arXiv:2301.03142v1 [stat.ML] 9 Jan 2023

# 带随机化奖励的基于模型的强化学习中的探索

Lingxiao Wang、PingLi

认知计算实验室
百度研究院
美国华盛顿州贝尔维尤市东北第8街10900号，邮编98004

## 摘要

[1]基于模型的强化学习（MBRL）因其样本效率高而得到广泛应用。然而，现有的最坏情况遗憾分析通常要求采用乐观规划，这在一般情况下并不现实。相比之下，受理论启发，实证研究采用模型集成，并在各种测试环境中取得了最先进的性能。理论与实证研究之间的这种偏差促使我们思考：随机化模型集成是否能保证乐观性，进而保证最优的最坏情况遗憾？本文从奖励随机化的角度对这一问题作出部分回答；奖励随机化是基于模型的强化学习中一种鲜少探索的探索方向。我们证明，在核化线性调节器（KNR）模型下，奖励随机化能够保证部分乐观性，进而就交互次数而言实现近乎最优的最坏情况遗憾。我们还将理论推广到广义函数逼近，并确定了奖励随机化实现可证明高效探索所需的条件。相应地，我们提出了奖励高效随机化的具体示例。据我们所知，我们的分析首次针对采用函数逼近的随机化基于模型的强化学习建立了最坏情况遗憾分析。

1. 本稿于2021年9月完成，当时两位作者均在百度认知计算实验室工作。

© .

## 1. 引言

强化学习（RL）（Sutton 和 Barto，2018年）旨在通过与环境反复交互来学习最优策略。基于模型的强化学习（MBRL）（Osband 和 Roy，2014年；Luo 等人，2019年；Ha 和 Schmidhuber，2018年；Luo 等人，2019年；Sun 等人，2019年；Kaiser 等人，2020年；Ayoub 等人，2020年；Kakade 等人，2020年）通过根据观测拟合环境，并从拟合得到的环境中获取策略，来实现这一目标。结合深度学习，基于模型的强化学习在现实任务中取得了巨大成功，包括电子游戏（Ha 和 Schmidhuber，2018年；Kaiser 等人，2020年）和控制任务（Watter 等人，2015年；Williams 等人，2015年；Chua 等人，2018年；Hafner 等人，2019年；Song 和 Sun，2021年）。

基于模型的强化学习成功的一个关键因素是样本效率。从理论分析的角度来看，样本效率由基于模型的强化学习的遗憾分析来刻画。先前的分析表明，结合探索策略后，基于模型的强化学习可实现近乎最优的 $\widetilde{\mathcal{O}}(\sqrt{T})$ 遗憾（Jaksch 等人，2010年；Ayoub 等人，2020年；Kakade 等人，2020年），其中 $T$ 是与环境交互的总次数。然而，此前有理论保证的高效探索通常采用乐观规划（Jaksch 等人，2010年；Luo 等人，2019年；Ayoub 等人，2020年；Kakade 等人，2020年）。这种探索策略需要：(i) 确定一个模型置信集 $\mathcal{D}$，以刻画模型估计中的不确定性；然后 (ii) 在 $\mathcal{D}$ 中的所有可能模型范围内搜索最优策略，进行乐观规划。乐观规划成功的关键在于面对不确定性原理时保持乐观（Jaksch 等人，2010年）。直观而言，乐观规划会促使智能体探索访问较少的区域，从而提高相应强化学习算法的样本效率。虽然步骤 (i) 可借助集成技术实现，但步骤 (ii) 通常无法实现，因为它需要在可能为连续空间的模型集合 $\mathcal{D}$ 上求解优化问题。作为替代方案，先前的实证研究（Chua 等人，2018年；Pathak 等人，2017年、2019年）通常借鉴乐观规划的思路以及基于 Thompson采样（TS）的算法研究（Osband 和 Roy，2014年）。一种常见的实证方法是利用模型集成来捕捉模型估计的不确定性，并进一步通过 TS（Chua 等人，2018年）或奖励构造（Pathak 等人，2017年、2019年）将此类集成用于规划。与乐观规划不同，这些方法通常不具备最坏情况遗憾保证。尽管如此，它们在各种测试环境中都达到了最先进的性能。理论与实践之间的这种差异促使我们提出以下问题：

随机化模型集成能否保证乐观主义，从而实现最优的最坏情况遗憾?

本文从奖励随机化这一角度，对上述问题给出了部分解答。奖励随机化是一种在基于模型的强化学习中相对较少研究的探索方法。我们首先在核化线性调节器（KNR）转移模型（Mania 等人，2022年；Kakade 等人，2020年；Song 和 Sun，2021年）及奖励函数已知的条件下展开分析。我们提出了 PlanEx，它通过反复使用拟合转移模型和随机化奖励函数进行规划来开展探索。我们进一步证明，PlanEx 可实现近乎最优的 $\widetilde{\mathcal{O}}(\sqrt{T})$ 遗憾。奖励随机化的一个关键观察是部分乐观性的概念（Russo，2019年；Zanette 等人，2020年），它确保在乐观原则下，有足够多的交互用于探索。受基于 KNR 转移模型的分析启发，我们将 PlanEx 扩展到具有校准模型的一般函数逼近情形（Curi 等人，2020年；Kidambi 等，2021），并提出了一种通用的奖励随机化设计原则。我们还提出了有效奖励随机化的具体示例

并从理论上展示了奖励随机化的有效性。我们强调，所提出的奖励随机化方法可以基于模型集成轻松实现。此外，奖励随机化具有很强的模块化特性，可以与各种最先进的基线方法结合使用。

贡献。我们的工作为所提出的问题提供了部分解决方案。具体而言，我们研究了奖励随机化，并提出了 PlanEx。我们的贡献如下。

- 我们提出了 PlanEx，这是一种用于基于模型的强化学习的全新探索算法，在一般函数参数化下可实现，并具有最坏情况遗憾保证。
- 我们证明，在 KNR 动力学下，PlanEx 的最坏情况遗憾接近最优。

据我们所知，我们的分析首次对采用函数逼近的随机化基于模型的强化学习进行了最坏情况遗憾分析。

相关工作。我们的工作与基于模型的强化学习和在线控制问题的遗憾分析密切相关（Osband 和 Roy，2014年；Luo 等人，2019年；Sun 等人，2019年；Lu 和 Roy，2019年；Ayoub 等人，2020年；Kakade 等人（2020）；Curi 等人，2020年；Agarwal 等人，2020b；Song 和 Sun，2021）。Ayoub 等人（2020年）提出了价值目标回归（VTR）算法，重点关注与强化学习相关的转移模型方面。Agarwal 等人（2020b）提出了 FLAMBE，这是一种在线性马尔可夫决策过程设定下具有理论效率保证的基于模型的强化学习算法（Jin 等人，2020年；Yang 和 Wang，2019年）。Kakade 等人（2020）提出了 LC3，这是一种基于 KNR 动力学的在线控制算法（Mania 等人，2022年），可实现最优最坏情况遗憾。Ayoub 等人（2020）和 Kakade 等人（2020）都采用乐观规划进行探索，但这种方法通常难以计算。相比之下，我们利用奖励随机化进行探索，同样可以实现最优最坏情况遗憾，而且计算起来非常可行。为实现计算可行的乐观规划，Curi 等人（2020）设计了 HUCRL，在规划优化中引入了额外的状态偏差变量。相比之下，随机化奖励规划不会在优化中引入额外变量。Luo 等人（2019）优化价值函数的下界，从而避免显式量化不确定性。近期研究也利用奖励项实现乐观规划（Kidambi 等，2021；Song 和 Sun，2021）。Song 和 Sun（2021）提出了 PC-MLP，该算法通过估计策略覆盖（Agarwal 等人，2020a）来构造奖励项，且计算上可行。相比之下，PC-MLP 需要额外采样来估计策略覆盖的协变量矩阵。因此，PC-MLP 无法实现最优的 $\widetilde{\mathcal{O}}(\sqrt{T})$-遗憾。相比之下，PlanEx 无需额外采样即可构造奖励项，并能实现 $\widetilde{\mathcal{O}}(\sqrt{T})$-遗憾。此外，为实现计算可行的实例化，PC-MLP 在模型拟合和奖励构造中采用了不同的特征，这与理论分析不一致。相比之下，在校准模型假设下，PlanEx 的实现与理论分析一致。以往研究也探索了采用函数逼近的高效无模型强化学习探索算法。例如，关于这方面的研究，参见 Jiang 等人（2017年）；Jin 等人（2020年）；Du 等人（2020年）；Wang 等人（2020年）；Cai 等人（2020年）；Agarwal 等人（2020a）；Modi 等人（2021年）及其中的参考文献。

我们的分析受到随机化强化学习算法最坏情况遗憾分析的最新进展启发（Russo，2019年；Pacchiano 等人，2020年；Zanette 等人，2020年；Ishfaq 等人，2021年）。我们的乐观性分析受到 Russo （2019年）和 Zanette 等人 （2020年）的启发。 Russo （2019年）首次针对表格型设定下的随机化最小二乘价值迭代（RLSVI）算法提出了最坏情况遗憾分析。 Zanette 等人 （2020年）将 RLSVI 的分析扩展到了截断线性

一般状态空间下的函数逼近。 Ishfaq 等人 （2021年）分析了同时采用线性函数近似和一般函数逼近的随机化 Q 学习。相比之下，我们关注随机化模型式强化学习算法。 Pacchiano 等人 （2020年）分析了同时采用奖励随机化和转移随机化的基于模型的强化学习的最坏情况遗憾。我们指出，Pacchiano 等人（2020年）和 Ishfaq 等人 （2021年）都要求在规划中为每个状态-动作对抽取多个样本，并进一步对所有随机化奖励函数取最大值。相比之下，我们每个时间步只需要一个样本，也无需进一步取最大值。此外，Pacchiano 等人（2020年）关注的是具有有限状态和动作空间的表格型设定，而我们考虑的是采用函数逼近的一般设定。

## 2. 背景

### 2.1. 强化学习

本文使用幕式马尔可夫决策过程（$\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h\in[H]}, \mathbb{P}$）对环境进行建模。其中， $\mathcal{S}$和 $\mathcal{A}$ 是状态空间， $H$ 是幕的长度， $r_h: \mathcal{S}\times\mathcal{A}\mapsto[0,1]$ 是针对 $h\in[H]$的有界奖励函数，而 $\mathbb{P}$ 是转移核，它定义了所有 $h\in[H]$ 和$(s_h, a_h)\in\mathcal{S}\times\mathcal{A}$的转移概率$s_{h+1}\sim\mathbb{P}(\cdot\,|\,s_h, a_h)$。交互过程。一个智能体

拥有一组策略 $\{\pi_h\}_{h\in[H]}$，并按如下方式与该环境交互。智能体从固定的初始状态 $s_1\in\mathcal{S}$ 开始。每当到达状态 $s_h\in\mathcal{S}$时，智能体便执行动作$a_h=\pi_h(s_h)$。随后，智能体收到奖励 $r_h(s_h, a_h)$。环境根据概率 $\mathbb{P}(\cdot\,|\,s_h, a_h)$转移至下一个状态 $s_{h+1}$。当智能体到达状态 $s_{H+1}$时，过程结束。

为描述期望累积奖励，对于每个策略 $\pi=\{\pi_h\}_{h\in[H]}$，我们引入动作价值函数 $\{Q_h^\pi\}_{h\in[H]}$，定义如下：

$$Q_h^\pi(s_h, a_h; \{r_h\}_{h\in[H]}, \mathbb{P}) = \sum_{\tau=h}^{H}\mathbb{E}\big[r_\tau(s_\tau, a_\tau)\,\big|\,s_h, a_h, \pi\big], \quad \forall h\in[H],\ (s_h, a_h)\in\mathcal{S}\times\mathcal{A}, \tag{1}$$

其中，对于所有 $\tau=h$，……，$H$，$a_\tau=\pi_\tau(s_\tau)$ 和 $s_{\tau+1}\sim\mathbb{P}(\cdot\,|\,s_\tau, a_\tau)$。类似地，我们定义价值函数 $\{V_h^\pi\}_{h\in[H]}$ 如下：

$$V_h^\pi(s_h; \{r_h\}_{h\in[H]}, \mathbb{P}) = \sum_{\tau=h}^{H}\mathbb{E}\big[r_\tau(s_\tau, a_\tau)\,\big|\,s_h, \pi\big], \quad \forall h\in[H],\ (s_h, a_h)\in\mathcal{S}\times\mathcal{A}. \tag{2}$$

我们将最优策略 $\pi^*=\{\pi_h^*\}_{h\in[H]}$ 定义为以下优化问题的最大化者：

$$\pi^* = \operatorname*{argmax}_\pi V_1^\pi(s_1; \{r_h\}_{h\in[H]}, \mathbb{P}). \tag{3}$$

相应地，我们将 $V^*$ 和 $Q^*$ 定义为与最优策略 $\pi^*$ 对应的价值函数和动作价值函数。

强化学习（RL）的目标是根据先前经验，依次选择策略 $\pi^k=\{\pi_h^k\}_{h\in[H]}$，以最大化智能体在交互过程中获得的期望累积奖励。等价地，目标是最小化以下遗憾：

$$R(K) = \sum_{k=1}^{K} V^*(s_1) - V^{\pi^k}(s_1), \tag{4}$$

其中， $K$ 是交互总次数， $s_1$ 是固定的初始状态。直观而言，遗憾 $R(K)$ 描述了交互过程中执行的策略与最优策略之间的偏差。

### 2.2. 在线非线性控制问题

W我们考虑具有如下状态转移动态的在线非线性控制问题 s,

$$s_{h+1} = f(s_h, a_h) + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2 \cdot I), \quad \forall h \in [H],\ (s_h, a_h) \in \mathcal{S} \times \mathcal{A}.$$

此处，函数 $f:\ \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$ 属于一个具有已知核的再生核希尔伯特空间（RKHS），且噪声 $\epsilon$ 在各次状态转移之间相互独立。以往研究也将此类状态转移称为核化非线性调节器（KNR）（Kakade 等人，2020年；Song 和 Sun，2021年）。本文沿用Mania 等人（2022年）；Kakade 等人 （2020年）；Song 和 Sun （2021年）的做法，并考虑将此类状态转移动态的原始形式作为强化学习问题的底层转移动态，其定义如下，

$$\begin{aligned} &s_{h+1} = f(s_h, a_h; W^*) + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2 \cdot I), \\ &f(s_h, a_h; W^*) = W^* \phi(s_h, a_h), \quad \forall h \in [H],\ (s_h, a_h) \in \mathcal{S} \times \mathcal{A}. \end{aligned} \tag{5}$$

此处， $\phi:\ \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_\phi}$ 是一个已知的特征嵌入。同时，状态空间 $\mathcal{S} \subseteq \mathbb{R}^{d_\mathcal{S}}$ 是欧几里得空间的一个子集，其维度为 $d_\mathcal{S}$，而 $W^* \in \mathbb{R}^{d_\mathcal{S} \times d_\phi}$ 是 KNR 转移动态的未知真实参数。

相应地，在下文中，我们用 $Q^\pi(\cdot, \cdot;\ \{r_h\}_{h\in[H]}, W)$ 和 $V^\pi(\cdot;\ \{r_h\}_{h\in[H]}, W)$ 表示策略 $\pi$ 在奖励函数 $\{r_h\}_{h\in[H]}$ 和由矩阵 $W \in \mathbb{R}^{d_\mathcal{S} \times d_\phi}$ 定义的转移动态下的价值函数。为简化分析，我们固定特征和参数的以下缩放方式。

假设1（归一化模型）我们假设 $\|\phi（s,\ a）\|_2 \le 1/\sqrt{H}$ 对所有$s,\ a）\ \in \mathcal{S} \times \mathcal{A}$。相应地，我们假设 $\|W^*\|_2 = \mathcal{O}（\sqrt{H}）$，其中 $W^*$ 是由（5）定义的 KNR 转移动态的真实参数。

类似的归一化假设也出现在Mania 等人（2022年）中。我们指出，假设1中的缩放假设只影响遗憾中 $H$ 的收敛速率，提出该假设是为了简化分析。

### 2.3. 针对未知转移动态的基于模型的强化学习

在基于模型的强化学习中，智能体通过迭代地根据收集到的数据拟合转移动态，并基于拟合后的转移动态进行最优规划，从而优化策略。对于每次迭代 $k$，基于模型的强化学习包括以下步骤。

- (i) 模型拟合。在此步骤中，智能体根据回放缓冲区 $\mathcal{D}_k$ 更新转移动态的参数 $W_k$ 。
- (ii) 规划。在此步骤中，智能体基于拟合得到的转移动态参数$W_k$ 进行最优规划。通过使用拟合模型进行规划，智能体更新策略$\pi^k$。

• (iii) 交互。在此步骤中，智能体使用策略 $\pi^k$与环境交互，并收集一条轨迹 $\iota_k = (s_1^k, a_1^k, \ldots, s_H^k, a_H^k, s_{H+1}^k)$。随后，智能体通过 $\mathcal{D}_{k+1} = \mathcal{D}_k \cup \iota_k$更新回放缓冲区。

接下来，我们提出以下假设，该假设认为我们可以调用规划预言机来处理步骤（ii）中的规划。

假设 2（规划预言机）我们假设可以调用预言机 Plan$(\cdot, \cdot, \cdot)$。对于任意输入的奖励函数 $\{r_h\}_{h\in[H]}$和转移动态的参数 $W$，该预言机都会返回最优策略 $\pi = \text{Plan}(s_1, \{r_h\}_{h\in[H]}, W)$。

备注 3（关于样本复杂度的备注）在实践中，针对拟合环境的规划通常由深度强化学习算法（Pathak 等人，2017年；Luo 等人，2019年；Pathak 等人，2019年；Song 和 Sun，2021年）或模型预测控制（Williams 等人，2015年；Chua 等人，2018年；Kakade 等人，2020年）来处理。我们指出，由于此类规划是在拟合环境中进行的，求解这类规划问题不会引发求解器样本复杂度方面的担忧。相比之下，在步骤（iii）中与真实环境交互时，则会产生此类样本复杂度方面的担忧。我们指出，探索的目标是以尽可能少的交互轮数 $K$获得近乎最优的策略。若采用式（4）中定义的遗憾 $R(K)$ 来衡量，探索的目标就是设计强化学习算法，使遗憾 $R(K)$ 随 $K$ 增长得尽可能慢。

## 3. 随机化奖励下的探索

本节提出 PlanEx，这是一种针对 KNR 动力学强化学习问题的、可证明高效且可实现的算法。接下来，我们将介绍 PlanEx 在第 $k$ 次迭代中各步骤的流程。

(i) 模型拟合。给定数据集 $\mathcal{D}_k = \{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)_{(h,\tau)\in[H]\times[k-1]}\}$，我们通过最小化给定 $(s_h, a_h)$ 时 $s_{h+1}$ 的预测误差来拟合转移参数 $W^k$。具体而言，我们最小化以下最小二乘损失：

$$W^k \leftarrow \operatorname*{argmin}_{W\in\mathbb{R}^{d_{\mathcal{S}}\times d_\phi}} \sum_{h=1}^{H}\sum_{\tau=1}^{k-1} \|s_{h+1}^\tau - W\phi(s_h^\tau, a_h^\tau)\|_2^2 + \lambda\cdot\|W\|_F^2, \tag{6}$$

其中，我们用 $\|\cdot\|_F$ 表示矩阵的弗罗贝尼乌斯范数。(6) 中的优化问题有如下显式解：

$$W^k \leftarrow \left(\sum_{h=1}^{H}\sum_{\tau=1}^{k-1} s_{h+1}^\tau \phi(s_h^\tau, a_h^\tau)^\top\right)\Lambda_k^{-1}, \quad \Lambda_k = \sum_{h=1}^{H}\sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I. \tag{7}$$

(ii) 规划。在规划阶段，我们旨在推导出一个与环境交互的策略 $\pi^k$。

在推导策略 $\pi^k$ 时，我们旨在实现两个目标。(a) 首先，策略 $\pi^k$ 应当充分利用我们对环境的了解，以优化累积奖励。(b) 其次，策略 $\pi^k$ 还应将我们对环境的不确定性纳入考量，并针对尚未探索的关键事件开展探索。为了在 (a) 和 (b) 之间取得适当平衡，我们需要量化对环境的不确定性。这种不确定性量化可以通过

矩阵 $\Lambda_k$ 在（7）中定义。具体而言，已知（），矩阵 $\Lambda_k$ 定义了以下置信区域

$$\mathcal{G}_k = \left\{W \in \mathbb{R}^{d_{\mathcal{S}} \times d_\phi} : \|(W - W^k)\Lambda_k^{1/2}\|_2^2 \le \beta_k\right\}.$$

对于适当设置的 $\beta_k$，已知 $W^* \in \mathcal{G}_k$ 以较高概率成立。在这种观测下，先前的尝试（Jaksch 等人，2010年；Kakade 等人，2020年）通过求出 $\pi$ 的最大化者 $V^\pi(s_1; \{r_h\}_{h\in[H]}, W)$，在 $W \in \mathcal{G}_k$上实现了利用与探索之间的平衡，但这在计算上难以处理。为提出一种计算上可行的替代方案，先前的经验方法利用集成模型，在有限观测下拟合模型时估计认知不确定性。

在这项工作中，我们研究将不确定性直接纳入奖励函数的方法。为此，我们引入以下扰动奖励函数，

$$r_{h,\xi}^k(s_h, a_h) = \left\{r_h(s_h, a_h) + \phi(s_h, a_h)^\top \xi_h^k\right\}^+, \quad \forall h \in [H],\ (s_h, a_h) \in \mathcal{S} \times \mathcal{A}, \tag{8}$$

其中，噪声 $\{\xi_h^k\}_{h\in[H]}$ 从高斯分布 $\xi_h^k \sim N(0, \sigma_k^2 \cdot \Lambda_k^{-1})$ 中独立采样。直观而言，智能体探索较少的区域中，此类噪声的方差较大；而在探索充分的区域中，方差较小。此外，我们对奖励进行截断，以确保奖励为正。

扰动奖励后，我们通过基于估计的状态转移和扰动奖励进行规划来更新策略，如下所示，

$$\pi^k = \{\pi_h^k\}_{h\in[H]} = \mathrm{Plan}\left(s_1, \{r_{h,\xi}^k\}_{h\in[H]}, W^k\right). \tag{9}$$

我们指出，式（8）中的奖励扰动是在规划之前进行的。式（8）中定义的扰动奖励在整个规划阶段保持不变。我们在算法 `PlanEx` 1中总结了 。

## 4. 理论分析

本节中，我们分析算法 `PlanEx` 1。我们的关键观察如下：

- PlanEx 中的奖励扰动使得至少在一定比例的交互中能够进行乐观规划，并且

- PlanEx 所保证的这种部分乐观性足以支持探索。

### 4.1. 部分乐观性

在下文中，我们将证明 PlanEx 具有部分乐观性。具体而言，以下引理成立。

引理 4（部分乐观性）在良好事件 $W^* \in \mathcal{G}_k = \{W \in \mathbb{R}^{d_{\mathcal{S}} \times d_\phi} : \|(W - W^k)\Lambda_k^{1/2}\|_2^2 \le \beta_k\}$ 下，对于适当选取的 $\sigma_k$，以下结论以至少 $\Phi(-1)$ 的概率成立：

$$V_1^*\left(s_1; \{r_h\}_{h\in[H]}, W^*\right) - V_1^{\pi^k}\left(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\right) \le 0, \tag{10}$$

其中 $\Phi(\cdot)$ 是标准高斯分布的累积分布函数。

算法 1：随机化奖励规划

要求：数据集 $\mathcal{D}$、奖励 $\{r_h\}_{h\in[H]}$。1：初始化：设置 $\Lambda_1 = \lambda \cdot I$。2：对于 $k = 1$、2、.....、$K$，执行 3：为所有 $h \in [H]$ 生成一组独立噪声 $\xi_h^k \sim N(0, \sigma_k^2 \cdot \Lambda_k^{-1})$。4：获取扰动奖励

$$r_{h,\xi}^k(s_h, a_h) = \{r_h(s_h, a_h) + \phi(s_h, a_h)^\top \xi_h^k\}^+, \quad \forall (s_h, a_h) \in \mathcal{S} \times \mathcal{A},\ h \in [H].$$

5：通过调用规划预言机，获得策略 $\pi^k$

$$\pi^k = \text{Plan}(s_1, \{r_{h,\xi}^k\}_{h\in[H]}, W^k).$$

6：执行 $\pi^k$，采样一条轨迹 $\tau^k = \{s_1^k, a_1^k, s_2^k, ..., s_H^k, a_H^k, s_{H+1}^k\}$。

7：更新数据集 $\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \cup \tau^k$。

8：更新模型和协变量矩阵

$$W^{k+1} \leftarrow \underset{W\in\mathbb{R}^{\mathcal{S}\times d}}{\text{argmin}} \sum_{h=1}^{H}\sum_{\tau=0}^{k} \|s_{h+1}^\tau - W\phi(s_h^\tau, a_h^\tau)\|_2^2 + \lambda \cdot \|W\|_2^2,$$

$$\Lambda_{k+1} \leftarrow \Lambda_k + \sum_{h=1}^{H} \phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top.$$

9：结束循环

证明 详见 §A.2了解详细证明。 ■

引理 4 保证至少有 $\Phi(-1)$ 的 PlanEx价值函数估计高估了我们希望得到的最优价值函数 $V_1^*(\cdot;\ \{r_h\}_{h\in[H]},\ W^*)$。因此，PlanEx中的随机化奖励保证至少有 $\Phi(-1)$ 的轨迹会在乐观原则下促进探索。直观而言，这种乐观主义成立的原因是：(i) 一方面，随机高斯扰动确保扰动奖励有足够大的概率高于真实奖励；(ii) 另一方面，良好事件 $\mathcal{G}_k$ 确保基于真实模型估计的价值函数 $(\{r_h\}_{h\in[H]},\ W^*)$ 与未受扰动时基于当前模型估计的价值函数 $(\{r_h\}_{h\in[H]},\ W^k)$ 之间的偏差不会太大。

### 4.2. 遗憾分析

我们强调，引理 4 中的乐观主义保证本身并不能保证最优遗憾。为了进行合理的探索，除了乐观主义之外，我们还需要确保扰动奖励引起的高估不会偏离真实奖励下的价值函数太远。在我们的工作中，我们通过将不确定性适当地纳入转移动态，来保证这种偏差不会过大。具体而言，回顾一下我们将奖励扰动定义如下

$$r_{h,\xi}^k(s_h,a_h)=\left\{r_h(s_h,a_h)+\phi(s_h,a_h)^\top\xi_h^k\right\}^+,\quad \forall h\in[H],\ (s_h,a_h)\in\mathcal{S}\times\mathcal{A},$$

其中，噪声 $\{\xi_h^k\}_{h\in[H]}$ 独立采样自高斯分布 $\xi_h^k\sim N(0,\ \sigma_k^2\cdot\Lambda_k^{-1})$。这种扰动引入了噪声 $\xi_h^k$，其方差随模型不确定性$\Lambda_k$而变化。这种奖励扰动确保以较高概率，扰动奖励下价值估计的偏差与式（7）中转移模型估计的误差成比例。因此，只要模型估计合理，例如最小化式（7）中的最小二乘误差，扰动奖励引起的高估就很小。具体而言，以下定理保证 PlanEx具有最优遗憾，其阶为 $K$。

定理5 设 $\lambda=1$ 和 $\sigma_k^2=H^3\cdot\beta_k/\sigma^2$ 满足 $\beta_k$ 附录 *A.1*中给出的条件。在假设*1* 和 *2*下，对 $K>1/\Phi(-1)$ 有

$$\mathbb{E}\left[R(K)\right]=\mathcal{O}\Big((d_{\mathcal{S}}+d_\phi)^{3/2}\cdot H^{7/2}\cdot\log^2(K)\cdot\sqrt{K}\Big),$$

其中，期望是针对*PlanEx*中的随机化奖励扰动和轨迹采样计算的。

证明 详细证明见 §A.4。 ■

我们指出，定理 5 中的交互次数 $K$ （与环境的交互）（Jiang等人，2017年）的速率在信息论意义上是最优的。我们还指出，与 LC3 等最优规划方法相比（Kakade 等人，2020年），PlanEx在$H$、 $d_\phi$ 和 $d_{\mathcal{S}}$方面存在额外的依赖项，这是由奖励的随机扰动引起的。此外，我们强调，与 PC-MLP 相比（Song 和 Sun，2021年），我们的算法在 $\mathcal{O}(\sqrt{K})$ 方面关于 $K$达到了最优依赖关系。样本效率更高，是因为 PlanEx无需额外采样来计算策略覆盖矩阵，而 PC-MLP 则需要这样做。

基于模型不确定性的探索。 我们指出，基于 Thompson采样的高概率乐观主义通常出现在强化学习随机化价值迭代的分析中（Russo，2019年；Zanette 等人，2020年）。相比之下，我们的工作将这一思想用于基于模型的探索。为了理解这一思想在基于模型的探索中的对应形式，我们指出，无论是基于模型还是无模型的探索，要设计可证明有效的探索方法，关键都在于将模型不确定性纳入价值函数及其对应的策略中。在 LSVI-UCB 等基于价值的方法中（Jin 等人，2020年），这种模型不确定性通过对目标价值函数进行回归来估计，回归对象为 $s_{h+1}$，对应于 （$s_h$， $a_h$），并作为奖励项纳入价值函数。在 UCRL 及其变体等基于模型的方法中（Jaksch 等人，2010年；Kakade 等人，2020年），这种模型不确定性由转移动态的置信区域刻画，并通过乐观规划纳入价值函数。此外，对于采用策略覆盖的算法（Song 和 Sun，2021年；Agarwal 等人，2020a），这种模型不确定性通过汇总当前策略的访问轨迹获得。我们的工作通过基于模型不确定性直接扰动奖励函数来具体实现这一思想，而这为所有探索算法提供了一种基础视角。

## 5. 一般函数逼近下的推广

从`PlanEx`的设计中可以得到一个关键观察：只要至少有固定比例的迭代用于乐观探索，就能保证充分探索。为进一步验证这一观察，接下来我们将`PlanEx`推广到一般函数逼近。我们在算法2中总结了该算法。为进行分析，我们假设转移动态的估计足够准确，并满足以下校准模型假设。

**算法 2** 随机化奖励规划

要求：奖励 $\{r_h\}_{h\in[H]}$。

1：初始化：将缓冲区 $\mathcal{D}_0$ 初始化为空集。初始化转移动态 $\mathbb{P}^1$。

2: 对于 $k = 1, 2, \ldots, K$ **do**

3: 生成随机化奖励 $\{r^k_{h,\xi}\}_{h\in[H]}$。

4: 通过调用规划预言机 $\pi^k = \text{Plan}(s_1, \{r^k_{h,\xi}\}_{h\in[H]}, \mathbb{P}^k)$，获得策略 $\pi^k$。

5: 执行 $\pi^k$ 以采样轨迹 $\tau^k = \{s^k_1, a^k_1, s^k_2, ..., s^k_H, a^k_H, s^k_{H+1}\}$.

6: 更新数据集 $\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \cup \tau^k$。

7: 根据数据集 $\mathcal{D}_k$更新转移动态 $\mathbb{P}^{k+1}$。

8: 结束循环

假设6（校准模型）令 $\mathbb{P}^k$ 为第 $k$次迭代中估计得到的转移动态。对于所有 $\delta > 0$ 和 $k \in [K]$，以下结论以至少 $1 - \delta$ 的概率成立：

$$\|\mathbb{P}(\cdot \mid s_h, a_h) - \mathbb{P}^k(\cdot \mid s_h, a_h)\|_1 \le \beta(\delta) \cdot \iota_k(s_h, a_h), \quad \forall k \in [K], (s_h, a_h) \in \mathcal{S} \times \mathcal{A}.$$

同时，对于所有 $k \in [K]$，均有 $\iota_k \le 1$。

这里，参数 $\beta(\delta)$ 刻画了集中不等式中的方差，其通常随$\log(1/\delta)$增长。一般函数逼近下的分析中也出现了类似假设（Curi 等人，

2020年；Kidambi 等人，2021年）。此外，我们指出，这一假设推广了多种常用的参数化模型，包括线性马尔可夫决策过程模型（Jin 等人，2020年）以及我们在前文采用的 KNR 模型（Kakade 等人，2020年）。相应地，我们针对强化学习问题提出以下复杂度度量，

$$I_K = \max_{\{\mathcal{D}_k\}_{k\in[K]}} \sum_{k=1}^{K}\sum_{h=1}^{H} \iota_k^2(s_h^k, a_h^k). \tag{11}$$

此处的最大值是在所有可能的数据集上取得的 $\{\mathcal{D}_k\}_{k\in[K]}$，这些数据集由满足以下条件的线上学习算法收集：$|\mathcal{D}_k| = H$ 对于所有 $k \in [K]$。我们指出，在分析具有一般函数逼近的基于模型的强化学习时，也会出现类似的复杂度度量（Curi 等人，2020；Kakade 等人，2020；Kidambi 等人，2021年）。

我们对奖励随机化提出以下条件，以确保充分探索。

**C**件 7（（乐观主义））对于随机化奖励函数 $\{r_{h,\xi}^k\}_{h\in[H]}$，有 $t$

$$\sum_{h=1}^{H} r_{h,\xi}^k(s_h, a_h) - r_h(s_h, a_h) \geq H \cdot \beta(\delta) \cdot \sum_{h=1}^{H} \iota_k(s_h, a_h),$$

以至少 $p_0$ 的概率，对所有轨迹 $\{(s_h,\ a_h)\}_{h\in[H]}$，该条件均成立 .

条件 8（（奖励集中性））对于所有 $\delta' > 0$，均有$|r_{h,\ \xi}^k - r_h| \leq C_r(\delta') \cdot \iota_k$，其成立概率至少为 $1-\delta'$ 对于所有 $(k,\ h) \in [K] \times [H]$，其中 $\iota_k$ 在假设6中定义。

奖励随机化条件背后的直觉。 我们指出，条件 7 和 8 是 `PlanEx`中奖励随机化取得成功的关键因素。一方面，条件 7确保固定的 $p_0$ 比例的评估结果会产生乐观价值函数。这种乐观的价值估计进一步使得遵循乐观原则进行探索成为可能。另一方面，条件 8中的集中条件确保以高概率，基于随机化奖励估计的价值函数不会偏离基于真实奖励评估的价值函数太多。

以下定理给出了算法 2 在假设 6下的遗憾上界。

定理 9（遗憾界）在假设 *6*下，对于满足条件 *7* 和 *8*的随机化奖励，算法 *2* 的遗憾有如下界：

$$\mathbb{E}\big[R(K)\big] = \mathcal{O}\Big(Poly\big(C_r(1/K), \beta(1/K), H\big) \cdot I_K \cdot \sqrt{K}\Big).$$

证明 详细证明见 §B 。 ■

我们指出，对于线性马尔可夫决策过程和 KNR 等常用的模型参数化方式，参数 $\beta(1/K)$ 通常随$\log(K)$增长。同时，对于设计得当的奖励随机化方案，项 $C_r(1/K)$ 也随 $\log(K)$增长。因此，定理 18 表明，算法 2 的遗憾界随 $\widetilde{\mathcal{O}}(I_K \cdot \sqrt{K})$增长，与此前基于模型的强化学习中探索的遗憾界相匹配（）。

5.1. 随机化奖励的设计

我们指出，在实践中，假设 $\{\iota_k\}_{k\in[K]}$ 6 中定义的模型不确定性可根据集成模型之间的分歧进行估计。因此，要实例化算法 2，还需要设计满足条件 7和 8的适当奖励随机化方案。接下来，我们将给出此类随机化奖励的示例。

示例 1（高斯扰动）令$r^k_{h,\ \xi}(s^k_h,\ a^k_h) = r(s^k_h,\ a^k_h) + \xi^k_h$，其中 $\{\xi^k_h\}_{(k,h)\in[K]\times[H]}$是从高斯分布中独立采样得到的 $N(0,\ \sigma_k\cdot\iota^2_k(s^k_h,\ a^k_h))$。在 §*B.2*中规定的规制条件下，随机化奖励 $\{r^k_{h,\xi}\}_{(k,h)\in[K]\times[H]}$ 满足条件*7* 和*8*。

示例 2（伯努利扰动）令 $r^k_{h,\xi}(s^k_h,a^k_h) = r(s^k_h,a^k_h) + \xi^k_h\cdot\sigma'_k\iota_k(s^k_h,a^k_h)$，其中$\xi^k_h = 1$ 以概率 $1/2$ 发生，且 $\xi^k_h = -1$ 以概率 $1/2$发生。对于 §B.2 中指定的参数 $\{\sigma'_k\}_{(k)\in[K]}$，随机化奖励 $\{r^k_{h,\xi}\}_{(k,h)\in[K]\times[H]}$ 满足条件7和条件8。

A 与基于奖励加成的方法比较。 我们指出，我们提出的随机化奖励与基于模型的强化学习的奖励 bonus 密切相关；后者源于近期基于模型的强化学习探索领域的进展（Kidambi 等人，2021年；Song 和 Sun，2021年）。这类基于奖励项的方法通常会估计模型不确定性 $\{\iota_k\}_{k\in[K]}$，然后设计融入该不确定性估计的奖励 bonus。事实上，可以将示例 1和2中的随机化奖励视为这类奖励 bonus 的随机化推广。具体而言，随机化奖励和奖励 bonus 都依据乐观原则实现探索。对于随机化奖励，我们在 §4.1 中研究的部分乐观性保证了这种探索。相比之下，对于奖励 bonus，这种探索则是通过直接施加奖励 bonus 来确保乐观性。

鉴于奖励 bonus 与随机化奖励之间的联系，人们可能会想问，为什么要使用随机化奖励？我们强调，添加奖励项实际上是一种悲观方法，旨在降低最坏情况遗憾。这种方法强制加入一个偏离真实奖励的奖励项，目的是在价值估计中引入偏差，以实现最小的最坏情况遗憾。相比之下，随机化奖励允许扰动围绕真实奖励分布，因此与真实奖励的偏差更小。此外，正如我们的工作所示，这种扰动也具有 $\widetilde{\mathcal{O}}(\sqrt{K})$ 阶的最坏情况遗憾保证。

## 参考文献

Yasin Abbasi-Yadkori、D´avid P´al 和 Csaba Szepesv´ari。《线性随机老虎机的改进算法》。载于《神经信息处理系统进展》（NIPS），第 2312–2320 页，格拉纳达，西班牙，2011年。

Alekh Agarwal、Mikael Henaff、Sham M. Kakade 和 Wen Sun。PC-PG：用于可证明策略梯度学习的策略覆盖定向探索。载于《神经信息处理系统进展》（NeurIPS），线上会议，2020a。

Alekh Agarwal、Sham M. Kakade、Akshay Krishnamurthy 和 Wen Sun。FLAMBE：低秩马尔可夫决策过程的结构复杂度与表征学习。载于《神经信息处理系统进展》（NeurIPS），线上会议，2020b。

Alex Ayoub、Zeyu Jia、Csaba Szepesv´ari、Mengdi Wang 和 Lin Yang。基于模型的强化学习与价值目标回归。载于《第37届国际机器学习大会论文集》（ICML），第463–474页，线上会议，2020年。

Qi Cai、Zhuoran Yang、Chi Jin 和 Zhaoran Wang。策略优化中的可证明高效探索。载于《第37届国际机器学习大会论文集》（ICML），第1283–1294页，线上会议，2020年。

Kurtland Chua、Roberto Calandra、Rowan McAllister 和 Sergey Levine。利用概率动力学模型，通过少量试验实现深度强化学习。载于《神经信息处理系统进展》（NeurIPS），第 4759–4770 页，´蒙特利尔，加拿大，2018年。

Sebastian Curi、Felix Berkenkamp 和 Andreas Krause。通过乐观策略搜索和规划实现高效的基于模型的强化学习。载于《神经信息处理系统进展》（NeurIPS），线上，2020年。

Simon S.Du、Sham M. Kakade、Ruosong Wang 和 Lin F. Yang。良好的表征是否足以实现样本高效的强化学习？载于《第8届国际学习表征大会论文集》（ICLR），亚的斯亚贝巴，埃塞俄比亚，2020年。

David Ha 和 J¨urgen Schmidhuber。世界模型。arXiv 预印本 arXiv:1803.10122，2018年。

Danijar Hafner、Timothy Lillicrap、Jimmy Ba 和 Mohammad Norouzi。《梦境控制：通过潜在想象学习行为》。arXiv 预印本 arXiv:1912.01603，2019年。

Haque Ishfaq、Qiwen Cui、Viet Nguyen、Alex Ayoub、Zhuoran Yang、Zhaoran Wang、Doina Precup 和 Lin Yang。《基于一般价值函数近似的强化学习随机探索》。载于《第38届国际机器学习会议论文集》（ICML），第4607–4616页，线上会议，2021年。

Thomas Jaksch、Ronald Ortner 和 Peter Auer。《强化学习的近最优遗憾界》。机器学习研究杂志，11：1563–1600，2010年。

Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal、John Langford 和 Robert E. Schapire。具有低贝尔曼秩的上下文决策过程是 PAC可学习的。载于第34届国际机器学习大会（ICML）论文集，第1704–1713页，澳大利亚悉尼，2017年。

Chi Jin、Zhuoran Yang、Zhaoran Wang 和 Michael I. Jordan。利用线性函数近似实现可证明高效的强化学习。载于学习理论会议（COLT）论文集，第2137–2143页，线上会议 [奥地利格拉茨], 2020。

Lukasz Kaiser、Mohammad Babaeizadeh、Piotr Milos、Blazej Osinski、Roy H. Campbell、Konrad Czechowski、Dumitru Erhan、Chelsea Finn、Piotr Kozakowski、Sergey Levine、Afroz Mohiuddin、RyanSepassi、George Tucker 和 Henryk Michalewski。《Atari 的基于模型的强化学习》。载于《第8届国际学习表征大会论文集》（ICLR），埃塞俄比亚亚的斯亚贝巴，2020年。

Sham M. Kakade、Akshay Krishnamurthy、Kendall Lowrey、Motoya Ohnishi 和 Wen Sun。《在线非线性控制的信息论遗憾界》。载于《神经信息处理系统进展》（NeurIPS），线上，2020年。

Rahul Kidambi、Jonathan Chang 和 Wen Sun。《MobILE：仅从观测中进行基于模型的模仿学习》。载于《神经信息处理系统进展》（NeurIPS），第28598–28611页，线上，2021年。

Xiuyuan Lu 和 Benjamin Van Roy。强化学习的信息论置信界。载于《神经信息处理系统进展》（NeurIPS），第 2458–2466 页，加拿大温哥华，2019年。

Yuping Luo、Huazhe Xu、Yuanzhi Li、Yuandong Tian、Trevor Darrell 和 Tengyu Ma。具有理论保证的基于模型的深度强化学习算法框架。载于第七届国际学习表征会议（ICLR）论文集，美国路易斯安那州新奥尔良，2019年。

Horia Mania、Michael I. Jordan 和 Benjamin Recht。具有保证的非线性系统辨识主动学习。《机器学习研究杂志》，23:32:1–32:30，2022年。

Aditya Modi、Jinglin Chen、Akshay Krishnamurthy、Nan Jiang 和 Alekh Agarwal。低秩马尔可夫决策过程中的无模型表征学习与探索。arXiv 预印本 arXiv:2102.07035，2021年。

Ian Osband 和 Benjamin Van Roy。基于模型的强化学习与 eluder 维度。在《神经信息处理系统进展》（NIPS）中，第1466–1474页，加拿大蒙特利尔，2014年。

Aldo Pacchiano、Philip Ball、Jack Parker-Holder、Krzysztof Choromanski 和 Stephen Roberts。基于模型的强化学习中的乐观主义。arXiv 预印本 arXiv:2006.11911，2020年。

Deepak Pathak、Pulkit Agrawal、Alexei A. Efros 和 Trevor Darrell。通过自监督预测实现好奇心驱动的探索。在《第34届国际机器学习大会》（ICML）论文集，第2778–2787页，澳大利亚悉尼，2017年8月6日至11日，2017年。

Deepak Pathak、Dhiraj Gandhi 和 Abhinav Gupta。通过分歧实现自监督探索。载于《第36届国际机器学习会议》（ICML）论文集，第 5062–5071 页，加利福尼亚州长滩，2019年。

Daniel Russo。通过随机化价值函数进行探索的最坏情形遗憾界。载于《神经信息处理系统进展》（NeurIPS），第 14410–14420 页，加拿大温哥华，2019年。

Yuda Song 和 Wen Sun。PC-MLP：采用策略覆盖引导探索的基于模型的强化学习。载于《第38届国际机器学习会议论文集》（ICML），第9801–9811页，线上会议，2021年。

Wen Sun、Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal 和 John Langford。上下文决策过程中的基于模型的强化学习：PAC界及相较于无模型方法的指数级改进。载于《学习理论会议论文集》（COLT），第2898–2933页，亚利桑那州凤凰城，2019年。

Richard S Sutton 和 Andrew G Barto。《强化学习：导论》。麻省理工学院出版社，2018年。

Mark Veraar。关于带权Khintchine不等式。《美国数学会会刊》，第4119–4121页，2010年。

Ruosong Wang、Simon S. Du、Lin F. Yang 和 Ruslan Salakhutdinov。关于采用线性函数近似的无奖励强化学习。载于《神经信息处理系统进展》（NeurIPS），线上，2020年。

Manuel Watter、Jost Tobias Springenberg、Joschka Boedecker 和 Martin A. Riedmiller。嵌入以实现控制：一种利用原始图像进行控制的局部线性潜在动力学模型。载于《神经信息处理系统进展》（NIPS），第2746–2754页，加拿大蒙特利尔，2015年。

Grady Williams、Andrew Aldrich 和 Evangelos Theodorou。使用协方差变量重要性采样的模型预测路径积分控制。arXiv预印本 arXiv:1509.01149，2015年。

LinYang 和 Mengdi Wang。使用线性可加特征的样本最优参数化Q学习。载于第36届国际机器学习会议（ICML）论文集，第6995–7004页，加利福尼亚州长滩，2019年。

Andrea Zanette、David Brandfonbrener、Emma Brunskill、Matteo Pirotta 和 Alessandro Lazaric。《随机化最小二乘价值迭代的频率学派遗憾界》。载于《第23届国际人工智能与统计会议》（AISTATS），第1954–1964页，线上 [意大利西西里岛巴勒莫], 2020。

## 附录 A. 主要结果的证明

本节给出第§4节主要结果的证明。

### A.1. 良好事件与参数

接下来，我们为分析定义以下良好事件。

定义 10（良好事件）我们定义以下良好事件，

$$\mathcal{G}_{W^k,good} = \{\|W^* - W^k\|^2_{\Lambda_k} \le \beta_k\}, \quad \mathcal{G}_{\xi^k,good} = \{\|\xi^k_h\|^2_{\Lambda^k} \le \beta_{k,\xi}, \quad \forall h \in [H]\},$$
$$\mathcal{G}_{\xi^k,opt} = \Big\{V_1^*\big(s_1;\{r_h\}_{h\in[H]},W^*\big) - V_1^{\pi^k}\big(s_1;\{r^k_{h,\xi}\}_{h\in[H]},W^k\big) \le 0\Big\}.$$

相应地，我们进一步定义

$$\mathcal{G}_{W,good} = \bigcap_{k\in[K]} \mathcal{G}_{W^k,good}, \qquad \mathcal{G}_{\xi,good} = \bigcap_{k\in[K]} \mathcal{G}_{\xi^k,good}.$$

接下来，我们遵循引理19 并设置 $\beta_k$ 如下

$$\beta_k = 2\lambda\cdot\|W^*\|_2^2 + 8\sigma^2\Big(d_{\mathcal{S}}\cdot\log(5) + 2\log(k) + \log(4) + \log\big(\det(\Lambda_k)/\det(\Lambda_0)\big)\Big),$$

其中 $\sigma$ 是定义式 (5) 中转移动态的噪声方差。相应地，我们将 $\sigma_k$ 在 `PlanEx`中的参数设置如下：

$$\sigma_k^2 = H^3\cdot\beta_k/\sigma^2,$$

同时，我们将定义 10 中的参数设为 $\beta_{k,\ \xi} = 2\sigma_k^2\cdot\log(KH/\delta)$。因此有$\mathbb{P}(\mathcal{G}_{\xi,\ \text{good}}) \ge 1-\delta$。

### A.2. 最优性

接下来，我们给出 `PlanEx` 的最优性分析，其思路受到 Russo (2019年) 和 Zanette 等人(2020年) 的启发。

引理 11（最优性概率）在良好事件 $\mathcal{G}_{W^k,\ \text{good}}$ 下，以至少 $\Phi(-1)$ 的概率，有

$$V_1^*\big(s_1;\{r_h\}_{h\in[H]},W^*\big) - V_1^{\pi^k}\big(s_1;\{r^k_{h,\xi}\}_{h\in[H]},W^k\big) \le 0. \tag{12}$$

换言之，有 $\mathbb{P}(\mathcal{G}_{\xi^k,\ \text{opt}}\ |\mathcal{G}_{W^k,\ \text{good}}) \ge \Phi(-1)$。

证明 注：

$$\begin{aligned}
&V_1^{\pi^k}\big(s_1;\{r^k_{h,\xi}\}_{h\in[H]},W^k\big) - V_1^*\big(s_1;\{r_h\}_{h\in[H]},W^*\big)\\
&\quad\ge V_1^{\pi^*}\big(s_1;\{r^k_{h,\xi}\}_{h\in[H]},W^k\big) - V_1^*\big(s_1;\{r_h\}_{h\in[H]},W^*\big),\\
&\quad\ge \phi(s_1,a_1)^\top\xi_1^k - \mathbb{E}\Big[V_2^*\big(s_2;\{r_h\}_{h\in[H]},W^*\big)\,\Big|\,s_1,a_1,W^*\Big]\\
&\qquad+\mathbb{E}\Big[V_2^{\pi^*}\big(s_2;\{r^k_{h,\xi}\}_{h\in[H]},W^k\big)\,\Big|\,s_1,a_1,W^k\Big],
\end{aligned} \tag{13}$$

其中第二个不等式成立，是因为 $\{r_1+\phi^\top\xi_1^k\}^+ - r_1 = \max\{\phi^\top\xi_1^k, -r_1\} \geq \phi^\top\xi_1^k$。这里我们将模型 $(\{r_h\}_{h\in[H]}, W^*)$ 下的最优策略记为 $\pi^*$，将最优动作记为 $a_1 a_1 = \pi^*(s_1)$。此外还有：

$$\begin{aligned}
&\mathbb{E}\Big[V_2^{\pi^*}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]}, W^k\big)\,\Big|\, s_1, a_1, W^k\Big] - \mathbb{E}\Big[V_2^*\big(s_2;\{r_h\}_{h\in[H]}, W^*\big)\,\Big|\, s_1, a_1, W^*\Big] \\
&= \underbrace{\mathbb{E}\Big[V_2^*\big(s_2;\{r_h\}_{h\in[H]}, W^*\big)\,\Big|\, s_1, a_1, W^k\Big] - \mathbb{E}\Big[V_2^*\big(s_2;\{r_h\}_{h\in[H]}, W^*\big)\,\Big|\, s_1, a_1, W^*\Big]}_{\text{(i)}} \\
&\quad + \underbrace{\mathbb{E}\Big[V_2^{\pi^*}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_2^*\big(s_2;\{r_h\}_{h\in[H]}, W^*\big)\,\Big|\, s_1, a_1, W^k\Big]}_{\text{(ii)}}. \qquad (14)
\end{aligned}$$

根据引理20 以及 $V_h^* \leq H$ 对所有 $h\in[H]$这一事实，我们将项 (i) 的绝对值上界估计如下：

$$|\text{(i)}| \leq H\cdot\|(W^k - W^*)\phi(s_1,a_1)\|_2/\sigma \leq H\cdot\|W^k - W^*\|_{\Lambda_l}\cdot\|\phi(s_1,a_1)\|_{\Lambda_k^{-1}}.$$

因此，在事件 $\mathcal{G}_{W^k}$, good) 下，还可以进一步得到

$$|\text{(i)}| \leq \sqrt{\beta_k H^2}/\sigma\cdot\|\phi(s_1,a_1)\|_{\Lambda_k^{-1}}. \qquad (15)$$

将 (15) 代入 (14)，并基于 (13) 和 (14) 中的类似计算进一步展开项 (ii)，我们得出

$$\begin{aligned}
&V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_1^*\big(s_1;\{r_h\}_{h\in[H]}, W^*\big) \\
&\geq \sum_{h=1}^H \mathbb{E}\big[\phi(s_h,a_h)^\top\xi_h^k - \sqrt{\beta_k H^2}/\sigma\cdot\|\phi(s_h,a_h)\|_{\Lambda_k^{-1}}\,\big|\, s_1, \pi^*, W^k\big].
\end{aligned}$$

注意，对于任意给定的轨迹 $\{(s_h,a_h)\}_{h\in[H]}$，有

$$\sum_{h=1}^H \phi(s_h,a_h)^\top\xi_h^k \sim N(0,\sigma_{k,H}^2), \quad \sigma_{k,H}^2 = \sigma_k^2\cdot\sum_{h=1}^H\|\phi(s_h,a_h)\|_{\Lambda_k^{-1}}^2.$$

根据设定 $\sigma_k^2 = H^3\cdot\beta_k/\sigma^2$ 和柯西-施瓦茨不等式，可得

$$\sigma_{k,H} = \sqrt{H^3\cdot\beta_k/\sigma^2\cdot\sum_{h=1}^H\|\phi(s_h,a_h)\|_{\Lambda_k^{-1}}^2} \geq \sqrt{\beta_k H^2}/\sigma\cdot\sum_{h=1}^H\|\phi(s_h,a_h)\|_{\Lambda_k^{-1}}.$$

因此，对于任意给定的轨迹 $\{(s_h,a_h)\}_{h\in[H]}$，以至少 $\Phi(-1)$ 的概率有

$$\sum_{h=1}^H \phi(s_h,a_h)^\top\xi_h^k \geq \sigma_{k,H} \geq \sqrt{\beta_k H^2}/\sigma\cdot\|\phi(s_h,a_h)\|_{\Lambda_k^{-1}}.$$

因此，对 $s_1, \pi^*, W^k$ 下的轨迹以及良好事件$\mathcal{G}_{W^k,}$ good 进行积分后，以至少 $\Phi(-1)$ 的概率有

$$V_1^{\pi^k}\left(s_1;\{r_{h,\xi}^k\}_{h\in[H]},W^k\right)-V_1^*\left(s_1;\{r_h\}_{h\in[H]},W^*\right)$$
$$\geq\sum_{h=1}^{H}\mathbb{E}\left[\phi(s_h,a_h)^\top\xi_h^k-\sqrt{\beta_k H^2}/\sigma^2\cdot\|\phi(s_h,a_h)\|_{\Lambda_k^{-1}}\,\middle|\,s_1,\pi^*,W^k\right]\geq 0.$$

至此，我们完成了引理11的证明。 ■

引理12（乐观界）对于 $K>1/\Phi(-1)$，有

$$\mathbb{E}\left[\sum_{k=1}^{K}V_1^*\left(s_1;\{r_h\}_{h\in[H]},W^*\right)-V_1^{\pi^k}\left(s_1;\{r_{h,\xi}^k\}_{h\in[H]},W^k\right)\,\middle|\,\mathcal{G}_{W,good},\mathcal{G}_{\xi,good}\right]=\widetilde{\mathcal{O}}(\sqrt{K}).$$

证明 下文中，令 $\delta=1/K$ 属于定义 10中定义的良好事件。固定任意一个$k\in[K]$，并对以下差值给出上界:

$$\Delta_k=V_1^*\left(s_1;\{r_h\}_{h\in[H]},W^*\right)-V_1^{\pi^k}\left(s_1;\{r_{h,\xi}^k\}_{h\in[H]},W^k\right).$$

我们构造一个噪声集 $\{\widetilde{\xi}_h^k\}_{h\in[H]}$，它是噪声集$\{\xi_h^k\}_{h\in[H]}$的同分布独立副本。相应地，我们在定义 $\mathcal{G}_{\widetilde{\xi}、\text{good}}$ 和 $\mathcal{G}_{\widetilde{\xi}、\text{opt}}$ 中定义良好事件 10。我们进一步定义最优价值函数 $V_1^{\widetilde{\pi}^k}(s_1;\{r_{h,\widetilde{\xi}}^k\}_{h\in[H]},\ W^k)$ 在扰动奖励集$\{r_h+\phi^\top\widetilde{\xi}_h^k\}_{h\in[H]}$和状态转移 $W^k$下。因此，由引理 11 可得

$$\begin{aligned}\Delta_k&=V_1^*\left(s_1;\{r_h\}_{h\in[H]},W^*\right)-V_1^{\pi^k}\left(s_1;\{r_{h,\xi}^k\}_{h\in[H]},W^k\right)\\&\leq\mathbb{E}\left[V_1^{\widetilde{\pi}^k}\left(s_1;\{r_{h,\widetilde{\xi}}^k\}_{h\in[H]},W^k\right)\,\middle|\,\mathcal{G}_{\widetilde{\xi},\text{good}},\mathcal{G}_{\widetilde{\xi},\text{opt}}\right]-V_1^{\pi^k}\left(s_1;\{r_{h,\xi}^k\}_{h\in[H]},W^k\right).\end{aligned}\tag{16}$$

我们进一步定义良好事件 $\mathcal{G}_{\xi,\ \text{good}}$ 下与最小扰动对应的价值函数如下:

$$\{\underline{\xi}_h^k\}_{h\in[H]}=\operatorname*{argmin}_{\|\xi_h^k\|_{\Lambda_k}^2\leq\beta_{k,\xi}}\max_{\pi}V_1^{\pi}\left(s_1;\{r_h+\phi^\top\xi_h^k\}_{h\in[H]}^+,W^k\right).\tag{17}$$

我们定义 $r_{h,\ \underline{\xi}}^k=\{r_h+\phi^\top\underline{\xi}_h^k\}^+$ 相应的扰动奖励。我们进一步定义 $\underline{\pi}^k$ 模型 $(\{r_{h,\underline{\xi}}^k\}_{h\in[H]},\ W^k)$ 的相应最优策略。因此，在良好事件 $\mathcal{G}_{\xi,\ \text{good}}$下，有

$$\Delta_k\leq\mathbb{E}\left[V_1^{\widetilde{\pi}^k}\left(s_1;\{r_{h,\widetilde{\xi}}^k\}_{h\in[H]},W^k\right)-V_1^{\underline{\pi}^k}\left(s_1;\{r_{h,\underline{\xi}}^k\}_{h\in[H]},W^k\right)\,\middle|\,\mathcal{G}_{\widetilde{\xi},\text{good}},\mathcal{G}_{\widetilde{\xi},\text{opt}}\right].\tag{18}$$

同时，注意在良好事件 $\mathcal{G}_{\widetilde{\xi}^k,}$good 下，我们有

$$V_1^{\widetilde{\pi}^k}\left(s_1;\{r_{h,\widetilde{\xi}}^k\}_{h\in[H]},W^k\right)\geq V_1^{\underline{\pi}^k}\left(s_1;\{r_{h,\underline{\xi}}^k\}_{h\in[H]},W^k\right).$$

为简化记号，下文中我们将 $V_1^{\widetilde{\pi}^k} = V_1^{\widetilde{\pi}^k}(s_1;\ \{r^k_{h,\widetilde{\xi}}\}_{h\in[H]}, W^k)$ 和 $V_1^{\underline{\pi}^k} = V_1^{\underline{\pi}^k}(s_1;\ \{r^k_{h,\underline{\xi}}\}_{h\in[H]}, W^k)$ 分别记为.....。有

$$\begin{aligned}\mathbb{E}_{\widetilde{\xi}}\big[V_1^{\widetilde{\pi}^k} - V_1^{\underline{\pi}^k} \,\big|\, \mathcal{G}_{\widetilde{\xi},\text{good}}\big] &= \mathbb{E}_{\widetilde{\xi}}\big[V_1^{\widetilde{\pi}^k} - V_1^{\underline{\pi}^k} \,\big|\, \mathcal{G}_{\widetilde{\xi}^k,\text{good}}, \mathcal{G}_{\widetilde{\xi}^k,\text{opt}}\big]\cdot \mathbb{P}(\mathcal{G}_{\widetilde{\xi}^k,\text{opt}} \,|\, \mathcal{G}_{\widetilde{\xi}^k,\text{good}}) \\ &\quad + \mathbb{E}_{\widetilde{\xi}}\big[V_1^{\widetilde{\pi}^k} - V_1^{\underline{\pi}^k} \,\big|\, \mathcal{G}_{\widetilde{\xi}^k,\text{good}}, \mathcal{G}^c_{\widetilde{\xi}^k,\text{opt}}\big]\cdot \mathbb{P}(\mathcal{G}^c_{\widetilde{\xi}^k,\text{opt}} \,|\, \mathcal{G}_{\widetilde{\xi}^k,\text{good}}) \\ &\geq \mathbb{E}_{\widetilde{\xi}}\big[V_1^{\widetilde{\pi}^k} - V_1^{\underline{\pi}^k} \,\big|\, \mathcal{G}_{\widetilde{\xi}^k,\text{good}}, \mathcal{G}_{\widetilde{\xi}^k,\text{opt}}\big]\cdot \mathbb{P}(\mathcal{G}_{\widetilde{\xi}^k,\text{opt}} \,|\, \mathcal{G}_{\widetilde{\xi}^k,\text{good}}). \end{aligned} \tag{19}$$

与此同时，由引理 11可知，在 $\mathcal{G}_{W,\ \text{good}}$ 下，

$$\mathbb{P}(\mathcal{G}_{\widetilde{\xi}^k,\text{opt}} \,|\, \mathcal{G}_{\widetilde{\xi}^k,\text{good}}) \geq \mathbb{P}(\mathcal{G}_{\widetilde{\xi}^k,\text{opt}} \cap \mathcal{G}_{\widetilde{\xi}^k,\text{good}}) \geq 1 - \mathbb{P}(\mathcal{G}^c_{\widetilde{\xi}^k,\text{opt}}) - \mathbb{P}(\mathcal{G}^c_{\widetilde{\xi}^k,\text{good}}) \geq \Phi(-1) - \delta. \tag{20}$$

进一步将（19）和（20）代入（18），可得

$$\begin{aligned}\Delta_k &\leq \big(\Phi(-1)-\delta\big)^{-1}\cdot \mathbb{E}_{\widetilde{\xi}}\big[V_1^{\widetilde{\pi}^k} - V_1^{\underline{\pi}^k} \,\big|\, \mathcal{G}_{\widetilde{\xi}^k,\text{good}}\big] \\ &= \big(\Phi(-1)-\delta\big)^{-1}\cdot \mathbb{E}_{\xi}\big[V_1^{\pi^k} - V_1^{\underline{\pi}^k} \,\big|\, \mathcal{G}_{\xi^k,\text{good}}\big], \end{aligned} \tag{21}$$

其中，我们利用了噪声集 $\{\widetilde{\xi}^k_h\}_{h\in[H]}$ 是噪声集 $\{\xi^k_h\}_{h\in[H]}$ 的同分布独立副本这一事实。

现在只需在良好事件 $\mathcal{G}_{\xi,\ \text{good}}$ 和 $\mathcal{G}_{W,\ \text{good}}$ 下对差值 $V_1^{\pi^k} - V_1^{\underline{\pi}^k}$ 给出上界。注意，

$$\begin{aligned}&V_1^{\pi^k}\big(s_1;\{r^k_{h,\xi}\}_{h\in[H]}, W^k\big) - V_1^{\underline{\pi}^k}\big(s_1;\{r^k_{h,\underline{\xi}}\}_{h\in[H]}, W^k\big) \\ &\quad\leq V_1^{\pi^k}\big(s_1;\{r^k_{h,\xi}\}_{h\in[H]}, W^k\big) - V_1^{\pi^k}\big(s_1;\{r^k_{h,\underline{\xi}}\}_{h\in[H]}, W^k\big), \end{aligned} \tag{22}$$

这成立是因为 $\underline{\pi}^k$ 是模型 $(\{r^k_{h,\underline{\xi}}\}_{h\in[H]},\ W^k)$的最优解。与此同时，在（22）中，通过加上并减去真实模型 $(\{r_h\}_{h\in[H]},\ W^*)$ 下 $\pi^k$ 的价值函数 $V_1^{\pi^k}(s_1;\ \{r_h\}_{h\in[H]},\ W^*)$，可得

$$\begin{aligned}&V_1^{\pi^k}\big(s_1;\{r^k_{h,\xi}\}_{h\in[H]}, W^k\big) - V_1^{\underline{\pi}^k}\big(s_1;\{r^k_{h,\underline{\xi}}\}_{h\in[H]}, W^k\big) \\ &\quad\leq \underbrace{V_1^{\pi^k}\big(s_1;\{r^k_{h,\xi}\}_{h\in[H]}, W^k\big) - V_1^{\pi^k}\big(s_1;\{r_h\}_{h\in[H]}, W^*\big)}_{\text{(iii)}} \\ &\qquad + \underbrace{V_1^{\pi^k}\big(s_1;\{r_h\}_{h\in[H]}, W^*\big) - V_1^{\pi^k}\big(s_1;\{r^k_{h,\underline{\xi}}\}_{h\in[H]}, W^k\big)}_{\text{(iv)}}. \end{aligned} \tag{23}$$

接下来，我们分别在良好事件 $\mathcal{G}_{\xi、\ \text{good}}$ 和 $\mathcal{G}_{W、\ \text{good}}$ 下，对（23）中的项 (iii) 和 (iv) 给出上界。

上界　　项 (iii) 的 d。 上界的推导与引理13的证明类似。注意到

$$\begin{aligned}
\text{(iii)} &= V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big) - V_1^{\pi^k}\big(s_1;\{r_h\}_{h\in[H]},W^*\big)\\
&= \max\{\phi(s_1,a_1)^\top \xi_1^k, -r_1(s_1,a_1)\} + \phi(s_1,a_1)^\top \xi_h^k + \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big)\,\Big|\, s_1,\pi^k,W^k\Big]\\
&\quad - \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_h\}_{h\in[H]},W^*\big)\,\Big|\, s_1,\pi^k,W^*\Big]\\
&\le \|\phi(s_1,a_1)\|_{\Lambda_k^{-1}}\cdot\|\xi_1^k\|_{\Lambda_k} + \Delta_{\xi,1}^k \qquad (24)\\
&\quad + \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big) - V_2^{\pi^k}\big(s_2;\{r_h\}_{h\in[H]},W^*\big)\,\Big|\, s_1,\pi^k,W^*\Big],
\end{aligned}$$

其中，不等式由柯西-施瓦茨不等式以及 $r_1 \ge 0$ 这一事实得出。这里我们定义 $a_1 = \pi^k(s_1)$，并且

$$\Delta_{\xi,1}^k = \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big)\,\Big|\, s_1,\pi^k,W^k\Big] - \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big)\,\Big|\, s_1,\pi^k,W^*\Big].$$

注意，在良好事件 $\mathcal{G}_{\xi,\ \text{good}}$ 下，对所有 $(s_h,a_h)\in\mathcal{S}\times\mathcal{A}$ 以及 $h\in[H]$，都有

$$\begin{aligned}
r_{h,\xi}^k(s_h,a_h) &\le r(s_h,a_h) + |\phi(s_h,a_h)^\top \xi_h^k| \le 1 + \|\phi(s_h,a_h)\|_{\Lambda_k^{-1}}\cdot\|\xi_h^k\|_{\Lambda_k}\\
&\le 1 + \sqrt{\beta_{k,\xi}/(\lambda^2 H)}.
\end{aligned}$$

因此，在良好事件 $\mathcal{G}_{\xi,\ \text{good}}$ 下，有

$$0 \le V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big) \le H\cdot\Big(1+\sqrt{\beta_{k,\xi}/(\lambda^2 H)}\Big) = H + \sqrt{\beta_{k,\xi}H}/\lambda. \qquad (25)$$

根据引理20，在良好事件 $\mathcal{G}_{\xi,\ \text{good}}$ 以及 $\mathcal{G}_{W,\ \text{good}}$ 下，进一步有

$$\begin{aligned}
\Delta_{\xi,1}^k &= \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big)\,\Big|\, s_1,\pi^k,W^k\Big] - \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big)\,\Big|\, s_1,\pi^k,W^*\Big].\\
&\le \big(H+\sqrt{\beta_{k,\xi}H}/\lambda\big)\cdot\sigma^{-1}\cdot\|(W^k - W^*)\phi(s_1,a_1)\|_2\\
&\le \big(H+\sqrt{\beta_{k,\xi}H}/\lambda\big)\cdot\sigma^{-1}\cdot\|W^k - W^*\|_{\Lambda_k}\cdot\|\phi(s_1,a_1)\|_{\Lambda_k^{-1}}\\
&\le \big(H+\sqrt{\beta_{k,\xi}H}/\lambda\big)\cdot\sigma^{-1}\cdot\sqrt{\beta_k}\cdot\|\phi(s_1,a_1)\|_{\Lambda_k^{-1}}. \qquad (26)
\end{aligned}$$

这里，第一个不等式由引理20 以及式（25）中的界得出；第二个不等式由柯西-施瓦茨不等式得出；第三个不等式则由定义 $\mathcal{G}_{W,\ \text{good}}$ 10 中良好事件的定义得出。将式（26）和良好事件 $\mathcal{G}_{\xi,\ \text{good}}$ 的定义代入式（24），可得

$$\begin{aligned}
\text{(iii)} &\le C_k\cdot\|\phi(s_1,a_1)\|_{\Lambda_k^{-1}} \qquad (27)\\
&\quad + \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big) - V_2^{\pi^k}\big(s_2;\{r_h\}_{h\in[H]},W^*\big)\,\Big|\, s_1,\pi^k,W^*\Big],
\end{aligned}$$

其中我们定义

$$C_k = \big(H+\sqrt{\beta_{k,\xi}H}/\lambda\big)\cdot\sigma^{-1}\cdot\sqrt{\beta_k} + \sqrt{\beta_{k,\xi}}.$$

通过进一步其展开过程（27），我们得出结论：在良好事件 $\mathcal{G}_{\xi,\text{good}}$ 和 $\mathcal{G}_{W,\text{good}}$ 下，我们有 e

$$\text{(iii)} \le \mathbb{E}\left[\sum_{h=1}^{H} C_k \cdot \|\phi(s_h,a_h)\|_{\Lambda_k^{-1}} \,\middle|\, s_1,\pi^k,W^*,\mathcal{G}_{\xi^k,\text{good}},\mathcal{G}_{W^k,\text{good}}\right]. \tag{28}$$

上界 项 (iv) 的上界。项 (iv) 的上界与项 (iii) 的上界类似。注意 at

$$\begin{aligned}
\text{(iv)} &= V_1^{\pi^k}\big(s_1;\{r_h\}_{h\in[H]},W^*\big) - V_1^{\pi^k}\big(s_1;\{r_{h,\underline{\xi}}^k\}_{h\in[H]},W^k\big)\\
&= \min\big\{-\phi(s_1,a_1)^\top \underline{\xi}_h^k, r_1(s_1,a_1)\big\} + \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_h\}_{h\in[H]},W^*\big)\,\Big|\, s_1,\pi^k,W^*\Big]\\
&\quad - \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\underline{\xi}}^k\}_{h\in[H]},W^k\big)\,\Big|\, s_1,\pi^k,W^k\Big]\\
&\le -\phi(s_1,a_1)^\top \underline{\xi}_h^k + \Delta_{\underline{\xi},1}^k\\
&\quad + \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_h\}_{h\in[H]},W^*\big) - V_2^{\pi^k}\big(s_2;\{r_{h,\underline{\xi}}^k\}_{h\in[H]},W^k\big)\,\Big|\, s_1,\pi^k,W^*\Big],
\end{aligned}$$

其中我们定义

$$\Delta_{\underline{\xi},1}^k = \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^*\big)\,\Big|\, s_1,\pi^k,W^k\Big] - \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},W^k\big)\,\Big|\, s_1,\pi^k,W^k\Big].$$

根据式（17）中最小扰动的定义，有 $\|\underline{\xi}_h^k\|_{\Lambda_k} \le \beta_\xi$ 对所有 $(h,k) \in [H]\times[K]$。因此，我们有

$$-\phi(s_1,a_1)^\top \underline{\xi}_h^k \le \beta_{k,\xi}\cdot\|\phi(s_1,a_1)\|_{\Lambda_k^{-1}}.$$

其余计算与 (iii) 中的几乎完全相同。为简洁起见，我们省略计算，并得出结论：在良好事件 $\mathcal{G}_{\xi,\text{good}}$ 和 $\mathcal{G}_{W,\text{good}}$ 下，我们有

$$\text{(iv)} \le \mathbb{E}\left[\sum_{h=1}^{H} C_k \cdot \|\phi(s_h,a_h)\|_{\Lambda_k^{-1}} \,\middle|\, s_1,\pi^k,W^*,\mathcal{G}_{\xi^k,\text{good}},\mathcal{G}_{W^k,\text{good}}\right], \tag{29}$$

其中我们定义

$$C_k = \big(H + \sqrt{\beta_{k,\xi}H}/\lambda\big)\cdot\sigma^{-1}\cdot\sqrt{\beta_k} + \sqrt{\beta_{k,\xi}}. \tag{30}$$

因此，将 (28) 和 (29) 代入 (23)，我们得出结论：

$$\begin{aligned}
\mathbb{E}[\Delta_k \,|\, \mathcal{G}_{W,\text{good}},\mathcal{G}_{\xi,\text{good}}] &\le \big(\Phi(-1)-\delta\big)^{-1}\cdot\mathbb{E}_\xi\big[V_1^{\pi^k} - V_1^{\pi^k} \,\big|\, \mathcal{G}_{W^k,\text{good}},\mathcal{G}_{\xi^k,\text{good}}\big]\\
&\le \big(\Phi(-1)-\delta\big)^{-1}\cdot C_{\max}\cdot\mathbb{E}\left[\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|_{\Lambda_k^{-1}}\,\middle|\,\mathcal{G}_{W^k,\text{good}},\mathcal{G}_{\xi^k,\text{good}}\right],
\end{aligned}$$

其中我们定义

$$C_{\max} = C_K = \big(H + \sqrt{\beta_{K,\xi}H}/\lambda\big)\cdot\sigma^{-1}\cdot\sqrt{\beta_K} + \sqrt{\beta_{K,\xi}} \ge C_k, \quad \forall k\in[K],$$

期望是针对以下对象的轨迹求取的： $\pi^k$ 其状态转移由$W^*$定义。回顾一下，我们设 $\lambda = 1$, $\delta = 1/K$，且 $\sigma_k = H^3 \cdot \beta_k/\sigma^2$。因此，根据假设1，我们得到

$$\begin{aligned}\beta_K &= \mathcal{O}\big(H + d_{\mathcal{S}} + \log(K) + d_\phi \cdot \log(K)\big),\\ \beta_{K,\xi} &= 2\sigma_K^2 \cdot \log(K/\delta) = \mathcal{O}\big(H^4 \cdot \log(KH) + (d_{\mathcal{S}} + d_\phi) \cdot H^3 \cdot \log^2(KH)\big).\end{aligned}$$

因此，经过计算，我们得到

$$C_{\max} = \mathcal{O}\big((d_{\mathcal{S}} + d_\phi) \cdot H^3 \cdot \log^{3/2}(KH)\big).$$

最后，有

$$\begin{aligned}&\mathbb{E}\Bigg[\sum_{k=1}^K V_1^*\big(s_1; \{r_h\}_{h\in[H]}, W^*\big) - V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) \,\Bigg|\, \mathcal{G}_{W,\text{good}}, \mathcal{G}_{\xi,\text{good}}\Bigg]\\ &\quad\le \big(\Phi(-1) - 1/K\big)^{-1} \cdot C_{\max} \cdot \mathbb{E}\Bigg[\sum_{k=1}^K \sum_{h=1}^H \|\phi(s_h^k, a_h^k)\|_{\Lambda_k^{-1}} \,\Bigg|\, \mathcal{G}_{W,\text{good}}, \mathcal{G}_{\xi,\text{good}}\Bigg]\\ &\quad\le \big(\Phi(-1) - 1/K\big)^{-1} \cdot C_{\max} \cdot \mathbb{E}\Bigg[\bigg(HK \cdot \sum_{k=1}^K \sum_{h=1}^H \|\phi(s_h^k, a_h^k)\|_{\Lambda_k^{-1}}^2\bigg)^{1/2} \,\Bigg|\, \mathcal{G}_{W,\text{good}}, \mathcal{G}_{\xi,\text{good}}\Bigg]\\ &\quad= \mathcal{O}\Big((d_{\mathcal{S}} + d_\phi)^{3/2} \cdot H^{7/2} \cdot \log^2(KH) \cdot \sqrt{K}\Big),\end{aligned}$$

其中，期望是相对于 $\{\pi^k\}_{k\in[K]}$ 在由 $W^*$定义的真实转移下的轨迹 $W^*$，最后一个不等式由引理 21 以及 $\log\det(\Lambda_{K+1}) \le d_\phi \cdot \log K$这一事实得出。因此，我们完成了引理 12的证明。■

### A.3. 估计误差

引理13（估计误差界）对于 $K > 1/\Phi(-1)$，有

$$\mathbb{E}\Bigg[\sum_{k=1}^K V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_1^{\pi^k}\big(s_1; \{r_h\}_{h\in[H]}, W^*\big) \,\Bigg|\, \mathcal{G}_{W,good}, \mathcal{G}_{\xi,good}\Bigg] = \widetilde{\mathcal{O}}(\sqrt{K}).$$

证明 注意到

$$\begin{aligned}&V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_1^{\pi^k}\big(s_1; \{r_h\}_{h\in[H]}, W^*\big)\\ &\quad= \max\{\phi(s_1, a_1)^\top \xi_1^k, -r_1(s_1, a_1)\} + \phi(s_1, a_1)^\top \xi_h^k + \mathbb{E}\Big[V_2^{\pi^k}\big(s_2; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) \,\Big|\, s_1, \pi^k, W^k\Big]\\ &\qquad- \mathbb{E}\Big[V_2^{\pi^k}\big(s_2; \{r_h\}_{h\in[H]}, W^*\big) \,\Big|\, s_1, \pi^k, W^*\Big]\\ &\quad\le \|\phi(s_1, a_1)\|_{\Lambda_k^{-1}} \cdot \|\xi_1^k\|_{\Lambda_k} + \Delta_{\xi,1}^k \qquad (31)\\ &\qquad+ \mathbb{E}\Big[V_2^{\pi^k}\big(s_2; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_2^{\pi^k}\big(s_2; \{r_h\}_{h\in[H]}, W^*\big) \,\Big|\, s_1, \pi^k, W^*\Big],\end{aligned}$$

其中，不等式由柯西-施瓦茨不等式以及事实 $r_1 \geq 0$ 得出。这里我们定义 $a_1 = \pi^k(s_1)$，并且

$$\Delta_{\xi,1}^k = \mathbb{E}\Big[V_2^{\pi^k}\big(s_2; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) \,\Big|\, s_1, \pi^k, W^k\Big] - \mathbb{E}\Big[V_2^{\pi^k}\big(s_2; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) \,\Big|\, s_1, \pi^k, W^*\Big].$$

注意，在良好事件 $\mathcal{G}_{\xi,\ \text{good}}$ 下，对所有 $(s_h, a_h) \in \mathcal{S}\times\mathcal{A}$ 和 $h \in [H]$，都有

$$\begin{aligned} r_{h,\xi}^k(s_h, a_h) &\leq r(s_h, a_h) + |\phi(s_h, a_h)^\top \xi_h^k| \leq 1 + \|\phi(s_h, a_h)\|_{\Lambda_k^{-1}} \cdot \|\xi_h^k\|_{\Lambda_k} \\ &\leq 1 + \beta_\xi/\lambda. \end{aligned}$$

因此，在良好事件 $\mathcal{G}_{\xi,\ \text{good}}$ 下，有

$$0 \leq V_2^{\pi^k}\big(s_2; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) \leq H \cdot (1 + \beta_\xi/\lambda). \tag{32}$$

根据引理20，在良好事件 $\mathcal{G}_{\xi,\ \text{good}}$ 和 $\mathcal{G}_{W,\ \text{good}}$ 下，还可以得到

$$\begin{aligned} \Delta_{\xi,1}^k &= \mathbb{E}\Big[V_2^{\pi^k}\big(s_2; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) \,\Big|\, s_1, \pi^k, W^k\Big] - \mathbb{E}\Big[V_2^{\pi^k}\big(s_2; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) \,\Big|\, s_1, \pi^k, W^*\Big]. \\ &\leq H \cdot (1 + \beta_\xi/\lambda)/\sigma \cdot \|(W^k - W^*)\phi(s_1, a_1)\|_2 \\ &\leq H \cdot (1 + \beta_\xi/\lambda)/\sigma \cdot \|W^k - W^*\|_{\Lambda_k} \cdot \|\phi(s_1, a_1)\|_{\Lambda_k^{-1}} \\ &\leq H \cdot (1 + \beta_\xi/\lambda)/\sigma \cdot \beta_k \cdot \|\phi(s_1, a_1)\|_{\Lambda_k^{-1}}. \end{aligned} \tag{33}$$

这里第一个不等式由引理20 以及式（32）中的界推出，第二个不等式由柯西-施瓦茨不等式推出，第三个不等式则由定义$\mathcal{G}_{W,\ \text{good}}$ 中的良好事件定义推出（定义10）。将式（33）和良好事件定义$\mathcal{G}_{\xi,\ \text{good}}$代入式（31），可得

$$\begin{aligned} V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_1^{\pi^k}\big(s_1; \{r_h\}_{h\in[H]}, W^*\big) &\leq C_k \cdot \|\phi(s_1, a_1)\|_{\Lambda_k^{-1}} \\ + \mathbb{E}\Big[V_2^{\pi^k}\big(s_2; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_2^{\pi^k}\big(s_2; \{r_h\}_{h\in[H]}, W^*\big) \,\Big|\, s_1, \pi^k, W^*\Big], \end{aligned}$$

其中 $C_k$ 为 d 定义于（30）。进一步展开（27）并对 $k \in [K]$ 求和，我们得出

$$\begin{aligned} &\sum_{k=1}^K V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_1^{\pi^k}\big(s_1; \{r_h\}_{h\in[H]}, W^*\big) \\ &\qquad \leq \mathbb{E}\Big[\sum_{k=1}^K \sum_{h=1}^H C_k \cdot \|\phi(s_h^k, a_h^k)\|_{\Lambda_k^{-1}} \,\Big|\, \mathcal{G}_{W^k,\text{good}}, \mathcal{G}_{\xi^k,\text{good}}\Big], \end{aligned}$$

其中，期望是相对于 $\{\pi^k\}_{k\in[K]}$ 在由 $W^*$定义的真实转移下生成的轨迹而言。其余计算与引理12相同。我们省略计算，并得出

$$\mathbb{E}\Big[\sum_{k=1}^K V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_1^{\pi^k}\big(s_1; \{r_h\}_{h\in[H]}, W^*\big) \,\Big|\, \mathcal{G}_{W,\text{good}}, \mathcal{G}_{\xi,\text{good}}\Big] = \widetilde{\mathcal{O}}(\sqrt{K}).$$

■

### A.4. 遗憾分析

定理 14（期望遗憾界）我们按照 § $\lambda = 1$ 中的设置取 $\sigma_k$，如A.1所示，对于 $K > 1/\Phi(-1)$, 有:

$$\mathbb{E}\big[R(K)\big] = \mathcal{O}\Big((d_{\mathcal{S}} + d_{\phi})^{3/2} \cdot H^{7/2} \cdot \log^2(KH) \cdot \sqrt{K}\Big).$$

证明：有

$$\mathbb{E}\big[R(K)\big] = \mathbb{E}\Bigg[\sum_{k=1}^{K} V_1^*\big(s_1; \{r_h\}_{h\in[H]}, W^*\big) - V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) + V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_1^{\pi^k}\big(s_1; \{r_h\}_{h\in[H]}, W^*\big)\Bigg].$$

因此，我们有

$$\mathbb{E}\big[R(K)\big] \le \text{Opt} + \text{Est} + V_{\max} \cdot \sum_{k=1}^{K}\big(\mathbb{P}(\mathcal{G}^c_{W^k,\text{good}})\big) + V_{\max} \cdot K \cdot \mathbb{P}(\mathcal{G}^c_{\xi,\text{good}}), \tag{34}$$

其中，我们定义

$$\text{Opt} = \mathbb{E}\Bigg[\sum_{k=1}^{K} V_1^*\big(s_1; \{r_h\}_{h\in[H]}, W^*\big) - V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) \,\Big|\, \mathcal{G}_{W,\text{good}}, \mathcal{G}_{\xi,\text{good}}\Bigg],$$

$$\text{Est} = \mathbb{E}\Bigg[\sum_{k=1}^{K} V_1^{\pi^k}\big(s_1; \{r_{h,\xi}^k\}_{h\in[H]}, W^k\big) - V_1^{\pi^k}\big(s_1; \{r_h\}_{h\in[H]}, W^*\big) \,\Big|\, \mathcal{G}_{W,\text{good}}, \mathcal{G}_{\xi,\text{good}}\Bigg].$$

将引理 12和 13中 Opt 和 Est 的界，以及 $V_{\max} = H$和 $\delta = 1/T$ 代入（34），可得

$$\mathbb{E}\big[R(K)\big] = \mathcal{O}\Big((d_{\mathcal{S}} + d_{\phi})^{3/2} \cdot H^{7/2} \cdot \log^2(KH) \cdot \sqrt{K}\Big).$$

因此，我们完成了定理 14的证明。 ■

## 附录 B. §5中结果的证明

本节中，我们给出 §5 中结果的证明。

### B.1. 遗憾分析

与 §A 中的证明类似，我们将良好事件 $\mathcal{G}_{\text{good}}$ 定义如下:

$$\mathcal{G}_{\text{good}} = \big\{|r_h - r_{h,\xi}^k| \le C_r(\delta') \cdot \iota_k, \quad \|\mathbb{P}(\cdot \,|\, s,a) - \mathbb{P}^k(\cdot \,|\, s,a)\|_1 \le \beta_k(\delta) \cdot \iota_k(s,a), \\ \forall (k,h) \in [K] \times [H],\ (s,a) \in \mathcal{S} \times \mathcal{A}\big\}.$$

在假设 6 下，对于满足条件 7 的随机化奖励，有 $\mathbb{P}(\mathcal{G}_{\text{good}}) \ge 1 - \delta' - \delta$。

引理 15（最优性概率）在假设 6 以及良好事件 $\mathcal{G}_{good}$下，对于满足条件 7 的随机化奖励，有

$$V_1^*\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big)\le V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\ge 1/2$$

概率至少为 $p_0$。

证明 证明与引理11的证明类似。注意

$$\begin{aligned}
&V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)-V_1^*\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big)\\
&\quad\ge V_1^{\pi^*}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)-V^*\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big)\\
&\quad= r_{h,\xi}^k\big(s_1,\pi^*(s_1)\big)-r_h\big(s_1,\pi^*(s_1)\big)+\mathbb{E}\Big[V_2^{\pi^*}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\,\Big|\,s_1,\pi^*,\mathbb{P}^k\Big]\\
&\qquad-\mathbb{E}\Big[V_2^*\big(s_2;\{r_h\}_{h\in[H]},\mathbb{P}\big)\,\Big|\,s_1,\pi^*,\mathbb{P}\Big],
\end{aligned}\tag{35}$$

其中，我们用 $\pi^*$ 表示环境由 $(\{r_h\}_{h\in[H]},\mathbb{P})$ 定义时的最优策略，第一个不等式由环境由 $(\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k)$ 定义时 $\pi^k$ 的最优性得出。同时，有

$$\begin{aligned}
&\mathbb{E}\Big[V_2^{\pi^*}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\,\Big|\,s_1,\pi^*,\mathbb{P}^k\Big]-\mathbb{E}\Big[V_2^*\big(s_2;\{r_h\}_{h\in[H]},\mathbb{P}\big)\,\Big|\,s_1,\pi^*,\mathbb{P}\Big]\\
&\quad=\underbrace{\mathbb{E}\Big[V_2^*\big(s_2;\{r_h\}_{h\in[H]},\mathbb{P}\big)\,\Big|\,s_1,\pi^*,\mathbb{P}^k\Big]-\mathbb{E}\Big[V_2^*\big(s_2;\{r_h\}_{h\in[H]},\mathbb{P}\big)\,\Big|\,s_1,\pi^*,\mathbb{P}\Big]}_{\text{(i)}}\\
&\qquad+\underbrace{\mathbb{E}\Big[V_2^{\pi^*}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)-V_2^*\big(s_2;\{r_h\}_{h\in[H]},\mathbb{P}\big)\,\Big|\,s_1,\pi^*,\mathbb{P}^k\Big]}_{\text{(ii)}}.
\end{aligned}\tag{36}$$

根据假设 6，H¨older 不等式以及 $V_1^*\le H$，在良好事件$\mathcal{G}_{\text{good}}$ 下，有

$$|\text{(i)}|\le\mathbb{E}\big[H\cdot\|(\mathbb{P}-\mathbb{P}^k)(\cdot\,|\,s_1,a_1)\|_1\,\big|\,s_1,\pi^*\big]\le\mathbb{E}\big[H\cdot\beta(\delta)\cdot\iota_k(s_1,a_1)\,\big|\,s_1,\pi^*\big].$$

进一步展开（36）中的项 (ii)，由（35）可得

$$\begin{aligned}
&V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)-V_1^*\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big)\\
&\quad\ge\mathbb{E}\Big[\sum_{h=1}^H r_{h,\xi}^k(s_1,a_1)-r_h(s_1,a_1)-H\cdot\beta_k\sum_{h=1}^H\iota_k(s_h,a_h)\,\Big|\,s_1,\pi^*,\mathbb{P}^k\Big].
\end{aligned}$$

因此，在条件7中的乐观条件下，有

$$V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)-V_1^*\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big)\ge 0$$

以至少 $p_0-\delta$ 的概率。 ■

**Lemma 16**（乐观界）对于 $\delta' \le p_0$，以至少 $1-\delta-\delta'$ 的概率成立 $t$

$$\mathbb{E}\Big[\sum_{k=1}^{K} V_1^*\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big) - V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\,\Big|\,\mathcal{G}_{good}\Big] = \mathcal{O}\big(C(\delta,\delta')\cdot I_K\cdot\sqrt{K}\big),$$

其中，我们定义 $C(\delta,\delta') = C_r(\delta') + H\cdot(1+C_r(\delta'))\cdot\beta(\delta)$。

证明 证明过程与引理 12 §A.2中的证明类似。我们定义如下最小扰动价值函数，

$$V_1^{\underline{\pi}^k} = \operatorname*{argmin}_{\widetilde{r}\in\mathcal{G}_{\text{good}}} \operatorname*{argmax}_{\pi} V_1^{\pi}\big(s_1;\{\widetilde{r}_h\}_{h\in[H]},\mathbb{P}^k\big).$$

按照引理 12证明中的相同计算，我们得到

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{k=1}^{K} V_1^*\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big) - V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\,\Big|\,\mathcal{G}_{\text{good}}\Big] \\
&\quad\le (p_0-\delta)^{-1}\cdot\underbrace{\mathbb{E}\Big[\sum_{k=1}^{K} V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big) - V_1^{\underline{\pi}^k}\big(s_1;\{\widetilde{r}_h\}_{h\in[H]},\mathbb{P}^k\big)\,\Big|\,\mathcal{G}_{\text{good}}\Big]}_{\Delta_k}. \qquad (37)
\end{aligned}$$

进一步加上并减去 $V^{\pi^k}(s_1;\{r_h\},\mathbb{P})$，可得

$$\begin{aligned}
\Delta_k = &\underbrace{\mathbb{E}\Big[\sum_{k=1}^{K} V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big) - V^{\pi^k}(s_1;\{r_h\},\mathbb{P})\,\Big|\,\mathcal{G}_{\text{good}}\Big]}_{\text{(iii)}} \\
&+\underbrace{\mathbb{E}\Big[\sum_{k=1}^{K} V^{\pi^k}(s_1;\{r_h\},\mathbb{P}) - V_1^{\underline{\pi}^k}\big(s_1;\{\widetilde{r}_h\}_{h\in[H]},\mathbb{P}^k\big)\,\Big|\,\mathcal{G}_{\text{good}}\Big]}_{\text{(iv)}}. \qquad (38)
\end{aligned}$$

因此，经过与 §A.2中的类似计算，在良好事件 $\mathcal{G}_{\text{good}}$下，还可得

$$V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big) - V^{\pi^k}(s_1;\{r_h\},\mathbb{P}) \le \mathbb{E}\Big[C(\delta,\delta')\sum_{h=1}^{H}\iota_k(s_h,a_h)\,\Big|\,s_1,\pi^k,\mathbb{P}\Big],$$

其中，我们定义

$$C(\delta,\delta') = C_r(\delta') + H\cdot\big(1+C_r(\delta')\big)\cdot\beta(\delta).$$

这里我们利用以下事实：$\|V_h^{\pi^k}(\cdot;\ \{r_{h,\xi}^k\}_{h\in[H]},\ \mathbb{P}^k)\|_\infty \le H\cdot(1+C_r(\delta'))$ 在良好事件$\mathcal{G}_{\text{good}}$下。根据事实 38，式（38）中的项 **(iv)** 也满足相同的界 $\widetilde{r}\in\mathcal{G}_{\text{good}}$。因此，通过

将式（37）代入式（37），我们得到

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{k=1}^{K} V_1^*\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big) - V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\,\Bigg|\,\mathcal{G}_{\text{good}}\Bigg] \\
&\quad\le C(\delta,\delta')\cdot(p_0-\delta)^{-1}\cdot\mathbb{E}\Bigg[\sum_{k=1}^{K}\sum_{h=1}^{H} C_k\cdot\sigma_k(s_h,a_h)\,\Bigg|\,\mathcal{G}_{\text{good}}\Bigg] \\
&\quad\le C(\delta,\delta')\cdot(p_0-\delta)^{-1}\cdot\sqrt{HK}\cdot\mathbb{E}\Bigg[\Bigg(\sum_{h=1}^{H}\iota_k^2(s_h,a_h)\Bigg)^{1/2}\,\Bigg|\,\mathcal{G}_{\text{good}}\Bigg],
\end{aligned}$$

进一步代入 $I_K$ 在（11）中的定义，可得

$$\mathbb{E}\Bigg[\sum_{k=1}^{K} V_1^*\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big) - V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\,\Bigg|\,\mathcal{G}_{\text{good}}\Bigg] = \mathcal{O}\big(C(\delta,\delta')\cdot I_K\cdot\sqrt{K}\big),$$

由此完成引理16的证明。 ■

引理17（估计误差界）有

$$\mathbb{E}\Bigg[\sum_{k=1}^{K} V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big) - V_1^{\pi^k}\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big)\,\Bigg|\,\mathcal{G}_{good}\Bigg] = \mathcal{O}\big(C(\delta,\delta')\cdot I_K\cdot\sqrt{K}\big).$$

证明 在良好事件 $\mathcal{G}_{\text{good}}$ 下，有

$$\begin{aligned}
&V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big) - V_1^{\pi^k}\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big) \\
&\quad\le C_r\cdot\sigma^k(s_1,a_1) + \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\,\Big|\,s_1,\pi^k,\mathbb{P}^k\Big] \\
&\qquad - \mathbb{E}\big[V_2^{\pi^k}\big(s_2;\{r_h\}_{h\in[H]},\mathbb{P}\big)\,\big|\,s_1,\pi^k,\mathbb{P}\big] \\
&\quad\le C_r\cdot\sigma^k(s_1,a_1) + \Delta_{\xi,1}^k \\
&\qquad + \mathbb{E}\big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big) - V_2^{\pi^k}\big(s_2;\{r_h\}_{h\in[H]},\mathbb{P}\big)\,\big|\,s_1,\pi^k,\mathbb{P}\big],
\end{aligned}\tag{39}$$

其中，我们定义 $a_1=\pi_1^k(s_1)$ 和

$$\Delta_{\xi,1}^k = \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\,\Big|\,s_1,\pi^k,\mathbb{P}^k\Big] - \mathbb{E}\Big[V_2^{\pi^k}\big(s_2;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big)\,\Big|\,s_1,\pi^k,\mathbb{P}\Big].$$

在良好事件 $\mathcal{G}_{\text{good}}$ 下，由 H¨older 不等式可得

$$\Delta_{\xi,1}^k \le H\cdot\big(1+C_r(\delta')\big)\cdot\beta_k\cdot\iota_k(s_1,a_1).\tag{40}$$

将（40）代入（39），并进一步展开（39），可得在良好事件 $\mathcal{G}_{\text{good}}$ 下，

$$\begin{aligned}
&V_1^{\pi^k}\big(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\big) - V_1^{\pi^k}\big(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\big) \\
&\quad\le C(\delta,\delta')\cdot\mathbb{E}\Bigg[\sum_{h=1}^{H}\iota_k(s_h,a_h)\,\Big|\,s_1,\pi^k,\mathbb{P}\Bigg],
\end{aligned}\tag{41}$$

其中定义 $C(\delta,\delta') = C_r(\delta') + H\cdot(1+C_r(\delta'))\cdot\beta(\delta)$。因此，将（41）对 $k\in[K]$ 求和，可得

$$\mathbb{E}\left[\sum_{k=1}^{K} V_1^{\pi^k}\left(s_1;\{r_{h,\xi}^k\}_{h\in[H]},\mathbb{P}^k\right) - V_1^{\pi^k}\left(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\right)\,\middle|\,\mathcal{G}_{\text{good}}\right]$$
$$\le C(\delta,\delta')\cdot\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\sigma_k(s_h^k,a_h^k)\,\middle|\,\mathcal{G}_{\text{good}}\right] \le C(\delta,\delta')\cdot\sqrt{HK}\cdot I_K,$$

其中 $I_K$ 在（11）中定义。因此，引理 17 的证明得证。 ■

定理18（遗憾界）设 $\delta=\delta'=1/K$。在假设 6 下，对于满足条件 7和8 的随机化奖励，有

$$\mathbb{E}\big[R(K)\big] = \mathcal{O}\left(\Big(C_r(1/K) + H\cdot\big(1+C_r(1/K)\big)\cdot\beta(1/K)\Big)\cdot\sqrt{H^2K}\cdot I_K\right).$$

证明 结合引理 16和17，可得

$$\mathbb{E}\left[\sum_{k=1}^{K} V_1^*\left(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\right) - V_1^{\pi^k}\left(s_1;\{r_h\}_{h\in[H]},\mathbb{P}^k\right)\,\middle|\,\mathcal{G}_{\text{good}}\right] = \mathcal{O}\big(C(\delta,\delta')\cdot I_K\cdot\sqrt{K}\big). \quad (42)$$

同时，在假设 6 下，对于满足条件 8 的奖励，有 $\mathbb{P}(\mathcal{G}_{\text{good}})1-\delta-\delta'$。因此，对于 $\delta=\delta'=1/K$，有

$$\mathbb{E}\left[\sum_{k=1}^{K} V_1^*\left(s_1;\{r_h\}_{h\in[H]},\mathbb{P}\right) - V_1^{\pi^k}\left(s_1;\{r_h\}_{h\in[H]},\mathbb{P}^k\right)\,\middle|\,\mathcal{G}_{\text{good}}^c\right] \le 2K\cdot(\delta+\delta') = 4. \quad (43)$$

结合（42）和（43），可得

$$\mathbb{E}\big[R(K)\big] = \mathcal{O}\big(C(1/K,1/K)\cdot\sqrt{H^2K}\cdot I_K\big)$$
$$= \mathcal{O}\left(\Big(C_r(1/K) + H\cdot\big(1+C_r(1/K)\big)\cdot\beta(1/K)\Big)\cdot\sqrt{H^2K}\cdot I_K\right),$$

由此完成定理 18 的证明。 ■

B.2 示例验证

接下来，我们将验证示例1和2满足条件7 和8。

示例3（高斯奖励）设

$$r_{h,\xi}^k(s_h,a_h) \sim N\big(r_h(s_h,a_h), H\cdot\beta(\delta)\cdot\sigma_k^2(s_h,a_h)\big),$$

这些样本是在 $(s_h,a_h)\in\mathcal{S}\times\mathcal{A}$ 上独立抽取的。其概率满足

$$\sum_{h=1}^{H} r_{h,\xi}^k(s_h,a_h) - r_h(s_h,a_h) \ge H\cdot\beta(\delta)\sum_{h=1}^{H}\sigma_k(s_h,a_h)$$

大于以下事件的概率,

$$N\left(0, H\cdot\sum_{h=1}^{H}\sigma_k^2(s_h,a_h)\right)\geq\sqrt{H\cdot\beta(\delta)\sum_{h=1}^{H}\sigma_k^2(s_h,a_h)},$$

其成立的概率至少为 $\Phi(-1)$。此外，如果 $r/\sigma_k$ 是关于 $(s,a)\in\mathcal{S}\times\mathcal{A}$ 的利普希茨函数，且 $\mathcal{S}\times\mathcal{A}$ 在相同度量下的覆盖数为 $\mathcal{C}_{\mathcal{S}\times\mathcal{A}}$，则还成立：$|r_{h,\xi}^k-r_h|\leq\sqrt{H\cdot\beta(\delta)}\cdot\log(\mathcal{C}_{\mathcal{S}\times\mathcal{A}}\cdot HK/\delta')\cdot\sigma_k$，其成立的概率至少为 $1-\delta'$。因此，条件7和8均满足。

示例 4（伯努利奖励）设

$$r_{h,\epsilon}^k(s_h,a_h)\sim r(s_h,a_h)+2\sqrt{H}\cdot\beta(\delta)\cdot\sigma_k(s_h,a_h)\cdot\epsilon_h,$$

其中 $\epsilon_h=1$ 以概率 $1/2$ 取值，且 $\epsilon_h=-1$ 以概率 $1/2$取值。对于任意轨迹 $\tau$，由欣钦不等式（*Veraar*，2010年）可得

$$\sum_{h=1}^{H}\sqrt{H}\cdot\beta(\delta)\cdot\sigma_k(s_h,a_h)\cdot\epsilon_h/S(\tau)\geq 1/2$$

其成立的概率至少为 $p_0=3/16$，其中我们定义

$$S(\tau)=H\sum_{h=1}^{H}\beta^2(\delta)\cdot\sigma_k^2(s_h,a_h).$$

因此，有

$$\begin{aligned}\sum_{h=1}^{H}r_{h,\epsilon}^k(s_h,a_h)-r(s_h,a_h)&\geq 2\sum_{h=1}^{H}\sqrt{H}\cdot\beta(\delta)\cdot\sigma_k(s_h,a_h)\cdot\epsilon_h\\&\geq S(\tau)\geq\sum_{h=1}^{H}\beta(\delta)\cdot\sigma_k(s_h,a_h)\end{aligned}$$

以至少 $p_0=3/16$ 的概率。此外，有 $|r_{h,\epsilon}^k(s_h,a_h)-r(s_h,a_h)|\leq 2\sqrt{H}\cdot\beta(\delta)\cdot\sigma_k(s_h,a_h)$。因此，条件 7 和 8 均得到满足。

## 附录C. 辅助引理

引理19（自归一化过程的集中不等式（Abbasi-Yadkori 等人，2011年；Kakade 等人，2020年））以下结论成立：

$$\beta_k=2\lambda\cdot\|W^*\|_2^2+8\sigma^2\Big(d_{\mathcal{S}}\cdot\log(5)+2\log(k)+\log(4)+\log\big(\det(\Lambda_k)/\det(\Lambda_0)\big)\Big)$$

*that*

$$\sum_{k=0}^{\infty}\mathbb{P}(\mathcal{G}_{W^k,good}^c)=\sum_{k=0}^{\infty}\mathbb{P}\big(\|W^k-W^*\|_{\Lambda_k}^2\geq\beta_k\big)\leq 1/2.$$

证明 详见Kakade 等人（2020年），其中给出了详细证明。 ■

引理20（两个高斯分布下的期望差（Kakade 等人，2020年））设 $z_1 \sim N(\mu_1, \sigma^2)$ 和 $z_2 \sim N(\mu_2, \sigma^2)$ 是两个高斯随机变量。设 $g$ 是一个正的可测函数，则有

$$\mathbb{E}_{z_1 \sim N(\mu_1,\sigma^2)}\big[g(z_1)\big] - \mathbb{E}_{z_2 \sim N(\mu_1,\sigma^2)}\big[g(z_2)\big] \leq \min\{\|\mu_1 - \mu_2\|_2/\sigma, 1\} \cdot \sqrt{\mathbb{E}_{z_1 \sim N(\mu_1,\sigma^2)}\big[g^2(z_1)\big]}.$$

证明 详见Kakade 等人（2020年），其中给出了详细证明。 ■

引理 21（椭圆势引理（Kakade 等人，2020年））设 $\|\phi_h^k\|_2 \leq 1/\sqrt{H}$ 对所有 $(k,\ h) \in [K] \times [H]$。设 $\Lambda_1 = I$ 且 $\Lambda_{k+1} = \Lambda_k + \sum_{h=1}^H \phi_h^k (\phi_h^k)^\top$。则有

$$\sum_{k=1}^K \sum_{h=1}^H \|\phi_h^k\|^2_{\Lambda_k^{-1}} \leq 2\log\big(\det(\Lambda_{K+1}) \cdot \det(\Lambda_0)^{-1}\big).$$

证明 注意到 wehave $\Lambda_k \succ \Lambda_1 = I$。因此有

$$0 \leq \sum_{h=1}^H (\phi_h^k)^\top \Lambda_k^{-1} \phi_h^k \leq \sum_{h=1}^H \|\phi_h^k\|_2^2 \leq 1, \quad \forall k \in [K].$$

同时，由于对于 $x \in [0,1],\ x \leq 2\log(1+x)$，我们有

$$2\log\left(1 + \sum_{h=1}^H (\phi_h^k)^\top \Lambda_k^{-1} \phi_h^k\right) \geq \sum_{h=1}^H (\phi_h^k)^\top \Lambda_k^{-1} \phi_h^k. \tag{44}$$

另一方面，由矩阵行列式引理可得

$$\log\det(\Lambda_{k+1}) - \log\det(\Lambda_k) = \log\det\Bigg(I + \underbrace{\Lambda_k^{-1/2} \sum_{h=1}^H (\phi_h^k)(\phi_h^k)^\top \Lambda_k^{-1/2}}_{\Xi_k}\Bigg).$$

设 $\Xi_k$ 的特征值为 $\{\sigma_i\}_{i\in[d]}$。对于所有 $i \in [d]$，均有 $\sigma_i > 0$，并且

$$\log\det(I + \Xi_k) = \log\Pi_{i\in[d]}(1+\sigma_i) \geq \log\left(1 + \sum_{i\in[d]} \sigma_i\right) = \log\det\big(1 + \mathrm{Tr}(\Xi_k)\big).$$

因此，我们有

$$\begin{aligned}
\log\det(\Lambda_{k+1}) - \log\det(\Lambda_k) &= \log\det(I + \Xi_k) \geq \log\det\big(1 + \mathrm{Tr}(\Xi_k)\big) \\
&= \log\left(1 + \sum_{h=1}^H \mathrm{Tr}\big(\Lambda_k^{-1/2}(\phi_h^k)(\phi_h^k)^\top \Lambda_k^{-1/2}\big)\right) \\
&= \log\left(1 + \sum_{h=1}^H (\phi_h^k)^\top \Lambda_k^{-1} \phi_h^k\right).
\end{aligned} \tag{45}$$

结合（44）和（45），我们得出

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi_h^k\|^2_{\Lambda_k^{-1}} \le \sum_{k=1}^{K}\log\det(\Lambda_{k+1}) - \log\det(\Lambda_k) = 2\log\big(\det(\Lambda_{K+1})\cdot\det(\Lambda_0)^{-1}\big),$$

由此完成引理20的证明。 ■