从策略梯度到 TRPO 与 PPO

一份逐步可验证的数学推导讲义

本讲义从马尔可夫决策过程的定义出发，完整推导策略梯度定理、性能差异引理、信任域方法的代理目标及其误差界、Fisher 信息矩阵的三种等价形式、信任域二次规划的精确解与自然梯度，最后给出 TRPO 的工程实现与 PPO 的简化路径。
与常见笔记相比，本讲义补上了三处通常被跳过的关键论证：**(i)** 性能差异引理的完整证明；**(ii)** “忽略状态分布漂移”并非近似而是一阶精确的严格理由；**(iii)** Fisher 信息矩阵与 KL 散度 Hessian 相等的证明，并说明为何正向与反向 KL 给出同一个二阶结构。
文中第 10 节给出四组数值实验，对上述结论逐条做了独立验证；所有图表均由这些计算直接生成。

全部推导自主完成，可自由使用与修改

目录

1 记号与问题设定

定义 1.1 (折扣 MDP)**.** 一个折扣马尔可夫决策过程由六元组 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma)$ 给出：$\mathcal{S}$ 为状态空间，$\mathcal{A}$ 为动作空间，$P(s' \mid s, a)$ 为状态转移核，$r(s, a)$ 为奖励函数，ρ_0 为初始状态分布，$\gamma \in [0, 1)$ 为折扣因子。

策略 $\pi_\theta(a \mid s)$ 由参数 $\theta \in \mathbb{R}^n$ 决定（例如神经网络权重）。轨迹 $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \dots)$ 的生成概率为

$$p_\theta(\tau) = \rho_0(s_0) \prod_{t \geq 0} \pi_\theta(a_t \mid s_t)\, P(s_{t+1} \mid s_t, a_t). \tag{1}$$

优化目标是折扣累积回报的期望

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta}\Big[\sum_{t \geq 0} \gamma^t r_t\Big] = \mathbb{E}_{s_0 \sim \rho_0}\big[V^{\pi_\theta}(s_0)\big]. \tag{2}$$

三个价值函数：

$$V^\pi(s) = \mathbb{E}_\pi\Big[\textstyle\sum_{t \geq 0} \gamma^t r_t \,\Big|\, s_0 = s\Big], \tag{3}$$

$$Q^\pi(s, a) = \mathbb{E}_\pi\Big[\textstyle\sum_{t \geq 0} \gamma^t r_t \,\Big|\, s_0 = s, a_0 = a\Big], \tag{4}$$

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s). \tag{5}$$

定义 1.2 (折扣状态访问测度)**.** $\rho_\pi(s) = \displaystyle\sum_{t \geq 0} \gamma^t \Pr(s_t = s \mid \pi)$。

注 1.3. ρ_π 是未归一化的测度，总质量为 $1/(1-\gamma)$。本讲义所有含 ρ_π 的等式都在这个约定下成立，这样可以避免反复出现 $1/(1-\gamma)$ 因子。很多笔记在这里含糊其辞，导致后续常数对不上。

2 策略梯度定理

2.1 对数导数技巧

难点在于：$J(\theta)$ 是对分布 p_θ 求期望，而分布本身依赖 θ，不能直接把 ∇_θ 挪进期望号。解决办法是下面这个恒等式。

引理 2.1 (得分函数恒等式)**.** 对任意在 θ 处可微且支撑集不依赖 θ 的概率密度 p_θ，

$$\nabla_\theta p_\theta(x) = p_\theta(x)\, \nabla_\theta \log p_\theta(x), \qquad \mathbb{E}_{x \sim p_\theta}\big[\nabla_\theta \log p_\theta(x)\big] = 0.$$

证明. 第一式由链式法则 $\nabla \log p = \nabla p / p$ 移项即得。第二式：

$$\mathbb{E}_{x \sim p_\theta}[\nabla_\theta \log p_\theta(x)] = \int p_\theta(x) \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} \mathrm{d}x = \nabla_\theta \int p_\theta(x) \mathrm{d}x = \nabla_\theta 1 = 0. \qquad \square$$

为什么第二式如此重要

$\mathbb{E}[\text{得分}] = 0$ 是后文三处关键结论的共同来源：基线的无偏性（命题 2.4）、代理目标的一阶精确性（定理 5.1）、以及 Fisher 信息矩阵等于 KL 的 Hessian（定理 6.2）。把它单独提出来，后面就不必重复论证。

2.2 基本形式

定理 2.2 (策略梯度，轨迹形式)**.**

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim p_\theta}\Big[\Big(\sum_{t\ge 0}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\Big) R(\tau)\Big], \qquad R(\tau) = \sum_{t\ge 0}\gamma^t r_t.$$

证明. 由引理 2.1，

$$\nabla_\theta J = \int \nabla_\theta p_\theta(\tau) R(\tau)\mathrm{d}\tau = \int p_\theta(\tau)\nabla_\theta \log p_\theta(\tau) R(\tau)\mathrm{d}\tau = \mathbb{E}_\tau\big[\nabla_\theta \log p_\theta(\tau) R(\tau)\big].$$

对 (1) 取对数：

$$\log p_\theta(\tau) = \log \rho_0(s_0) + \sum_t \log \pi_\theta(a_t \mid s_t) + \sum_t \log P(s_{t+1} \mid s_t, a_t).$$

第一项与第三项均不含 θ（环境是给定的，我们只能调策略），求导后为零，故 $\nabla_\theta \log p_\theta(\tau) = \sum_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)$。 □

这一步的物理含义

环境动力学 P 被消掉了——这正是策略梯度可以无模型使用的根本原因：我们从不需要知道 P，只需要能从环境中采样。

2.3 三个精化

定理 2.2 的估计量方差极大，实用形式需要三步改进。

(1) 因果性（reward-to-go）。 t 时刻的动作不可能影响 $t' < t$ 时刻已经拿到的奖励，这部分相关性纯属噪声。

命题 2.3. 对任意 $t' < t$，$\mathbb{E}\big[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\,\gamma^{t'} r_{t'}\big] = 0$。因而

$$\nabla_\theta J = \mathbb{E}\Big[\sum_{t\ge 0}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\sum_{t'\ge t}\gamma^{t'} r_{t'}\Big].$$

证明. 用全期望公式，先对 a_t 在给定历史 $\mathcal{F}_t = (s_0, a_0, \ldots, s_t)$ 下取条件期望。$r_{t'}$（$t' < t$）是 $\mathcal{F}_t$ 可测的，可提到条件期望外：

$$\mathbb{E}\big[\nabla \log \pi_\theta(a_t \mid s_t)\gamma^{t'} r_{t'}\big] = \mathbb{E}\Big[\gamma^{t'} r_{t'}\underbrace{\mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}\big[\nabla \log \pi_\theta(a_t \mid s_t)\big]}_{=0\ (\text{引理 } 2.1)}\Big] = 0.$$ □

(2) 基线（baseline）。

命题 2.4 (基线的无偏性)**.** 对任意只依赖状态的函数 $b : \mathcal{S} \to \mathbb{R}$，$\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\big[\nabla_\theta \log \pi_\theta(a \mid s)\, b(s)\big] = 0$。

证明. $b(s)$ 对 a 而言是常数，提到期望外，再用引理 2.1 即得。 □

取 $b(s) = V^\pi(s)$，权重就从 Q^π 变成 $A^\pi = Q^\pi - V^\pi$：期望不变（无偏），但方差通常大幅下降，因为减掉了“状态本身好坏”这部分与动作无关的共同成分。

(3) 状态分布形式。

定理 2.5 (策略梯度定理)**.**

$$\boxed{\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_{\pi_\theta},\, a\sim\pi_\theta}\big[\nabla_\theta \log \pi_\theta(a \mid s)\, A^{\pi_\theta}(s,a)\big].}$$

一个几乎所有实现都存在的偏差

定理 2.5 中 ρ_π 带有 γ^t 权重，即时刻 t 的样本应按 γ^t 加权。而主流实现（含 PPO 的标准实现）对所有时刻等权处理。因此实践中的“策略梯度”严格来说不是 J 的梯度，而是另一个目标的梯度。这一偏差在 $\gamma \to 1$ 时可忽略，是被广泛接受的工程折中，但面试中被追问时应当能说清楚。

3 优势函数与 GAE

3.1 两条基本性质

命题 3.1. (i) $A^\pi(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\big[r(s,a) + \gamma V^\pi(s') - V^\pi(s)\big]$；
(ii) $\mathbb{E}_{a\sim\pi(\cdot|s)}\big[A^\pi(s,a)\big] = 0$ 对每个 s 成立。

证明. (i) 把 Q^π 的贝尔曼方程 $Q^\pi(s,a) = \mathbb{E}_{s'}[r + \gamma V^\pi(s')]$ 代入定义。
(ii) $\mathbb{E}_{a\sim\pi}[Q^\pi(s,a)] = V^\pi(s)$ 是 V 的定义，故 $\mathbb{E}_{a\sim\pi}[A^\pi] = V^\pi(s) - V^\pi(s) = 0$。 □

性质 (ii) 说明优势函数是零中心化的：它只衡量相对好坏，不含绝对价值。这条性质在定理 5.1 中会起决定性作用。

3.2 GAE 的推导

实践中 Q 与 V 都是期望值，拿不到，只能估计。记 TD 残差

$$\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t). \tag{6}$$

若 $V = V^\pi$ 精确，则由命题 3.1(i) 有 $\mathbb{E}[\delta_t^V] = A^\pi(s_t,a_t)$，即 δ_t 是无偏估计。定义 k 步估计

$$\hat{A}_t^{(k)} = \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}^V = -V(s_t) + \sum_{l=0}^{k-1} \gamma^l r_{t+l} + \gamma^k V(s_{t+k}). \tag{7}$$

k 越大越接近蒙特卡洛（偏差小、方差大），k 越小越依赖 V（偏差大、方差小）。

命题 3.2 (GAE 的闭式)**.** 以 $(1-\lambda)\lambda^{k-1}$ 为权重对 $\hat{A}^{(k)}$ 做指数加权平均，得

$$\boxed{\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} := (1-\lambda)\sum_{k\geq 1} \lambda^{k-1}\hat{A}_t^{(k)} = \sum_{l\geq 0} (\gamma\lambda)^l\, \delta_{t+l}^V.}$$

证明. 把定义展开，按 δ_{t+l} 归并同类项。δ_{t+l} 出现在所有 $k \geq l+1$ 的项中，其总系数为

$$(1-\lambda)\gamma^l \sum_{k\geq l+1} \lambda^{k-1} = (1-\lambda)\gamma^l \cdot \frac{\lambda^l}{1-\lambda} = (\gamma\lambda)^l.$$ □

两个端点值得记住：$\lambda = 0$ 时 $\hat{A}_t = \delta_t$（单步 TD）；$\lambda = 1$ 时 $\hat{A}_t = \sum_{l\geq 0}\gamma^l r_{t+l} - V(s_t)$（蒙特卡洛，与 V 的偏差无关）。

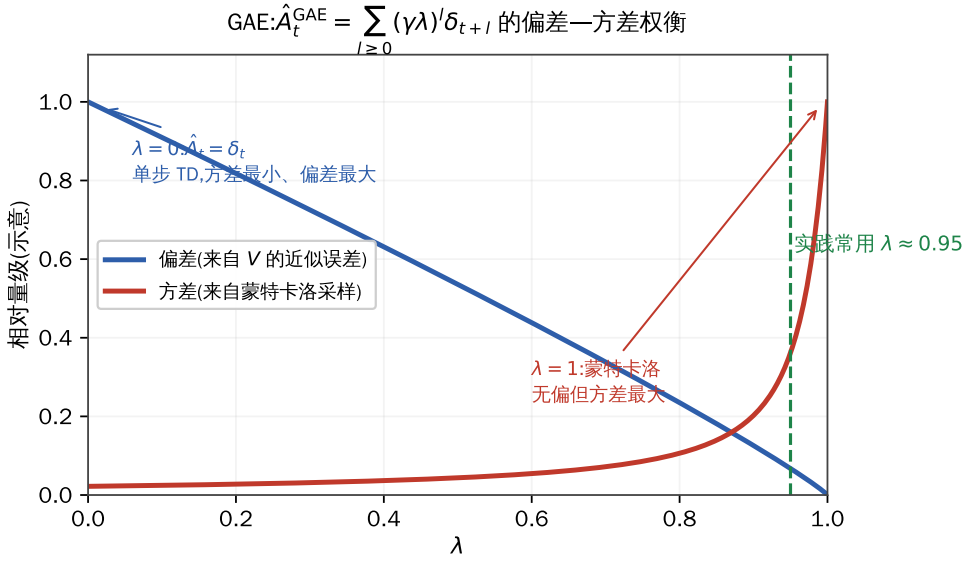

图 1　GAE 中 λ 的作用：在 V 的近似偏差与蒙特卡洛采样方差之间连续插值。

4　性能差异引理：一切的起点

下面这条引理是整个信任域方法族的地基。很多笔记直接引用它的结论，但它其实有一个只用望远镜求和（telescoping）就能完成的漂亮证明。

定理 4.1 (性能差异引理，Kakade–Langford 2002)**.** 对任意两个策略 π 与 π'，

$$\boxed{J(\pi') - J(\pi) = \mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t\ge 0}\gamma^t A^{\pi}(s_t,a_t)\Big] = \sum_{s}\rho_{\pi'}(s)\sum_{a}\pi'(a\mid s)A^{\pi}(s,a).}$$

证明. 在 π' 生成的轨迹上展开 A^{π}（命题 3.1(i)）：

$$\begin{aligned}\mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t\ge 0}\gamma^t A^{\pi}(s_t,a_t)\Big] &= \mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t\ge 0}\gamma^t\big(r_t+\gamma V^{\pi}(s_{t+1})-V^{\pi}(s_t)\big)\Big]\\ &= \underbrace{\mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t\ge 0}\gamma^t r_t\Big]}_{=J(\pi')} + \mathbb{E}_{\tau\sim\pi'}\Big[\sum_{t\ge 0}\big(\gamma^{t+1}V^{\pi}(s_{t+1})-\gamma^t V^{\pi}(s_t)\big)\Big].\end{aligned}$$

第二个求和望远镜相消，只剩首项（$\gamma<1$ 且 V^{π} 有界时尾项趋于零）：

$$\mathbb{E}_{\tau\sim\pi'}\Big[-V^{\pi}(s_0)\Big] = -\mathbb{E}_{s_0\sim\rho_0}\big[V^{\pi}(s_0)\big] = -J(\pi).$$

两式相加即得结论。第二个等号只是把轨迹期望按状态访问频率重写。 □

引理在说什么

新策略比旧策略好多少，完全由新策略走过的状态上、旧策略的优势函数决定。注意其中的错配：期望对 π' 取，而优势函数是 π 的。这个错配正是所有困难的来源——我们手上只有 π 采集的数据。

5 代理目标：为什么可以“忽略状态分布漂移”

定理 4.1 无法直接优化：$\rho_{\pi'}$ 依赖于我们尚未确定的 π'，要评估它就得用 π' 重新采样，而这正是我们想避免的。于是把 $\rho_{\pi'}$ 换成 $\rho_{\pi_{\theta_{\text{old}}}}$，定义**代理目标**

$$L_{\theta_{\text{old}}}(\theta) := J(\theta_{\text{old}}) + \sum_s \rho_{\pi_{\theta_{\text{old}}}}(s) \sum_a \pi_\theta(a \mid s) A^{\pi_{\theta_{\text{old}}}}(s,a). \tag{8}$$

用重要性采样把对 π_θ 的求和换成可用旧数据估计的形式：

$$\sum_a \pi_\theta(a \mid s) A(s,a) = \sum_a \pi_{\theta_{\text{old}}}(a \mid s) \frac{\pi_\theta(a \mid s)}{\pi_{\theta_{\text{old}}}(a \mid s)} A(s,a) = \mathbb{E}_{a \sim \pi_{\theta_{\text{old}}}} \left[\frac{\pi_\theta(a \mid s)}{\pi_{\theta_{\text{old}}}(a \mid s)} A(s,a) \right], \tag{9}$$

于是

$$L_{\theta_{\text{old}}}(\theta) = J(\theta_{\text{old}}) + \mathbb{E}_{s \sim \rho_{\pi_{\theta_{\text{old}}}},\, a \sim \pi_{\theta_{\text{old}}}} \left[\frac{\pi_\theta(a \mid s)}{\pi_{\theta_{\text{old}}}(a \mid s)} A^{\pi_{\theta_{\text{old}}}}(s,a) \right]. \tag{10}$$

5.1 一阶精确性：这不是近似

替换 $\rho_{\pi'} \to \rho_{\pi_{\theta_{\text{old}}}}$ 常被描述为“假设状态分布变化忽略不计”。这个说法容易让人以为它是一个粗糙的近似。实际上它比这强得多：

定理 5.1 (代理目标与真实目标一阶重合)**.**

$$L_{\theta_{\text{old}}}(\theta_{\text{old}}) = J(\theta_{\text{old}}), \qquad \nabla_\theta L_{\theta_{\text{old}}}(\theta)\big|_{\theta=\theta_{\text{old}}} = \nabla_\theta J(\theta)\big|_{\theta=\theta_{\text{old}}}.$$

证明. **零阶**。在 $\theta = \theta_{\text{old}}$ 处 (8) 的第二项为 $\sum_s \rho_{\pi_{\theta_{\text{old}}}}(s) \sum_a \pi_{\theta_{\text{old}}}(a \mid s) A^{\pi_{\theta_{\text{old}}}}(s,a)$，由命题 3.1(ii) 内层和对每个 s 都为零，故 $L(\theta_{\text{old}}) = J(\theta_{\text{old}})$。

一阶。由定理 4.1，$J(\theta) = J(\theta_{\text{old}}) + \sum_s \rho_{\pi_\theta}(s)\, h_\theta(s)$，其中 $h_\theta(s) := \sum_a \pi_\theta(a \mid s) A^{\pi_{\theta_{\text{old}}}}(s,a)$。对 θ 求导并用乘积法则：

$$\nabla_\theta J(\theta)\Big|_{\theta_{\text{old}}} = \underbrace{\sum_s \big(\nabla_\theta \rho_{\pi_\theta}(s)\big)\Big|_{\theta_{\text{old}}} h_{\theta_{\text{old}}}(s)}_{\text{(I)}} + \underbrace{\sum_s \rho_{\pi_{\theta_{\text{old}}}}(s)\, \nabla_\theta h_\theta(s)\Big|_{\theta_{\text{old}}}}_{\text{(II)}}.$$

关键在于 $h_{\theta_{\text{old}}}(s) = \mathbb{E}_{a \sim \pi_{\theta_{\text{old}}}}[A^{\pi_{\theta_{\text{old}}}}(s,a)] = 0$ 对每一个 s 成立（命题 3.1(ii)），因此 (I) $= 0$——无论 $\nabla_\theta \rho_{\pi_\theta}$ 多大。而 (II) 恰是 $\nabla_\theta L_{\theta_{\text{old}}}(\theta)|_{\theta_{\text{old}}}$。 □

正确的理解方式

状态分布的导数 $\nabla_\theta \rho_{\pi_\theta}$ 一点也不小——策略稍变，访问的状态分布可以剧烈改变。它之所以不出现在一阶项里，是因为它被优势函数的零均值性质整体抹掉了。所以“忽略状态分布漂移”造成的误差是 $O(\|\theta - \theta_{\text{old}}\|^2)$ 而非 $O(\|\theta - \theta_{\text{old}}\|)$，这才是信任域方法成立的真正基础：只要步子足够小，代理目标就可信。

5.2 误差有多大：单调改进定理

一阶重合只说明误差是二阶的，还需要一个可计算的界来决定步子该多小。

定理 5.2 (单调改进下界，Schulman et al. 2015)**.** 记 $\varepsilon = \max_{s,a} |A^{\pi_{\theta_{\text{old}}}}(s,a)|$，$\alpha = \max_s D_{\text{TV}}\big(\pi_{\theta_{\text{old}}}(\cdot \mid s), \pi_\theta(\cdot \mid s)\big)$，则

$$J(\theta) \;\geq\; L_{\theta_{\text{old}}}(\theta) - \frac{4\varepsilon\gamma}{(1-\gamma)^2}\alpha^2.$$

结合 Pinsker 不等式 $D_{\text{TV}}(p,q)^2 \leq \frac{1}{2} D_{\text{KL}}(p\|q)$，可写成 KL 形式

$$J(\theta) \;\geq\; \underbrace{L_{\theta_{\text{old}}}(\theta) - C\, D_{\text{KL}}^{\max}(\pi_{\theta_{\text{old}}}, \pi_\theta)}_{=:M(\theta)}, \qquad C = \frac{2\varepsilon\gamma}{(1-\gamma)^2}.$$

证明思路. 构造 $\pi_{\theta_{\text{old}}}$ 与 π_θ 的一个耦合：在每个状态上以概率 $\geq 1-\alpha$ 让两者取相同动作。则两条轨迹在前 t 步完全一致的概率至少为 $(1-\alpha)^t$。把 $J(\theta)-L_{\theta_{\text{old}}}(\theta)$ 按“首次分歧时刻”分解，每个分歧事件贡献至多 ε 的误差、发生概率至多 $1-(1-\alpha)^t \leq t\alpha$，再对 t 按 γ^t 求和并使用 $\sum_t t\gamma^t = \gamma/(1-\gamma)^2$，即得上述常数。完整论证见原文附录。 □

推论 5.3 (单调不降)**.** 若取 $\theta_{\text{new}} = \arg\max_\theta M(\theta)$，则 $J(\theta_{\text{new}}) \geq M(\theta_{\text{new}}) \geq M(\theta_{\text{old}}) = J(\theta_{\text{old}})$，即回报单调不降。

证明. 第一个不等号是定理 5.2；第二个由 θ_{new} 的最优性；第三个等号因为 $D_{\text{KL}}(\pi_{\theta_{\text{old}}} \| \pi_{\theta_{\text{old}}}) = 0$ 且 $L(\theta_{\text{old}}) = J(\theta_{\text{old}})$（定理 5.1）。 □

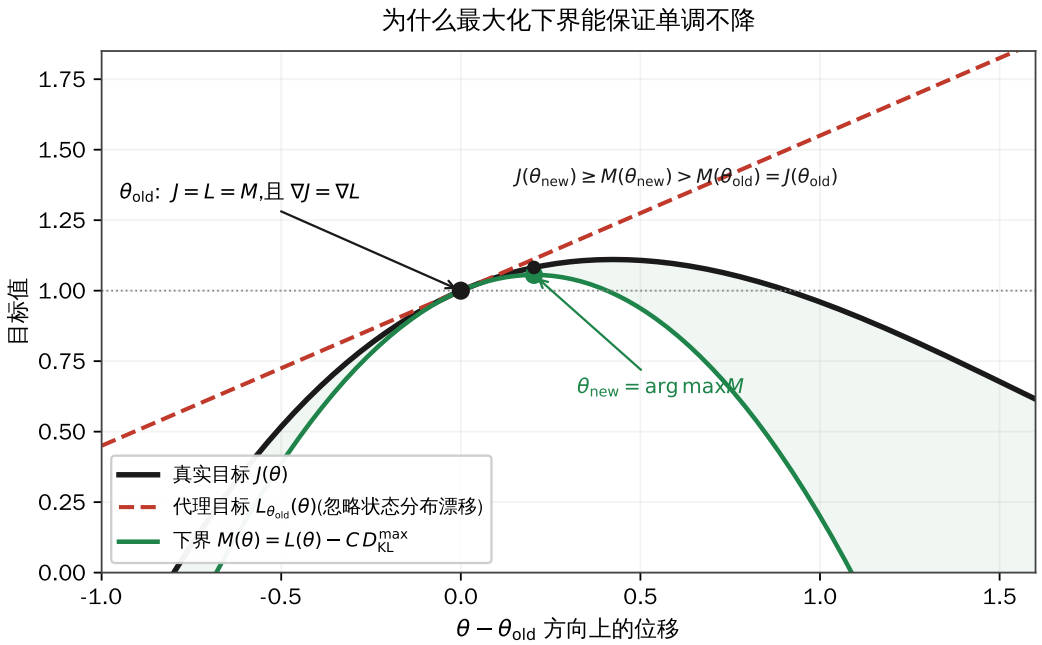

图 2 最小化–最大化（MM）结构。代理目标 L 与 J 在 θ_{old} 处零阶、一阶重合但会越走越偏；减去 KL 惩罚后得到处处不高于 J 的下界 M，最大化 M 即保证 J 不降。

为什么实践中不直接用这个定理

理论常数 $C = 2\varepsilon\gamma/(1-\gamma)^2$ 极其保守：取 $\gamma = 0.99$ 即有 $(1-\gamma)^{-2} = 10^4$，按此惩罚项走出的步长小到无法训练。所以 TRPO 把它改造为硬约束：不再惩罚 KL，而是限制 $\bar{D}_{\text{KL}} \leq \delta$，$\delta$ 作为超参数（典型取 0.01）。同时把 $D_{\text{KL}}^{\max}$ 换成对 $\rho_{\pi_{\theta_{\text{old}}}}$ 的平均 KL，因为最大值在连续状态空间中无法估计。这两步都牺牲了定理的严格保证，换取可用性——理论给的是方向，不是配方。

于是得到 TRPO 的优化问题：

$$\max_\theta \ \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t\right] \qquad \text{s.t.} \quad \bar{D}_{\text{KL}}(\theta_{\text{old}}, \theta) := \hat{\mathbb{E}}_t\left[D_{\text{KL}}\left(\pi_{\theta_{\text{old}}}(\cdot \mid s_t) \,\|\, \pi_\theta(\cdot \mid s_t)\right)\right] \leq \delta. \tag{11}$$

为什么约束策略而不是参数

参数与输出行为之间是高度非线性的关系：同样大小的 $\|\Delta\theta\|$，在某些方向上几乎不改变策略，在另一些方向上可以把动作分布彻底改写。而定理 5.2 的误差界是用分布距离表述的，所以约束必须直接施加在行为上，界才成立。一个额外的好处是：KL 约束对参数的重参数化不变（见 7.1 节），而 $\|\Delta\theta\| \leq r$ 依赖于你恰好怎么写这个网络。

6 两次泰勒展开与 Fisher 信息矩阵

问题 (11) 仍难直接求解。TRPO 的做法是在 θ_{old} 附近做局部近似：目标取一阶、约束取二阶。

6.1 目标的一阶展开

记 $L(\theta)$ 为 (10) 的第二项。由定理 5.1, $L(\theta_{\text{old}}) = 0$（去掉常数 $J(\theta_{\text{old}})$ 后）且 $\nabla_\theta L|_{\theta_{\text{old}}} = \nabla_\theta J|_{\theta_{\text{old}}} =: g$, 即

$$L(\theta_{\text{old}} + \Delta\theta) \approx g^\top \Delta\theta, \qquad g = \mathbb{E}\big[\nabla_\theta \log \pi_\theta(a \mid s)\big|_{\theta_{\text{old}}} A^{\pi_{\theta_{\text{old}}}}(s,a)\big]. \tag{12}$$

可以直接验证：$\nabla_\theta \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}\big|_{\theta_{\text{old}}} = \frac{\nabla_\theta \pi_\theta}{\pi_{\theta_{\text{old}}}}\big|_{\theta_{\text{old}}} = \nabla_\theta \log \pi_\theta\big|_{\theta_{\text{old}}}$，即**朴素策略梯度就是代理目标的梯度**。

6.2 Fisher 信息矩阵的三种等价形式

定义 6.1. $F(\theta) := \mathbb{E}_{x\sim p_\theta}\big[\nabla_\theta \log p_\theta(x)\, \nabla_\theta \log p_\theta(x)^\top\big]$。

定理 6.2 (三形式等价)**.** 在正则条件下（p_θ 二阶可微、支撑集不依赖 θ、可交换求导与积分），

$$\underbrace{\mathbb{E}\big[\nabla \log p_\theta\, \nabla \log p_\theta^\top\big]}_{\text{(a) 外积}} = \underbrace{-\,\mathbb{E}\big[\nabla^2 \log p_\theta\big]}_{\text{(b) 对数似然 Hessian 的负期望}} = \underbrace{\nabla^2_{\theta'} D_{\mathrm{KL}}\big(p_\theta \,\|\, p_{\theta'}\big)\Big|_{\theta'=\theta}}_{\text{(c) KL 的 Hessian}}.$$

证明. (a)=(b). 对 $\nabla \log p = \nabla p / p$ 再求一次导：

$$\nabla^2 \log p_\theta = \frac{\nabla^2 p_\theta}{p_\theta} - \frac{\nabla p_\theta\, \nabla p_\theta^\top}{p_\theta^2} = \frac{\nabla^2 p_\theta}{p_\theta} - \nabla \log p_\theta\, \nabla \log p_\theta^\top.$$

两边对 $x \sim p_\theta$ 取期望，第一项 $\mathbb{E}\big[\nabla^2 p_\theta / p_\theta\big] = \int \nabla^2 p_\theta \,\mathrm{d}x = \nabla^2 \int p_\theta \mathrm{d}x = \nabla^2 1 = 0$，于是 $\mathbb{E}[\nabla^2 \log p_\theta] = -F(\theta)$。

(b)=(c). 令 $f(\theta') = D_{\mathrm{KL}}(p_\theta \| p_{\theta'}) = \mathbb{E}_{x\sim p_\theta}\big[\log p_\theta(x) - \log p_{\theta'}(x)\big]$。注意期望是对固定的 p_θ 取的，与 θ' 无关，故可直接对 θ' 求导：

$$\nabla_{\theta'} f = -\mathbb{E}_{x\sim p_\theta}\big[\nabla_{\theta'} \log p_{\theta'}(x)\big], \qquad \nabla^2_{\theta'} f = -\mathbb{E}_{x\sim p_\theta}\big[\nabla^2_{\theta'} \log p_{\theta'}(x)\big].$$

在 $\theta' = \theta$ 处代入 (b) 即得 $\nabla^2_{\theta'} f|_{\theta'=\theta} = F(\theta)$。顺带地，$\nabla_{\theta'} f|_{\theta'=\theta} = -\mathbb{E}[\nabla \log p_\theta] = 0$（引理 2.1）。 □

推论 6.3 (KL 的二阶展开)**.**

$$\boxed{D_{\mathrm{KL}}\big(p_{\theta_{\text{old}}} \,\|\, p_{\theta_{\text{old}}+\Delta\theta}\big) = \tfrac{1}{2}\Delta\theta^\top F(\theta_{\text{old}})\Delta\theta + O(\|\Delta\theta\|^3).}$$

零阶项为 0（同分布），一阶项为 0（得分零均值）——这就是**为什么约束必须展开到二阶**：展开到一阶的话约束退化为 $0 \le \delta$，完全失效。而目标函数在一阶就有非零信息，故只需一阶。

正向 KL 还是反向 KL

定理 6.2(c) 用的是正向 KL $D_{\mathrm{KL}}(p_\theta \| p_{\theta'})$。可以证明反向 KL $D_{\mathrm{KL}}(p_{\theta'} \| p_\theta)$ 在 $\theta' = \theta$ 处的 Hessian 也等于 F。两者只在三阶及以上才分道扬镳，因此在信任域这种局部近似里，KL 写哪个方向不影响结果。第 10 节的实验 1 对此做了数值确认（图 5 右）。

为什么目标不也展开到二阶

数学上当然可以，但目标的 Hessian $H_L = \mathbb{E}\big[(\nabla^2 \log \pi_\theta) A + \nabla \log \pi_\theta \nabla \log \pi_\theta^\top A\big]$ 带有优势函数 A 作权重，因而 **(i)** 不保证半正定——最大化一个非凹的二次近似会让优化器认为“越远越好”，数值上极不稳定；**(ii)** 每个样本的 A 差异很大，估计方差远高于 F（F 只依赖策略本身，与奖励无关）。所以 TRPO 的分工是：目标的一阶信息决定往哪走，*KL* 的二阶结构决定能走多远。

注 6.4 (策略场景下的 F)**.** 在 RL 中 F 是对状态取平均的条件 Fisher：$F = \mathbb{E}_{s\sim\rho_{\pi_{\theta_{\text{old}}}}}\big[\mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(\cdot|s)}[\nabla \log \pi\, \nabla \log \pi^\top]\big]$。由外积形式立刻可知 $F \succeq 0$；但它常常奇异（例如 softmax 参数化下全 1 方向落在零空间中），因此实现里必须加阻尼 $F + \alpha I$，$\alpha \sim 10^{-1} \sim 10^{-3}$。

7 信任域二次规划的精确解

把两次展开代回 (11)，得到标准二次规划

$$\max_{\Delta\theta}\; g^{\top}\Delta\theta \qquad \text{s.t.} \quad \tfrac{1}{2}\Delta\theta^{\top}F\Delta\theta \le \delta. \tag{13}$$

定理 7.1 (自然梯度步)**.** 设 $F \succ 0$ 且 $g \neq 0$。则 (13) 的唯一最优解为

$$\boxed{\Delta\theta^{\star} = \sqrt{\frac{2\delta}{g^{\top}F^{-1}g}}\,F^{-1}g,} \qquad \text{最优值} \quad g^{\top}\Delta\theta^{\star} = \sqrt{2\delta\, g^{\top}F^{-1}g}.$$

证明. 目标线性、可行域为紧凸椭球，故最优解存在且必在边界上（否则沿 g 方向可继续增大）。构造拉格朗日函数 $\mathcal{L}(\Delta\theta,\lambda) = g^{\top}\Delta\theta - \lambda\big(\tfrac{1}{2}\Delta\theta^{\top}F\Delta\theta - \delta\big)$，稳定性条件

$$\nabla_{\Delta\theta}\mathcal{L} = g - \lambda F\Delta\theta = 0 \;\Longrightarrow\; \Delta\theta = \tfrac{1}{\lambda}F^{-1}g \quad (\lambda > 0).$$

代入取等的约束 $\tfrac{1}{2}\Delta\theta^{\top}F\Delta\theta = \delta$：

$$\frac{1}{2\lambda^{2}}\,g^{\top}F^{-1}FF^{-1}g = \frac{g^{\top}F^{-1}g}{2\lambda^{2}} = \delta \;\Longrightarrow\; \lambda = \sqrt{\frac{g^{\top}F^{-1}g}{2\delta}}.$$

回代得 $\Delta\theta^{\star}$。由于 $F \succ 0$，KKT 条件对该凸可行域是充分的，且解唯一。 □

方向 $F^{-1}g$ 称为**自然梯度**；步长因子 $\sqrt{2\delta/(g^{\top}F^{-1}g)}$ 由信任域半径决定。第 10 节实验 2 用数值优化独立验证了该闭式解。

7.1 几何意义与重参数化不变性

普通梯度上升 $\theta \leftarrow \theta + \alpha g$ 隐含一个假设：在参数空间中沿各方向移动相同的欧氏距离 $\|\Delta\theta\|_2$，对策略造成的改变是可比的。这个假设是错的——参数空间是平的，策略空间是弯的。

在概率分布流形上，衡量 π_{θ} 与 $\pi_{\theta+\Delta\theta}$ 差异的自然度量不是 $\|\Delta\theta\|_2^2$，而是二阶 KL，即推论 6.3 中的 $\Delta\theta^{\top}F\Delta\theta$。$F$ 在这里扮演**黎曼度量张量**的角色，F^{-1} 则把欧氏梯度“矫正”到流形的曲率上。

命题 7.2 (自然梯度的协变性)**.** 设重参数化 $\varphi = h(\theta)$ 可微可逆，雅可比 $\mathbf{J} = \partial\varphi/\partial\theta$。则

$$F_{\varphi}^{-1}g_{\varphi} = \mathbf{J}\,\big(F_{\theta}^{-1}g_{\theta}\big).$$

即自然梯度按切向量的方式变换，因而两种参数化下走出的策略变化完全相同。普通梯度不具备该性质。

证明. 链式法则给出 $\nabla_{\theta} = \mathbf{J}^{\top}\nabla_{\varphi}$，故 $g_{\theta} = \mathbf{J}^{\top}g_{\varphi}$，即 $g_{\varphi} = \mathbf{J}^{-\top}g_{\theta}$。同理 $F_{\theta} = \mathbb{E}[\mathbf{J}^{\top}\nabla_{\varphi}\log p\,(\mathbf{J}^{\top}\nabla_{\varphi}\log p)^{\top}] = \mathbf{J}^{\top}F_{\varphi}\mathbf{J}$，即 $F_{\varphi} = \mathbf{J}^{-\top}F_{\theta}\mathbf{J}^{-1}$。于是

$$F_{\varphi}^{-1}g_{\varphi} = \mathbf{J}F_{\theta}^{-1}\mathbf{J}^{\top}\cdot\mathbf{J}^{-\top}g_{\theta} = \mathbf{J}F_{\theta}^{-1}g_{\theta}.$$ □

这条性质是自然梯度真正的卖点：它使算法对“你恰好怎么写这个网络”不敏感。第 10 节实验 3 给出了数值确认。

8 TRPO 的工程实现

理论解 $\Delta\theta^{\star}$ 直接算不出来：F 是 $n \times n$ 矩阵，n 是网络参数量（10^6 起步），显式构造需 $O(n^2)$ 存储、求逆需 $O(n^3)$ 时间。TRPO 用三个技巧绕开。

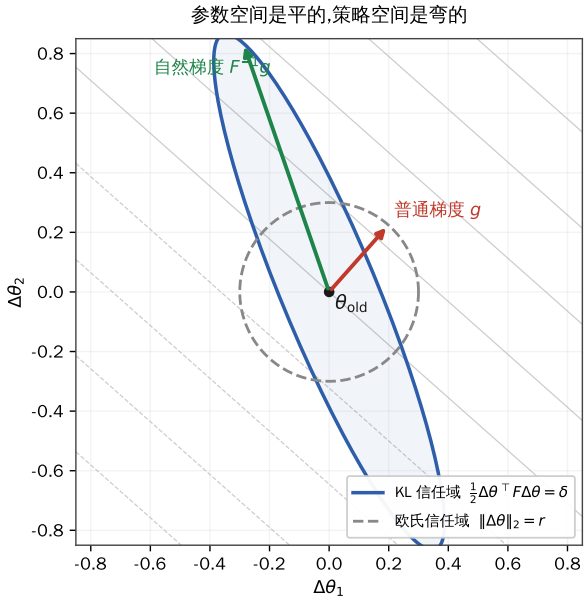

图 3 信任域几何。当 F 病态时，KL 信任域是一个被拉长的椭球：沿椭球长轴方向参数可以走很远而策略几乎不变，沿短轴方向则相反。普通梯度 g（红）指向目标上升最快的欧氏方向，自然梯度 $F^{-1}g$（绿）指向在给定策略变化预算下收益最大的方向。

(1) Fisher 向量积（FVP）：永不构造 F。 由定理 6.2(c)，$F = \nabla_\theta^2 \bar{D}_{\text{KL}}$，因此对任意向量 v，

$$Fv = \nabla_\theta\Big(\big(\nabla_\theta \bar{D}_{\text{KL}}\big)^\top v\Big), \tag{14}$$

即先算 KL 的梯度、与 v 做内积得到标量、再反传一次。代价约为两次反向传播，存储 $O(n)$。

(2) 共轭梯度法：解线性方程而非求逆。 用 CG 迭代求解 $Fx = g$，得 $x \approx F^{-1}g$。CG 只需要 FVP，通常 $10\sim20$ 次迭代即足够，总代价 $O(kn)$。实现中对 F 加阻尼 $F + \alpha I$ 以应对奇异性与数值噪声。

(3) 线性搜索：把近似拉回现实。 由于 CG 只给出近似解、且两次泰勒展开本身有误差，按 $\Delta\theta^\star$ 走完整步未必真的满足 $\bar{D}_{\text{KL}} \le \delta$，也未必真的提升了目标。于是做指数回退搜索：令 $\theta_j = \theta_{\text{old}} + \beta^j \Delta\theta^\star$（$\beta = 0.5$，$j = 0, 1, 2, \dots$），取第一个同时满足

$$\bar{D}_{\text{KL}}(\theta_{\text{old}}, \theta_j) \le \delta \qquad \text{且} \qquad L(\theta_j) > 0$$

的 j；若若干次后仍不满足则本轮不更新。这一步是“信任域”三个字在代码中真正的落点。

一个容易写错的数值细节

计算步长时应当用 $\sqrt{2\delta/(x^\top Fx)}\,x$ 而不是 $\sqrt{2\delta/(g^\top x)}\,x$。在精确算术下两者相同（$x = F^{-1}g \Rightarrow x^\top Fx = g^\top F^{-1}g = g^\top x$），但 CG 只给近似解时，$x^\top Fx$ 用一次 FVP 即可得到且始终非负，数值上更稳。

9 从 TRPO 到 PPO

TRPO 每一步都要跑 CG 加线性搜索，实现复杂、与 dropout/参数共享等结构不易兼容。PPO 的出发点是：能不能只用一阶方法，近似达到“不让策略变化太大”的效果?

记概率比 $r_t(\theta) = \dfrac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$，PPO 的截断目标为

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t\Big[\min\big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big]. \tag{15}$$

命题 9.1 (截断的单边性)**.**

$$\min\left(r\hat{A}, \mathrm{clip}(r, 1-\epsilon, 1+\epsilon)\hat{A}\right) = \begin{cases} \hat{A}\cdot\min(r,\,1+\epsilon), & \hat{A} > 0, \\ \hat{A}\cdot\max(r,\,1-\epsilon), & \hat{A} < 0. \end{cases}$$

即：$\hat{A} > 0$ 时只在 $r > 1+\epsilon$ 一侧截断，$\hat{A} < 0$ 时只在 $r < 1-\epsilon$ 一侧截断。

证明. $\hat{A} > 0$ 时 min 作用在 r 与 $\mathrm{clip}(r)$ 上，而 $\mathrm{clip}(r) \le r$ 当且仅当 $r > 1+\epsilon$；$\hat{A} < 0$ 时不等号反向。 □

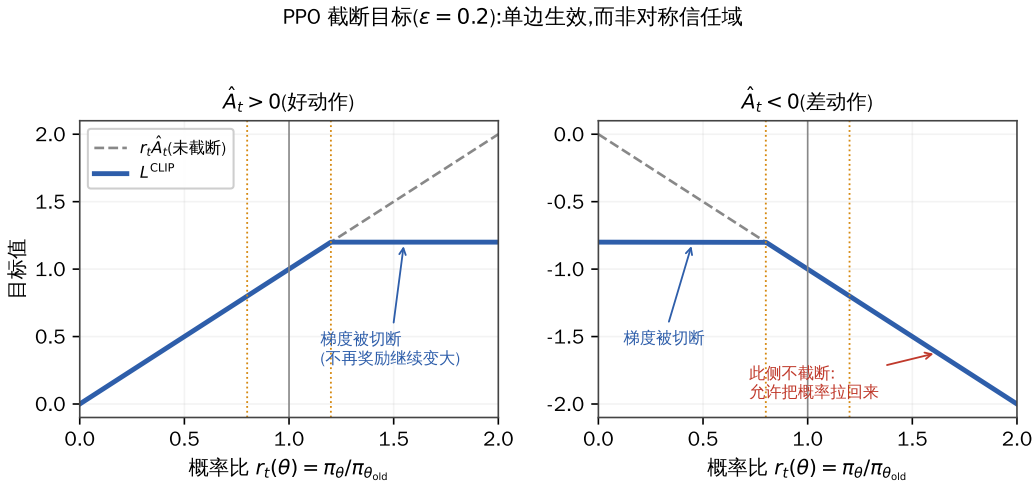

图 4 PPO 截断目标。斜率变为零处梯度被切断，优化器不再有动力把 r 推得更远；但另一侧始终保持斜率，允许把已经偏离过头的概率拉回来。

PPO 常见误解

误解一：“clip 给出了 KL 的上界。”不成立。截断只约束单个样本的概率比，且只是让梯度归零，并未施加任何惩罚力。在同一批数据上做多个 epoch 的更新时，$\bar{D}_{\mathrm{KL}}$ 完全可以累积到远超 δ 的水平。这正是主流实现要额外加 *KL early stopping*（监控 $\bar{D}_{\mathrm{KL}}$，超阈值就提前结束本轮 epoch）的原因。

误解二：“截断区之外梯度为零，所以样本被丢弃了。”只是该样本在该侧的一阶导为零；若参数更新后 r 回到区间内，它又会重新产生梯度。

误解三：“PPO 是 TRPO 的严格近似。”两者只在动机上相承。PPO 没有定理 5.2 意义上的单调改进保证，它的成功更多来自工程上的稳健性与实现简洁。

另一变体 PPO-Penalty 直接把 KL 放进目标：$L = \hat{\mathbb{E}}_t[r_t\hat{A}_t] - \beta\,\bar{D}_{\mathrm{KL}}$，并按 $\bar{D}_{\mathrm{KL}}$ 与目标值的比较自适应调整 β。这一形式反而更贴近定理 5.2 的原始结构。

	Vanilla PG	**TRPO**	**PPO-Clip**
步长控制	学习率 α（手调）	KL 硬约束 δ	概率比截断 ϵ
用到的曲率	无	KL 的二阶（F）	无（一阶）
每步代价	1 次反传	CG（k 次 FVP）+ 线搜	1 次反传
数据复用	单次	单次	多 epoch
理论保证	无	近似单调改进	无

表 1 三种方法的定位对照。

注 9.2 (延伸: GRPO). 在大模型 RLHF 场景中, 训练一个与策略同规模的价值网络代价高昂。GRPO 的思路是对同一 prompt 采样一组 G 个回答，用组内奖励的标准化值 $\hat{A}_i = (R_i - \text{mean}(R))/\text{std}(R)$ 直接替代优势函数，从而完全去掉 critic。它继承了 PPO 的截断结构，差别只在优势函数的估计方式——这也说明本讲义前半部分（策略梯度、优势、信任域）才是真正通用的骨架。

10 数值验证

理论推导应当可被独立检验。以下四组实验直接对前文结论做数值确认。

10.1 实验 1：KL 的二阶展开与两个方向的等价性

取 $K = 6$ 的 softmax 策略 $p = \text{softmax}(\theta)$，其 Fisher 矩阵有闭式 $F = \text{diag}(p) - pp^\top$。沿随机方向 d（已投影掉 F 的零空间）计算正向与反向 KL，与 $\frac{1}{2}\epsilon^2 d^\top F d$ 比较：

ϵ	正向 KL	反向 KL	$\frac{1}{2}\epsilon^2 d^\top F d$	正向残差 $/\epsilon^3$	反向残差 $/\epsilon^3$
10^{-1}	7.8655×10^{-4}	7.9037×10^{-4}	7.8259×10^{-4}	0.0040	0.0078
10^{-2}	7.8300×10^{-6}	7.8340×10^{-6}	7.8259×10^{-6}	0.0041	0.0081
10^{-3}	7.8263×10^{-8}	7.8267×10^{-8}	7.8259×10^{-8}	0.0041	0.0082
10^{-4}	7.8259×10^{-10}	7.8260×10^{-10}	7.8259×10^{-10}	0.0043	0.0079

表 2 残差除以 ϵ^3 后收敛到有限常数，说明两个方向的 KL 残差都恰为 $O(\epsilon^3)$，即二者在 θ_{old} 处的 Hessian 均为 F。这验证了定理 6.2 与推论 6.3。

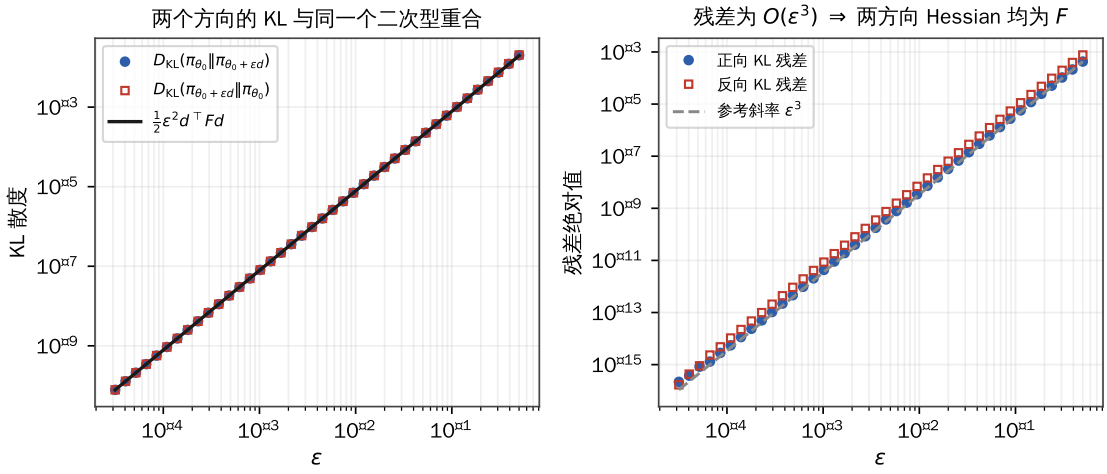

图 5 左：两个方向的 KL 与同一条二次曲线重合。右：残差斜率与 ϵ^3 参考线平行，确认三阶收敛。

10.2 实验 2：信任域二次规划闭式解

取 $n = 8$、随机正定 F、随机 g、$\delta = 0.01$，将定理 7.1 的闭式解与 SLSQP 数值优化结果比较：

闭式解目标值 $g^\top \Delta\theta^\star$	0.2120802592
数值优化目标值	0.2120802592
理论最优值 $\sqrt{2\delta\, g^\top F^{-1} g}$	0.2120802592
两解相对误差	8.11×10^{-7}
闭式解处约束值 $\frac{1}{2}\Delta\theta^\top F \Delta\theta$	0.0100000000 $(= \delta)$

约束在最优解处取等，互补松弛条件成立，与证明中“最优解必在边界”的论断一致。

10.3 实验 3：自然梯度的重参数化不变性

取随机可逆雅可比 $\mathbf{J}$，比较 $F_\varphi^{-1} g_\varphi$ 与 $\mathbf{J} F_\theta^{-1} g_\theta$：

$$\frac{\left\| F_\varphi^{-1} g_\varphi - \mathbf{J} F_\theta^{-1} g_\theta \right\|}{\left\| F_\varphi^{-1} g_\varphi \right\|} = 8.40 \times 10^{-14} \quad \text{（机器精度）},$$

而作为对照，普通梯度的相应偏差为 1.94（量级为 1，完全不成立）。这验证了命题 7.2。

11 要点回顾与自测

11.1 逻辑链条

1. 得分函数零均值（引理 2.1）$\Rightarrow$ 策略梯度定理（定理 2.5）。
2. 望远镜求和 $\Rightarrow$ 性能差异引理（定理 4.1）：改进量由新策略的状态分布与旧策略的优势决定。
3. 换掉状态分布得代理目标 L；优势零均值 $\Rightarrow$ L 与 J 一阶重合（定理 5.1），故误差是二阶的。
4. 二阶误差可被 KL 控制（定理 5.2）$\Rightarrow$ 最大化下界保证单调不降 $\Rightarrow$ 工程上改写为硬约束 (11)。
5. 目标一阶展开 + KL 二阶展开（F 为 Hessian，定理 6.2）$\Rightarrow$ 二次规划 (13)。
6. KKT $\Rightarrow$ 自然梯度步（定理 7.1）$\Rightarrow$ FVP + CG + 线搜落地。
7. 用截断替代整套信任域机制 $\Rightarrow$ PPO（丢掉理论保证，换来实现简洁）。

11.2 自测题

1. 为什么策略梯度的推导中环境转移概率 P 会消失？这对算法意味着什么？
2. 加入基线为什么不改变梯度的期望？V^π 是不是方差最小的基线？（提示：不是。最优基线是 $\mathbb{E}[\|\nabla \log \pi\|^2 Q] / \mathbb{E}[\|\nabla \log \pi\|^2]$。）
3. 证明性能差异引理。若把 γ 取到 1，证明中的哪一步会出问题？
4. “忽略状态分布变化”造成的误差是几阶的？给出理由。为什么这个理由依赖于优势函数而非 Q 函数？
5. 为什么 KL 约束要展开到二阶，而目标只需一阶？如果反过来做会发生什么？
6. Fisher 信息矩阵为什么一定半正定？它什么时候奇异？工程上如何处理？
7. 推导 $\Delta\theta^\star = \sqrt{2\delta / (g^\top F^{-1} g)} F^{-1} g$，并说明为何约束一定取等号。
8. TRPO 中的线性搜索在补救什么误差？去掉它会怎样？
9. PPO 的截断是否给出了 KL 的上界？若不是，实践中靠什么防止策略跑偏？
10. 若把 γ 从 0.99 改成 0.999，定理 5.2 中的常数 C 变化多少倍？这说明了什么？

参考文献

Kakade & Langford, *Approximately Optimal Approximate Reinforcement Learning*, ICML 2002.
Kakade, *A Natural Policy Gradient*, NIPS 2001.
Amari, *Natural Gradient Works Efficiently in Learning*, Neural Computation 1998.
Schulman et al., *Trust Region Policy Optimization*, ICML 2015.
Schulman et al., *High-Dimensional Continuous Control Using Generalized Advantage Estimation*, ICLR 2016.
Schulman et al., *Proximal Policy Optimization Algorithms*, arXiv:1707.06347, 2017.
Achiam et al., *Constrained Policy Optimization*, ICML 2017.
Nota & Thomas, *Is the Policy Gradient a Gradient?*, AAMAS 2020.