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 $|L_\pi - \eta| <= X$， 就能在 $\nabla L_\pi >= X$ 前提下做到 $\nabla \eta >= 0$

理论实现

- 给出 $|L_\pi - \eta| <= X$ 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
  - π -> θ
  - 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

$$\eta(\tilde{\pi}) \geq L_{\pi}(\tilde{\pi}) - \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2}$$

$$C = \frac{4\epsilon\gamma}{(1-\gamma)^2}$$

$$\alpha = D_{TV}^{max}(\pi_{old}, \pi_{new})$$

$$D_{TV}(p||q)^2 \leq D_{KL}(p||q)$$

$$D_{KL}^{max}(\pi, \tilde{\pi}) = max D_{KL}(\pi(.|s)||\tilde{\pi}(.|s))$$

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

TRPO算法 —— 工程实现 —— 优化目标θ化

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

- π -> θ

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 $\boxed{D_{\mathrm{KL}}(\theta \,\|\, \tilde{\theta}) := D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\tilde{\theta}})}$. We

- 惩罚系数C导致更新太慢，优化目标近似为

$$\begin{aligned} &\underset{\theta}{\text{maximize}}\ L_{\theta_{\text{old}}}(\theta) \\ &\text{subject to } D_{\mathrm{KL}}^{\max}(\theta_{\text{old}}, \theta) \leq \delta. \end{aligned}$$

δ是个超参数，论文的实验用的是0.01

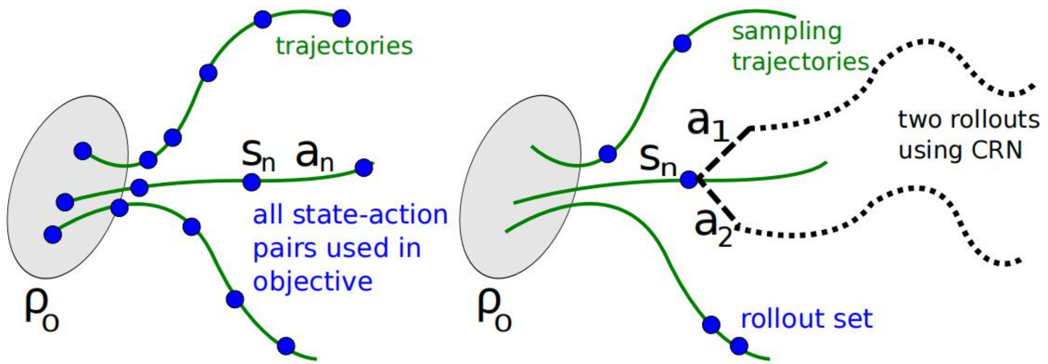
- maxKL散度不好算，近似成平均KL散度，优化目标近似为：

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

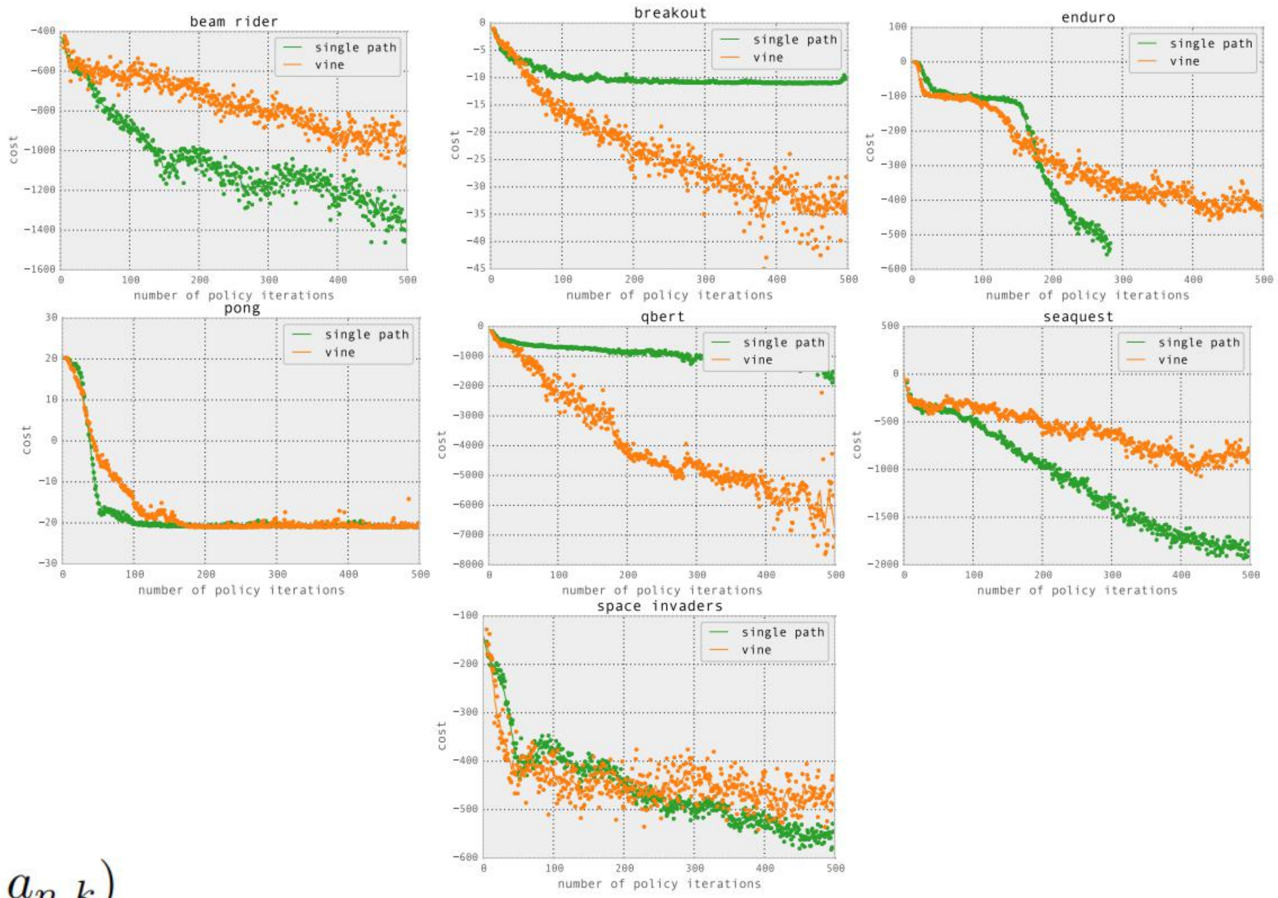
实验表明结果类似

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



结果对比



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

TRPO算法 —— 工程实现 —— 基于蒙特卡洛采样求解

求和变期望

$$L_{\theta_{old}}(\theta) = \eta(\theta_{old}) + \sum_s \rho_{\theta_{old}}(s) \sum_a \pi_\theta(a \mid s) A_{\theta_{old}}(s,a)$$

$$= \eta(\theta_{old}) + \frac{1}{1-\gamma} \mathbb{E}_{s\sim\rho_{\theta_{old}}}\left[\sum_a \pi_\theta(a \mid s) A_{\theta_{old}}(s,a)\right]$$

$$= \eta(\theta_{old}) + \frac{1}{1-\gamma} \mathbb{E}_{s\sim\rho_{\theta_{old}}} \mathbb{E}_{a\sim q}\left[\frac{\pi_\theta(a|s)}{q_(a|s)} A_{\theta_{old}}(s,a)\right]$$

状态访问分布归一化因子

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

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

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

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

$$\max_\theta L_{\theta_{old}}(\theta)$$

$$s.t.\ \bar{D}_{KL}^{\rho_{\theta_{old}}}(\theta_{old}, \theta) \le \delta$$

$$\Rightarrow$$

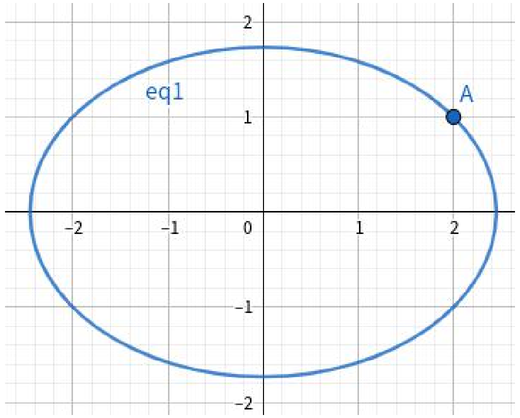
$$\max_\theta \mathbb{E}_{s\sim\rho_{\theta_{old}}, a\sim q}\left[\frac{\pi_\theta(a|s)}{q_(a|s)} Q_{\theta_{old}}(s,a)\right]$$

$$s.t.\ \mathbb{E}_{s\sim\rho_{\theta_{old}}}[D_{KL}(\pi_{\theta_{old}}(.|s)||\pi_\theta(.|s))] \le \delta$$

TRPO算法 —— 工程实现 —— 参数空间与目标空间

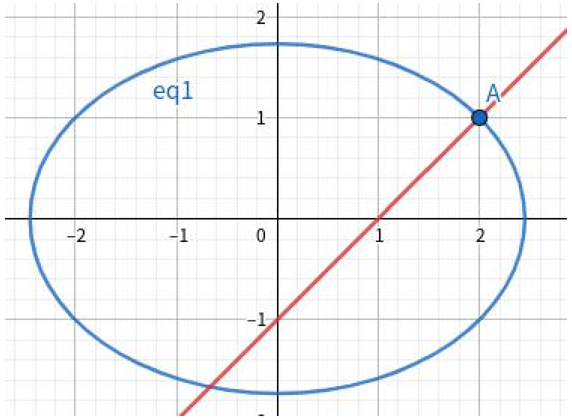
e.g. 求 $z = x^2 + 2y^2$ 最小值

参数空间

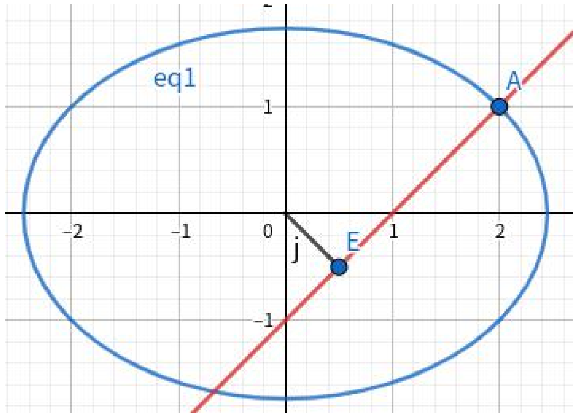


$x^2 + 2y^2 = 6$

A: (2, 1)



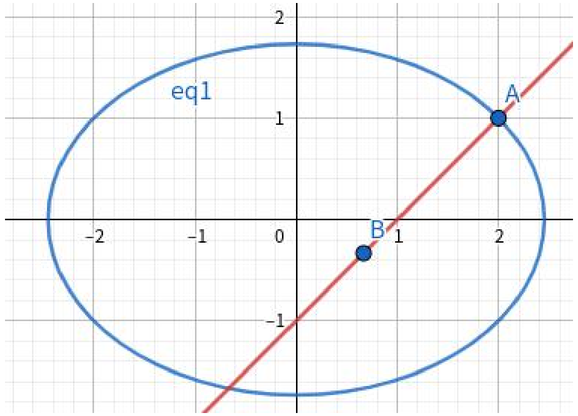
l_1: $y = x - 1$



E: (1/2, -1/2)

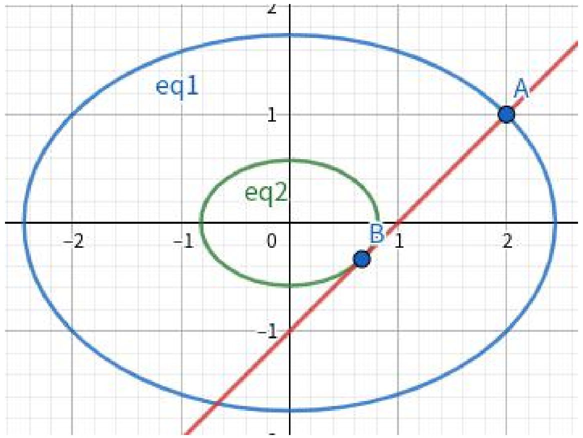
z = 3/4

目标空间



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

z = 2/3



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

TRPO算法 —— 工程实现 —— 优化更新算法 —— 泰勒展开

$$\max_{\theta} \mathbb{E}_{s\sim\rho_{\theta_{old}},a\sim q}\left[\frac{\pi_\theta(a|s)}{q_(a|s)}Q_{\theta_{old}}(s,a)\right]$$
$$s.t.\ \mathbb{E}_{s\sim\rho_{\theta_{old}}}[D_{KL}(\pi_{\theta_{old}}(.|s)||\pi_\theta(.|s))] \le \delta$$

$$\max_{\theta} l(\theta) \qquad s.t. \quad kl(\theta) \le \delta$$

$$l(\theta) \approx l(\theta_{old}) + \triangledown_\theta^T g_l$$

$$kl(\theta) \approx kl(\theta_{old}) + \triangledown_\theta^T g_{kl} + \frac{1}{2}\triangledown_\theta^T H_{kl} \triangledown_\theta = \frac{1}{2}\triangledown_\theta^T H_{kl}\triangledown_\theta$$

$$\max_{\theta} \triangledown_\theta^T g_l \qquad s.t. \quad \frac{1}{2}\triangledown_\theta^T H_{kl}\triangledown_\theta \le \delta$$

TRPO算法 —— 工程实现 —— 优化更新算法 —— KKT条件

$$\max_{\theta} \nabla_\theta^T g_l \qquad s.t. \quad \frac{1}{2} \nabla_\theta^T H_{kl} \nabla_\theta \le \delta$$

KKL条件

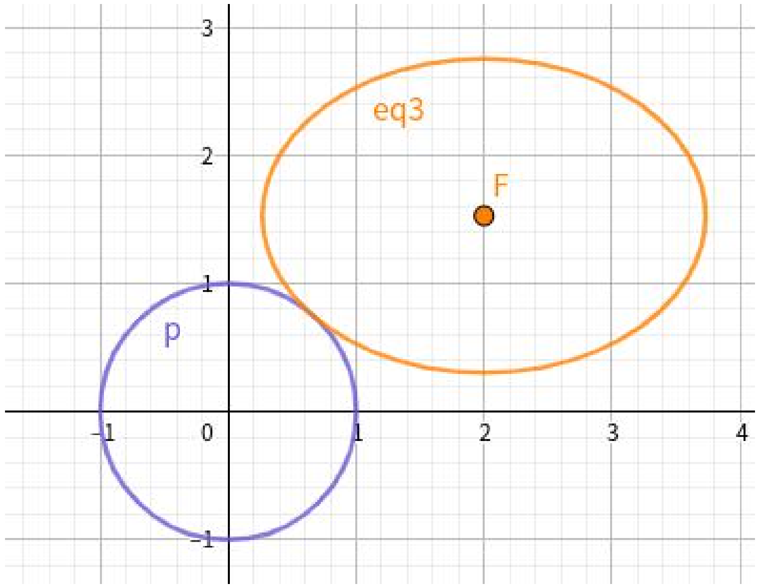
$$\mathcal{L}(\nabla_\theta, \lambda) = \nabla_\theta^T g_l - \lambda(\frac{1}{2}\nabla_\theta^T H_{kl}\nabla_\theta - \delta)$$

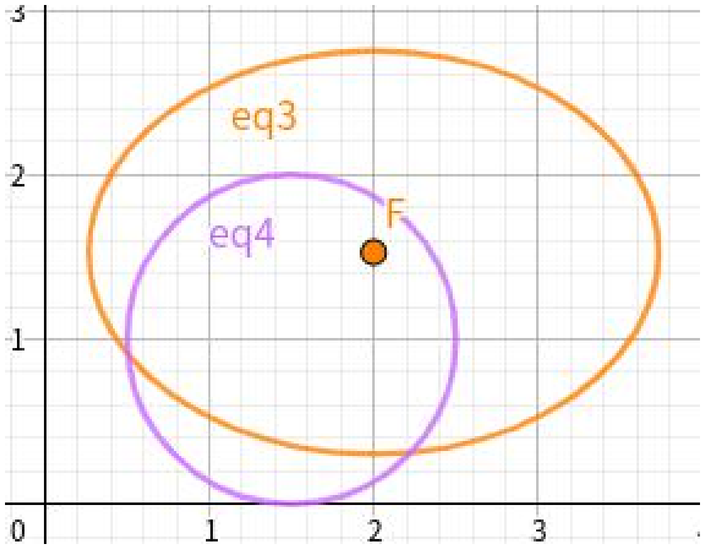
$$\frac{\partial \mathcal{L}}{\partial \nabla_\theta} = g_l - \lambda H_{kl}\nabla_\theta = 0 \qquad \frac{1}{2} \nabla_\theta^T H_{kl} \nabla_\theta \le \delta$$

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

$$\nabla_\theta^* = \frac{1}{\lambda} H_{kl}^{-1} g_l$$

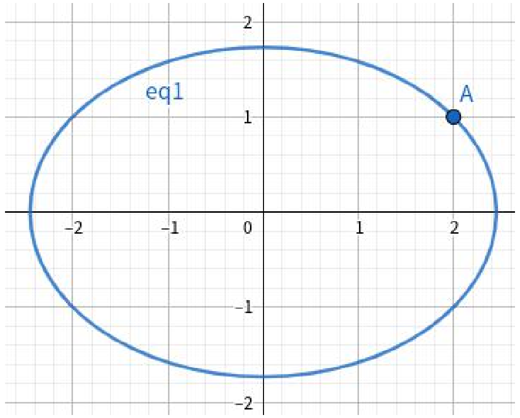
（自然梯度）



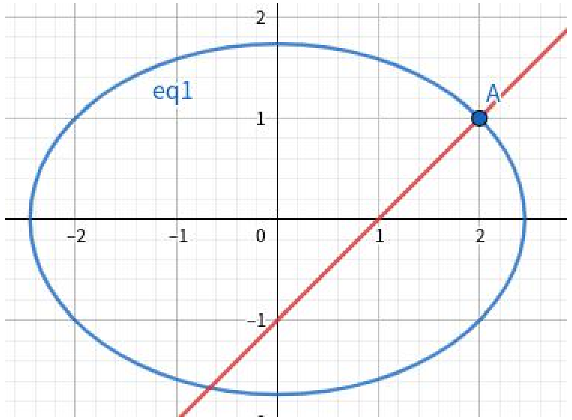


TRPO算法 —— 工程实现 —— 优化更新算法 —— 共轭梯度法求解

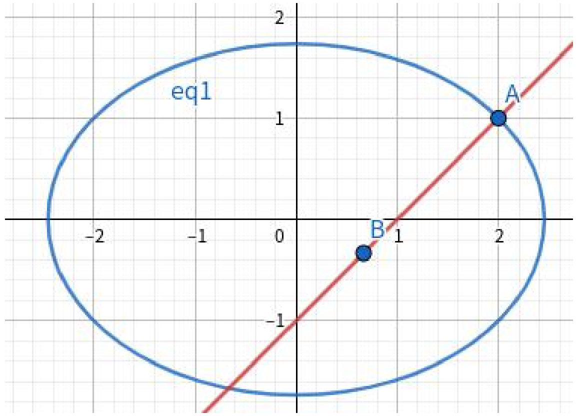
共轭方向法：在关于H共轭方向优化，类似按每个坐标轴方向优化　　e.g. 求 $z = x^2 + 2y^2$ 最小值



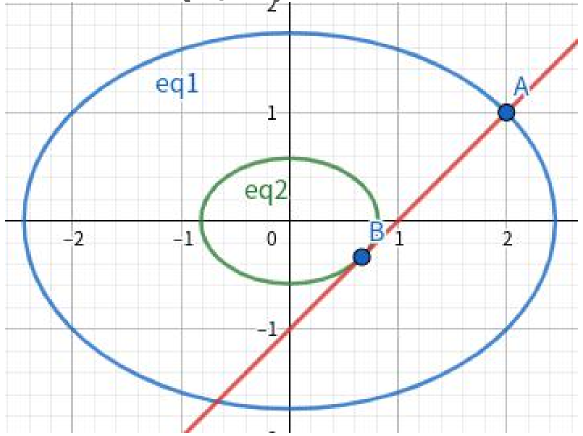
$x^2 + 2y^2 = 6$
A: (2, 1)



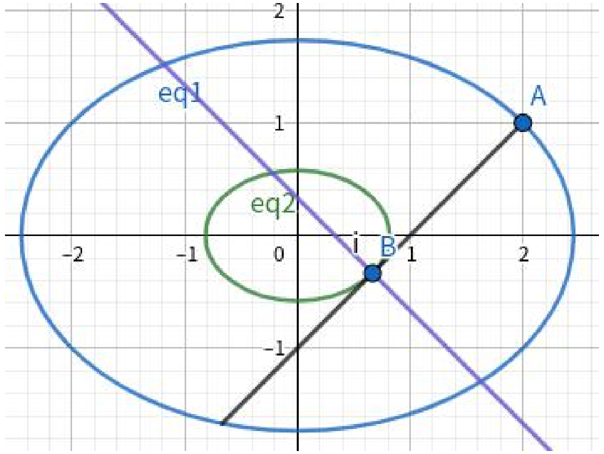
l_1: $y = x - 1$



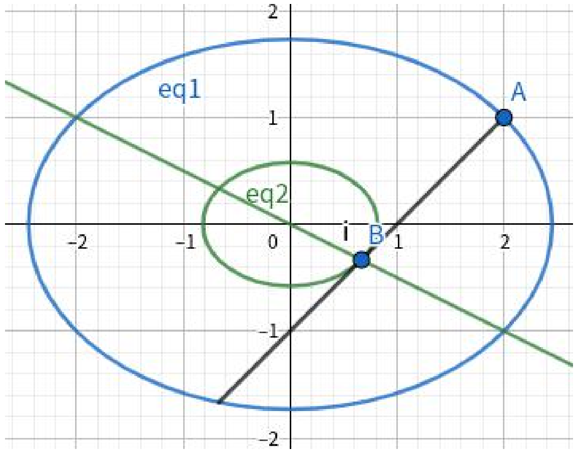
B: (2/3, -1/3)
$z = 2/3$



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



l_2: $y = -x + 1/3$



l_3: $y = -0.5x$

二次型矩阵
J: [[1, 0],
     [0, 2]]

A点梯度d：(4,4)

dJ：(4, 8)

垂直方向：(8, -4)
也就是(1, -0.5)

TRPO算法 —— 工程实现 —— 优化更新算法 —— 共轭梯度法求解

$$H_{kl}x = g_l$$

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

$$H_{kl}x = g_l$$

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

$$\alpha_k = \frac{r_k^T r_k}{p_k^T H p_k} \qquad x_{k+1} = x_k + \alpha_k p_k \qquad r_{k+1} = r_k - \alpha_k H p_k$$

$$\beta_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k} \qquad p_{k+1} = r_{k+1} + \beta_k p_k$$

HVP（Hessian-vector product）

$$\nabla_\theta (p^T \cdot \nabla_\theta KL) = H \cdot p$$

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 $\delta = \bar{D}_{KL} \approx \frac{1}{2}(\beta s)^T H(\beta s) = \frac{1}{2}\beta^2 s^T H s$ $\beta = \sqrt{2\delta / s^T H s}$

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