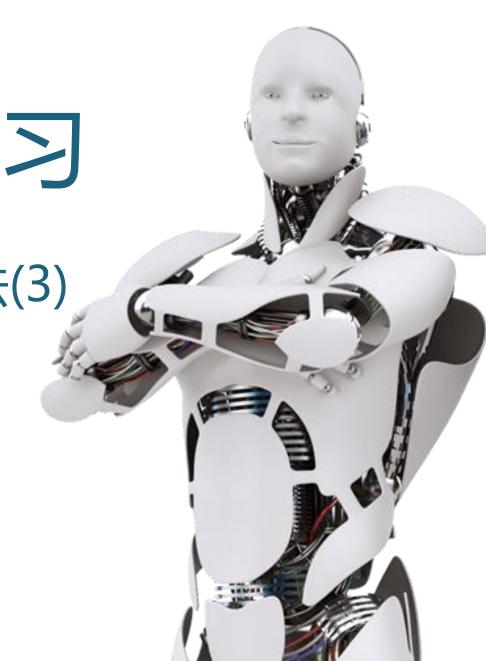
彪哥带你学强化学习

16.深入理解TRPO算法(3)

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

讲师: 韩路彪



TRPO算法 —— 整体结构

理论

- 每训练一步,整体回报都有提升
- 目标函数η,近似函数L
- 如果能保证 | L_π η | <= X , 就能在 ▽ L_π >= X 前提下做到 ▽η >=0
 理论实现
- 给出 | L_π η | <= X 里边的上限X

工程实现

• 实现每训练一步η都有提升

TRPO算法 —— 工程实现

工程实现的内容结构

- TV散度转KL散度
- 优化目标θ化
 - $\pi \rightarrow \theta$
 - 惩罚系数C导致更新太慢
 - maxKL散度不好算,近似成平均KL散度
- 如何基于蒙特卡洛采样求解
 - 采样的方法
 - 求和变期望
- 优化更新算法
 - 自然梯度
 - 泰勒展开
 - KKT条件
 - 共轭梯度法 (优化目标函数)
 - 优化方向
 - 优化步长

TRPO算法 —— 工程实现 —— TV散度转KL散度

$$egin{aligned} \eta(ilde{\pi}) &\geq L_{\pi}(ilde{\pi}) - rac{4lpha^2\gamma\epsilon}{(1-\gamma)^2} \ C &= rac{4\epsilon\gamma}{(1-\gamma)^2} \ lpha &= D_{TV}^{max}ig(\pi_{old},\pi_{new}ig) \ D_{TV}(p||q)^2 &\leq D_{KL}(p||q) \ D_{KL}^{max}ig(\pi, ilde{\pi}ig) &= maxD_{KL}ig(\pi(.\,|s)|| ilde{\pi}(.\,|s)ig) \end{aligned}$$

$$\eta(ilde{\pi}) \geq L_{\pi}(ilde{\pi}) - CD_{KL}^{max}(\pi, ilde{\pi})$$

TRPO算法 —— 工程实现 —— 优化目标θ化

$$\eta(ilde{\pi}) \geq L_{\pi}(ilde{\pi}) - CD_{KL}^{max}(\pi, ilde{\pi})$$

• $\pi \rightarrow \theta$

Since we consider parameterized policies $\pi_{\theta}(a|s)$ with parameter vector θ , we will overload our previous notation to use functions of θ rather than π , e.g. $\eta(\theta) := \eta(\pi_{\theta})$, $L_{\theta}(\tilde{\theta}) := L_{\pi_{\theta}}(\pi_{\tilde{\theta}})$, and $D_{\mathrm{KL}}(\theta \parallel \tilde{\theta}) := D_{\mathrm{KL}}(\pi_{\theta} \parallel \pi_{\tilde{\theta}})$ We

• 惩罚系数C导致更新太慢,优化目标近似为

$$\underset{\theta}{\operatorname{maximize}} \, L_{\theta_{\operatorname{old}}}(\theta)$$

subject to $D_{\mathrm{KL}}^{\mathrm{max}}(\theta_{\mathrm{old}}, \theta) \leq \delta$.

δ是个超参数,论文的实验用的是0.01

• maxKL散度不好算,近似成平均KL散度,优化目标近似为:

$$\underset{\theta}{\operatorname{maximize}} L_{\theta_{\operatorname{old}}}(\theta)$$

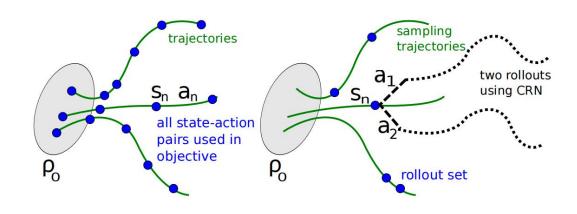
subject to $\overline{D}_{\mathrm{KL}}^{\rho_{\theta_{\mathrm{old}}}}(\theta_{\mathrm{old}}, \theta) \leq \delta$.

实验表明结果类似



TRPO算法 —— 工程实现 —— 基于蒙特卡洛采样求解

采样方法: 单路经采样、藤蔓采样



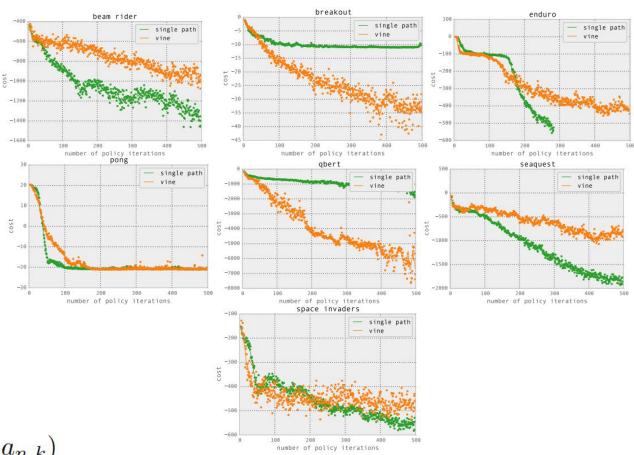
藤蔓采样:主干的每个状态再采样 k 个action

- 连续问题用π采样效果较好
- 离散问题用均匀采样效果较好

小型、离散:
$$L_n(\theta) = \sum_{k=1}^K \pi_{\theta}(a_k|s_n)\hat{Q}(s_n,a_k)$$

大型、连续:
$$L_n(\theta) = \frac{\sum_{k=1}^K \frac{\pi_{\theta}(a_{n,k}|s_n)}{\pi_{\theta_{\text{old}}}(a_{n,k}|s_n)} \hat{Q}(s_n, a_{n,k})}{\sum_{k=1}^K \frac{\pi_{\theta}(a_{n,k}|s_n)}{\pi_{\theta_{\text{old}}}(a_{n,k}|s_n)}}$$

结果对比





TRPO算法 —— 工程实现 —— 基于蒙特卡洛采样求解

求和变期望

$$L_{ heta_{old}}(heta) = \eta(heta_{old}) + \sum_s
ho_{ heta_{old}}(s) \sum_a \pi_{ heta}(a \mid s) A_{ heta_{old}}(s,a)$$

$$= \eta(heta_{old}) + rac{1}{1-\gamma} \mathbb{E}_{s \sim
ho_{ heta_{old}}} \Bigg[\sum_a \pi_{ heta}(a \mid s) A_{ heta_{old}}(s,a) \Bigg].$$

$$= \! \eta(heta_{old}) + rac{1}{1-\gamma} \mathbb{E}_{s \sim
ho_{ heta_{old}}} \mathbb{E}_{a \sim q} igg[rac{\pi_{ heta}(a|s)}{q_(a|s)} A_{ heta_{old}}(s,a) igg]$$

状态访问分布归一化因子

$$\sum_{t=0}^{\infty} \sum_{s_t \sim S} \gamma^t P_{\pi}(s_t)$$

$$=\sum_{t=0}^{\infty} \gamma^t \sum_{s_t \sim S} P_{\pi}(s_t)$$

$$=\sum_{t=0}^{\infty}\gamma^t$$

$$= rac{1}{1-\gamma}$$

$$\max_{ heta} L_{ heta_{old}}(heta)$$

$$s.\,t.\,ar{D}_{KL}^{
ho_{ heta old}}(heta_{old}, heta) \leq \delta$$

$$\max_{ heta} \mathbb{E}_{s \sim
ho_{ heta_{old}}, a \sim q} \left\lfloor rac{\pi_{ heta}(a|s)}{q_(a|s)} Q_{ heta_{old}}(s,a)
ight
floor$$

$$s.\,t.\,\mathbb{E}_{s\sim
ho_{ heta_{old}}}[D_{KL}(\pi_{ heta_{old}}(.\,|s)||\pi_{ heta}(.\,|s))] \leq \delta$$

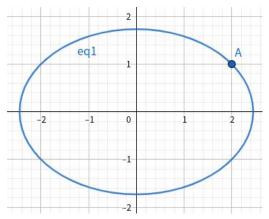


TRPO算法 —— 工程实现 —— 参数空间与目标空间

e.g. 求
$$z = x^2 + 2y^2$$
 最小值

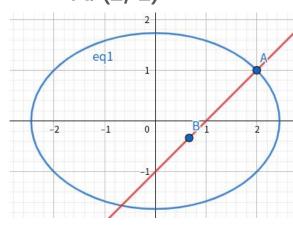
参数空间

目标空间



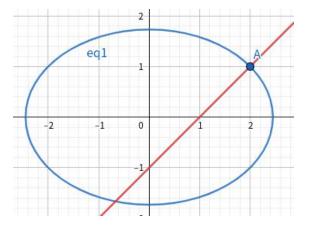
$$x^2 + 2y^2 = 6$$

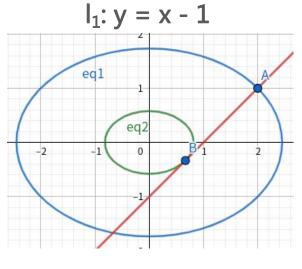
A: (2, 1)



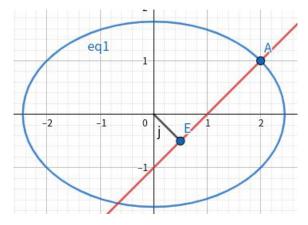
B:
$$(2/3, -1/3)$$

z = $2/3$





$$x^2 + 2y^2 = 2/3$$



E: (1/2, -1/2)z = 3/4

TRPO算法 —— 工程实现 —— 优化更新算法 —— 泰勒展开

$$\max_{ heta} \mathbb{E}_{s \sim
ho_{ hetaold}, a \sim q} \left[rac{\pi_{ heta}(a|s)}{q_(a|s)} Q_{ heta_{old}}(s,a)
ight]$$

$$s.\,t.\,\mathbb{E}_{s\sim
ho_{ heta ld}}[D_{KL}(\pi_{ heta_{old}}(.\,|s)||\pi_{ heta}(.\,|s))] \leq \delta$$

$$\max_{\theta} l(\theta)$$

 $\max_{ heta} l(heta) \qquad s.\,t. \quad kl(heta) \leq \delta$

$$l(heta) pprox l(heta_{old}) + igtriangledown_{ heta}^T g_l$$

$$kl(heta)pprox kl(heta_{old})+igtriangledown_{ heta}^Tg_{kl}+rac{1}{2}igtriangledown_{ heta}^TH_{kl}igtriangledown_{ heta}=rac{1}{2}igtriangledown_{ heta}^TH_{kl}igtriangledown_{ heta}$$

$$\max_{ heta} igtriangledown_{ heta}^T g_l$$

 $\max_{ heta} igtriangledown_{ heta}^T g_l \qquad s.\ t. \quad rac{1}{2} igtriangledown_{ heta}^T H_{kl} igtriangledown_{ heta} \leq \delta$



TRPO算法 —— 工程实现 —— 优化更新算法 —— KKT条件

$$\max_{ heta} igtriangledown_{ heta}^T g_l$$

$$\max_{ heta} igtriangledown_{ heta}^T g_l \qquad s.\,t. \quad rac{1}{2} igtriangledown_{ heta}^T H_{kl} igtriangledown_{ heta} \leq \delta$$

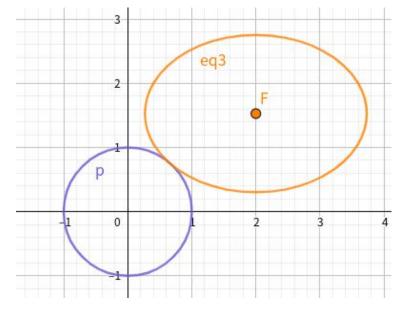
KKL条件

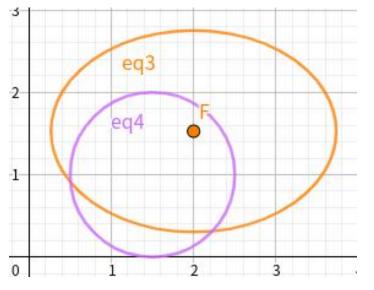
$$\mathcal{L}(
abla_{ heta}, \lambda) =
abla_{ heta}^T g_l - \lambda (rac{1}{2}
abla_{ heta}^T H_{kl}
abla_{ heta} - \delta)$$

$$rac{\partial \mathcal{L}}{\partial
abla_{ heta}} = g_l - \lambda H_{kl}
abla_{ heta} = 0 \quad rac{1}{2} igtharpooldown_{ heta}^T H_{kl} igtharpooldown_{ heta} \leq \delta$$

$$\lambda \geq 0$$
 $\lambda (\frac{1}{2} \nabla_{\theta}^T H_{kl} \nabla_{\theta} - \delta) = 0$

$$abla_{ heta}^* = rac{1}{\lambda} H_{kl}^{-1} g_l$$
 (自然梯度)

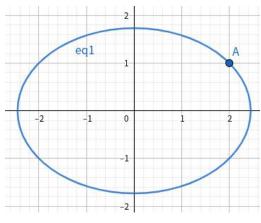


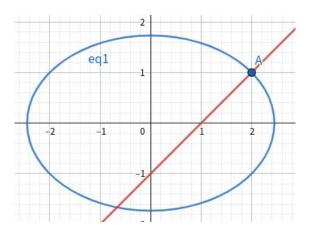


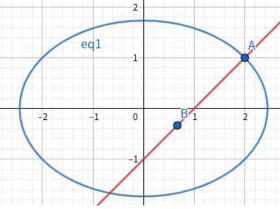
TRPO算法 —— 工程实现 —— 优化更新算法 —— 共轭梯度法求解

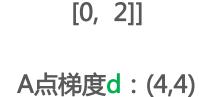
共轭方向法:在关于H共轭方向优化,类似按每个坐标轴方向优化

e.g. 求 $z = x^2 + 2y^2$ 最小值



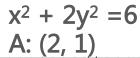




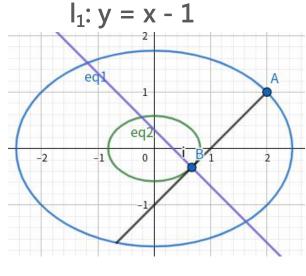


二次型矩阵

J: [[1, 0],

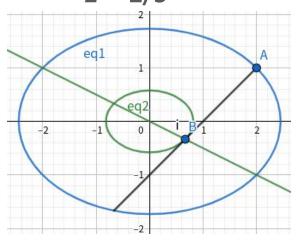


eq1



B:
$$(2/3, -1/3)$$

z = $2/3$



垂直方向:(8,-4) 也就是(1,-0.5)

 $x^2 + 2y^2 = 2/3$

$$I_2$$
: $y = -x + 1/3$

$$I_3$$
: y = -0.5x



$$H_{kl}x=g_l$$

共轭梯度法条件:H 需要对称正定

H是KL散度的二阶导数,等于Fisher信息矩阵,对称半正定

实际计算中,浮点数误差会使得 H 近似正定 另外,实际算法中有最大迭代次数,也能得到一个近似解

优化思路:关键就是找方向和步长

方向:按共轭梯度法迭代,找到共轭方向

步长:

- 1. 用KL约束计算最大步长
- 2. 用最大步长作为初始步长做线性搜索,直到满足条件(目标函数有提升,且kl散度小于阈值)
- 3. 如果搜索次数达到阈值后没找到符合条件的参照,那么参数不做更新

TRPO算法 —— 工程实现 —— 优化更新算法 —— 共轭梯度法求解 —— 方向

$$H_{kl}x=g_l$$

方向:按共轭梯度法迭代,找到方向 $s \approx H_{kl}^{-1} g_l$

初始化: $x_0=0$, 残差: $r_0=g-Hx_0=g$, 搜索方向: $p_0=r_0$

迭代: k=1,2,3... 直到残差r<阈值 或者达到最大迭代次数后停止

$$egin{aligned} lpha_k = rac{r_k^T r_k}{p_k^T H p_k} & x_{k+1} = x_k + lpha_k p_k & r_{k+1} = r_k - lpha_k H p_k \end{aligned}$$

$$eta_k = rac{r_{k+1}^T r_{k+1}}{r_k^T r_k} \quad p_{k+1} = r_{k+1} + eta_k p_k$$

HVP (Hessian-vector product)

$$abla_{ heta}(p^T \cdot
abla_{ heta} KL) = H \cdot p$$



TRPO算法 —— 工程实现 —— 优化更新算法 —— 共轭梯度法求解 —— 步长

$$H_{kl}x=g_l$$

步长:

1. 用KL约束计算最大步长
$$\delta=ar{D}_{KL}pproxrac{1}{2}(eta s)^TH(eta s)=rac{1}{2}eta^2 s^THs$$
 $eta=\sqrt{2\delta/s^THs}$

- 2. 用最大步长作为初始步长做线性搜索,直到满足条件(目标函数有提升,且kl散度小于阈值) $\alpha^n\beta$ n=0,1,2...
- 3. 如果搜索次数达到阈值后没找到符合条件的参照,那么参数不做更新