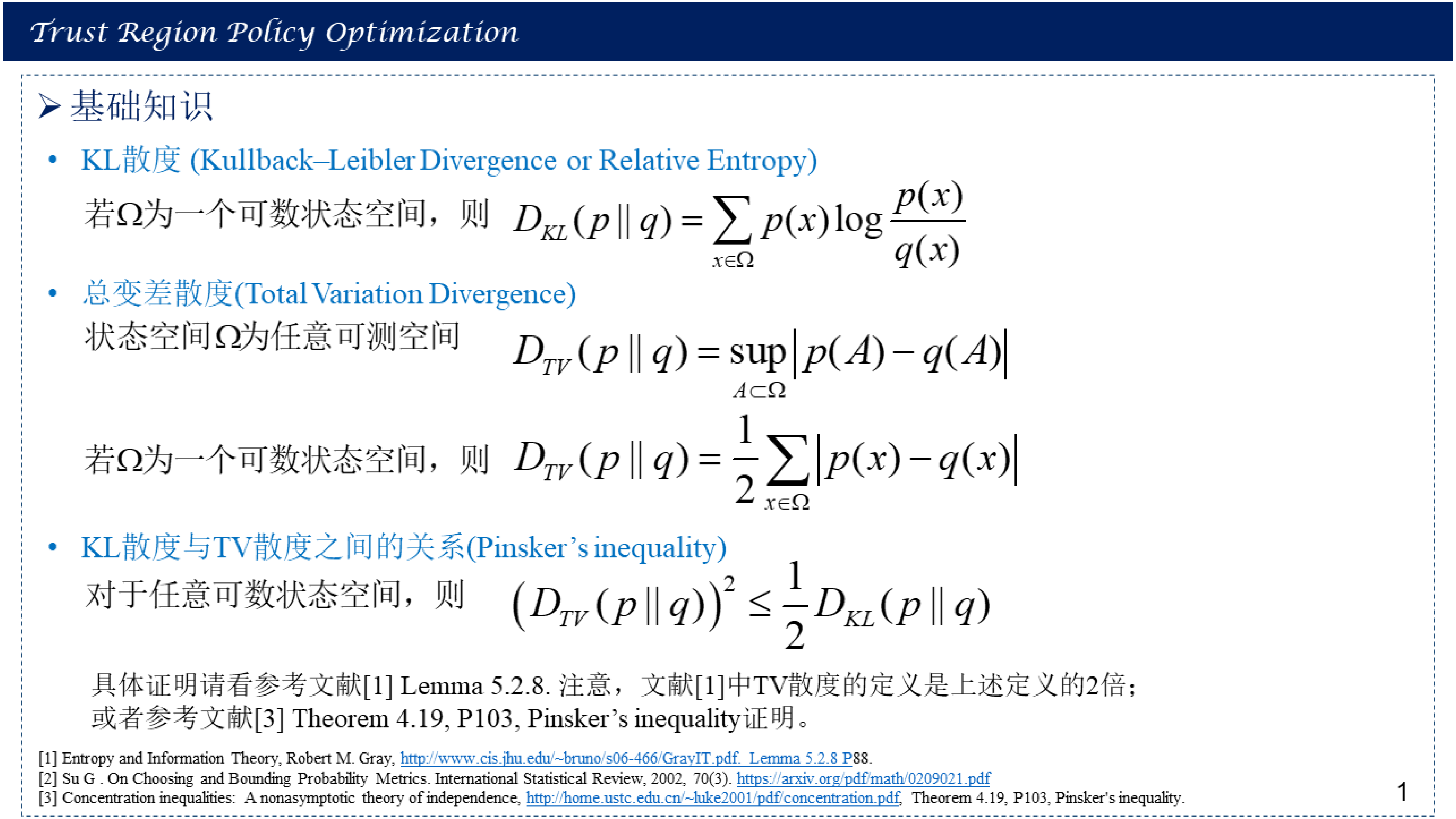
信赖域策略优化(Trust Region Policy Optimization, TRPO)

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

这篇博文是John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.的阅读笔记，用来介绍TRPO策略优化方法及其一些公式的推导。TRPO是一种基于策略梯度的强化学习方法，除了定理1没推导之外，其他公式的来龙去脉都进行了详细介绍，为后续进一步深入研究其他强化学习方法提供基础。更多强化学习内容，请看：随笔分类 - Reinforcement Learning。

1. 基础知识

KL散度 (Kullback-Leibler Divergence or Relative Entropy)，总变差散度(Total Variation Divergence)，以及KL散度与TV散度之间的关系(Pinsker’s inequality)



Trust Region Policy Optimization

➢ 基础知识

- KL散度 (Kullback–Leibler Divergence or Relative Entropy)

若Ω为一个可数状态空间，则 $$D_{KL}(p\|q)=\sum_{x\in\Omega}p(x)\log\frac{p(x)}{q(x)}$$

- 总变差散度(Total Variation Divergence)

状态空间Ω为任意可测空间 $$D_{TV}(p\|q)=\sup_{A\subset\Omega}\left|p(A)-q(A)\right|$$

若Ω为一个可数状态空间，则 $$D_{TV}(p\|q)=\frac{1}{2}\sum_{x\in\Omega}\left|p(x)-q(x)\right|$$

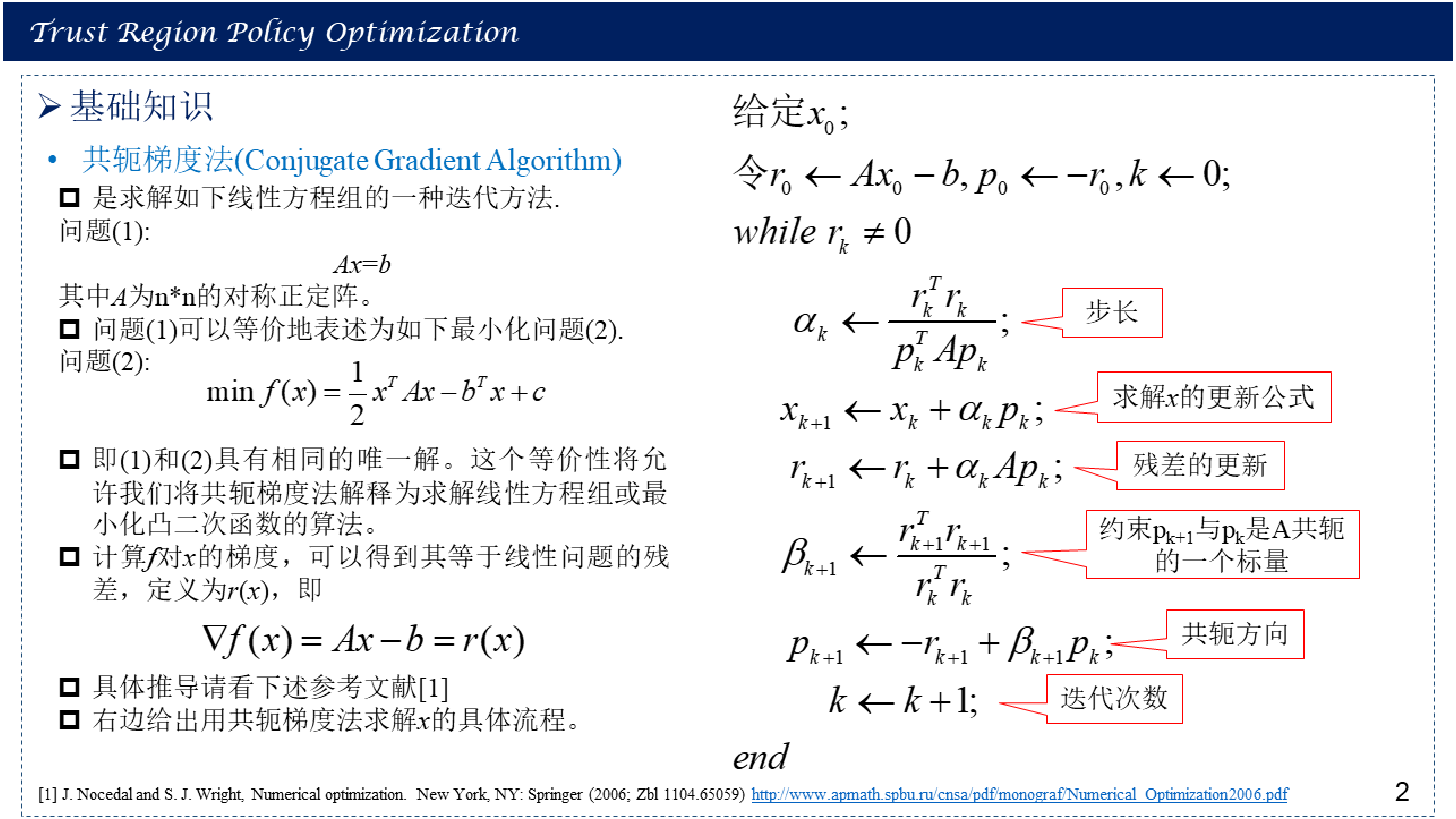
- KL散度与TV散度之间的关系(Pinsker’s inequality)

对于任意可数状态空间，则 $$\left(D_{TV}(p\|q)\right)^2\le\frac{1}{2}D_{KL}(p\|q)$$

具体证明请看参考文献[1] Lemma 5.2.8. 注意，文献[1]中TV散度的定义是上述定义的2倍；
或者参考文献[3] Theorem 4.19, P103, Pinsker’s inequality证明。

[1] Entropy and Information Theory, Robert M. Gray, http://www.cis.jhu.edu/~bruno/s06-466/GrayIT.pdf. Lemma 5.2.8 P88.
[2] Su G . On Choosing and Bounding Probability Metrics. International Statistical Review, 2002, 70(3). https://arxiv.org/pdf/math/0209021.pdf
[3] Concentration inequalities: A nonasymptotic theory of independence, http://home.ustc.edu.cn/~luke2001/pdf/concentration.pdf, Theorem 4.19, P103, Pinsker's inequality.

共轭梯度法(Conjugate Gradient Algorithm)



➢ 基础知识

- 共轭梯度法(Conjugate Gradient Algorithm)

□ 是求解如下线性方程组的一种迭代方法.

问题(1):

$$Ax=b$$

其中A为n*n的对称正定阵。

□ 问题(1)可以等价地表述为如下最小化问题(2).

问题(2):

$$\min f(x)=\frac{1}{2}x^TAx-b^Tx+c$$

□ 即(1)和(2)具有相同的唯一解。这个等价性将允许我们将共轭梯度法解释为求解线性方程组或最小化凸二次函数的算法。

□ 计算f对x的梯度，可以得到其等于线性问题的残差，定义为$r(x)$，即

$$\nabla f(x)=Ax-b=r(x)$$

□ 具体推导请看下述参考文献[1]

□ 右边给出用共轭梯度法求解x的具体流程。

```
给定$x_0$;
令$r_0 \leftarrow Ax_0 - b, p_0 \leftarrow -r_0, k \leftarrow 0$;
while $r_k \neq 0$
    $\alpha_k \leftarrow \frac{r_k^T r_k}{p_k^T A p_k}$;    步长
    $x_{k+1} \leftarrow x_k + \alpha_k p_k$;    求解x的更新公式
    $r_{k+1} \leftarrow r_k + \alpha_k A p_k$;    残差的更新
    $\beta_{k+1} \leftarrow \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$;    约束$p_{k+1}$与$p_k$是A共轭的一个标量
    $p_{k+1} \leftarrow -r_{k+1} + \beta_{k+1} p_k$;    共轭方向
    $k \leftarrow k+1$;    迭代次数
end
```

[1] J. Nocedal and S. J. Wright, Numerical optimization. New York, NY: Springer (2006; Zbl 1104.65059) http://www.apmath.spbu.ru/cnsa/pdf/monograf/Numerical_Optimization2006.pdf

新旧策略期望折扣奖励差

➢ 基础知识

$\mathcal{S}$是状态集, $\mathcal{A}$是动作集, $P:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$是转移概率分布

$r:\mathcal{S}\to\mathbb{R}$是回报函数, $\gamma\in(0,1)$是折扣率, $\rho_0:\mathcal{S}\to\mathbb{R}$是初始状态 s_0的分布.

令$\pi:\mathcal{S}\times\mathcal{A}\to[0,1]$是随机策略, 期望折扣奖励

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

其中$s_0\sim\rho_0(s_0), a_t\sim\pi(a_t\,|\,s_t), s_{t+1}\sim P(s_{t+1}\,|\,s_t,a_t)$

$$Q_\pi(s_t,a_t)=\mathbb{E}_{s_{t+1}}\left[r(s_t)+\gamma V_\pi(s_{t+1})\right]$$

状态-动作值函数$Q_\pi(s_t,a_t)=\mathbb{E}_{s_{t+1},a_{t+1},\cdots\sim\pi}\left[\sum_{l=0}^{\infty}\gamma^l r(s_{t+l})\right]$

值函数$V_\pi(s_t)=\mathbb{E}_{a_t,s_{t+1},a_{t+1},\cdots\sim\pi}\left[\sum_{l=0}^{\infty}\gamma^l r(s_{t+l})\right]$

优势函数$A_\pi(s_t,a_t)=Q_\pi(s_t,a_t)-V_\pi(s_t)$

用以评价当前动作值函数相比于平均值的大小。如果优势函数大于零，则说明该动作比平均动作好，如果优势函数小于零，则说明当前动作不如平均动作好

新策略$\tilde{\pi}$与旧策略π期望折扣奖励之差 $\boxed{\eta(\tilde{\pi})-\eta(\pi)=\mathbb{E}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right]}$

其中$s_0\sim\rho_0(s_0), a_t\sim\tilde{\pi}(a_t\,|\,s_t)$

John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.

Trust Region Policy Optimization

➢ 基础知识

上述等式验证(这里是从结果反推)：

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

$$=\mathbb{E}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t)+\gamma V_\pi(s_{t+1})-V_\pi(s_t)\right)\right]$$

$$=\mathbb{E}_{s_0,a_0,\cdots\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t)+\left(\gamma V_\pi(s_1)+\gamma^2 V_\pi(s_2)+\cdots-V_\pi(s_0)-\gamma V_\pi(s_1)-\gamma^2 V_\pi(s_2)-\cdots\right)\right]$$

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

$$=-\mathbb{E}_{s_0}\left[\mathbb{E}_{a_0,s_1,a_1,\cdots\sim\pi}\left[\sum_{l=0}^{\infty}\gamma^l r(s_l)\right]\right]+\eta(\tilde{\pi})=-\mathbb{E}_{s_0,a_0,s_1,a_1,\cdots\sim\pi}\left[\sum_{l=0}^{\infty}\gamma^l r(s_l)\right]+\eta(\tilde{\pi})$$

$$=\eta(\tilde{\pi})-\eta(\pi)$$

初始s_0与策略π无关，由环境决定，因此$s_0\sim\pi$等价于$s_0\sim\tilde{\pi}$

用到的式子

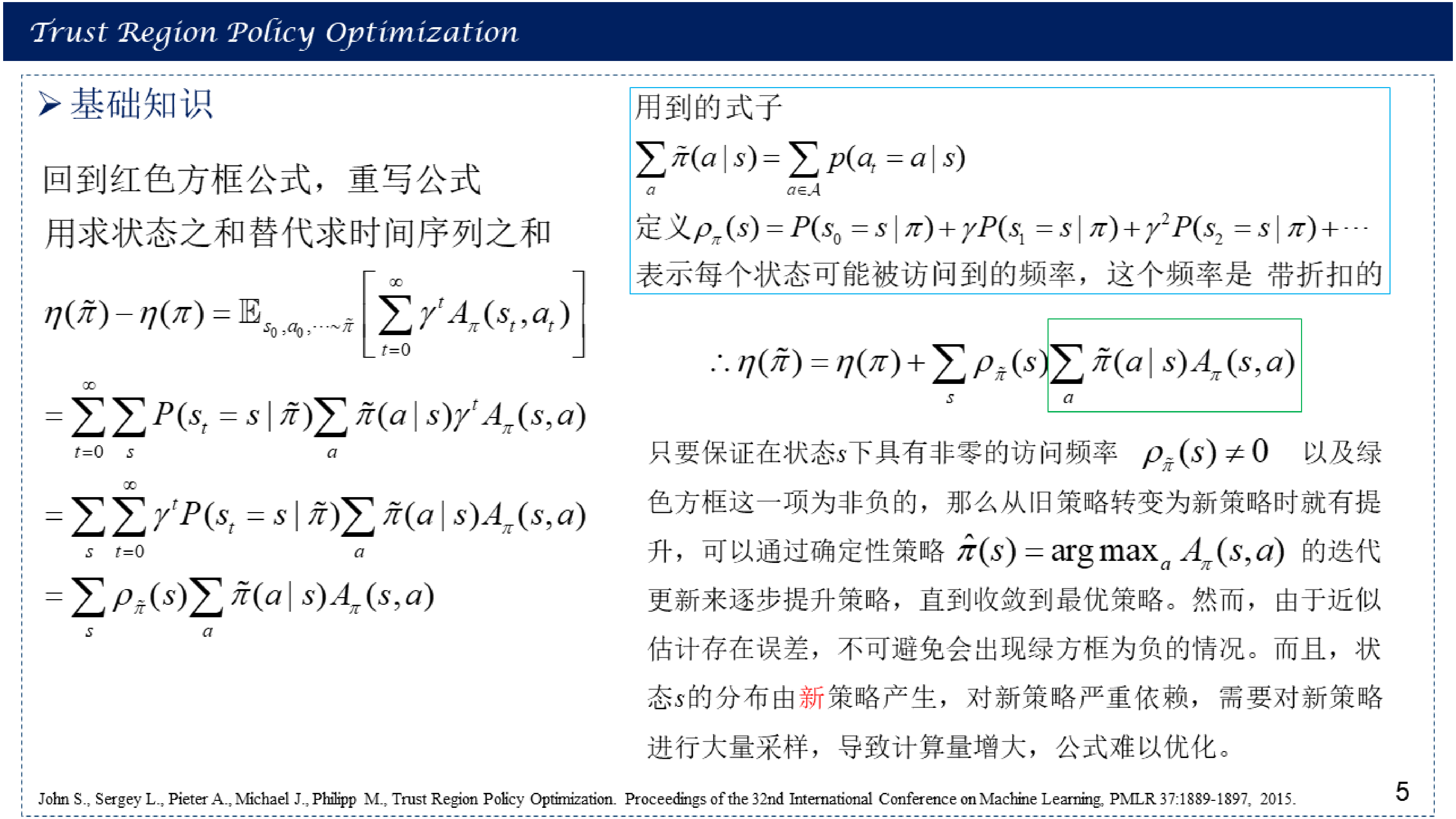
$$A_\pi(s_t,a_t)=Q_\pi(s_t,a_t)-V_\pi(s_t)$$

$$Q_\pi(s_t,a_t)=\mathbb{E}_{s_{t+1}}\left[r(s_t)+\gamma V_\pi(s_{t+1})\right]$$

$$V_\pi(s_t)=\mathbb{E}_{a_t,s_{t+1},a_{t+1},\cdots\sim\pi}\left[\sum_{l=0}^{\infty}\gamma^l r(s_{t+l})\right]$$

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

John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.



Trust Region Policy Optimization

➢ 基础知识

回到红色方框公式，重写公式

用求状态之和替代求时间序列之和

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

$$=\sum_{t=0}^{\infty}\sum_{s}P(s_t=s\,|\,\tilde{\pi})\sum_{a}\tilde{\pi}(a\,|\,s)\gamma^t A_\pi(s,a)$$

$$=\sum_{s}\sum_{t=0}^{\infty}\gamma^t P(s_t=s\,|\,\tilde{\pi})\sum_{a}\tilde{\pi}(a\,|\,s)A_\pi(s,a)$$

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

用到的式子

$$\sum_{a}\tilde{\pi}(a\,|\,s)=\sum_{a\in A}p(a_t=a\,|\,s)$$

定义$\rho_\pi(s)=P(s_0=s\,|\,\pi)+\gamma P(s_1=s\,|\,\pi)+\gamma^2 P(s_2=s\,|\,\pi)+\cdots$

表示每个状态可能被访问到的频率，这个频率是 带折扣的

$$\therefore\eta(\tilde{\pi})=\eta(\pi)+\sum_{s}\rho_{\tilde{\pi}}(s)\boxed{\sum_{a}\tilde{\pi}(a\,|\,s)A_\pi(s,a)}$$

只要保证在状态s下具有非零的访问频率 $\rho_{\tilde{\pi}}(s)\neq 0$ 以及绿色方框这一项为非负的，那么从旧策略转变为新策略时就有提升，可以通过确定性策略 $\hat{\pi}(s)=\arg\max_a A_\pi(s,a)$ 的迭代更新来逐步提升策略，直到收敛到最优策略。然而，由于近似估计存在误差，不可避免会出现绿方框为负的情况。而且，状态s的分布由新策略产生，对新策略严重依赖，需要对新策略进行大量采样，导致计算量增大，公式难以优化。

John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.

2．η的局部近似

➢ η的局部近似

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

- 注意，这里采用的是旧策略对应的折扣访问频率，忽略由于策略变化而导致的访问频率的变化。当新旧访问频率很接近时，使用旧策略所对应的访问频率来代替新的访问频率，要求就是新旧策略的更新幅度不要太大，两者要相似。这样，新策略下的L就近似为原先的η。
- 因此，只要对L进行优化即可。假设有一个参数化策略$\pi_{\theta}(a|s)$，那么对于任何参数值θ_0，有

$$L_{\pi_{\theta_0}}(\pi_{\theta_0})=\eta(\pi_{\theta_0}),\ \nabla_{\theta}L_{\pi_{\theta_0}}(\pi_{\theta})\Big|_{\theta=\theta_0}=\nabla_{\theta}\eta(\pi_{\theta})\Big|_{\theta=\theta_0}$$

红框式子证明

$$L_{\pi_{\theta_0}}(\pi_{\theta_0})=\eta(\pi_{\theta_0})$$

$$L_{\pi_{\theta_0}}(\pi_{\theta_0})=\eta(\pi_{\theta_0})+\sum_{s}\rho_{\pi_{\theta_0}}(s)\sum_{a}\pi_{\theta_0}(a|s)A_{\pi_{\theta_0}}(s,a)$$

其中$\sum_{a}\pi_{\theta_0}(a|s)=1$

$$\sum_{a}\pi_{\theta_0}(a|s)A_{\pi_{\theta_0}}(s,a)=\sum_{a}\pi_{\theta_0}(a|s)\left(Q_{\pi_{\theta_0}}(s,a)-V_{\pi_{\theta_0}}(s)\right)=\sum_{a}\pi_{\theta_0}(a|s)Q_{\pi_{\theta_0}}(s,a)-V_{\pi_{\theta_0}}(s)=0$$

$$\therefore L_{\pi_{\theta_0}}(\pi_{\theta_0})=\eta(\pi_{\theta_0})$$

John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.

Trust Region Policy Optimization

➢ η的局部近似

红框式子证明

$$\boxed{\nabla_\theta L_{\pi_{\theta_0}}(\pi_\theta)\Big|_{\theta=\theta_0} = \nabla_\theta \eta(\pi_\theta)\Big|_{\theta=\theta_0}}$$

用到的技巧：变量替换，导数加法运算

$$L_{\pi_{\theta_0}}(\pi_\theta) = \eta(\pi_{\theta_0}) + \sum_s \rho_{\pi_{\theta_0}}(s) \sum_a \pi_\theta(a \mid s) A_{\pi_{\theta_0}}(s,a)$$

$$\eta(\pi_\theta) = \eta(\pi_{\theta_0}) + \sum_s \rho_{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) A_{\pi_{\theta_0}}(s,a)$$

$$\therefore \nabla_\theta L_{\pi_{\theta_0}}(\pi_\theta) = \sum_s \rho_{\pi_{\theta_0}}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s) A_{\pi_{\theta_0}}(s,a)$$

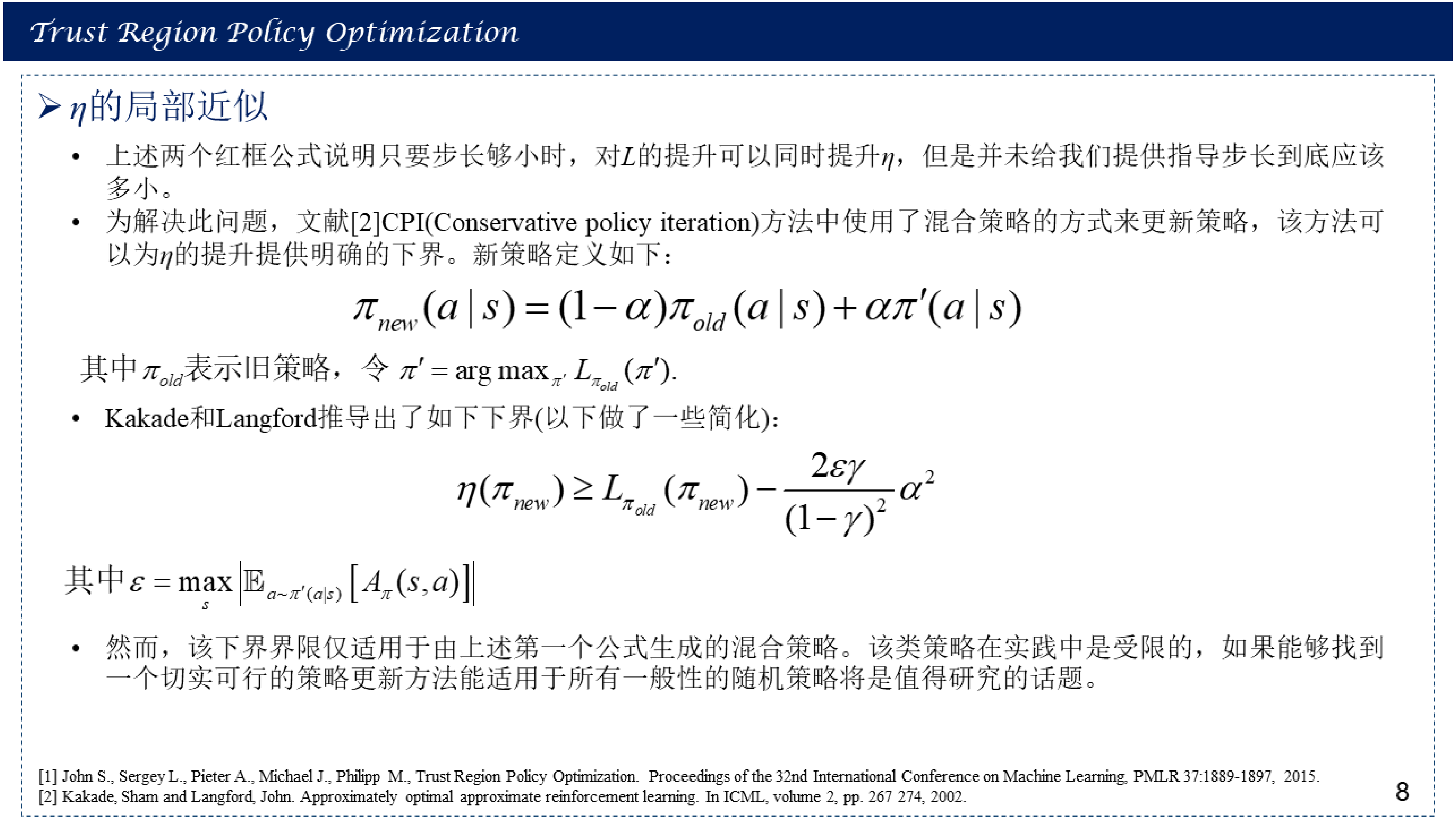
$$\nabla_\theta \eta(\pi_\theta) = \nabla_\theta \left(\sum_s \rho_{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) A_{\pi_{\theta_0}}(s,a) \right)$$

$$= \sum_s \nabla_\theta \rho_{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) A_{\pi_{\theta_0}}(s,a) + \sum_s \rho_{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s) A_{\pi_{\theta_0}}(s,a)$$

当$\theta = \theta_0$时, $\sum_a \pi_\theta(a \mid s) A_{\pi_{\theta_0}}(s,a) = 0$

$$\therefore \nabla_\theta \eta(\pi_\theta) = \sum_s \rho_{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s) A_{\pi_{\theta_0}}(s,a) \qquad \therefore \nabla_\theta L_{\pi_{\theta_0}}(\pi_\theta)\Big|_{\theta=\theta_0} = \nabla_\theta \eta(\pi_\theta)\Big|_{\theta=\theta_0}$$

John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.



Trust Region Policy Optimization

➢ η的局部近似

- 上述两个红框公式说明只要步长够小时，对L的提升可以同时提升η，但是并未给我们提供指导步长到底应该多小。
- 为解决此问题，文献[2]CPI(Conservative policy iteration)方法中使用了混合策略的方式来更新策略，该方法可以为η的提升提供明确的下界。新策略定义如下：

$$\pi_{new}(a|s)=(1-\alpha)\pi_{old}(a|s)+\alpha\pi'(a|s)$$

其中π_{old}表示旧策略，令 $\pi'=\arg\max_{\pi'}L_{\pi_{old}}(\pi')$.

- Kakade和Langford推导出了如下下界(以下做了一些简化)：

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

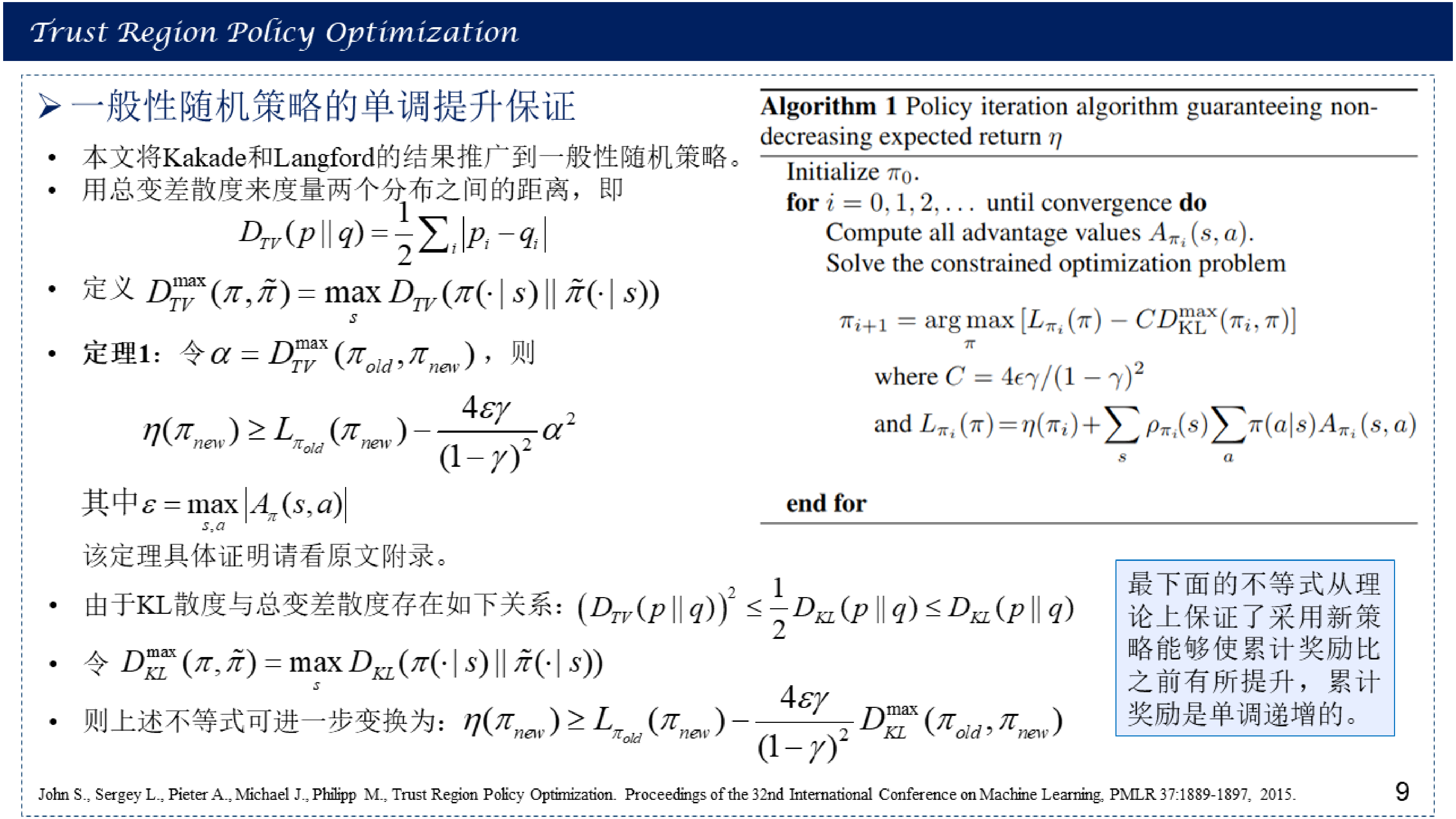
其中$\varepsilon=\max_s\left|\mathbb{E}_{a\sim\pi'(a|s)}\left[A_\pi(s,a)\right]\right|$

- 然而，该下界界限仅适用于由上述第一个公式生成的混合策略。该类策略在实践中是受限的，如果能够找到一个切实可行的策略更新方法能适用于所有一般性的随机策略将是值得研究的话题。

[1] John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.
[2] Kakade, Sham and Langford, John. Approximately optimal approximate reinforcement learning. In ICML, volume 2, pp. 267 274, 2002.

8

3．一般性随机策略的单调提升保证



Trust Region Policy Optimization

➢ 一般性随机策略的单调提升保证

- 本文将Kakade和Langford的结果推广到一般性随机策略。
- 用总变差散度来度量两个分布之间的距离，即

$$D_{TV}(p\|q)=\frac{1}{2}\sum_i|p_i-q_i|$$

- 定义 $D_{TV}^{\max}(\pi,\tilde{\pi})=\max_s D_{TV}(\pi(\cdot|s)\|\tilde{\pi}(\cdot|s))$
- **定理1：** 令 $\alpha=D_{TV}^{\max}(\pi_{old},\pi_{new})$，则

$$\eta(\pi_{new})\ge L_{\pi_{old}}(\pi_{new})-\frac{4\varepsilon\gamma}{(1-\gamma)^2}\alpha^2$$

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

该定理具体证明请看原文附录。

Algorithm 1 Policy iteration algorithm guaranteeing non-decreasing expected return η

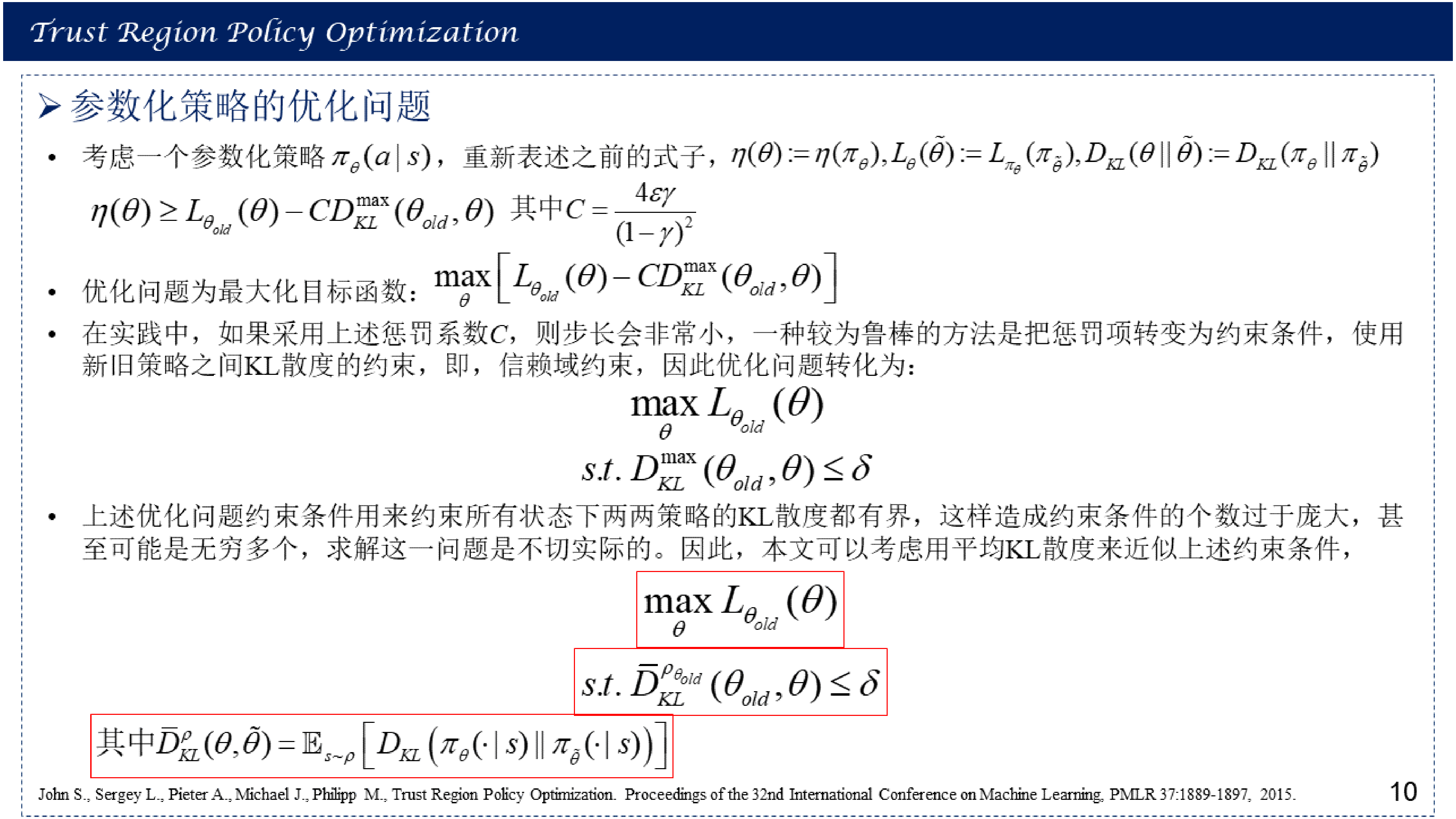
```
Initialize $\pi_0$.
for $i = 0, 1, 2, \dots$ until convergence do
    Compute all advantage values $A_{\pi_i}(s,a)$.
    Solve the constrained optimization problem

        $\pi_{i+1} = \arg\max_\pi [L_{\pi_i}(\pi) - CD_{\mathrm{KL}}^{\max}(\pi_i,\pi)]$
            where $C = 4\epsilon\gamma/(1-\gamma)^2$
            and $L_{\pi_i}(\pi) = \eta(\pi_i) + \sum_s \rho_{\pi_i}(s) \sum_a \pi(a|s) A_{\pi_i}(s,a)$

end for
```

- 由于KL散度与总变差散度存在如下关系：$(D_{TV}(p\|q))^2\le\frac{1}{2}D_{KL}(p\|q)\le D_{KL}(p\|q)$
- 令 $D_{KL}^{\max}(\pi,\tilde{\pi})=\max_s D_{KL}(\pi(\cdot|s)\|\tilde{\pi}(\cdot|s))$
- 则上述不等式可进一步变换为：$\eta(\pi_{new})\ge L_{\pi_{old}}(\pi_{new})-\frac{4\varepsilon\gamma}{(1-\gamma)^2}D_{KL}^{\max}(\pi_{old},\pi_{new})$

最下面的不等式从理论上保证了采用新策略能够使累计奖励比之前有所提升，累计奖励是单调递增的。

John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.

4．参数化策略的优化问题



Trust Region Policy Optimization

➢ 参数化策略的优化问题

- 考虑一个参数化策略$\pi_\theta(a|s)$，重新表述之前的式子，$\eta(\theta):=\eta(\pi_\theta), L_\theta(\tilde{\theta}):=L_{\pi_\theta}(\pi_{\tilde{\theta}}), D_{KL}(\theta\|\tilde{\theta}):=D_{KL}(\pi_\theta\|\pi_{\tilde{\theta}})$

  $\eta(\theta)\ge L_{\theta_{old}}(\theta)-CD_{KL}^{\max}(\theta_{old},\theta)$ 其中$C=\dfrac{4\varepsilon\gamma}{(1-\gamma)^2}$

- 优化问题为最大化目标函数：$\max\limits_\theta\left[L_{\theta_{old}}(\theta)-CD_{KL}^{\max}(\theta_{old},\theta)\right]$
- 在实践中，如果采用上述惩罚系数C，则步长会非常小，一种较为鲁棒的方法是把惩罚项转变为约束条件，使用新旧策略之间KL散度的约束，即，信赖域约束，因此优化问题转化为：

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

$$s.t.\ D_{KL}^{\max}(\theta_{old},\theta)\le\delta$$

- 上述优化问题约束条件用来约束所有状态下两两策略的KL散度都有界，这样造成约束条件的个数过于庞大，甚至可能是无穷多个，求解这一问题是不切实际的。因此，本文可以考虑用平均KL散度来近似上述约束条件，

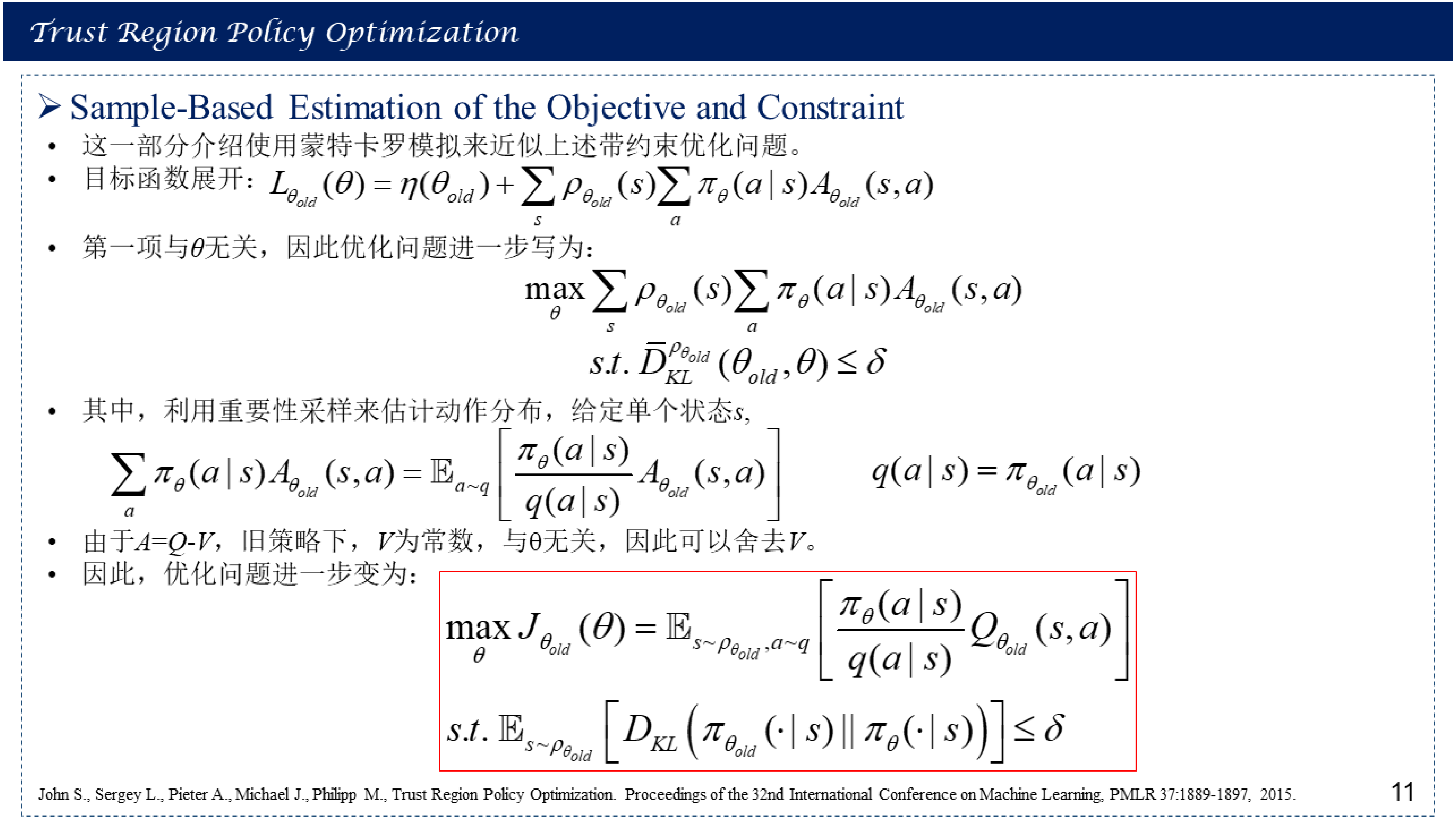
$$\boxed{\max_\theta L_{\theta_{old}}(\theta)}$$

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

$\boxed{其中\bar{D}_{KL}^{\rho}(\theta,\tilde{\theta})=\mathbb{E}_{s\sim\rho}\left[D_{KL}\left(\pi_\theta(\cdot|s)\|\pi_{\tilde{\theta}}(\cdot|s)\right)\right]}$

John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.

5. Sample-Based Estimation of the Objective and Constraint



Trust Region Policy Optimization

➢ Sample-Based Estimation of the Objective and Constraint

- 这一部分介绍使用蒙特卡罗模拟来近似上述带约束优化问题。
- 目标函数展开：$L_{\theta_{old}}(\theta)=\eta(\theta_{old})+\sum_{s}\rho_{\theta_{old}}(s)\sum_{a}\pi_{\theta}(a\,|\,s)A_{\theta_{old}}(s,a)$
- 第一项与θ无关，因此优化问题进一步写为：

$$\max_{\theta}\sum_{s}\rho_{\theta_{old}}(s)\sum_{a}\pi_{\theta}(a\,|\,s)A_{\theta_{old}}(s,a)$$

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

- 其中，利用重要性采样来估计动作分布，给定单个状态*s*,

$$\sum_{a}\pi_{\theta}(a\,|\,s)A_{\theta_{old}}(s,a)=\mathbb{E}_{a\sim q}\left[\frac{\pi_{\theta}(a\,|\,s)}{q(a\,|\,s)}A_{\theta_{old}}(s,a)\right]\qquad q(a\,|\,s)=\pi_{\theta_{old}}(a\,|\,s)$$

- 由于*A*=*Q*-*V*，旧策略下，*V*为常数，与θ无关，因此可以舍去*V*。
- 因此，优化问题进一步变为：

$$\boxed{\begin{aligned}&\max_{\theta}J_{\theta_{old}}(\theta)=\mathbb{E}_{s\sim\rho_{\theta_{old}},a\sim q}\left[\frac{\pi_{\theta}(a\,|\,s)}{q(a\,|\,s)}Q_{\theta_{old}}(s,a)\right]\\&s.t.\ \mathbb{E}_{s\sim\rho_{\theta_{old}}}\left[D_{KL}\left(\pi_{\theta_{old}}(\cdot\,|\,s)\,\|\,\pi_{\theta}(\cdot\,|\,s)\right)\right]\le\delta\end{aligned}}$$

John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.

11

6．约束优化问题的求解

➢约束优化问题的求解

- 泰勒展开式

$$f(x)=\sum_{n=0}^{\infty}\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n=f(x_0)+f'(x_0)(x-x_0)+\frac{1}{2}f''(x_0)(x-x_0)^2+\frac{f^{(3)}(\varepsilon)}{3!}(x-x_0)^3$$

- 对上述目标函数采用线性近似(类似于求解$Ax=y$中的x)，对约束条件采用二次近似

$$J_{\theta_{old}}(\theta)\approx\nabla_\theta J_{\theta_{old}}(\theta)\Big|_{\theta=\theta_{old}}\cdot(\theta-\theta_{old})$$

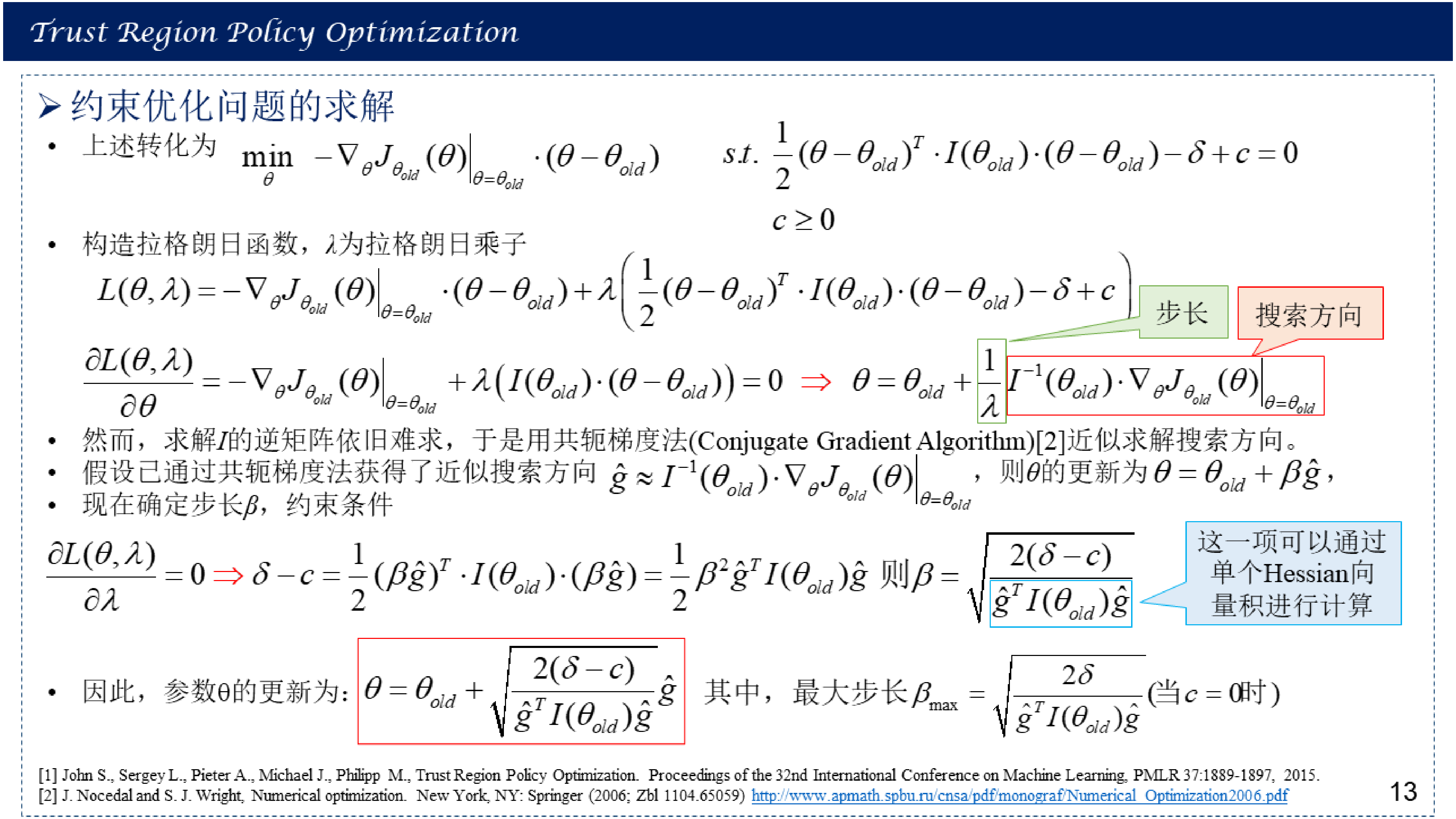
$$\mathbb{E}_{s\sim\rho_{\theta_{old}}}\left[D_{KL}\left(\pi_{\theta_{old}}(\cdot\,|\,s)\,\|\,\pi_\theta(\cdot\,|\,s)\right)\right]\approx\frac{1}{2}(\theta-\theta_{old})^T\cdot I(\theta_{old})\cdot(\theta-\theta_{old})$$

其中矩阵$I(\theta_{old})$中的每一个数值 $I_{ij}(\theta_{old})=\left.\dfrac{\partial^2\mathbb{E}_{s\sim\rho_{\theta_{old}}}\left[D_{KL}\left(\pi_{\theta_{old}}(\cdot\,|\,s)\,\|\,\pi_\theta(\cdot\,|\,s)\right)\right]}{\partial\theta_i\partial\theta_j}\right|_{\theta=\theta_{old}}$

- 因此，约束优化问题近似为：

$$\boxed{\begin{aligned}&\max_\theta\ \nabla_\theta J_{\theta_{old}}(\theta)\Big|_{\theta=\theta_{old}}\cdot(\theta-\theta_{old})\\&s.t.\ \frac{1}{2}(\theta-\theta_{old})^T\cdot I(\theta_{old})\cdot(\theta-\theta_{old})\le\delta\end{aligned}}$$

John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.



Trust Region Policy Optimization

➢ 约束优化问题的求解

- 上述转化为 $\min\limits_{\theta} \ -\nabla_{\theta} J_{\theta_{old}}(\theta)\Big|_{\theta=\theta_{old}} \cdot(\theta-\theta_{old}) \qquad s.t.\ \frac{1}{2}(\theta-\theta_{old})^{T}\cdot I(\theta_{old})\cdot(\theta-\theta_{old})-\delta+c=0$
  $c\geq 0$
- 构造拉格朗日函数，λ为拉格朗日乘子

$$L(\theta,\lambda)=-\nabla_{\theta} J_{\theta_{old}}(\theta)\Big|_{\theta=\theta_{old}}\cdot(\theta-\theta_{old})+\lambda\left(\frac{1}{2}(\theta-\theta_{old})^{T}\cdot I(\theta_{old})\cdot(\theta-\theta_{old})-\delta+c\right)$$

$$\frac{\partial L(\theta,\lambda)}{\partial\theta}=-\nabla_{\theta} J_{\theta_{old}}(\theta)\Big|_{\theta=\theta_{old}}+\lambda\big(I(\theta_{old})\cdot(\theta-\theta_{old})\big)=0 \Rightarrow \theta=\theta_{old}+\boxed{\frac{1}{\lambda}}\boxed{I^{-1}(\theta_{old})\cdot\nabla_{\theta} J_{\theta_{old}}(\theta)\Big|_{\theta=\theta_{old}}}$$

步长

搜索方向

- 然而，求解I的逆矩阵依旧难求，于是用共轭梯度法(Conjugate Gradient Algorithm)[2]近似求解搜索方向。
- 假设已通过共轭梯度法获得了近似搜索方向 $\hat{g}\approx I^{-1}(\theta_{old})\cdot\nabla_{\theta} J_{\theta_{old}}(\theta)\Big|_{\theta=\theta_{old}}$，则$\theta$的更新为$\theta=\theta_{old}+\beta\hat{g}$，
- 现在确定步长β，约束条件

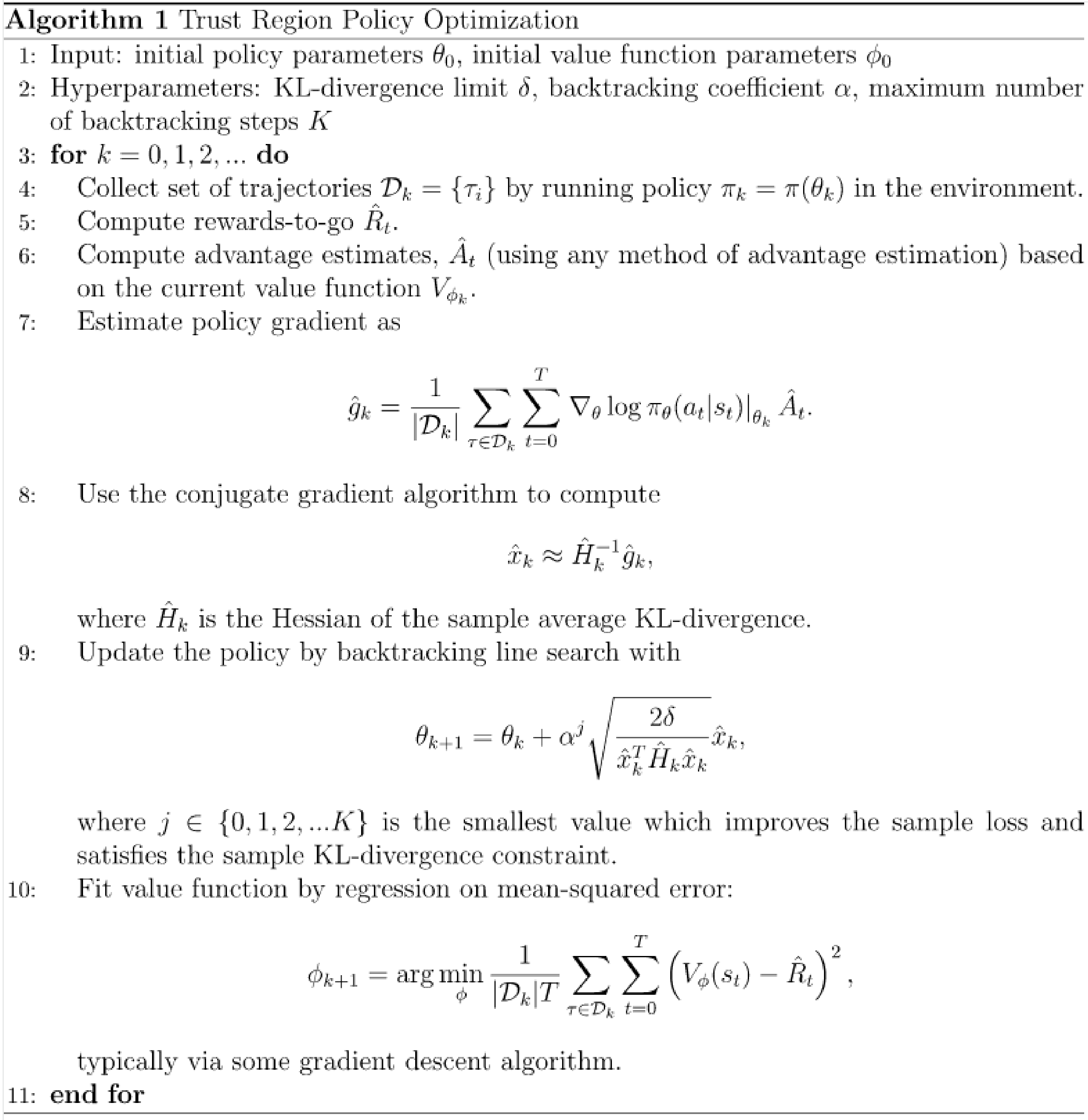
$$\frac{\partial L(\theta,\lambda)}{\partial\lambda}=0 \Rightarrow \delta-c=\frac{1}{2}(\beta\hat{g})^{T}\cdot I(\theta_{old})\cdot(\beta\hat{g})=\frac{1}{2}\beta^{2}\hat{g}^{T}I(\theta_{old})\hat{g} \text{ 则}\beta=\sqrt{\frac{2(\delta-c)}{\boxed{\hat{g}^{T}I(\theta_{old})\hat{g}}}}$$

这一项可以通过单个Hessian向量积进行计算

- 因此，参数θ的更新为: $\boxed{\theta=\theta_{old}+\sqrt{\frac{2(\delta-c)}{\hat{g}^{T}I(\theta_{old})\hat{g}}}\hat{g}}$ 其中，最大步长$\beta_{\max}=\sqrt{\frac{2\delta}{\hat{g}^{T}I(\theta_{old})\hat{g}}}$(当$c=0$时)

[1] John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.
[2] J. Nocedal and S. J. Wright, Numerical optimization. New York, NY: Springer (2006; Zbl 1104.65059) http://www.apmath.spbu.ru/cnsa/pdf/monograf/Numerical_Optimization2006.pdf

7. 算法总体流程



Algorithm 1 Trust Region Policy Optimization

```
Input: initial policy parameters $\theta_0$, initial value function parameters $\phi_0$
Hyperparameters: KL-divergence limit $\delta$, backtracking coefficient $\alpha$, maximum number of backtracking steps $K$
for $k = 0, 1, 2, ...$ do
    Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.
    Compute rewards-to-go $\hat{R}_t$.
    Compute advantage estimates, $\hat{A}_t$ (using any method of advantage estimation) based on the current value function $V_{\phi_k}$.
    Estimate policy gradient as
        $\hat{g}_k = \frac{1}{|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\big|_{\theta_k} \hat{A}_t.$
    Use the conjugate gradient algorithm to compute
        $\hat{x}_k \approx \hat{H}_k^{-1} \hat{g}_k,$
    where $\hat{H}_k$ is the Hessian of the sample average KL-divergence.
    Update the policy by backtracking line search with
        $\theta_{k+1} = \theta_k + \alpha^j \sqrt{\frac{2\delta}{\hat{x}_k^T \hat{H}_k \hat{x}_k}} \hat{x}_k,$
    where $j \in \{0, 1, 2, ...K\}$ is the smallest value which improves the sample loss and satisfies the sample KL-divergence constraint.
    Fit value function by regression on mean-squared error:
        $\phi_{k+1} = \arg\min_\phi \frac{1}{|\mathcal{D}_k| T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left( V_\phi(s_t) - \hat{R}_t \right)^2,$
    typically via some gradient descent algorithm.
end for
```

8. 参考文献

[1] John S., Sergey L., Pieter A., Michael J., Philipp M., Trust Region Policy Optimization. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1889-1897, 2015.

[2] Entropy and Information Theory, Robert M. Gray, http://www.cis.jhu.edu/~bruno/s06-466/GrayIT.pdf. Lemma 5.2.8 P88.

[3] Su G . On Choosing and Bounding Probability Metrics. International Statistical Review, 2002, 70(3). https://arxiv.org/pdf/math/0209021.pdf

[4] Concentration inequalities: A nonasymptotic theory of independence, http://home.ustc.edu.cn/~luke2001/pdf/concentration.pdf, Theorem 4.19, P103, Pinsker's inequality.

[5] J. Nocedal and S. J. Wright, Numerical optimization. New York, NY: Springer (2006; Zbl 1104.65059) http://www.apmath.spbu.ru/cnsa/pdf/monograf/Numerical_Optimization2006.pdf

[6] Kakade, Sham and Langford, John. Approximately optimal approximate reinforcement learning. In ICML, volume 2, pp. 267 274, 2002.

[7] Trust Region Policy Optimization https://spinningup.openai.com/en/latest/algorithms/trpo.html