目标更新频率在 Q 学习中的作用 g

Simon Weissmann[1], Tilman Aach[1],Benedikt Wille[1], Sebastian Kassing[2], 和 Leif D¨oring[1]

[1]曼海姆大学数学研究所，德国曼海姆 68138，{simon.weissmann, tilman.aach, benedikt.wille, leif.doering}@uni-mannheim.de[2]伍珀塔尔大学数学与信息学系，德国伍珀塔尔 42119，kassing@uni-wuppertal.de

arXiv:2602.03911v1 [cs.LG] 3 Feb 2026

摘要

目标网络更新频率（目标更新频率）是（深度）Q 学习中的核心稳定机制。然而，人们对如何选择目标更新频率仍缺乏充分认识，往往仅将其视为另一个可调超参数，而非一项有原则的设计决策。本文从近似动态规划的角度，对表格型 Q-learning 中的目标固定进行了理论分析。我们将周期性目标更新表述为一种嵌套优化方案，其中每次外层迭代都会应用一个不精确的贝尔曼最优算子，并由通用内循环优化器近似求得。严谨的理论结果为异步采样情形提供了有限时间收敛性分析，并将内循环具体化为随机梯度下降。我们的结果明确刻画了目标更新周期所导致的偏差—方差权衡，并展示了如何为这一关键超参数设置最优值。我们证明，恒定的目标更新计划并非最优，会带来样本复杂度的对数级额外开销，而通过自适应计划可以完全避免这种开销。分析表明，最优目标更新频率会随着学习过程的推进而呈几何级数增长。

1 引言

强化学习（RL）是一个强大的序贯决策框架，使智能体能够通过与环境交互来学习最优行为 [26]。许多成功的强化学习算法都植根于动态规划（DP），它利用价值函数的递归结构来求解马尔可夫决策过程（MDP） [22]。然而，经典动态规划要求显式掌握转移动态，在大型或连续状态空间中也难以处理，因此催生了可扩展至现实问题的近似动态规划（ADP）方法。基于价值的近似动态规划方法是强化学习算法的核心类别之一，旨在学习最优价值函数，并据此推断最优策略。在近似动态规划方法中，Q 学习的变体 [30] 尤为重要，它提供了一种无模型方法，通过迭代应用贝尔曼最优算子T^*来学习最优动作价值函数。它的吸引力在于简单：时序差分更新通过自举估计价值，无需显式的环境模型。尽管 Q 学习在表格情形下具有收敛保证，但在函数逼近（尤其是使用神经网络）时并不稳定，这是由近似、自举和离策略学习构成的致命三元组所致。深度 Q 网络（DQN）通过经验回放和目标网络 [20]缓解了这些问题。目标网络是一个定期固定的

Q 网络副本，有助于使目标值与当前参数去相关。在表格形式下，固定目标 Q 学习定义为

$$Q(s,a) \leftarrow Q(s,a) + \alpha\big(r + \gamma \max_{a'} Q^{-}(s',a') - Q(s,a)\big),$$

其中，固定目标矩阵 Q^{-} 会定期被 Q 覆写，例如每隔 $K = 1000$ 步覆写一次。我们将目标更新频率 K 简称为 TUF。从近似动态规划的角度看，固定目标会改变对贝尔曼最优算子的近似方式。标准 Q 学习使用一次随机梯度下降（SGD）更新来最小化贝尔曼误差的平方，而参数为 K 的 TUF 则对应于在同一目标函数上连续执行 K 次 SGD 迭代（见下文第 3 节）。因此，我们将 TUF 的选择视为一个多层次设计问题，并探讨选择固定 TUF 是否最优。

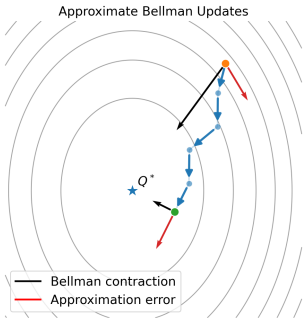
能否通过系统地改变 当前估计与自举目标之间的时间间隔，改善 Q 学习的收敛性？

我们的答案是肯定的。训练过程中，TUF 应按几何级数递增；保持 K 不变并非最优。在讨论理论之前，我们先来看图 1 所示的两个示例实验。我们考察一个精细设计的表格型网格世界示例和 Lunar Lander。对于前者，我们采用带目标固定的表格型 Q-learning（步长恒定的标准 Q-learning 对应 TUF=1）。对于后者，我们使用深度 Q 网络，并以随机梯度下降（SGD）替代 ADAM，以便与我们的理论结果保持一致。更多细节见附录 A。
图中表明，较短的目标更新频率有时会完全失效，中等频率能够快速学习，但会使学习饱和，而较大的频率学习速度较慢，却能够成功完成学习。深度 Q 学习的相关研究也报告了类似的性能对目标更新频率选择的敏感性 [15, 13, 25, 2]。

从近似动态规划的角度来看，贝尔曼算子的压缩性质提供了一个直观解释，第四节将对此进行严格论证：

（i）在早期的贝尔曼迭代中，当当前价值估计值远离最优不动点 Q^*时， T^* 的压缩性会推动估计值朝正确方向移动，即使 T^*存在较大的近似误差。

（ii）随着学习的推进，当当前估计值 Q接近最优 Q^*时，近似误差的容许空间会缩小。要持续收敛，就需要对 T^* 进行越来越精确的近似，以避免近似误差占据主导地位。



本文的理论分析表明，如何将这一直觉迁移到目标更新频率的选择中
频率。以下是我们的主要理论结果：该结果表明，几何递增的目标更新频率能够严格改善带有延迟目标更新的表格型 Q-learning 的样本复杂度中的误差依赖关系。第 5 节给出了一个更详细的版本，其探索条件限制更少。

定理 1.1（非正式）。 假设奖励有界，且状态-动作访问概率有下界 $\xi > 0$。若要通过 t 次更新达到准确率 $\mathbb{E}[\|Q_t - Q^*\|_\infty] < \varepsilon$，则需要以下数量的样本：

(i) 采用最优常数目标更新频率时，达到准确率 ε 所需的样本复杂度为

$$\mathcal{O}\Big(\frac{\log((1-\gamma)^{-1}\varepsilon^{-1})}{\xi^2(1-\gamma)^4\varepsilon^2\log(2(1+\gamma)^{-1})}\Big).$$

(ii) 当目标更新频率以几何级数（阶数为 $\gamma^{-\frac{2}{3}n}$）递增时，可达到准确率 ε，且样本复杂度为

$$\mathcal{O}\Big(\frac{1}{\xi^2(1-\gamma)^5\varepsilon^2}\Big).$$

$\mathcal{O}$ 记号表示关于 ε 的渐近关系，且不涉及 $|\mathcal{S}|$、$|\mathcal{A}|$ 和 γ 中的其他常数 .

该定理表明，若目标更新频率按 $\gamma^{-\frac{2}{3}n}$ 的方式递增，样本复杂度便能得到可证明的改善。注意，随着 $\gamma\to 1$，$\log(2(1+\gamma)^{-1})^{-1}\sim 2(1-\gamma)^{-1}$。假设采用满足 $\xi=\frac{1}{|\mathcal{S}||\mathcal{A}|}$ 的（理论上的）均匀状态–动作采样分布，则样本复杂度中的主导常数与 $|\mathcal{S}||\mathcal{A}|$ 呈二次关系。

虽然 Q 学习的复杂度结果是最坏情形下的结果（对任意马尔可夫决策过程都成立），而对数级ε改进看似微不足道，但实际情况更有意思。图 1 最下排绘出了采用递增目标更新频率时的平滑学习曲线。结果表明，训练的稳定性显著提升。增大目标更新频率有助于较小的初始目标更新频率，但不会损害较大的初始目标更新频率。

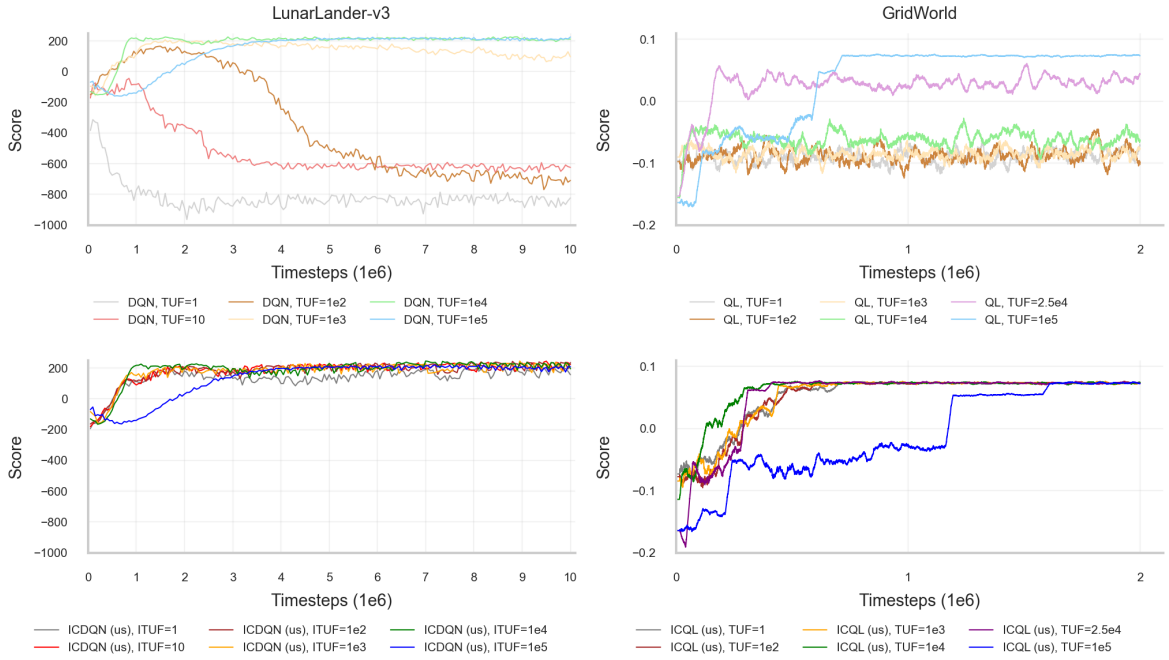


图 1：上排：Lunar Lander（算法：深度 Q 网络，采用不同的固定目标更新频率 (TUF)，使用随机梯度下降，并采用自定义学习率调度方案，在每个目标冻结区间内从 0.01 线性递减至 0.0001）和附录 A.2 中的网格世界（算法：采用目标冻结的 Q 学习，以及由我们的理论指导的学习率 $1/(1+k/(2|\mathcal{S}||\mathcal{A}|))$，在每个目标冻结区间内以 $|\mathcal{S}||\mathcal{A}|=52$ 进行调整）。下排：相同的环境和算法，但目标更新频率递增（ITUF），其初始值设为上排中的固定目标更新频率。

备注 1.2。Q 值高估是 Q 学习中一个众所周知的问题。标准 Q 学习（$TUF=1$）无法收敛的原因，是由于步长常数较大，初期出现了严重高估（见下文图4）。从这个角度看，优化目标更新频率可视为通过更精确地近似贝尔曼算子，进一步帮助减少高估。

本文组织如下。我们首先在第2节讨论相关工作。第3节汇总基本符号，并说明目标值定期更新的 Q 学习的循环结构，这种方法有时称为周期性Q学习。周期性Q学习通过内层随机梯度下降近似 T^*Q_n来运行外层值迭代$Q_{n+1} = T^*Q_n$。在第 4 节，我们分析内层循环和外层循环中累积的误差。第 5 节将二者结合起来，推导出最优目标更新频率。第 6 节总结我们的分析，并展望未来研究方向。更多技术细节和补充实验见附录。

2 相关工作

Q 学习及相关算法的收敛性质已得到广泛研究，早期研究便已证明其在异步随机逼近下的收敛性 [27]。复杂度结果举例如下： [3], 开创性的上界$_{2\ln(1/\varepsilon)}\tilde{\mathcal{O}}(^{|\mathcal{A}|\,|\mathcal{S}|}{}_{\varepsilon^2}{}^{1-\gamma})$、 [9], 的通用下界 $\Omega(^{|\mathcal{A}|\,|\mathcal{S}|}_{\varepsilon^2})$, 以及 $\tilde{\mathcal{O}}(\ ^{|\mathcal{A}|\,|\mathcal{S}|}_{\varepsilon^4})$复杂度的分阶段Q学习 [17]。与我们工作最接近的是 [18], ，作者研究了采用恒定 TUE 的表格型Q 学习，旨在阐明深度 Q 网络的目标网络冻结机制。他们推导出了样本复杂度$\tilde{\mathcal{O}}($ $^{|\mathcal{A}|\,|\mathcal{S}|}{}_{\varepsilon^2}{}^{3}$)。具体而言，对于 $\mathbb{E}[|Q_t - Q^*|_\infty] < \varepsilon$，若要通过最优的均匀状态-动作探索达到 ε准确率，则至少需要 $_{\varepsilon^2}{}^{2048}\ {}^{|\mathcal{A}||\mathcal{S}|\ 3}_{4\quad\ ()(1-\gamma)\ln(\gamma^{-1})}\ln\left(\ {}^{4}_{1-\gamma()\varepsilon}\right)$ 个样本。相比之下，我们的细致分析将对 $|\mathcal{S}||\mathcal{A}|$的依赖在指数上降低了一个阶。此外，几何递增的目标更新频率消除了对数级的 ε依赖。

我们将目标冻结解释为一种嵌套优化方案，这与经典的近似动态规划（ADP）密切相关；在近似动态规划中，贝尔曼算子的应用并不精确。人们已在价值迭代和策略迭代中对近似贝尔曼更新下的误差传播进行了广泛研究 [5, 21, 1]。由于目标网络的提出晚于近似动态规划领域的大多数研究成果，我们的分析无法纳入标准结果之中也就不足为奇了。具体而言，我们的内循环分析是专为随机梯度下降优化量身定制的。
目标网络最初是在深度 Q 网络中引入的，目的是稳定函数逼近下的学习 [20]，此后又被改进并应用于连续控制和演员-评论家方法 [19, 11]。 [10] 在假设每次迭代都精确求解底层贝尔曼回归问题的前提下，对深度 Q 网络进行了统计分析。利用双重 Q 学习等方法缓解过高估计偏差的做法可追溯至 [28] ，而将其与深度学习场景中的目标网络冻结相结合则可追溯至[29]。虽然我们没有针对这两个主题的交叉领域提供理论分析，但进一步研究这一方向或许很有意义。

我们的循环视角和不断提高的目标更新频率，与多时间尺度随机逼近有关 [6], ，其中不同组件以不同速度演化。近期，快慢系统框架已应用于监督学习中的 ADAM 优化器 [7, 8]。对双时间尺度算法的经典分析通常是在时间尺度完全

分离的渐近状态下进行。相比之下，我们基于有限时间误差界推导出一种显式且近乎最优的增长计划。

3 P预备知识：从基础到目标更新频率

基础知识：对于理论结果，我们考虑由元组 $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ 定义的马尔可夫决策过程，其中$\mathcal{S}$ 是有限状态空间， $\mathcal{A}$ 是有限动作空间， $P\colon \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ 是状态转移概率函数， $r\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ 是奖励函数，$\gamma\in[0,1)$ 是折扣因子。策略 $\pi\colon \mathcal{S}\to[0,1]^{\mathcal{A}}$将状态映射为所有动作上的概率分布，其目标是找到一个最优策略π^*，使期望折扣奖励最大化。策略 π的动作价值函数 $Q^\pi\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$定义为 $Q^\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) \mid s_0=s, a_0=a, a_t\sim\pi(s_t)]$。最优动作价值函数 $Q^*(s,a) = \max_\pi Q^\pi(s,a)$ 可表征为贝尔曼最优算子 $T^*\colon \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}\to\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ 的唯一不动点，该算子定义如下：

$$(T^*Q)(s,a) = \mathbb{E}_{s',r}\big[r(s,a) + \gamma\max_{a'} Q(s',a')\big].$$

算子 T^* 是上确界范数下的 γ压缩映射，因此反复应用$Q_{n+1} = T^*Q_n$ 会收敛到其唯一的不动点 Q^*。Q^* 之所以重要，是因为与 Q^* 对应的贪心策略是最优的。背景理论请参阅 [22]。

Q 学习、深度 Q 网络和目标冻结：在无模型设定中，这个问题会变得更具挑战性，因为此时只能获取 T^* 的估计值；在非表格情形中， $|\mathcal{S}|$ 和/或 $|\mathcal{A}|$ 也可能很大甚至无穷。近似动态规划通过近似 T^*并使用参数化函数逼近器 Q_θ来应对这一挑战： $\mathcal{S}\times\mathcal{A}\to\mathbb{R}$，通常涉及神经网络，即所谓的 Q 网络。在这种情况下，参数 $\theta\in\mathbb{R}^d$是神经网络中可学习的权重和偏置。完整的表格参数化$Q_\theta(s,\ a) = \theta_{s,\ a}$ （其中 $d = |\mathcal{S}|\cdot|\mathcal{A}$ ）将表格情形视为一种特殊情况。ADP 方法并不精确计算 T^*Q_θ，而是通过最小化损失函数来近似贝尔曼算子。一般更新使用均方贝尔曼误差（MSBE）：

$$\mathcal{L}(\theta)(s,a) = \frac{1}{2}\mathbb{E}_{s',r}\big[\big(Q_\theta(s,a) - \tau\big)^2\big],$$

回归目标为 $\tau = r(s,\ a) + \gamma\max_{a'} Q_\theta(s',\ a')$。随后使用优化器沿着降低均方贝尔曼误差的方向迭代更新参数。例如，标准 Q 学习

$$Q(s,a) \leftarrow Q(s,a) + \alpha\big(r + \gamma\max_{a'} Q(s',a') - Q(s,a)\big),$$

是使用表格型参数化并采用步长为α 的随机梯度下降，仅执行一步最小化时对均方贝尔曼误差进行最小化的特殊情况。众所周知，通过优化均方贝尔曼误差来学习存在严重问题，参见例如 [23]。Q网络往往会对当前的状态-动作对过拟合，且几乎无法泛化。为稳定学习，深度 Q 网络（DQN）[20] 引入了两项关键改进：经验回放，将状态转移存储在缓冲区中 $\mathcal{D}$；以及一个参数固定的目标网络θ^-。修改后的损失函数为

$$\overline{\mathcal{L}}(\theta) = \frac{1}{2}\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\big[\,(Q_\theta(s,a) - \tau)^2\,\big], \tag{1}$$

其中 $\tau = r(s,\ a) + \gamma$最大值$_{a'}\, Q_{\theta^-}(s',\ a')$。目标参数 θ^- 每隔 K 步使用 θ 进行更新。对于表格型 Q 学习，目标冻结会导致使用两个矩阵并进行如下更新

$$Q(s,a) \leftarrow Q(s,a) + \alpha\big(r + \gamma \max_{a'} Q^-(s',a') - Q(s,a)\big),$$

目标矩阵 Q^- 每隔 K 步便由 Q 覆盖。

目标冻结的循环视角：在我们的分析中，我们将周期性目标冻结视为一种循环方法：通过外层循环进行值迭代（贝尔曼更新），并通过内循环利用随机梯度下降（目标值冻结时的 Q 学习更新）近似值迭代结果 T^*Q_n，相关内容总结于伪代码（算法 1）中。我们用 K_n 表示外层循环第 n 步所使用的内层优化步数，并比较以下方案：仅使用一个内层优化步骤 $K_n = 1$（Q 学习）、使用固定的 $K_n = K$ 步（深度 Q 网络），以及本文提出的按几何级数递增的优化步数 K_n（增长率为 $\gamma^{-\frac{2}{3}n}$）。由于 K_n与深度 Q 网络的目标更新频率密切相关，我们将 K_n 称为目标更新频率（目标更新频率）。回顾统计学中的一个简单事实：对于 $X \in L^2(\Omega\ \mathcal{F},\ \mathbb{P})$，标量 $\mathbb{E}[X]$ 是二次风险的唯一极小值

$$\mathbb{E}[X] \ = \ \arg\min_{\theta\in\mathbb{R}} \ \mathbb{E}\big[(X-\theta)^2\big]\,. \tag{2}$$

因此，可以通过对损失 $f(\theta) := \mathbb{E}[(X-\theta)^2]$ 应用任意（随机）优化算法来计算期望。在第 n 次贝尔曼更新近似开始时，算法取一个外层迭代值 Q_n，并在内循环求解回归问题时将其固定。状态-动作对 $(s_{n,k},a_{n,k})$ 根据某个（可能依赖于历史记录的）分布 $\pi_{n,k}$ 进行采样。对于任意状态-动作对 (s,a)，定义随机变量$X_n(s,a) := r + \gamma \max_{a'\in\mathcal{A}} Q_n(s',a')$，其中 $(r,s') \sim p(\cdot \mid s,a)$，使得 $(T^*Q_n)(s,a) = \mathbb{E}[X_n(s,a)]$。根据 (2)，我们可得到等价的回归表征

$$(T^*Q_n)(s,a) = \arg\min_{\theta\in\mathbb{R}} \ \frac{1}{2}\,\mathbb{E}\big[\big(X_n(s,a)-\theta\big)^2\big]\,. \tag{3}$$

为求解 (3)，使用模拟器生成的样本运行随机优化器 K_n 步。由此得到一系列迭代值 $(Q_n^{(k)})_{k=0}^{K_n}$，其满足

$$Q_n^{(k+1)} = \Psi_n^{(k)}(Q_n^{(k)})\,, \tag{4}$$

其中，$\Psi_n^{(k)}$：$\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ 表示由第 k 次内层迭代中的随机优化器所诱导的（随机）可测更新算子。$\Psi_n^{(k)}$中的随机性源于第 k 次迭代所用的模拟器样本。这种一般性还允许我们对异步更新方案进行建模，具体实例将在第 4.2 节中给出。我们用一个随机算子 $\widetilde{T}_n$：$\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ 来表示整个内循环的作用，该算子通过以下方式隐式定义：

$$\widetilde{T}_n Q_n^{(0)} \ := \ Q_n^{(K_n)}. \tag{5}$$

$\widetilde{T}_n Q_n^{(0)}$作为 T^*Q_n 的近似。随后，下一次外层迭代初始化为 $Q_{n+1}^{(0)} := Q^{(K_n)n}$.在接下来的分析中，令 $(\mathcal{F}_{n,k})_{n\ge0,\,0\le k\le K_n}$ 表示截至第 n 次外层循环的第 k 次内层迭代为止，由所有样本和内部随机性生成的滤子。

算法 1 周期性 Q-学习（表格型），可变目标更新频率

输入：目标更新频率 $\{K_n\}_{n=1}^{N}$ **初始化** Q **：** $\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ **任意设置，** $Q^{-}=Q$ **对于从** $n=1$ **到** N **的外层贝尔曼迭代，执行 对于从** $k=1$ **到** K_n **的内层迭代，执行 观测当前状态** s **选择，例如** $a\sim\varepsilon$**-贪心(**$Q(s,\cdot)$**)** **执行动作** a**，观测** r, s' **计算** $\delta=r+\gamma\max_{a'\in\mathcal{A}}Q^{-}(s',a')$ **更新** $Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha\,\delta$ **内层迭代结束** $Q^{-}\leftarrow Q$ **（更新延迟目标值）外层贝尔曼迭代结束**

4 摘要收敛性分析

4.1 外循环分析

我们首先给出外层循环的一般收敛性结果。为此，只需量化内循环对贝尔曼算子的近似精度。我们定义（条件）算子近似误差

$$\eta_n \;:=\; \mathbb{E}\Big[\big\|\widetilde{T}_n Q_n^{(0)} - T^{*}Q_n^{(0)}\big\|_{\infty}\;\Big|\;\mathcal{F}_{n,0}\Big], \tag{6}$$

这将成为我们内循环误差分析的主要对象。给定这些近似误差 (η_n)，我们推导出外层循环误差的递归估计。重要的是，本节的分析仅基于算子近似误差序列 (η_n)，不对内循环中均方贝尔曼误差的最小化方式作任何假设。这是一个通用框架，可应用于任何均方贝尔曼误差优化方法（例如随机梯度下降、ADAM、RMSProp）。

命题 4.1 (外层循环的近似压缩)。 设 T^{*} 为 $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ 上的贝尔曼最优算子， Q^{*} 表示其唯一不动点。设 $\{Q_n^{(0)}\}_{n\ge0}$ 由(4)–(5)生成，并定义

$$E_n \;:=\; \|Q_n^{(0)} - Q^{*}\|_{\infty}. \tag{7}$$

那么，对所有 $n\ge0$,

$$\mathbb{E}[E_{n+1}\,|\,\mathcal{F}_{n,0}] \;\le\; \gamma\,E_n + \eta_n, \tag{8}$$

其中 η_n 由(6)定义。此外，若 $\sum_{n=0}^{\infty}\eta_n<\infty$ 几乎必然成立，则 $E_n\to0$ 也几乎必然成立，其中 $n\to\infty$。

证明。 利用不动点恒等式 $T^{*}Q^{*}=Q^{*}$、三角不等式以及压缩性，可得

$$\begin{aligned}E_{n+1}=\|Q_{n+1}^{(0)}-Q^{*}\|_{\infty}=\|\tilde{T}_n^{(K_n)}Q_n^{(0)}-T^{*}Q^{*}\|_{\infty}&\le\|\tilde{T}_n^{(K_n)}Q_n^{(0)}-T^{*}Q_n^{(0)}\|_{\infty}+\|T^{*}Q_n^{(0)}-T^{*}Q^{*}\|_{\infty}\\&\le\|\tilde{T}_n^{(K_n)}Q_n^{(0)}-T^{*}Q_n^{(0)}\|_{\infty}+\gamma\|Q_n^{(0)}-Q^{*}\|_{\infty}.\end{aligned}$$

对给定$\mathcal{F}_{n,0}$ 取条件期望，得到 (8)。注意 (E_n) 与 (η_n) 均为非负，且均为 $(\mathcal{F}_{n,0})$-适应的。对于几乎必然收敛这一结论，将下文推论 4.2 所给出的 Robbins–Siegmund 定理的一种改编形式应用于非负适应过程 $\{E_n\}_{n\geq 0}$，并利用 (8) 以及几乎必然成立的假设 $\sum_{n\geq 0}\eta_n < \infty$ 。 □

上述证明依赖于对 Robbins–Siegmund 定理的一项重要改造 [24], ；该定理是分析随机逼近方法的经典工具；例如，参见 [16] 中的推论 B.7。

推论 4.2。 设$(\Omega\ \mathcal{A}, \mathcal{F}, \mathbb{P})$ 为一个滤过概率空间，$(Z_k)_{k\in\mathbb{N}}$、$(A_k)_{k\in\mathbb{N}}$、$(B_k)_{k\in\mathbb{N}}$ 以及$(D_k)_{k\in\mathbb{N}}$ 为非负且关于 $\mathcal{F}$ 适应的随机过程，并满足

$$\sum_{k=0}^{\infty} A_k < \infty, \quad \sum_{k=0}^{\infty} B_k < \infty \quad and \quad \sum_{k=0}^{\infty} D_k = \infty$$

几乎必然成立。此外，假设

$$\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \leq Z_k(1 + A_k - D_k) + B_k.$$

那么，对于 $k \to \infty$， Z_k 几乎必然收敛至0 。

有了命题 4.1，我们便可清楚地区分压缩与准确率的影响:

(i) 外循环压缩 仅由 γ决定；(ii) 内循环近似误差 η_n，它取决于采样过程、目标更新频率 K_n 以及所选的优化器。

4.2 内循环：随机梯度下降迭代的分析

本节将第 4.1 节的抽象框架具体化为均方贝尔曼误差的内循环优化器采用随机梯度下降的情形。我们推导非渐近误差界，以衡量在异步采样下每个内循环近似贝尔曼更新的准确度。

假设 4.3。 我们作出以下假设:

- 持续探索：每一步中，每个状态-动作对被选中的概率至少为$\xi > 0$。状态-动作探索分布 $\pi_{n,k}$ 可以依赖于历史记录。
- 奖励方差由某个 $\sigma^2 > 0$ 所界定。

注 4.4。 举例来说，读者可以想到 ε-贪心或均匀探索。可以用一些较弱的采样条件替代持续探索假设。例如，可以假设平均而言，每个状态-动作对在有限时间窗口内至少被访问一次；这一假设可用于分析异步随机逼近方案 *[4]*。与仅要求每个状态–动作对被无限次更新（例如 *[32],* 中所考虑的情况）相比，这一条件更强，但弱于均匀状态-动作访问。

固定索引 n 和目标更新频率 K_n。如前所述，贝尔曼目标在内循环开始时固定，并定义为 $Q_n^{(0)}$。在内层迭代 k 中，从 $\pi_{n,k}$ 中采样一个状态-动作对 $(s_{n,k},a_{n,k})$，并异步更新表格型 Q 矩阵，如下所示

$$\begin{aligned} Q_n^{(k+1)}(s,a) &= \Psi_n^{(k)}(Q_n^{(k)}, s_{n,k}, a_{n,k})(s,a) \\ &:= Q_n^{(k)}(s,a) + \mathbb{1}_{\{(s,a)=(s_{n,k},a_{n,k})\}} \alpha_n^{(k)} ((\hat{T}Q_n^{(0)})(s,a) - Q_n^{(k)}(s,a)), \end{aligned}$$

其中，对于 $(r_{n,k},s'_{n,k})$，$\sim p(\cdot \mid s_{n,k},a_{n,k})$

$$(\hat{T}Q_n^{(0)})(s,a) = r_{n,k} + \gamma \max_{a' \in \mathcal{A}} Q_n^{(0)}(s'_{n,k}, a')$$

是在给定 $(s_{n,k}, a_{n,k})$ 条件下，贝尔曼最优算子 $(T^*Q_n^{(0)})(s,a)$ 的无偏估计量 $=(s,a)$。完成 K_n 个内层步骤后，下一个外层迭代设为 $Q_{n+1}^{(0)} := Q^{(K_n)n}$.我们定义迭代内循环近似误差

$$e_n^{(k)} := \mathbb{E}\big[\|Q_n^{(k)} - T^*Q_n^{(0)}\|_2^2 \mid \mathcal{F}_{n,0}\big] .$$

为获得显式的收敛速率，我们现在将步长调度限定为 $\frac{1}{k}$，这是二次目标函数上随机梯度下降的常用调度方式，并且可实现最快的收敛速率 [14]。

定理 4.5 （内循环收敛速率）。 假设假设 *4.3* 成立，令 $s := \frac{2}{\xi}$ 并选择$\alpha_n^{(k)} = \xi^{2}{}_{(k+s)}$ · 那么，对于所有~~$k \geq 0$~~,

$$e_n^{(k+1)} \;\leq\; \frac{c_1 \|Q_n^{(0)} - Q^*\|_\infty^2 + c_2}{k+s} , \tag{9}$$

其中

$$c_1 := \Big(\tfrac{2}{\xi} + 1\Big)|\mathcal{S}||\mathcal{A}|(1+\gamma)^2 + \Big(\tfrac{16}{\xi^2} + \tfrac{8}{\xi}\Big)\gamma^2, \quad c_2 := \Big(\tfrac{8}{\xi^2} + \tfrac{4}{\xi}\Big)\big(\sigma^2 + 2\gamma^2\|Q^*\|_\infty^2\big) .$$

完整证明推迟至附录 B.1。在命题 B.1 中，我们首先采用标准的随机逼近论证，为内循环误差推导出如下形式的单步递推

$$e_n^{(k+1)} \leq (1 - \xi\alpha_n^{(k)})e_n^{(k)} + (\alpha_n^{(k)})^2(2\sigma^2 + 4\gamma^2 e_n^{(0)} + 4\gamma^2\|Q^*\|_\infty^2), \tag{10}$$

对任意步长 $\alpha_n^{(k)} \in (0,1]$ 均成立。主要技术难点在于控制异步更新机制，在该机制下每次迭代仅更新单个状态–动作对。这导致阶为 $(1 - \xi\alpha_n^{(k)})$ 的有效压缩，反映了最小访问概率 ξ。第二步，我们用归纳法证明，对于步长选择 $\alpha_n^{(k)} = \xi^{2}{}_{(k+s)}$，递推关系 (10) 给出了 (9) 中所述的内循环收敛界。

4.3 几乎必然收敛

定理 4　　.5 表明，条件贝尔曼算子近似误差的表现如　　s

$$\eta_n \le \frac{\sqrt{c_1}E_n + \sqrt{c_2}}{\sqrt{K_n}}\,.$$

结合命题 4.1，我们可推得

$$\mathbb{E}\big[E_{n+1} \mid \mathcal{F}_{n,0}\big] \le \gamma E_n + \frac{\sqrt{c_1}\,E_n + \sqrt{c_2}}{\sqrt{K_n}}\,. \tag{11}$$

如果内循环误差可求和，则外层循环的压缩效应占主导，从而得到收敛。下一条结果的证明推迟至附录 B.2。

推论 4.6。 假设 K_n 满足 $\sum_{n=0}^{\infty} \frac{1}{\sqrt{K_n}} < \infty$。那么，在定理 *4.5* 的设定下，当 $n \to \infty$ 时，我们几乎必然有 $\|Q_n^{(0)} - Q^*\|_\infty \to 0$ 。

5 最优 TUF 设计

在本节中，我们探讨如何选择目标更新频率（K_n），以尽可能降低达到特定目标精度所需的总计算成本。这项任务可视为一个多层优化问题，其灵感源自多层蒙特卡洛方法 [12] ，并且最近应用于 [31], 中的逆问题；其证明技巧与我们的分析密切相关。下文假设，获取迭代点 $Q_n^{(0)}$的计算成本与采样状态—动作对的数量成正比。具体而言，我们将成本定义为

$$\mathrm{cost}(Q_n^{(0)}) = \mathrm{cost}(Q_{n-1}^{(K_{n-1})}) := \sum_{j=0}^{n-1} K_j$$

我们将其称为 $Q_n^{(0)}$的样本复杂度。给定精度 $\varepsilon > 0$后，我们的总体目标是选择 N 和（K_n）$_{n=0}^{N-1}$，以求解约束优化问题

$$\min_{N,(K_n)_{n=0}^{N-1}} \ \mathrm{cost}(Q_N^{(0)}) \quad \text{s.t.} \quad e_N := \mathbb{E}[E_N] \le \varepsilon\,. \tag{12}$$

为此，我们引入有效压缩因子 $\mu \in$ （γ，1），并要求目标更新频率足够大，使得对所有 $n \in \mathbb{N}_0$,

$$0 < \gamma + \sqrt{\frac{c_1}{K_n}} \le \mu < 1\,. \tag{13}$$

更准确地说，我们要求 $\min_n K_n \ge K_{\min}(\mu) :=$ $\frac{c_1}{(\mu-\gamma)^2}$。因此，对方程 **(11)** 两边取期望，并对 $n \in \mathbb{N}$ 进行迭代，可得

$$e_n \le \mu^n e_0 + \sum_{j=0}^{n-1} \mu^{n-1-j} \sqrt{\frac{c_2}{K_j}}, \quad n \in \mathbb{N}. \tag{14}$$

在后续分析中，我们将通过式 (14) 分解并要求满足以下条件，以强制实现 $e_N \le \varepsilon$

$$\mu^N e_0 \le \frac{\varepsilon}{2} \quad \text{and} \quad \sum_{j=0}^{N-1} \mu^{N-1-j} \sqrt{\frac{c_2}{K_j}} \le \frac{\varepsilon}{2}$$

这将外循环的数量固定为 $N(\varepsilon) \ge \left\lceil \frac{\log(\varepsilon/(2e_0))}{\log(\mu)} \right\rceil$，并将设计问题简化为选择 $K_j(\varepsilon)$ $j = 0, \ldots,$ $N(\varepsilon) - 1$。因此，我们考虑所谓的准最优（固定/可变）TUF（见定义 C.1），当 $\varepsilon \to 0$ 时无法在渐近意义上进一步改进（例如，通过不对称地拆分式 (14) 中的误差项）。为简单起见，我们允许 TUF（K_j）取实数值。

固定 TUF。 我们首先考虑最简单的设计，即在各个内循环中将内层随机梯度下降步数保持不变。在这种情况下，我们最终需要在约束 $e_N \le \varepsilon$ 下最小化 $\mathsf{cost}(Q_N^{(0)}) = NK$。下面的定理量化了外层迭代次数与每个内循环成本之间由此产生的权衡。完整证明见附录C。

定理 5.1 (固定 TUF)。设定 $\mu = \mu(\gamma) = \frac{1+\gamma}{2}$。那么，选择

$$N(\varepsilon) = \left\lceil \frac{\log(\frac{\varepsilon}{2e_0})}{\log(\mu)} \right\rceil, \quad K(\varepsilon) = \frac{4c_2}{\varepsilon^2} \cdot \Big(\frac{1-\mu^{N(\varepsilon)}}{1-\mu} \Big)^2$$

定义了一个准最优的固定 *TUF*。如果奖励由1 界定，则有

$$\mathrm{cost}(Q^{(0)}_{N(\varepsilon)}) = \mathcal{O}\Big(\frac{\log((1-\gamma)^{-1}\varepsilon^{-1})}{(1-\gamma)^4 \xi^2 \log(2(1+\gamma)^{-1})\varepsilon^2} \Big).$$

固定 TUF 的样本复杂度对 ε 呈对数依赖，这是由贝尔曼算子的压缩导致的；同时，对 ε 呈二次依赖，这是由内循环中贝尔曼算子的近似导致的。

注 5.2。 注意，固定 *TUF* 下样本复杂度的主导常数渐近地按

$$\frac{1}{(1-\gamma)^4 \log(2(1+\gamma)^{-1})} \sim \frac{1}{(1-\gamma)^5}$$

缩放为 $\gamma \to 1$。此外，上述结果表明，选择 $N(\varepsilon)$、 $K(\varepsilon)$ 定义了一个ˆ ˆ准最优的固定 *TUF*。证明通过展示任意满足以下条件的对 N、 K 验证了这一点：

$$\mu^{\hat{N}} e_0 + \sqrt{\frac{c_2}{\hat{K}}} \sum_{j=0}^{\hat{N}} \mu^{N-1-j} \le \varepsilon$$

必须满足

$$N(2\varepsilon)K(2\varepsilon) \le \hat{N}\hat{K}\,.$$

因此，在误差界 (14) 下，固定 *TUF* 复杂度在对 $1-\gamma$、 ξ 和 ε 的依赖方面无法进一步改善。

最优 TUF 设计。固定 TUF 将所有外层迭代一视同仁，尽管它们的作用有很大差异。在训练初期，当 e_n 较大时，压缩项 γe_n 主导了动态过程，只需对贝尔曼算子进行粗略近似。相反，在学习后期，迭代点逐渐接近不动点，进展受限于贝尔曼算子的近似准确率。这一观察表明，应在各内循环之间非均匀地分配计算开销：初期只需较短的内循环，而要达到较高准确率，则需要逐步提高近似准确率（因此也需要更长的内循环）。现在，我们通过在固定目标精度 ε 下优化 TUF 序列 (K_n)，将这一思路形式化。

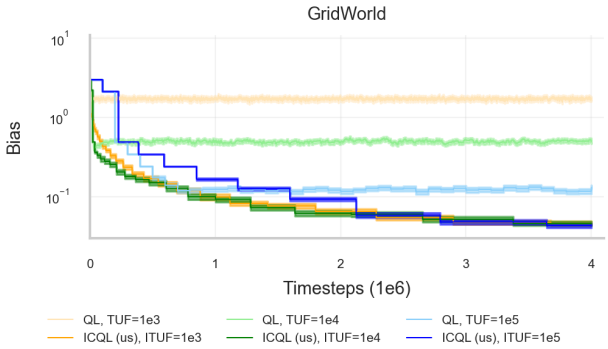


图 2：附录 A.2 中网格世界上的 Q 学习，采用定理 5.3 给出的最优递增目标更新频率（递增目标更新频率），并与针对 $K \in \{1000, 10000, 100000\}$ 采用固定目标更新频率（目标更新频率）的 Q 学习进行比较。阴影带表示 50 个随机种子上的不确定性 95% 置信区间：较小的固定 TUF 会因粗糙的贝尔曼算子近似而导致早期平台期，而较大的固定 TUF 则因过于精确的内循环近似在早期迭代中占主导而收敛缓慢。我们的理论所预测的几何递增调度通过平衡压缩与贝尔曼近似误差避免了这一权衡，从而同时实现快速的初始进展和渐近收敛。

我们的最终定理给出了求解代价最小化问题 (12) 的递增目标更新频率调度的显式构造。我们在附录C中提供了详细证明。

定理 5.3. 设 $\mu = \mu(\gamma) = \frac{1+\gamma}{2}$。则该选择

$$N(\varepsilon) = \left\lceil \frac{\log(\frac{\varepsilon}{2e_0})}{\log(\mu)} \right\rceil, \quad K_j(\varepsilon) = C_\varepsilon \mu^{\frac{2}{3}(N(\varepsilon)-1-j)}, \quad C_\varepsilon = \frac{4c_2}{\varepsilon^2}\left(\frac{1-\mu^{\frac{2}{3}N(\varepsilon)}}{1-\mu^{\frac{2}{3}}}\right)^2, \tag{15}$$

对于 $j = 0, \ldots, N(\varepsilon) - 1$，定义了一个准最优目标更新频率调度。若奖励有界于1，则有

$$\mathrm{cost}(Q^{(0)}_{N(\varepsilon)}) = \mathcal{O}\Big(\frac{1}{\xi^2(1-\gamma)^5\varepsilon^2}\Big).$$

该定理还表明，选择使目标更新频率以速率 $\gamma^{-\frac{2}{3}n}$ 几何增长，可通过去除对数因子 $|\log(\frac{1}{(1-\gamma)\varepsilon})|$来改进样本复杂度。尽管从最坏情况的理论视角来看，这相对于固定目标更新计划仅构成对数级改进，但下文的数值实验表明，由此产生的收益在实践中可能相当显著。

作为该定理的示例，图2将不同的恒定目标更新频率与最优递增目标更新频率进行了比较。该图展示了偏差 $||Q_n - Q^*||_\infty$，其以最优方式快速收敛至零。我们观察到，恒定目标更新频率可以快速降低偏差，但最终会饱和。相比之下，递增的目标更新频率可以同样快速地降低偏差且不会饱和（这是由日益精确的贝尔曼近似所导致的）。

注 5.4. 与固定 *TUF* 情形类似，为证明选择 $N(\varepsilon)$、 $K_j(\varepsilon)$、 $j = \hat{} \hat{} \hat{}1$、. . . 、 $N(\varepsilon) - 1$ 是准最优的，我们推导出任何选择 N、 K_j、 $j = 0$、. . . 、 $N - 1$ 只要满足

$$\mu^{\hat{N}} e_0 + \sum_{j=0}^{\hat{N}-1} \mu^{\hat{N}-1-j} \sqrt{\frac{c_2}{\hat{K}_j}} \leq \varepsilon$$

都必须服从

$$\sum_{j=0}^{N(2\varepsilon)-1} K_j(2\varepsilon) \leq \sum_{j=0}^{\hat{N}-1} \hat{K}_j \,.$$

因此，在误差界(14) 下，可变 *TUF* 复杂度就其对于 $1 - \gamma$、 ξ 和 ε 的依赖而言无法再改进。

为方便读者，我们在附录 A.1 中补充了所得算法的详细表述。我们将该算法称为递增周期 *Q* 学习（ICQL），因为它采用理论上最优的几何增长目标更新频率。递增 TUF 的思想可以很容易地集成到标准深度 Q 网络实现中，正如我们在图 1 的 Lunar Lander 仿真中所做的那样。

6 贡献与未来工作

本文从近似动态规划的原则性视角审视目标网络，将周期性目标固定解释为一种嵌套的贝尔曼近似方案。我们表明，目标更新频率控制着贝尔曼算子近似中的偏差—方差权衡，并支配收敛行为。对于具有固定目标值的表格型 Q-learning，我们在异步采样以及内循环的随机梯度下降更新（即均方贝尔曼误差最小化）下，推导了更精细的非渐近收敛界。我们的分析表明，使用固定 TUF 可被证明是次优的。我们推导出一个显式的几何递增 $\gamma^{-\frac{2}{3}n}$-TUF 调度，它是近最优的。与最佳固定 TUF 选择相比，所得调度消除了样本复杂度中对 ε 的对数依赖。特别地，对于采用最优递增目标更新频率的表格型 Q-learning，我们得到了 $\mathcal{O}((1-\gamma)^{-5}\xi^{-2}\varepsilon^{-2})$ 的样本复杂度界。作为结果的示例，我们考虑了网格世界，以及使用随机梯度下降误差最小化的深度 Q 网络 Lunar Lander（以契合理论）。结果表明，与先前文章的启发式发现一致，提高目标网络更新频率有助于稳定学习过程。

这篇理论文章的动机来自一个具体的实际问题：Q 学习与深度 Q 网络中目标网络更新频率的选择，以及这一设计决策缺乏理论指导。我们的工作为未来研究开辟了若干方向。(i) 为进一步弥合与现代深度强化学习实践之间的差距，有必要超越随机梯度下降，并分析采用 ADAM 等自适应方法的内循环优化。在本工作中，内循环被建模为将随机梯度下降应用于由贝尔曼回归产生的二次目标函数。扩展

到 ADAM 的分析因此需要对其在（随机）二次目标函数上行为的理论认识，而这仍是一个活跃的研究领域。(ii) 本文提出的近似贝尔曼压缩视角自然地表明，可以超越预先确定的TUF调度加以扩展。命题 4.1 中的外循环递推表明，进展并不直接取决于内循环迭代的次数，而取决于由此得到的贝尔曼近似精度 η_n。这一观察启发了精度触发的目标更新：一旦内循环达到规定的逼近精度，便立即更新目标网络，而不是在固定步数之后才更新。具体而言，与其预先固定TUF K_n，不如为 η_n 指定一系列精度阈值，并在达到期望准确率后终止内循环。在表格型设置中，我们概述了一种在内循环期间在线估计 η_n 的实用机制，从而无需额外的预言机访问即可实现此类提前停止规则。该精度触发方案的细节，以及首个说明性的网格世界实验，见附录D。

参考文献

[1] Andr´as Antos、Csaba Szepesv´ari 和 R´emi Munos。基于贝尔曼残差最小化和单条样本路径的拟合策略迭代：学习近似最优策略。*MachineLearning*，71(1):89–129，2008年。[2] Kavosh Asadi、Rasool Fakoor 和 Shoham Sabach。重置深度强化学习中的优化器：一项实证研究。载于第三十七届神经信息处理系统会议，2023年。[3] Mohammad Gheshlaghi Azar、Ian Osband 和 R´emi Munos。强化学习的极小极大遗憾界。国际机器学习会议，第263–272页，2017年。[4] C.L. Beck 和 R. Srikant。常数步长Q学习的误差界。*Systems & ControlLetters*，61(12):1203–1208，2012年。[5] Dimitri P. Bertsekas 和 John N. Tsitsiklis。神经动态规划。，优化与神经计算丛书第3卷。Athena Scientific，1996年。[6]Vivek S. Borkar。双时间尺度随机逼近。*Syst. Control Lett.*，29(5):291–294，1997年2月。[7] Steffen Dereich 和 Arnulf Jentzen。ADAM 优化器的收敛速率。*arXiv*预印本 *arXiv:2407.21078*，2024年。[8] Steffen Dereich、Arnulf Jentzen 和 Sebastian Kassing。Adam算法的常微分方程近似：一般设置与过参数化设置。*arXiv*预印本 *arXiv:2511.04622*，2025年。[9] Eyal Even-Dar 和 Yishay Mansour。Q学习的学习率。*J. Mach. Learn. Res.*，5:1–25，2003年12月。[10] Jianqing Fan、Zhaoran Wang、Yuchen Xie 和 Zhuoran Yang。深度Q学习的理论分析。载于 Alexandre M. Bayen、Ali Jadbabaie、George Pappas、Pablo A. Parrilo、Benjamin Recht、Claire Tomlin 和 Melanie Zeilinger（编），第二届动力学与控制学习会议论文集，第120卷，机器学习研究论文集，第486–489页。PMLR，2020年6月10日至11日。

[11] Scott Fujimoto、Herke van Hoof 和 David Meger。解决演员-评论家方法中的函数逼近误差。载于 Jennifer G. Dy 和 Andreas Krause（编），第*35*届国际机器学习会议，*ICML 2018*，斯德哥尔摩¨，瑞典斯德哥尔摩，*2018*年*7*月*10*日至*15*日，第80卷，机器学习研究论文集，第1582–1591页。PMLR，2018年。[12] Michael B Giles。多层蒙特卡洛路径模拟。运筹学，56(3):607–617，2008年。[13] J. Fernando Hernandez-Garcia 和 Richard S. Sutton。理解多步深度强化学习：对深度 Q 网络目标值的系统研究。*arXiv*预印本，arXiv:1901.07510，2019年。[14] Arnulf Jentzen 和 Philippe von Wurstemberger。随机梯度下降优化算法的误差下界：针对缓慢和快速衰减学习率的尖锐收敛速率。《复杂性杂志》，57:101438，2020年。[15] Weihang Jia、Jinhong Li 和 Yujie Zhao。基于目标值网络参数动态更新的DQN算法。载于*2021*年*IEEE*第四届计算机与通信工程技术国际会议（*CCET*），第285–289页，2021年。[16] Sebastian Kassing、Simon Weissmann 和 Leif D¨oring。控制流动：具有衰减正则化的随机梯度下降的稳定性与收敛性。*arXiv*预印本，arXiv:2505.11434，2025年。[17] Michael Kearns 和 Satinder Singh。Q学习与间接算法的有限样本收敛速率。《神经信息处理进展》，11，1999年4月。[18] Donghwan Lee 和 Niao He。周期性Q学习。载于 Alexandre M. Bayen、Ali Jadbabaie、George Pappas、Pablo A. Parrilo、Benjamin Recht、Claire Tomlin 和 Melanie Zeilinger（编），第二届动力学与控制学习会议论文集，第120卷，机器学习研究论文集，第582–598页。PMLR，2020年6月10日至11日。[19] Timothy P. Lillicrap、Jonathan J. Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。基于深度强化学习的连续控制。载于 Yoshua Bengio 和 Yann LeCun（编），第四届国际学习表征会议，*ICLR 2016*，波多黎各圣胡安，*2016*年*5*月*2*日至*4*日，会议论文集，2016年。[20] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A. Rusu、Joel Veness、Marc G. Bellemare、Alex Graves、Martin Riedmiller、Andreas K. Fidjeland、Georg Ostrovski、Stig Pe- tersen、Charles Beattie、Amir Sadik、Ioannis Antonoglou、Helen King、Dharshan Kumaran、Daan Wierstra、Shane Legg 和 Demis Hassabis。通过深度强化学习实现人类水平控制。《自然》，518(7540):529–533，2015年2月。[21] R´emi Munos。近似值迭代的误差界。载于第*20*届全国人工智能会议论文集——第*2*卷，AAAI’05，第1006–1011页。AAAI出版社，2005年。[22] Martin L Puterman。马尔可夫决策过程：离散随机动态规划。John Wiley & Sons，2014年。

[23] Martin Riedmiller. 神经拟合 Q 迭代——数据高效神经强化学习方法的初步经验. 载于第*16*届欧洲机器学习会议论文集，ECML'05，第317–328页，2005年.[24] H. Robbins与D. Siegmund. 非负几乎超鞅的收敛定理及其若干应用**研究由美国国立卫生研究院资助项目5-r01-gm-16895-03及美国海军研究办公室资助项目n00014-67-a-0108-0018支持. 见Jagdish S. Rustagi编，统计优化方法，第233–257页. 学术出版社，1971年.[25] Miaoping Sun、Zequan Yang、Xunhua Dai、Xiaohong Nian、Hongyun Xiong与Haibo Wang. 一种基于矩估计的深度强化学习目标网络自适应更新方法. 《神经处理快报》，55(5):5515–5537，2023年10月.[26] Richard S. Sutton与Andrew G. Barto. 《强化学习：导论》. 麻省理工学院出版社，第二版，2018年.[27] John N. Tsitsiklis. 异步随机逼近与Q 学习. *Machine Learning*，16(3):185–202，1994年9月.[28] Hado Van Hasselt. 双重 Q 学习. 见J. Lafferty、C. Williams、J. Shawe-Taylor、R. Zemel与A. Culotta编，《神经信息处理系统进展》，第23卷. Curran Associates, Inc.，2010年.[29] Hado Van Hasselt、Arthur Guez与David Silver. 结合双重 Q 学习的深度强化学习. *AAAI* 人工智能会议论文集，30(1)，2016年.[30] Christopher J. C. H. Watkins与Peter Dayan. Q 学习. *Machine Learning*，8(3):279–292，1992年5月.[31] Simon Weissmann、Ashia Wilson与Jakob Zech. 面向逆问题的多层优化. 见Po-Ling Loh与Maxim Raginsky编，第三十五届学习理论会议论文集，机器学习研究论文集第178卷，第5489–5524页. PMLR，2022年7月2–5日.[32] Huizhen Yu与Dimitri P. Bertsekas. 论随机最短路径问题中Q 学习迭代序列的有界性. 《运筹学数学》，38(2):209–227，2013年.

A 算法细节

A.1 递增周期 Q 学习

本节给出在内循环中使用随机梯度下降的具体算法实例。该设定对应于表格情形下的经典周期性 Q 学习。受第4节和第5节精细收敛性分析的启发，我们建议随时间增加两次目标同步之间执行的随机梯度下降更新次数。早期内循环采用较短的目标冻结期，以高效利用贝尔曼算子的压缩性质；而后期内循环则逐步提高目标更新频率，以降低近似误差并避免收敛饱和。

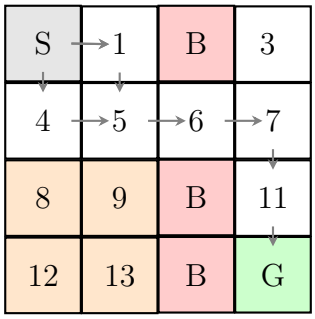
所得算法称为递增周期 *Q* 学习 （ICQL），与标准周期性 Q 学习的唯一区别在于 TUF 的选择（K_n）；这些 TUF 根据定理 5.3 中推导出的理论最优调度按几何方式增长。

算法 2 递增周期 Q 学习（ICQL）

1：输入：初始 Q**-矩阵** $Q_0^{(0)}$**；折扣因子** $\gamma \in$ **(0, 1)；初始 TUF** K_0**；目标更新次数** N**；步长序列** $\alpha_n^{(k)}$**2：输出：** $Q_T^{(0)} \approx Q^*$**3：对于** $n = 0$**，1，...，** $N-1$ **执行 4：设置 TUF** $K_n = \left\lceil K_0\, \gamma^{-\frac{2}{3}n} \right\rceil$**5：初始化内层迭代** $Q_n^{(0)}$**6：对于** $k = 0$**，1，...，** $K_n - 1$ **执行 7：观测当前状态** s**8：选择例如** $a \sim \varepsilon$**-贪心(**$Q_n^{(k)}$**(**$s, \cdot$**)) 9：执行动作** a**，观测** r, s'**10：计算贝尔曼目标：** $\tau \leftarrow r + \gamma \max_{a'} Q_n^{(0)}$**(**$s', a'$**) 11：更新** $Q_n^{(k+1)}$**(**s, a**)** $\leftarrow Q_n^{(k)}$**(**s, a**)** $+ \alpha_n^{(k)} \big(\tau - Q_n^{(k)}$**(**$s, a$**)**$\big)$**12：结束循环 13：设置** $Q_{n+1}^{(0)} \leftarrow Q_n^{(K_n)}$**) 14：结束循环**

A.2 网格世界

我们在随机的 4×4 网格世界中评估所提出的 ICQL（图 3）。网格世界环境包含 52 个状态-动作对，因为执行会导致智能体离开网格的动作时，智能体会停留在原地。离开每个状态时都会获得奖励。网格中的每一步都会产生少量负期望奖励，促使智能体寻找通往目标的最短路径。难点在于该区域内奖励的方差很高，这可能会误导学习算法，使其高估通往该区域的动作价值。尽管该区域有可能获得较高的正奖励，但由于其期望奖励低于默认状态，最优路径会避开高方差区域。评估期间，智能体从左上角的格子出发，执行 7 步；这恰好足以到达目标状态并继续离开该状态，从而获得奖励。当折扣因子设为 $\gamma = 0.7$ 时，起始状态下执行动作 $a \in \{$下，右$\}$ 的最优 Q 值为 0.0735；当折扣因子为 $\gamma = 0.9$ 时，最优 Q 值为 0.4615；当折扣因子为 $\gamma = 0.95$ 时，最优 Q 值为 0.6551。所有实验均



状态类型	奖励分布	Mean	方差
起始与默认	$\{-0.08, 0.05\}$, $p = 0.5$	-0.015	0.0018225
Goal	$\{0.5, 1.5\}$, $p = 0.5$	1.0	0.25
随机区域	$\{-2.1, 2.0\}$, $p = 0.5$	-0.05	4.2025
Bomb	确定性： -3	-3	0

图 3：用于数值实验的网格世界环境。智能体从状态 S（灰色）出发，目标是在避开炸弹状态 B（红色）的同时抵达目标状态 G（绿色）。橙色状态表示高方差区域。灰色箭头表示最优路径。

严格按照我们的理论设定进行。探索以均匀方式遍历所有状态-动作对，从而得到 $\xi = 1/|S||A| = 1/52$。多层版本的学习率根据定理 4.5 设置为

$$\alpha_k = \frac{1}{1 + \frac{1}{2 \cdot 52} k}, \quad k = 0, 1, 2, \ldots.$$

图4展示了 $\gamma \in \{0.9, 0.95\}$ 的网格世界实验结果（分别对应上、下两行），包括固定目标更新频率 Q-learning （ QL ）和所提出的递增周期方法（ ICQL ）在不同初始目标更新频率（ ITUF ）下的偏差（左列）以及评估期间取得的得分（右列）。回顾一下，偏差定义为每个外层循环索引 n 处，周期起始目标矩阵与最优动作价值函数之间的 ℓ_∞ 距离，即 Bias(n) := $\left\| Q_n^{(0)} - Q^* \right\|_\infty$。图 1 和图 2 中的网格世界示例将折扣因子设为 $\gamma = 0.7$。对于$\gamma = 0.9$，固定目标更新频率 QL 表现出明显的调度敏感性：较小的目标更新频率（例如，目标更新频率 $= 10^3$，10_4）会先出现早期下降，随后稳定在相对较大的误差水平；而只有足够大的冻结时长（此处目标更新频率为 $= 10^5$）才能进入约 10^{-1} 的低偏差区间。ICQL 的偏差下降得更稳定，并且在所比较的设置中达到最低的最终误差。具体而言，较大 ITUF 下可见的阶梯式下降反映了周期结构，表明其在固定调度所出现的平台期之后仍持续改善。得分曲线也呈现相同趋势：达到低偏差的配置能够取得较高且稳定的回报，而偏差较大的固定目标更新频率运行则始终停留在低性能区间。

对于 $\gamma = 0.95$，这一影响更为显著，因为较弱的贝尔曼压缩会放大目标近似误差。因此，对于 TUF $= 10^3$ ，固定 TUF 的 Q 学习始终存在较大偏差，而对于 TUF $= 10^4$，其会稳定在较大的误差水平；只有采用 TUF $= 10^5$ ，才能实现较小的偏差和较高的得分。ICQL 仍表现出明显更强的鲁棒性。尽管较小的初始化（ITUF $= 10^3$）仍可能产生较大的残余误差，但中等到较大的初始化（ITUF $\in \{10^4, 10^5\}$）会带来持续的偏差降低以及相应的得分提升，其中 ITUF $= 10^5$ 达到了表现最佳的状态。与图 2 中的 $\gamma = 0.7$ 偏差图相比，在后者中，所有 ICQL 变体都收敛到较低误差，而固定 TUF 的 Q 学习通常会在取决于配置的水平上进入平台期；这些结果表明，在训练过程中增加有效的目标冻结时长，可以缓解稳定性–准确性权衡，并随着折扣因子增大而提升鲁棒性。

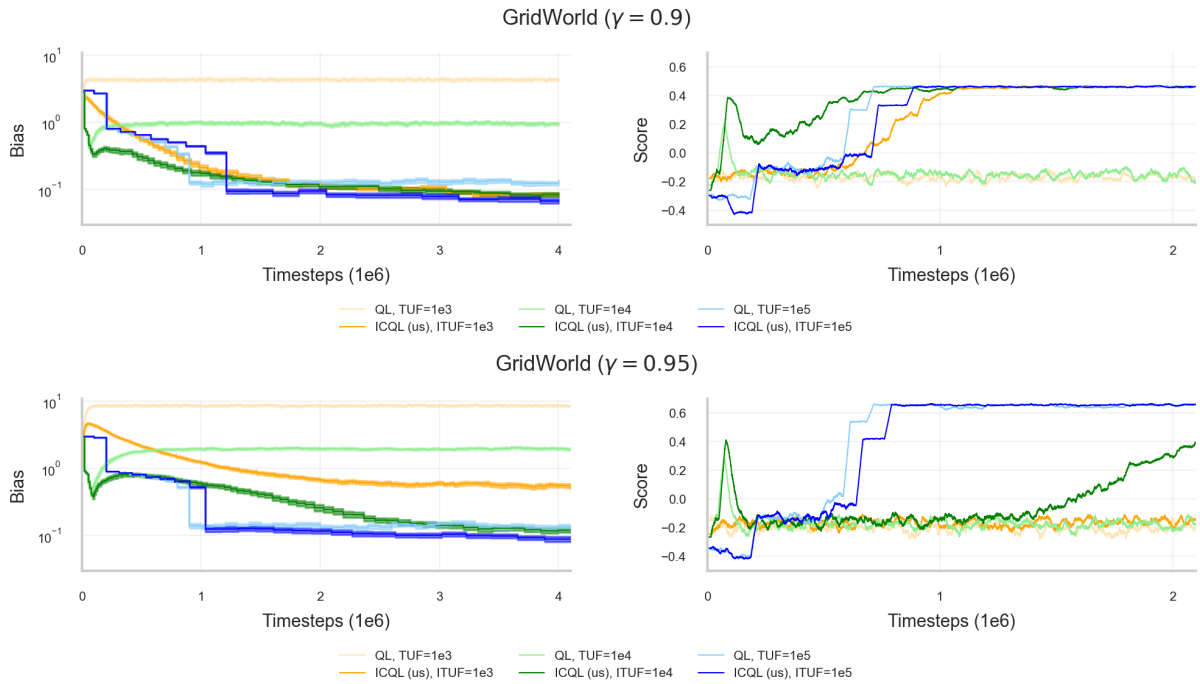


图4：附录 A.2 中网格世界上的 Q 学习，采用定理 5.3 中的最优递增目标更新频率（ITUF），并与采用固定目标更新频率（TUF）的 Q 学习进行比较，适用于 $K \in \{1000, 10000, 100000\}$。顶行的折扣因子设为$\gamma = 0.9$，下方设为 $\gamma = 0.95$。第一列中的阴影带表示基于 50 个随机种子的 95% 置信区间：较小的固定 TUF 会因贝尔曼算子近似较粗糙而导致较早进入平台期，而较大的固定 TUF 收敛缓慢，因为过于精确的内循环近似主导了早期迭代。理论预测的几何递增调度通过平衡压缩与贝尔曼近似误差，避免了这一权衡，既实现了初期快速进展，又实现了渐近收敛，从而在评估过程中选择最优路径。

B 精细化随机梯度下降分析：第 4 节的证明

对于子　　在后续分析中，我们将周期 n第 k 个内层步骤之前的历史记录记为

$$\{Q\}_n^{(k)} := \{Q_i^{(j)} : i < n,\ 0 \le j \le K_i\} \cup \{Q_n^{(j)} : 0 \le j \le k\},$$
$$\{(s,a)\}_n^{(k)} := \{(s_{i,j}, a_{i,j}) : i < n,\ 0 \le j \le K_i\} \cup \{(s_{n,j}, a_{n,j}) : 0 \le j \le k\}.$$

并定义滤过

$$\mathcal{F}_{n,k} := \sigma\big(\{Q\}_n^{(k)},\ \{(s,a)\}_n^{(k-1)}\big), \quad \mathcal{F}_{n,k}^{s,a} := \sigma\big(\mathcal{F}_{n,k},\ (s_{n,k}, a_{n,k}) = (s,a)\big).$$

注意，$\mathcal{F}_{n,0}$ 包含截至 n个周期开始时的所有信息。特别地，$Q_n^{(0)}$是$\mathcal{F}_{n,0}$-可测的。此外，$\mathcal{F}_{n,k}$ 包含截至 k个周期中第 n个随机梯度下降步骤的所有信息，但不包含当前采样状态—动作对（$s_{n,k}$，$a_{n,k}$）。以 $\mathcal{F}^{s,a}_{n,k}$为条件，

固定当前状态–动作对 $(s_{n,\,k},\ a_{n,\,k}) = (s,\ a)$，并且随机性仅存在于 $(r_{n,\,k},\ s'_{n,\,k})$ 中。因此，

$$\mathbb{E}\Big[\hat{T}Q_n^{(0)}(s,a) \;\Big|\; \mathcal{F}_{n,k}^{s,a}\Big] = (T^*Q_n^{(0)})(s,a). \tag{16}$$

定义 $I^{s,a}_{n,k} := \mathbb{1}_{\{(s_{n,k},a_{n,k})=(s,a)\}}$，它表示内循环 n 的第 k 步采样到的状态–动作对是否等于 (s,a)。它仅取决于当前抽样是否等于 (s,a)，因此关于 $\mathcal{F}^{s,a}_{n,k}$ 可测。我们允许采样分布使用 $\mathcal{F}_{n,k}$ 中包含的信息，使得在滤子 $\mathcal{F}_{n,k}$ 条件下，第 k 步采样到 (s,a) 的概率为 $\pi_{n,k}(s,a)$。由于 $(s_{n,k},a_{n,k})$ 是根据 $\pi_{n,k}$ 采样的，指示函数 $\mathbb{1}_{\{(s_{n,k},a_{n,k})=(s,a)\}}$ 以概率 $\pi_{n,k}(s,a)$ 取值 1，否则取值 0。因此，

$$\mathbb{E}\left[I^{s,a}_{n,k} \mid \mathcal{F}_{n,k}\right] = \pi_{n,k}(s,a). \tag{17}$$

B.1 内循环收敛

现在，我们分析内层随机梯度下降循环对单次贝尔曼更新的近似效果。以下命题给出了内循环的单步递推，并将压缩项与随机采样所引起的方差项清晰地区分开来。

命题 B.1（内循环的一步递推关系）。 假设假设 *4.3* 成立，令 $0 < \alpha_n^{(k)} \le 1$并定义

$$e_n^{(k)} := \mathbb{E}\big[\|Q_n^{(k)} - T^*Q_n^{(0)}\|_2^2 \mid \mathcal{F}_{n,0}\big]\,.$$

那么，对于所有的 $n \ge 0,\ k = 0,\ldots,K_n - 1$，内循环误差满足

$$e_n^{(k+1)} \;\le\; \big(1 - \xi\alpha_n^{(k)}\big)e_n^{(k)} + (\alpha_n^{(k)})^2 C_n\,,$$

其中 $C_n :\, = 2\sigma^2 + 4\gamma^2\big(\|Q_n^{(0)} - Q^*\|_\infty^2 + \|Q^*\|_\infty^2\big)\,.$

证明。 我们定义逐坐标误差

$$\Delta_k(s,a) \;:=\; Q_n^{(k)}(s,a) - (T^*Q_n^{(0)})(s,a)$$

并写作

$$\begin{aligned}\mathbb{E}\left[\|Q_n^{(k+1)} - T^*Q_n^{(0)}\|_2^2 \mid \mathcal{F}_{n,0}\right] &= \sum_{s,a}\mathbb{E}\left[(Q_n^{(k+1)}(s,a) - T^*Q_n^{(0)}(s,a))^2 \mid \mathcal{F}_{n,0}\right] \\ &= \sum_{s,a}\mathbb{E}\left[\Delta_{k+1}(s,a)^2 \mid \mathcal{F}_{n,0}\right]\,.\end{aligned}$$

利用内循环更新，我们有

$$Q_n^{(k+1)}(s,a) = Q_n^{(k)}(s,a) + \mathbb{1}_{\{(s,a)=(s_{n,k},a_{n,k})\}}\alpha_n^{(k)}((\hat{T}Q_n^{(0)})(s,a) - Q_n^{(k)}(s,a))\,.$$

定义 $\mathbb{1}_{\{(s,a)=(s_{n,k},a_{n,k})\}} = I^{s,\,a}_{n,\,k}$，后，误差项可改写为

$$\begin{aligned}\Delta_{k+1}(s,a) &= \Delta_k(s,a) + I^{s,a}_{n,k}\,\alpha_n^{(k)}\Big((\hat{T}Q_n^{(0)})(s,a) - Q_n^{(k)}(s,a)\Big) \\ &= \Delta_k(s,a) + I^{s,a}_{n,k}\,\alpha_n^{(k)}\Big((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a) - \Delta_k(s,a)\Big)\,.\end{aligned}$$

展开平方项，对 $\mathcal{F}_{n,\ k}$ 取条件期望，并利用 $\Delta_k(s,\ a)$ 和 $\alpha_n^{(k)}$ 的 $\mathcal{F}_{n,\ k}$-可测性，可得

$$\begin{aligned}\mathbb{E}\left[\Delta_{k+1}(s,a)^2 \mid \mathcal{F}_{n,k}\right] &= \mathbb{E}\left[\Delta_k(s,a)^2 \mid \mathcal{F}_{n,k}\right] \\ &\quad + 2\mathbb{E}\left[I_{n,k}^{s,a}\alpha_n^{(k)}\,\Delta_k(s,a)\Big((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a) - \Delta_k(s,a)\Big) \mid \mathcal{F}_{n,k}\right] \\ &\quad + \mathbb{E}\left[I_{n,k}^{s,a}(\alpha_n^{(k)})^2\Big((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a) - \Delta_k(s,a)\Big)^2 \mid \mathcal{F}_{n,k}\right] \\ &= \Delta_k(s,a)^2 \\ &\quad + 2\alpha_n^{(k)}\,\Delta_k(s,a)\mathbb{E}\left[I_{n,k}^{s,a}\Big((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a) - \Delta_k(s,a)\Big) \mid \mathcal{F}_{n,k}\right] \\ &\quad + (\alpha_n^{(k)})^2\mathbb{E}\left[I_{n,k}^{s,a}\Big((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a) - \Delta_k(s,a)\Big)^2 \mid \mathcal{F}_{n,k}\right].\end{aligned}$$

现在我们分别处理每个 条件期望项，然后再将它们合并 ：

项 1：首先，我们通过以下方式进一步分解该期望：

$$\begin{aligned}&\mathbb{E}\left[I_{n,k}^{s,a}\Big((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a) - \Delta_k(s,a)\Big) \mid \mathcal{F}_{n,k}\right] \\ &= \underbrace{\mathbb{E}\left[I_{n,k}^{s,a}((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a)) \mid \mathcal{F}_{n,k}\right]}_{(a)} - \underbrace{\mathbb{E}\left[I_{n,k}^{s,a}\Delta_k(s,a) \mid \mathcal{F}_{n,k}\right]}_{(b)}.\end{aligned}$$

对于 (a)，我们利用 $\mathcal{F}_{n,k} \subset \mathcal{F}^{s}_{,n}{}^{a}_{k,}$ 的塔式性质，并利用 $\mathcal{F}^{s}_{,n}{}^{a}_{k,}$ 对 $I^{s}_{,n}{}^{a}_{k,}$ 的可测性，得到

$$\begin{aligned}&\mathbb{E}\big[I_{n,k}^{s,a}((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a)) \mid \mathcal{F}_{n,k}\big] \\ &\qquad = \mathbb{E}\big[\mathbb{E}\big[I_{n,k}^{s,a}((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a)) \mid \mathcal{F}_{n,k}^{s,a}\big] \mid \mathcal{F}_{n,k}\big] \\ &\qquad = \mathbb{E}\big[I_{n,k}^{s,a}\mathbb{E}\big[((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a)) \mid \mathcal{F}_{n,k}^{s,a}\big] \mid \mathcal{F}_{n,k}\big].\end{aligned}$$

根据单样本贝尔曼目标的无偏性，我们有

$$\mathbb{E}\left[((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a)) \mid \mathcal{F}_{n,k}^{s,a}\right] = 0,$$

因此 (a) 等于零。对于 (b)，我们利用 (,) 和 ,(,) 关于
$\mathcal{F}_{n\,k}$ $\Delta_k\ s\ a$ $\mathbb{E}\big[I_{n,k}^{s,a} \mid \mathcal{F}_{n,k}\big] =$, 的可测性，推导得出
$\pi_{n\,k}\ s\ a$

$$\mathbb{E}\left[I_{n,k}^{s,a}\Delta_k(s,a) \mid \mathcal{F}_{n,k}\right] = \Delta_k(s,a)\mathbb{E}\left[I_{n,k}^{s,a} \mid \mathcal{F}_{n,k}\right] = \Delta_k(s,a)\pi_{n,k}(s,a).$$

结合 (a) 和 (b)，我们验证了

$$\mathbb{E}\left[I_{n,k}^{s,a}\Big((\hat{T}Q_n^{(0)})(s,a) - (T^*Q_n^{(0)})(s,a) - \Delta_k(s,a)\Big) \mid \mathcal{F}_{n,k}\right] = -\Delta_k(s,a)\pi_{n,k}(s,a).$$

项 2：我们再次拆分期望，并应用与之前相同的论证

$$
\begin{aligned}
&\mathbb{E}\left[I_{n,k}^{s,a}\Big((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a)-\Delta_k(s,a)\Big)^2 \mid \mathcal{F}_{n,k}\right]\\
&=\mathbb{E}\left[I_{n,k}^{s,a}((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a))^2 \mid \mathcal{F}_{n,k}\right]\\
&\quad-2\mathbb{E}\left[I_{n,k}^{s,a}\Delta_k(s,a)((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a)) \mid \mathcal{F}_{n,k}\right]\\
&\quad+\mathbb{E}\left[I_{n,k}^{s,a}\Delta_k(s,a)^2 \mid \mathcal{F}_{n,k}\right]\\
&=\underbrace{\mathbb{E}\left[I_{n,k}^{s,a}((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a))^2 \mid \mathcal{F}_{n,k}\right]}_{(c)}+\underbrace{\mathbb{E}\left[I_{n,k}^{s,a}\Delta_k(s,a)^2 \mid \mathcal{F}_{n,k}\right]}_{(d)}.
\end{aligned}
$$

对于（c），我们再次利用塔式性质以及 $\mathcal{F}_{n,k}\subseteq\mathcal{F}^{s,a}_{n,k}$ 和 $\mathcal{F}^{s,a}_{n,k}$ 对 $I^{s,a}_{n,k}$ 的可测性，推导得

$$
\begin{aligned}
&\mathbb{E}\big[I_{n,k}^{s,a}((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a))^2 \mid \mathcal{F}_{n,k}\big]\\
&\qquad=\mathbb{E}\big[\mathbb{E}\big[I_{n,k}^{s,a}((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a))^2 \mid \mathcal{F}_{n,k}^{s,a}\big] \mid \mathcal{F}_{n,k}\big]\\
&\qquad=\mathbb{E}\big[I_{n,k}^{s,a}\mathbb{E}\big[((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a))^2 \mid \mathcal{F}_{n,k}^{s,a}\big] \mid \mathcal{F}_{n,k}\big].
\end{aligned}
$$

展开平方项可得

$$
\begin{aligned}
\mathbb{E}\left[((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a))^2 \mid \mathcal{F}_{n,k}^{s,a}\right]&=\mathrm{Var}\big((\hat{T}Q_n^{(0)})(s,a) \mid \mathcal{F}_{n,k}^{s,a}\big)\\
&=\mathrm{Var}\big(R(s,a)+\gamma\max_{a'}Q_n^{(0)}(S',a') \mid \mathcal{F}_{n,k}^{s,a}\big)\\
&\le 2\,\mathrm{Var}(R(s,a) \mid \mathcal{F}_{n,k}^{s,a})+2\gamma^2\,\mathrm{Var}\big(\max_{a'}Q_n^{(0)}(S',a') \mid \mathcal{F}_{n,k}^{s,a}\big)\\
&\le 2\sigma^2+2\gamma^2\|Q_n^{(0)}\|_\infty^2,
\end{aligned}
$$

其中，我们用到了 $\mathrm{Var}(X+Y \mid \mathcal{F}^{s,a}_{n,k}) \le 2\mathrm{Var}(X \mid \mathcal{F}^{s,a}_{n,k})+2\mathrm{Var}(Y \mid \mathcal{F}^{s,a}_{n,k})$，以及界 $\mathrm{Var}(Z \mid \mathcal{F}^{s,a}_{n,k}) \le \mathbb{E}[Z^2 \mid \mathcal{F}^{s,a}_{n,k}] \le \|Q_n^{(0)}\|_\infty^2$ for $Z:=\max_{a'}Q_n^{(0)}(S',a')$。因此，（c）可由以下式子界定

$$
\mathbb{E}\left[I_{n,k}^{s,a}((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a))^2 \mid \mathcal{F}_{n,k}\right]\le\big(2\sigma^2+2\gamma^2\|Q_n^{(0)}\|_\infty^2\big)\pi_{n,k}(s,a).
$$

对于（d），我们利用 $\mathcal{F}_{n,k}$ 对 $\Delta_k(s,a)^2$ 的可测性以及公式（17），推得

$$
\mathbb{E}\Big[I_{n,k}^{s,a}\Delta_k(s,a)^2 \mid \mathcal{F}_{n,k}\Big]=\Delta_k(s,a)^2\mathbb{E}\left[I_{n,k}^{s,a} \mid \mathcal{F}_{n,k}\right]=\Delta_k(s,a)^2\pi_{n,k}(s,a)\,.
$$

结合（c）和（d），可得

$$
\begin{aligned}
&\mathbb{E}\Big[I_{n,k}^{s,a}\Big((\hat{T}Q_n^{(0)})(s,a)-(T^*Q_n^{(0)})(s,a)-\Delta_k(s,a)\Big)^2 \mid \mathcal{F}_{n,k}\Big]\\
&\qquad\le\big(2\sigma^2+2\gamma^2\|Q_n^{(0)}\|_\infty^2\big)\pi_{n,k}(s,a)+\Delta_k(s,a)^2\pi_{n,k}(s,a).
\end{aligned}
$$

最后，结合项 2 和项 3，我们得到逐坐标误差的递推关系：

$$
\begin{aligned}
\mathbb{E}\left[\Delta_{k+1}(s,a)^2 \mid \mathcal{F}_{n,k}\right]&\le\Delta_k(s,a)^2-2\alpha_n^{(k)}\,\pi_{n,k}(s,a)\,\Delta_k(s,a)^2\\
&\quad+(\alpha_n^{(k)})^2\,\pi_{n,k}(s,a)\,\Delta_k(s,a)^2+(\alpha_n^{(k)})^2\,\pi_{n,k}(s,a)\big(2\sigma^2+2\gamma^2\|Q_n^{(0)}\|_\infty^2\big)\\
&=\Big(1-(2\alpha_n^{(k)}-(\alpha_n^{(k)})^2)\pi_{n,k}(s,a)\Big)\Delta_k(s,a)^2\\
&\quad+(\alpha_n^{(k)})^2\,\pi_{n,k}(s,a)\big(2\sigma^2+2\gamma^2\|Q_n^{(0)}\|_\infty^2\big).
\end{aligned}
$$

定义

$$E_n^{(k)} := \sum_{s,a} \Delta_k(s,a)^2 = \|Q_n^{(k)} - T^* Q_n^{(0)}\|_2^2$$

并对所有满足以下条件的 $(s,\ a)$，将这些逐坐标界求和：利用 $\sum_{s,\ a} \pi_{n,\ k}(s,\ a) = 1$其中 $\pi_{n,\ k}(s,\ a) \geq \xi$ 且 $\alpha_n^{(k)} \leq 1$，我们得到

$$\begin{aligned}
\mathbb{E}\big[E_n^{(k+1)} \mid \mathcal{F}_{n,k}\big] &= \sum_{s,a} \mathbb{E}\big[\Delta_{k+1}(s,a)^2 \mid \mathcal{F}_{n,k}\big] \\
&\leq \sum_{s,a} \Big(1 - (2\alpha_n^{(k)} - (\alpha_n^{(k)})^2)\pi_{n,k}(s,a)\Big)\Delta_k(s,a)^2 \\
&\quad + (\alpha_n^{(k)})^2 \sum_{s,a} \pi_{n,k}(s,a)\big(2\sigma^2 + 2\gamma^2 \|Q_n^{(0)}\|_\infty^2\big) \\
&= \sum_{s,a} \Delta_k(s,a)^2 - (2\alpha_n^{(k)} - (\alpha_n^{(k)})^2) \sum_{s,a} \pi_{n,k}(s,a)\Delta_k(s,a)^2 \\
&\quad + (\alpha_n^{(k)})^2\big(2\sigma^2 + 2\gamma^2 \|Q_n^{(0)}\|_\infty^2\big) \\
&\leq \Big(1 - \xi(2\alpha_n^{(k)} - (\alpha_n^{(k)})^2)\Big) \sum_{s,a} \Delta_k(s,a)^2 + (\alpha_n^{(k)})^2\big(2\sigma^2 + 2\gamma^2 \|Q_n^{(0)}\|_\infty^2\big) \\
&= \big(1 - \xi(2\alpha_n^{(k)} - (\alpha_n^{(k)})^2)\big) E_n^{(k)} + (\alpha_n^{(k)})^2 C_n,
\end{aligned}$$

其中，在倒数第二行中，我们用到了 $\sum_{s,\ a} \pi_{n,\ k}(s,\ a)\Delta_k(s,\ a)^2 \geq \xi \sum_{s,\ a} \Delta_k(s,\ a)^2$。此外，我们定义

$$C_n := 2\sigma^2 + 4\gamma^2(\|Q_n^{(0)} - Q^*\|_\infty^2 + \|Q^*\|_\infty^2),$$

它满足

$$2\sigma^2 + 2\gamma^2 \|Q_n^{(0)}\|_\infty^2 \ \leq\ C_n,$$

因为 $\|Q_n^{(0)}\|_\infty^2 \leq 2\big(\|Q_n^{(0)} - Q^*\|_\infty^2 + \|Q^*\|_\infty^2\big)$。注意，$Q_n^{(0)}$、$\|Q_n^{(0)}\|_\infty$，因而 C_n，都是 $\mathcal{F}_{n,\ 0}$-可测的。应用塔式性质，我们得到相同的标量递推关系，但条件为 $\mathcal{F}_{n,\ 0}$：

$$\mathbb{E}\Big[\|Q_n^{(k+1)} - T^* Q_n^{(0)}\|_2^2 \,\Big|\, \mathcal{F}_{n,0}\Big] \ \leq\ \big(1 - \xi(2\alpha_n^{(k)} - (\alpha_n^{(k)})^2)\big)\, \mathbb{E}\Big[\|Q_n^{(k)} - T^* Q_n^{(0)}\|_2^2 \,\Big|\, \mathcal{F}_{n,0}\Big] \ +\ (\alpha_n^{(k)})^2\, C_n.$$

为简化该界，注意到 $2\alpha_n^{(k)} - (\alpha_n^{(k)})^2 = \alpha_n^{(k)} + (\alpha_n^{(k)} - (\alpha_n^{(k)})^2)$，其中 $(\alpha_n^{(k)} - (\alpha_n^{(k)})^2) > 0$，因为 $0 < \alpha_n^{(k)} \leq 1$。因此，我们可以写成

$$1 - \xi(2\alpha_n^{(k)} - (\alpha_n^{(k)})^2) = 1 - \xi\alpha_n^{(k)} - \xi(\alpha_n^{(k)} - (\alpha_n^{(k)})^2)$$

这意味着

$$1 - \xi(2\alpha_n^{(k)} - (\alpha_n^{(k)})^2) \leq 1 - \xi\alpha_n^{(k)}\,.$$

利用这个界，递推关系可简化为

$$\mathbb{E}\Big[\|Q_n^{(k+1)} - T^* Q_n^{(0)}\|_2^2 \,\Big|\, \mathcal{F}_{n,0}\Big] \ \leq\ \big(1 - \xi\alpha_n^{(k)}\big)\, \mathbb{E}\Big[\|Q_n^{(k)} - T^* Q_n^{(0)}\|_2^2 \,\Big|\, \mathcal{F}_{n,0}\Big] \ +\ (\alpha_n^{(k)})^2\, C_n\,.$$

□

定理 *4.5* 的证明。令 $e_n^{(k)} := \mathbb{E}\Big[\|Q_n^{(k)} - T^* Q_n^{(0)}\|_2^2 \,\Big|\, \mathcal{F}_{n,0}\Big]$。首先，我们根据定义$\alpha_n^{(0)} = 1$，可知 $\alpha_n^{(k)} \le 1$ 对所有 $k \ge 0$均成立。因此，我们可以应用命题 B.1 得到递推关系

$$e_n^{(k+1)} \le \Big(1 - \alpha_n^{(k)}\xi\Big) e_n^{(k)} + (\alpha_n^{(k)})^2 C_n, \quad k \ge 0,$$

其中 $\alpha_n^{(k)} = \frac{2}{\xi(k+s)}$ 且 $s = \frac{2}{\xi}$。现在我们将证明，对于 $k \ge 1$，有 $e_n^{(k)} \le \frac{\Gamma}{k+s}$，其中

$$\Gamma := \frac{(s+1)\max(e_0, \frac{2C_n}{\xi})}{k+s}.$$

对于 $k = 1$，我们有

$$e_n^{(1)} \le (1 - \alpha_n^{(0)}\xi) e_n^{(0)} + (\alpha_n^{(0)})^2 C_n = (1 - \frac{2}{s}) e_n^{(0)} + \frac{4C_n}{\xi^2 s^2} \le \big(1 - \frac{1}{s}\big)\frac{\Gamma}{s+1} \le \frac{\Gamma}{s+1}.$$

现在，假设 对于某个 $k \ge 1$，上界 $e_n^{(k)} \le \frac{\Gamma}{k+s}$ 成立。利用基本展开式

$$\tfrac{\Gamma}{k+s+1} = \tfrac{\Gamma}{k+s} - \tfrac{\Gamma}{(k+s)(k+s+1)}$$

由此可得

$$\begin{aligned}
e_n^{(k+1)} \le (1 - \alpha_n^{(k)}\xi) e_n^{(k)} + (\alpha_n^{(k)})^2 c &= \Big(1 - \frac{2}{k+s}\Big)\frac{\Gamma}{k+s} + \frac{4c}{\xi^2}\frac{1}{(k+s)^2} \\
&= \frac{\Gamma}{k+1+s} + \frac{\Gamma}{(k+s)(k+1+s)} - \frac{2\Gamma}{(k+s)^2} + \frac{4c}{\xi^2}\frac{1}{(k+s)^2} \\
&= \frac{\Gamma}{k+1+s} + \frac{\frac{k+s}{k+s+1}\Gamma - 2\Gamma + \frac{4c}{\xi^2}}{(k+s)^2} \\
&\le \frac{\Gamma}{k+1+s} + \frac{-\Gamma + \frac{4c}{\xi^2}}{(k+s)^2} \le \frac{\Gamma}{k+1+s} + \frac{\frac{4c}{\xi^2} - \Gamma}{(k+s)^2} \le \frac{\Gamma}{k+1+s}.
\end{aligned}$$

这里，我们用到了

$$\Gamma = (s+1)\max\Big(e_n^{(0)}, \frac{4c}{s\xi^2}\Big) \ge (s+1)\frac{4c}{s\xi^2} \ge \frac{4c}{\xi^2},$$

作为 $s \ge 1$，这确保了 $\frac{4c}{\xi^2} - \Gamma \le 0$。我们可以直接应用归纳结果，得到

$$e_n^{(k)} \le \frac{(s+1)\max(e_n^{(0)}, \frac{2C_n}{\xi})}{k+s} \le \frac{(\frac{2}{\xi}+1)(e_n^{(0)} + \frac{2C_n}{\xi})}{k+s}, \quad \text{for all } k \ge 1.$$

回顾一下

$$C_n \;=\; 2\sigma^2 + 4\gamma^2\big(\|Q_n^{(0)} - Q^*\|_\infty^2 + \|Q^*\|_\infty^2\big).$$

我们可以通过以下方式给每个初始随机梯度下降误差设定上界：

$$\begin{aligned}
e_n^{(0)} = \|Q_n^{(0)} - Q^* + T^*Q^* - T^*Q_n^{(0)}\|_2^2 &\le |S||A| \cdot \|Q_n^{(0)} - Q^* + T^*Q^* - T^*Q_n^{(0)}\|_\infty^2 \\
&\le |S||A| \cdot \big(\|Q_n^{(0)} - Q^*\|_\infty^2 + \|T^*Q^* - T^*Q_n^{(0)}\|_\infty^2\big) \\
&\le |S||A|(1+\gamma)^2 \|Q_n^{(0)} - Q^*\|_\infty^2,
\end{aligned}$$

其中我们使用了三角不等式以及 T^* 的 γ-压缩性。最后，我们设定

$$c_1 \;:=\; \Big(\frac{2}{\xi}+1\Big)|S||A|(1+\gamma)^2+\Big(\frac{16}{\xi^2}+\frac{8}{\xi}\Big)\gamma^2, \qquad c_2 \;:=\; \Big(\frac{8}{\xi^2}+\frac{4}{\xi}\Big)\Big(\sigma^2+2\gamma^2\|Q^*\|_\infty^2\Big)$$

从而推导出

$$e_n^{(k+1)} \le \frac{c_1\,\|Q_n^{(0)}-Q^*\|_\infty^2 \;+\; c_2}{k+s}\,.$$

证明至此完成。 □

B.2 收敛结果

推论 *4.6* 的证明。 令 $\mathcal{F}:=(\mathcal{F}_{n,0})_{n\in\mathbb{N}_0}$ 为内循环开始时的滤子。利用$\|x\|_\infty \le \|x\|_2$ 的界以及詹森不等式，可得

$$\mathbb{E}\Big[\|Q_n^{(K_n)}-T^*Q_n^{(0)}\|_\infty \,\Big|\, \mathcal{F}_{n,0}\Big] \le \Big(\mathbb{E}\Big[\|Q_n^{(K_n)}-T^*Q_n^{(0)}\|_2^2 \,\Big|\, \mathcal{F}_{n,0}\Big]\Big)^{1/2}.$$

Plu在最终迭代点 $k=K_n$ 处代入定理 4.5 的内循环界，得 ds

$$\mathbb{E}\Big[\|Q_n^{(K_n)}-T^*Q_n^{(0)}\|_2^2 \,\Big|\, \mathcal{F}_{n,0}\Big] \le \frac{c_1\,\|Q_n^{(0)}-Q^*\|_\infty^2+c_2}{K_n+s} \le \frac{c_1\,E_n^2+c_2}{K_n}\,.$$

取平方根，应用次可加性，并加上 γE_n 项（该项关于 $\mathcal{F}_{n,0}$可测），得到

$$\mathbb{E}[E_{n+1}\,|\,\mathcal{F}_{n,0}] \;\le\; \gamma\,E_n \;+\; \frac{\sqrt{c_1}\,E_n+\sqrt{c_2}}{\sqrt{K_n}}\,. \tag{18}$$

因此，根据命题 4.1，我们有

$$\mathbb{E}\Big[\big\|Q_{n+1}^{(0)}-Q^*\big\|_\infty \,\Big|\, \mathcal{F}_{n,0}\Big] \;\le\; \gamma\,\big\|Q_n^{(0)}-Q^*\big\|_\infty \;+\; \frac{\sqrt{c_1}\,\|Q_n^{(0)}-Q^*\|_\infty+\sqrt{c_2}}{\sqrt{K_n}},$$

对所有 $n\ge 0$ 以及定理 4.5 中定义的常数 c_1、 c_2。令 $Z_n := \|Q_n^{(0)}-Q^*\|_\infty$，并定义确定性的非负序列

$$A_n := \frac{\sqrt{c_1}}{\sqrt{K_n}}, \qquad B_n := \frac{\sqrt{c_2}}{\sqrt{K_n}}, \qquad D_n := 1-\gamma \in (0,1).$$

于是，一步不等式可改写为

$$\mathbb{E}[Z_{n+1}\mid\mathcal{F}_{n,0}] \;\le\; Z_n\,(1+A_n-D_n) \;+\; B_n.$$

[1]假设 $\sum_{n=0}^{\infty} K_n^{-2} < \infty$。那么我们有

$$\sum_{n=1}^{\infty} A_n \;=\; \sum_{n=1}^{\infty}\frac{\sqrt{c_1}}{\sqrt{K_n}} \;<\; \infty,$$

类似地， $\sum_{n=1}^{\infty} B_n < \infty$，而

$$\sum_{n=1}^{\infty} D_n \;=\; \sum_{n=1}^{\infty} (1-\gamma) \;=\; \infty.$$

因此，适应的非负过程 $(Z_n)_{n\geq 0}$ 满足推论 4.2 的条件。因此， $Z_n \to 0$ 几乎必然成立，即我们已验证

$$\|Q_n^{(0)} - Q^*\|_\infty \;\longrightarrow\; 0 \qquad \text{a.s. as } n \to \infty.$$

□

B.3 辅助结果

定理 B.2 ([24] 中的定理 1)。 设 $(\Omega\ \mathcal{A}, \mathcal{F}, \mathbb{P})$ 为一个滤过概率空间，$(Z_k)_{k\in\mathbb{N}}$、$(A_k)_{k\in\mathbb{N}}$、$(B_k)_{k\in\mathbb{N}}$ 和 $(C_k)_{k\in\mathbb{N}}$ 是非负且 $\mathcal{F}$-适应的随机过程，并满足

$$\sum_{k=0}^{\infty} A_k < \infty \quad and \quad \sum_{k=0}^{\infty} B_k < \infty$$

几乎必然成立。此外，假设

$$\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \leq Z_k(1 + A_k) + B_k - C_k.$$

Then

(i) 存在一个几乎必然有限的随机变量 Z_∞ ，使得对于$k \to \infty$， $Z_k \to Z_\infty$ 几乎必然成立，

(ii) $\sum_{k=0}^{\infty} C_k < \infty$ 几乎必然成立。

C TUF 设计：第 5 节的详细内容与证明

C.1 准最优 TUF 调度

本节将更详细地介绍最优 TUF 设计的构造。回顾一下，我们希望选择 TUF（K_n），以便在给定目标精度的情况下，尽可能降低总计算成本。计算 $Q_n^{(K_n)}$ 的成本为

$$\mathrm{cost}(Q_n^{(K_n)}) = \sum_{j=1}^{n} K_j \,.$$

给定 $\varepsilon > 0$，目标是选择 N 和 $(K_n)_{n=0}^{N-1}$，以求解（松弛的）约束优化问题

$$\min_{N\in\mathbb{N},(K_n)_{n=0}^{N-1}\in\mathbb{R}^N} \mathrm{cost}(Q_{N-1}^{(K_N)}) \quad \text{s.t.} \quad e_N := \mathbb{E}[E_N] \leq \varepsilon \,. \tag{19}$$

可以定义有效压缩因子 $\mu \in (\gamma,\ 1)$，并确保最小 TUF 足够大，使得对于所有 $n \in \mathbb{N}_0$，

$$0 < \gamma + \sqrt{\frac{c_1}{K_n}} \le \mu < 1\,. \tag{20}$$

更准确地说，我们要求 $\min_n K_n \ge K_{\min}(\mu) := \frac{c_1}{(\mu-\gamma)^2}$。因此，对方程 (11) 两边取期望，并对 $n \in \mathbb{N}$ 进行迭代，得到

$$e_{n+1} \le \mu\, e_n + \sqrt{\frac{c_2}{K_n}}, \qquad n \in \mathbb{N}_0\,.$$

展开该递推关系可得，对于所有 $n \in \mathbb{N}$，$e_n \le \mu^n e_0 + \sum_{j=0}^{n-1} \mu^{n-1-j} \sqrt{\frac{c_2}{K_j}}$。从现在起，我们在以下条件下优化 (19)：

$$e_n := e_N(K_j : j = 0, \dots, N-1) := \mu^n e_0 + \sum_{j=0}^{n-1} \mu^{n-1-j} \sqrt{\frac{c_2}{K_j}}, \quad n \in \mathbb{N} \tag{21}$$

并附加约束 $\min_n K_n \ge K_{\min}(\mu) := \frac{c_1}{(\mu-\gamma)^2}$。为简单起见，我们允许 TUF (K_n) 取正实数。若考虑固定 TUF，则简记为 $e_N = e_N(K)$。如第 5 节所述，我们通过将误差约束 $e_N \le \varepsilon$ 分解为以下部分，来满足 (19) 中的约束：

$$\mu^N e_0 \le \frac{\varepsilon}{2} \quad \text{and} \quad \sum_{j=0}^{N-1} \mu^{N-1-j} \sqrt{\frac{c_2}{K_j}} \le \frac{\varepsilon}{2}\,.$$

这样做的优点是，所需外层循环的次数固定为 $N(\varepsilon) \ge \left\lceil \frac{\log(\varepsilon/(2e_0))}{\log(\mu)} \right\rceil$，并将设计问题简化为选择 $K_n(\varepsilon)\ \ n = 0, \dots, N(\varepsilon)$。虽然这样会得到一个更简单的约束优化问题，但事先并不清楚，不对称地拆分这两个误差项是否能降低复杂度。因此，我们先介绍最优性的概念。我们关注这样的 TUF 设计选择：当 $\varepsilon \to 0$ 时，在渐近意义下，它们只能通过常数因子得到改进。

定义 C.1（准最优目标更新频率调度）。对于 $N \in \mathbb{N}$ 以及目标更新频率 $(K_j,\ j = 0,\ \dots\dots,\ N-1) \subset \mathbb{R}_+$考虑 $e_N = e_N(K_j :\ j = 0,\ \dots\dots,\ N-1)$，其定义见 (19)。

(i) 令 $(N(\varepsilon)\ K(\varepsilon, N(\varepsilon)))_{\varepsilon>0}$ 为一族整数 $N(\varepsilon) \in \mathbb{N}$ 和实数 $K(\varepsilon, N(\varepsilon)) \in \mathbb{R}_+$，满足 $e_{N(\varepsilon)}(K(\varepsilon, N(\varepsilon))) \le \varepsilon$ 对所有充分小的 $\varepsilon > 0$ 均成立。若 $(N(\varepsilon)\ K(\varepsilon, N(\varepsilon)))_{\varepsilon>0}$ 满足以下条件，则称其为准最优固定目标更新频率

$$N(\varepsilon)\, K(\varepsilon, N(\varepsilon)) \in \mathcal{O}\big(\inf\big\{\hat{N}\hat{K} : e_{\hat{N}}(\hat{K}) \le \varepsilon,\ \hat{N} \in \mathbb{N},\ \hat{K} > 0\big\}\big) \quad \textit{as } \varepsilon \to 0\,.$$

(ii) 设 $(N(\varepsilon)\ K_j(\varepsilon, N(\varepsilon))\ 0 \le j \le N(\varepsilon) - 1)_{\varepsilon>0}$ 为一族整数 $N(\varepsilon) \in \mathbb{N}$ 以及序列 $(K_j(\varepsilon, N(\varepsilon))\ j = 0, \dots, N(\varepsilon) - 1) \subset \mathbb{R}_+$，满足 $e_{N(\varepsilon)}(K_j(\varepsilon, N(\varepsilon))\ \ j = 0, \dots, N(\varepsilon) - 1) \le \varepsilon$ 对所有充分小的 $\varepsilon > 0$ 均成立。若 $(N(\varepsilon)\ K_j(\varepsilon, N(\varepsilon))\ 0 \le j \le N(\varepsilon) - 1)_{\varepsilon>0}$ 是一个准最优的可变目标更新频率调度，则称之为

$$\sum_{j=0}^{N(\varepsilon)-1} K_j(\varepsilon, N(\varepsilon)) \in \mathcal{O}\Big(\inf\Big\{\sum_{j=0}^{\hat{N}-1} \hat{K}_j : e_{\hat{N}}(\hat{K}_j,\ j = 0, \dots, \hat{N}-1) \le \varepsilon,\ \hat{N} \in \mathbb{N},\ \hat{K}_j > 0\ \forall j < \hat{N}-1\Big\}\Big)$$

即 $\varepsilon \to 0$。

C.2 样本复杂度的证明

定理 *5.1* 的证明。 首先，我们省略式 (20) 中要求的下界 $K \geq K_{\min}$。固定周期长度为 K 时，式 (21) 给出

$$e_n = \mu^n e_0 + \frac{\sqrt{c_2}}{\sqrt{K}} \frac{1-\mu^n}{1-\mu}.$$

我们分别考虑这些约束

$$\mu^n e_0 \leq \tfrac{\varepsilon}{2} \quad \text{and} \quad \frac{\sqrt{c_2}}{\sqrt{K}} \frac{1-\mu^n}{1-\mu} \leq \frac{\varepsilon}{2}. \tag{22}$$

与定理 5 类似，在 [31], 中将这些约束取等号求解，并令 n 和 K，可得到以下取值:

$$N(\varepsilon) = \Big\lceil \frac{\log(\varepsilon/(2e_0))}{\log(\mu)} \Big\rceil, \quad K(\varepsilon) = \frac{4c_2}{\varepsilon^2} \cdot \Big(\frac{1-\mu^{N(\varepsilon)}}{1-\mu} \Big)^2$$

接下来令 $\mu = \frac{1+\gamma}{2} \in (\gamma, 1)$，并假设 $\varepsilon \leq (1-\gamma)\sqrt{\frac{c_2}{c_1}}$，便得

$$\frac{c_1}{(\mu-\gamma)^2} = \frac{4c_1}{(1-\gamma)^2} \leq \frac{4c_2(1-\gamma)^2}{\varepsilon^2(1-\mu)^2} = \frac{16c_2}{\varepsilon^2}$$

其中我们用到了 $\mu - \gamma = \frac{1}{2}(1-\gamma) = 1-\mu$。因此，我们有

$$K(\varepsilon) \geq K_{\min}(\mu) = \frac{c_1}{(\mu-\gamma)^2}$$

以及相应的计算成本

$$\mathrm{cost}(Q^{(0)}_{N(\varepsilon)}) = N(\varepsilon)\, K \leq \Big\lceil \frac{\log((2e_0)/\varepsilon)}{\log(\mu^{-1})} \Big\rceil \frac{4c_2}{\varepsilon^2} \left(\frac{1}{1-\mu} \right)^2 \leq \Big\lceil \frac{\log((2e_0)/\varepsilon)}{\log(2(1+\gamma)^{-1})} \Big\rceil \frac{16c_2}{\varepsilon^2} \left(\frac{1}{1-\gamma} \right)^2.$$

如果我们进一步假设 $R(s, a) \in [-1, 1]$，则可直接得出 $\sigma^2 \leq 1$、$\|Q^*\|_\infty \leq \frac{1}{1-\gamma}$ 和 $e_0 \leq \frac{2}{1-\gamma}$, 这意味着

$$c_2 \lesssim \frac{1}{\xi^2}\big(1 + \frac{\gamma^2}{(1-\gamma)^2}\big) = \frac{1}{\xi^2}\big(\frac{1-2\gamma+2\gamma^2}{(1-\gamma)^2}\big) \lesssim \frac{1}{\xi^2(1-\gamma)^2}. \tag{23}$$

由此得到的样本复杂度为

$$\mathrm{cost}(Q^{(0)}_{N(\varepsilon)}) = \mathcal{O}\Big(\frac{1}{(1-\gamma)^4 \xi^2 \log(2(1+\gamma)^{-1})}\, \varepsilon^{-2}\, \log \frac{4}{(1-\gamma)\varepsilon} \Big). \tag{24}$$

最后，我们需要论证其准最优性。假设我们选择

$$\mu^n e_0 \leq \lambda\varepsilon \quad \text{and} \quad \frac{\sqrt{c_2}}{\sqrt{K}} \frac{1-\mu^n}{1-\mu} \leq (1-\lambda)\varepsilon.$$

对于任意的 $\lambda \in (0, 1)$，而不是式 (22) 中带有 $\lambda = 1/2$ 的对称分解。此时，$\hat{}$ 最优选择将得到 $N(\varepsilon) = N(2\lambda\varepsilon) \geq N(2\varepsilon)$ 和 $K(\varepsilon) = K(2\lambda\varepsilon) \geq K(2\varepsilon)$，从而计算成本的下界为

$$\hat{N}(\varepsilon)\hat{K}(\varepsilon) \geq \Big\lceil \frac{\log(e_0/\varepsilon)}{\log(2(1+\gamma)^{-1})} \Big\rceil \frac{4c_2}{\varepsilon^2} \Big(\frac{1}{1-\gamma} \Big)^2.$$

更准确地说，我们有

$$\left\lceil \frac{\log(e_0/\varepsilon)}{\log(2(1+\gamma)^{-1})} \right\rceil \frac{4c_2}{\varepsilon^2}\Big(\frac{1}{1-\gamma}\Big)^2 \le \inf\left\{\hat{N}\hat{K} : e_{\hat{N}}(\hat{K}) \le \varepsilon,\ \hat{N}\in\mathbb{N},\ \hat{K}>0\right\}$$
$$\le \left\lceil \frac{\log((2e_0)/\varepsilon)}{\log(2(1+\gamma)^{-1})} \right\rceil \frac{16c_2}{\varepsilon^2}\Big(\frac{1}{1-\gamma}\Big)^2$$

对所有 $0<\varepsilon\le(1-\gamma)\sqrt{\frac{c_2}{c_1}}$。 □

定理 *5.3* 的证明。 由于我们不固定 TUF，因此先将式 (21) 进行误差分解，得到

$$\mu^N e_0 \le \frac{\varepsilon}{2} \quad \text{and} \quad \sum_{j=0}^{N-1} \mu^{N-1-j}\sqrt{\frac{c_2}{K_j}} \le \frac{\varepsilon}{2}.$$

忽略所需的最小 TUF，在 [31]中应用定理 8，并作如下对应

$$c=\mu, \qquad \ell_j = \frac{K_j}{c_2}, \qquad \alpha = \tfrac{1}{2}$$

可得到松弛约束问题的一个解

$$\min_{(K_n)_{n=0}^{N(\varepsilon)-1}} \operatorname{cost}(Q^{(0)}_{N(\varepsilon)}) \quad \text{s.t.} \quad \sum_{j=0}^{N(\varepsilon)-1} \mu^{N(\varepsilon)-1-j}\sqrt{\frac{c_2}{K_j}} \le \frac{\varepsilon}{2}\,.$$

该解可写为

$$N(\varepsilon) = \left\lceil \frac{\log(\frac{\varepsilon}{2e_0})}{\log(\mu)} \right\rceil$$

and

$$K_j(\varepsilon) = C_\varepsilon \mu^{\frac{2}{3}(N(\varepsilon)-1-j)} \quad \text{with} \quad C_\varepsilon = \frac{4c_2}{\varepsilon^2}\left(\frac{1-\mu^{\frac{2}{3}N(\varepsilon)}}{1-\mu^{\frac{2}{3}}}\right)^2.$$

接下来，我们寻求关于 $\varepsilon>0$ 的一个（充分）条件，使得

$$K_j(\varepsilon) \ge K_0(\varepsilon) > \frac{4c_2}{(1-\gamma)^2}.$$

代入定义并消去 $4c_2$ 后，得到

$$\frac{1}{\varepsilon^2}\left(\frac{1-\mu^{\frac{2}{3}N(\varepsilon)}}{1-\mu^{\frac{2}{3}}}\right)^2 \mu^{\frac{2}{3}(N(\varepsilon)-1)} > \frac{1}{(1-\gamma)^2}.$$

两边开平方（所有量均为非负数），得到等价不等式

$$\varepsilon < (1-\gamma)\,\mu^{\frac{N(\varepsilon)-1}{3}}\,\frac{1-\mu^{\frac{2}{3}N(\varepsilon)}}{1-\mu^{\frac{2}{3}}}. \tag{25}$$

根据 $N(\varepsilon)$ 的上取整定义，我们有如下夹逼关系

$$\mu^{N(\varepsilon)} \;\le\; \frac{\varepsilon}{2e_0} \;<\; \mu^{N(\varepsilon)-1}.$$

由于 $\mu \in (0,1)$，这意味着

$$\mu^{\frac{N(\varepsilon)-1}{3}} > \left(\frac{\varepsilon}{2e_0}\right)^{1/3}, \qquad 1-\mu^{\frac{2}{3}N(\varepsilon)} \ge 1-\left(\frac{\varepsilon}{2e_0}\right)^{2/3}.$$

将这些界代入 (25)，得到充分条件

$$\varepsilon < (1-\gamma)\left(\frac{\varepsilon}{2e_0}\right)^{1/3} \frac{1-\left(\frac{\varepsilon}{2e_0}\right)^{2/3}}{1-\mu^{2/3}}. \tag{26}$$

令 $x := \left(\frac{\varepsilon}{2e_0}\right)^{1/3} > 0$。则 $\varepsilon = 2e_0x^3$，且 (26) 变为

$$2e_0x^3 < (1-\gamma)\,x\,\frac{1-x^2}{1-\mu^{2/3}}.$$

除以 x 并令 $\nu := \frac{1-\gamma}{1-\mu^{2/3}} > 0$，得到

$$2e_0x^2 < \nu(1-x^2) \quad\Longleftrightarrow\quad (2e_0+\nu)x^2 < \nu \quad\Longleftrightarrow\quad x^2 < \frac{\nu}{2e_0+\nu}.$$

回到 ε，得到显式充分界

$$\varepsilon < 2e_0\left(\frac{\nu}{2e_0+\nu}\right)^{3/2}$$

这保证了 $K_0(\varepsilon) > 4c_2(1-\gamma)^2$。
总代价可通过应用几何级数计算如下

$$\mathrm{cost}(Q^{(0)}_{N(\varepsilon)}) \;=\; \sum_{j=0}^{N(\varepsilon)-1} \max(C_\varepsilon \mu^{\frac{2}{3}(N(\varepsilon)-1-j)}, K_{\min}) \;=\; \left(\frac{\varepsilon}{2\sqrt{c_2}}\right)^{-2}\left(\frac{1-\mu^{\frac{2}{3}N(\varepsilon)}}{1-\mu^{2/3}}\right)^3.$$

此外，我们使用 $\mu^{N(\varepsilon)} \le \varepsilon/(2e_0)$，使得 $\mu^{\frac{2}{3}N(\varepsilon)} \to 0$ 为 $\varepsilon \to 0$，以及 $1-\mu = \frac{1}{2}(1-\gamma)$，使得

$$\left(\frac{1-\mu^{\frac{2}{3}N(\varepsilon)}}{1-\mu^{2/3}}\right)^3 = \mathcal{O}\Big(\frac{1}{(1-\mu^{2/3})^3}\Big) = \mathcal{O}\Big(\frac{1}{(1-\mu)^3}\Big) = \mathcal{O}\Big(\frac{1}{(1-\gamma)^3}\Big),$$

其中倒数第二步使用了 $1-\mu^{2/3} \ge \frac{2}{3}(1-\mu)$，这是由于 $x \mapsto x^{2/3}$ 在 $[0,1]$ 上具有凹性。结合式 (23) 中的 $c_2 \lesssim \frac{1}{\xi^2}\big(\frac{1-2\gamma+2\gamma^2}{1-\gamma}\big) \lesssim \xi^{-2}(1-\gamma)^{-2}$，我们有

$$\mathrm{cost}(Q^{(0)}_{N(\varepsilon)}) = \mathcal{O}\Big(\frac{1}{\xi^2(1-\gamma)^5\varepsilon^2}\Big). \tag{27}$$

推导的频率更新具有准最优性的论证与证明定理 5.1 的论证一致，即

$$\begin{aligned}
&\sum_{j=0}^{N(2\varepsilon)-1} K_j(2\varepsilon) = \left(\frac{\varepsilon}{\sqrt{c_2}}\right)^{-2} \left(\frac{1-\gamma^{\frac{2}{3}N(2\varepsilon)}}{1-\gamma^{2/3}}\right)^3 \\
&\leq \inf\Big\{ \sum_{j=0}^{\hat{N}-1} \hat{K}_j : e_{\hat{N}}(\hat{K}_j,\ j=0,\dots,\hat{N}-1) \leq \varepsilon,\ \hat{N}\in\mathbb{N},\ \hat{K}_j > 0\ \forall j < \hat{N}-1 \Big\} \\
&\leq \sum_{j=0}^{N(\varepsilon)-1} K_j(\varepsilon) = \left(\frac{\varepsilon}{2\sqrt{c_2}}\right)^{-2} \left(\frac{1-\gamma^{\frac{2}{3}N(\varepsilon)}}{1-\gamma^{2/3}}\right)^3 .
\end{aligned}$$

□

D 展望：准确率触发的目标更新

下面，我们将详细介绍第 6 节中提到的准确率触发的目标更新。本文提出的贝尔曼压缩近似视角表明，可以自然地将方法扩展到预先设定的 TUF 之外。回顾外循环动态满足

$$\mathbb{E}[E_{n+1} \mid \mathcal{F}_{n,0}] \ \leq\ \gamma E_n + \eta_n,$$

其中， η_n 表示周期 n中的内循环近似误差。这一递推关系表明，外循环的进展主要取决于能否得到足够准确的贝尔曼算子近似，而非内循环迭代次数本身。受此观察启发，我们考虑准确率触发的目标更新：当内循环近似误差低于预设阈值时，同步目标网络。我们不再预先固定 TUF K_n ，而是指定一系列精度级别（ε_n）$_{n\geq 1}$ ，并在达到相应精度判据时立即终止内循环。这样，每个周期的计算开销便可根据贝尔曼算子近似的局部难度进行调整。

表格型设置中的在线估计。在表格情形下，我们概述一种启发式实用机制，用于在内循环中在线估计 η_n 。回顾一下，冻结目标值的表格型 Q-learning 可视为针对二次贝尔曼回归目标函数的随机梯度下降。时序差分误差 $\delta = r + \gamma \max_{a'} Q^{-}(s',\ a') - Q(s,\ a)$ 是相应单样本损失梯度的一个噪声样本。对于每个状态–动作对（$s,\ a$），我们维护观测到的时序差分误差的经验均值，并通过标准平均法在线更新。随后，我们将这些局部统计量汇总为一个全局停止判据

$$M \ =\ \frac{1}{|\mathcal{S}||\mathcal{A}|} \sum_{(s,a)} |\bar{\delta}(s,a)|,$$

只有当所有状态–动作对都表现出持续较小的平均时序差分误差时，该值才会较小。由于表格型贝尔曼回归问题是强凸的，并且其唯一极小值点由梯度消失刻画，因此条件 $M \approx 0$ 表明，整个状态–动作空间上的内循环优化问题都得到了近似求解。

为确保收敛，近似程度必须在各个周期内可求和，这是命题 4.1 所要求的。因此，一个简单的可行选择是多项式衰减序列$\varepsilon_n = n^{-p},\ p > 1,$ ，其满足 $\sum_{n\geq 1} \varepsilon_n < \infty$。在我们的实现中，我们使用具体实例$\varepsilon_n = n^{-2}$。

Alg算法 3周期性 Q-学习 g（表格型）以准确率 y-trigg触发的目标更新 get up更新（ATQL

）1：输入：最大 TUF $K_{\max}$**、最小 TUF** $K_{\min}$**、目标精度** $\{\varepsilon_n\}_{n=1}^{N}$**2：初始化** Q**：将** $\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ **任意设置，** $Q^- = Q$**3：对于从** $n=1$ **到** N **的外层贝尔曼迭代，执行 4：为所有** (s,a) $\in\mathcal{S}\times\mathcal{A}$ **初始化计数器** $c(s,a)\leftarrow 0$ **和均值** $\mu(s,a)\leftarrow 0$ **5：对于从** $k=1$ **到** $K_{\max}$ **的内层随机梯度下降迭代，执行 6：观测当前状态** s**，选择动作** a**，例如** $a\sim\varepsilon$**-贪心(**$Q(\cdot,\cdot)$**)，执行动作** a**，观测**r**、** s'**7：计算时序差分误差** $\delta\leftarrow r+\gamma\max_{a'\in\mathcal{A}} Q^-(s',a') - Q(s,a)$ **8：更新**$Q(s,a)\leftarrow Q(s,a)$ $+\alpha\,\delta$**9：** $c(s,a)\leftarrow c(s,a)+1$**10：** $\mu(s,a)\leftarrow\mu(s,a)+\frac{\delta-\mu(s,a)}{c(s,a)}$

11 $M\leftarrow\frac{1}{|\mathcal{S}\times\mathcal{A}|}\sum \tilde{s}\,\tilde{a}\ |\mu\ \tilde{s}\ \tilde{a}\ |$ **：（，）（，）：如果，然后 13：break 14：结束 if 15：**

12 $k\geq K_{\min}\,n$ $M\leq\varepsilon_n$

结束 for 16：$Q^-\leftarrow Q$**17：结束 for**

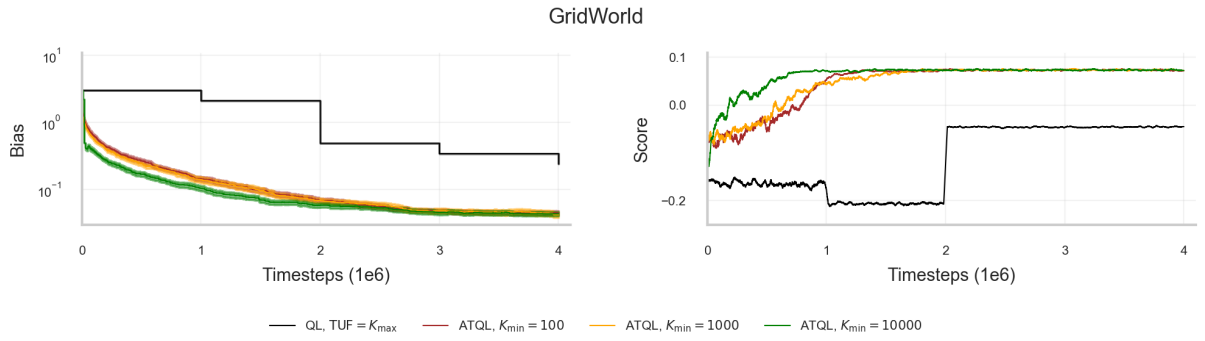


图 5：在附录 A.2 的网格世界环境中，采用精度触发的目标更新的 Q 学习，$K_{\min}\in\{100,1000,10000\}$ 和$K_{\max}=1e6$。目标精度由 $\varepsilon_n=\frac{1}{n^2}$ 选定，其依据是命题 4.1 中的收敛条件。作为对比，我们纳入了目标更新频率为 $=K_{\max}=1e6$ 的 Q 学习，以说明精度触发的目标更新所产生的影响。对于贝尔曼近似，不必要地过长的内循环会提前终止，从而改善收敛行为。