_i \log\frac{p_i}{q_i}$$

衡量两个概率分布的远近

表示用q来近似p时的信息损失

- TV散度

$$D_{\mathrm{TV}}(P||Q) = \frac{1}{2}\Sigma_i |p_i - q_i|$$

Pinsker不等式 $\|Q-P\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(Q\|P)}$

- 由其简化形式得到:

  $D_{TV}(p||q)^2 \leq D_{KL}(p||q)$
- 论文证明了L_π和η的误差下界

  $$\eta(\pi_{new}) \geq L_{\pi_{old}}(\pi_{new}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2$$
- 其中

  α为新旧两个策略在所有状态下的最大总变差散度

  $$\epsilon = \max_{s,a} |A_\pi(s,a)|$$

$$\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - C \cdot D_{KL}^{\max}(\pi, \tilde{\pi})$$

$$\eta(\tilde{\pi}) \geq L_{\pi}(\tilde{\pi}) - C \cdot D_{KL}^{\max}(\pi, \tilde{\pi})$$

- 不等式右边成为新策略的性能下界

  $$M(\pi) = L_{\pi}(\tilde{\pi}) - C \cdot D_{KL}^{\max}(\pi, \tilde{\pi})$$

- 只要我们提升$M(\pi)$

  或者说：最大化$M(\pi)$

  就能保证η的性能单调上升

- MM

  Minorization-Maximization

  每次迭代找到目标函数的一个下界函数

  不断求这个下界函数的最大值

置信域的引入

- C通常很大

  导致惩罚项权重过高

  每轮参数更新幅度过小

- 将惩罚项改为置信域

$$\underset{\theta}{\text{maximize}}\left[L_{\theta_{old}}(\theta)-C\cdot D_{KL}^{\max}(\theta_{old},\theta)\right]$$

- 变成：

$$\begin{aligned}\underset{\theta}{\text{maximize}} \quad & L_{\theta_{old}}(\theta)\\ \text{subject to} \quad & \overline{D}_{KL}^{\rho_{\theta_{old}}}(\theta_{old},\theta)\leq\delta\end{aligned}$$

最大KL散度变为平均KL散度

- 最大KL散度

  要求所有状态的KL散度都小于某个值

  难以实现

- 平均KL散度

  实际上约束的是

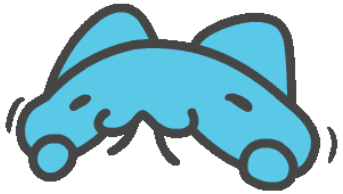
  旧策略访问到的状态的平均KL散度

$$\overline{D}_{\mathrm{KL}}^{\rho}(\theta_1, \theta_2) := \mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}(\pi_{\theta_1}(\cdot|s) \parallel \pi_{\theta_2}(\cdot|s))\right]$$

重要性采样

- 我们想计算$E_{X\sim q}[f(X)]$

  从q上采样，可能不是最优的



$$\begin{aligned}\mathbb{E}_{X\sim q}[f(X)] &= \int f(x)q(x)dx \\ &= \int f(x)\cdot\frac{q(x)}{p(x)}\cdot p(x)dx \\ &= \mathbb{E}_{X\sim p}\left[f(X)\cdot\frac{q(x)}{p(x)}\right]\end{aligned}$$

- $\frac{q(x)}{p(x)}$称为重要性权重

  从p(x)采样的样本修正到q(x)分布下的期望估计

重要性采样

$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad & L_{\theta_{old}}(\theta) \\ \text{subject to} \quad & \overline{D}_{KL}^{\rho_{\theta_{old}}}(\theta_{old}, \theta) \leq \delta \end{aligned}$$

$$L_{\theta_{\text{old}}}(\theta) = \mathbb{E}_{s\sim\rho_{\theta_{\text{old}}}, a\sim\pi_\theta}\left[A_{\pi_{\theta_{\text{old}}}}(s,a)\right]$$

- $L_{\theta_{old}}(\theta)$指的是新策略θ相比于旧策略θ_{old}的性能改进期望

  这里的a需要从新策略θ里采样动作

  可我们只能从旧策略θ_{old}里采样动作

  因此使用重要性采样:

$$\mathbb{E}_{a\sim\pi_\theta}[f(a)] = \mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}}\left[f(a)\cdot\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}\right]$$

$$L_{\theta_{\text{old}}}(\theta) = \mathbb{E}_{s\sim\rho_{\theta_{\text{old}}}, a\sim\pi_{\theta_{\text{old}}}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}\cdot A_{\pi_{\theta_{\text{old}}}}(s,a)\right]$$

近似

- TRPO在$\theta = \theta_{old}$附近做近似:
- 目标函数L与约束条件

  目标函数$L_{\theta_{old}}(\theta)$ 在$\theta = \theta_{old}$做一阶泰勒展开

  约束条件（平均KL散度）在$\theta = \theta_{old}$做二阶泰勒展开

$$L_{\theta_{old}}(\theta) \approx L_{\theta_{old}}(\theta_{old}) + \nabla_\theta L_{\theta_{old}}(\theta_{old}) \cdot (\theta - \theta_{old})$$

$$\overline{D}_{KL}(\theta_{old}, \theta_{old}) \quad \nabla_\theta \overline{D}_{KL}(\theta_{old}, \theta)\big|_{\theta=\theta_{old}}$$

$$\overline{D}_{KL}^{\rho_{\theta_{old}}}(\theta_{old}, \theta) \approx \frac{1}{2}\Delta\theta^T A \Delta\theta$$

- 这里的Δθ为参数的更新量

  A为平均KL散度在θ_{old}处的Hessian矩阵

海森矩阵

- 海森矩阵

  多元函数二阶偏导数构成的对称矩阵

  f(X)在X(0)处的海森矩阵为

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}_{X^{(0)}}$$

- A就是

$$\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \mathbb{E}_{s \sim \rho_\pi} \left[D_{\mathrm{KL}}(\pi(\cdot|s, \theta_{\mathrm{old}}) \parallel \pi(\cdot|s, \theta)) \right] \Big|_{\theta = \theta_{\mathrm{old}}}$$

问题转化为

$$\max \ g^T(\theta - \theta_{old})$$

$$s.t. \ \frac{1}{2}(\theta - \theta_{old})^T H(\theta - \theta_{old}) \leq \delta$$

- g就是刚才的$\nabla_\theta L_{\theta_{old}}(\theta_{old})$
- H就是刚才的海森矩阵