关于后验更新的奖励表示的说明

PedroA. Ortega[**]
Daios Technologies

现代控制和强化学习中的许多思想都将决策视为推断：从一个基线分布出发，并在收到“信号”时对其进行更新。我们探讨如何使这种说法成为字面意义上的描述，而非比喻。我们研究这样一种特殊情形：KL 正则化软更新恰好是单个固定概率模型中的贝叶斯后验，因此更新变量是一个真正传递信息的信道。在这种情形下，行为变化仅由该信道所携带的证据驱动：更新必须能够解释为对基线分布进行证据重加权。这得出了一个明确的识别结果：后验更新决定了会引发行为变化的相对、依赖上下文的激励信号，但无法唯一确定绝对奖励；绝对奖励仍存在特定上下文基线所造成的歧义。要求不同更新方向共用同一个延续价值，则会进一步引入一项相干性约束，将不同条件化顺序所对应的奖励描述联系起来。

arXiv:2602.02912v1 [cs.LG] 2 Feb 2026

关键词：*KL* 正则化控制、有界理性、作为优化器的后验、条件点态互信息、奖励可识别性。

1. 引言与贡献

KL 正则化（或有界理性）优化是最大熵控制、变分推断和正则化强化学习 [6, 15, 16, 21]中的标准方法。在本文中，我们聚焦于一个明确的识别问题：当我们将范围限制在这样一类 KL 正则化目标的子类时，情况会发生什么变化？该子类的优化器恰好是同一个整体联合分布下的贝叶斯后验条件分布。

这并非一般性质，而是对可接受的奖励/价值函数族施加的结构约束：在识别类中，依赖交互项必须与点互信息（PMI）一致。具体而言，除 KL 正则化外，我们还提出两个明确的假设。第一，后验识别：优化器等于同一个联合分布下的后验条件分布。第二，顺序无关性：在不同更新方向上始终一致地使用同一个价值函数，因此后续过程仅取决于信息事件（等价地说，取决于条件变量集合）。这些假设带来三项结果：

- **表示：** 后验识别逐点确定更新对数比；等价地，它为已识别交互项给出后验对数比表示（条件点互信息）。
- **可识别性界限：** 后验分布只能识别与上下文基线无关的奖励–价值交互项（一个规范类）；若没有基线约定，则无法唯一确定奖励。
- **相干性约束：** 顺序无关性对不同更新方向施加可交换性（可积性）约束，使不同条件化顺序对应的奖励参数化彼此关联。

本质上，后验识别得出一种点态表示结构：唯一兼容的奖励–价值交互项是后验对数比，即条件分布点互信息，并带有残余的上下文基线

通讯作者：*pedro@daios.ai*© 2026
DAIOS。版权所有

规范自由
度。

为何这很重要。 许多“控制即推断”和广义贝叶斯构造都会产生类似后验的 软更新 [1, 8, 11, 21]; 在此，我们单独考察这样一个边界情形：该更新实际上就是 某个整体 P下的贝叶斯条件分布。在这种情况下，观测到的条件分布只能识别驱动行为变化的交互项（点互信息信号），而无法仅凭自身识别绝对奖励水平，除非采用外部基线约定 [2, 9, 14]。随后，交换性条件刻画了何时可以在不同的条件化顺序之间一致地复用单个基于事件的延续价值 [22, 23]。

2. 主要结果与路线图

我们研究 KL 正则化更新，并探讨当优化器被识别为 单个整体联合分布下的贝叶斯后验条件分布时，会被强制要求具备哪些额外结构。我们先给出主要恒等式，然后按顺序推导这些恒等式。

（在所述假设下的恒等式汇总。） 当相应的假设成立时（定义见第 3 节），奖励/价值族逐点受到如下约束：

- **后验点互信息表示（公式 7）：** 后验识别要求奖励—价值交互项逐点等于后验对数比，即条件点互信息，

$$\alpha\big[r_z(x \mid y) + V(x,y,z) - V(y,z)\big] = i(x;z \mid y).$$

- **规范自由度（公式 8）：** 只有交互项是可识别的；奖励—价值分解在仅依赖上下文的基线范围内存在歧义。对于任意 $c(y,z)$，平移 $r_z(x \mid y) \mapsto r_z(x \mid y) + c(y,z)$ 和$V(y,z) \mapsto V(y,z) + c(y,z)$，都不会改变 $P(x \mid y,z)$。
- **交换性／可积性（公式 13）：** 要求不同更新方向共用单一价值函数，会使不同条件化顺序下的奖励参数化相互关联，

$$r_z(x \mid y) - V(y,z) = r_x(z \mid y) - V(x,y).$$

本文余下部分将推导这些恒等式，并阐释其对推断—优化统一、智能体设计的解读，以及基于观测到的条件分布进行奖励可识别的限度所具有的启示。

3. 设定

环境模型。 固定离散随机变量（或随机变量组） X、 Y、 Z上的联合分布 $P(x,y,z)$。之所以选取三个变量，是因为推导代数结果需要它们。所有条件／边缘记号均由同一个 P 导出。我们针对所需条件分布有定义的上下文逐点进行分析。

价值函数。 价值是信息状态（事件）的函数：存在一个事件函数 V，使得 $V(x,y) = \mathrm{V}(\{X=x, Y=y\})$， $V(x,y,z) = \mathrm{V}(\{X=x, Y=y, Z=z\})$，等等。因此，由于底层事件相同，这些函数具有置换不变性（例如 $V(x,y) = V(y,x)$，以及 $V(x,y,z) = V(z,y,x)$）。按照控制领域的惯例，优化问题将 $V(x,y,z)$（即所有随机变量都已确定时）视为外生给定的终端价值（边界条件），而将 $V(y,z)$ 等内部价值（即仍存在剩余不确定性时）视为推导所得。

条件点互信息。 对于任意在 P 下条件分布定义良好的 x, y, z ，定义条件点态互信息（PMI）

$$i(x; z \mid y) := \log \frac{P(x, z \mid y)}{P(x \mid y)\, P(z \mid y)} = \log \frac{P(x \mid y, z)}{P(x \mid y)}. \tag{1}$$

这是条件互信息的逐点类比，并继承了标准恒等式（例如， x、 z 在给定 y时的对称性） [3, 17]。

后文将使用的假设。

1. **后验识别（A1）：** 当我们优化一个参考分布由环境联合分布 P诱导的 KL 正则化目标时，还会将优化器识别为同一 P下的真实后验条件分布。

2. **更新方向间的一致性（A2）：** 当我们针对多个更新方向施加后验识别（例如，在给定（Y， Z）时更新 X，以及在给定（Y， X）时更新 Z），我们要求具备基于事件的一致性，这通过价值函数的置换不变性来体现：价值只取决于已知的信息，而不取决于更新方向或获取信息的顺序。

4. KL 正则化条件更新

固定 $\alpha > 0$。对于任意实值函数 $r_z(x \mid y)$ 和终端价值 $V(x, y, z)$，对每个上下文 (y, z)，定义 KL 正则化目标，

$$J_z(\tilde{P}(x \mid y)) := \sum_x \tilde{P}(x \mid y) \Big[r_z(x \mid y) - \frac{1}{\alpha} \log \frac{\tilde{P}(x \mid y)}{P(x \mid y)} + V(x, y, z) \Big], \tag{2}$$

其中， $\tilde{P}(x \mid y)$ 遍历相对于 $P(x \mid y)$ 绝对连续的分布。这里， $r_z(x \mid y)$ 对应于通过“应用 z”获得的、 x 在给定 y 时的奖励。目前，可以将 z 视为执行优化计算的工具变量。这种自由能／KL 惩罚形式的目标在信息论有界理性和变分 KL 对偶性中很常见 [4, 5, 15, 16, 21]。

(2) 的最大化解存在且唯一，并具有指数倾斜形式

$$\tilde{P}^\star(x \mid y) = \frac{P(x \mid y) \exp\Big\{ \alpha \big[r_z(x \mid y) + V(x, y, z) \big] \Big\}}{\sum_{\tilde{x}} P(\tilde{x} \mid y) \exp\Big\{ \alpha \big[r_z(\tilde{x} \mid y) + V(\tilde{x}, y, z) \big] \Big\}}. \tag{3}$$

此外，最优值等于对数归一化因子

$$V(y, z) = \max_{\tilde{P}(x \mid y)} J_z(\tilde{P}(x \mid y)) = \frac{1}{\alpha} \log \sum_x P(x \mid y) \exp\Big\{ \alpha \big[r_z(x \mid y) + V(x, y, z) \big] \Big\}. \tag{4}$$

KL 正则化下的这种指数倾斜与对数配分函数结构是经典结果（例如，相对熵控制、线性可解控制和变分自由能） [4, 5, 7, 19, 20]。我们将这一标准事实作为代数起点；表示与可识别性结果则始于施加后验识别。

5. 后验识别确定了点互信息形式

假设（后验识别）。我们将优化器 (3) 识别为同一环境联合分布 P 下的后验条件分布：

$$P(x \mid y, z) := \tilde{P}^{\star}(x \mid y). \tag{5}$$

这是一项额外的建模假设（并非 KL 正则化的一般性质）：一般而言，KL 正则化目标会产生“类似后验”的更新，但不会与单个固定联合分布 [1, 21] 的贝叶斯条件分布重合。因此，优化工具变量 $Z = z$ 成为一个真正的随机变量，我们以它为条件，并假定它存在于一个包罗万象的环境分布中。这一假设最自然的动机来自控制即推断的观点：在辅助模型 [8, 11], 中，最优策略被表示为后验（或变分后验），而此处则进一步强化为在一个环境联合分布 P下的精确识别。

已识别交互项。 在（5）的条件下，更新（3）推出逐点对数比值恒等式

$$\log \frac{P(x \mid y, z)}{P(x \mid y)} = \alpha \Big[r_z(x \mid y) + V(x, y, z) - V(y, z) \Big]. \tag{6}$$

利用（1），这等价于点互信息形恒等式

$$\alpha \Big[r_z(x \mid y) + V(x, y, z) - V(y, z) \Big] = i(x; z \mid y), \tag{7}$$

在正概率上下文上逐点成立。

可识别性与规范结构。 后验识别只能确定软更新所诱导的相对动作偏好；它本身不能确定“奖励”和“价值”的唯一划分。在实际比较中，在固定上下文（y, z）下比较两个动作 x和 x'时，更新取决于它们的差值

$$\big(r_z(x \mid y) + V(x, y, z)\big) - \big(r_z(x' \mid y) + V(x', y, z)\big),$$

因此，任何与 x 无关的组合信号偏移都不会影响最终的条件分布。等价地，对于任意函数 $c(y, z)$，变换

$$r_z(x \mid y) \mapsto r_z(x \mid y) + c(y, z), \qquad V(y, z) \mapsto V(y, z) + c(y, z) \tag{8}$$

不会改变优化器（因而也不会改变 $P(x \mid y, z)$）。因此，奖励只能在仅依赖上下文的基线意义下被识别：后验分布确定了 $r_z(x \mid y) + V(x, y, z)$ 在 x 内的差异，但若无额外的校准约定，就无法确定其绝对水平。这是策略不变的奖励变换及现代（熵正则化）逆向强化学习中奖励可识别性障碍的一种依赖上下文的对应形式 [2, 9, 14]。

6. 以 PMI 重述的目标

方程（6）–（7）表明，可以将 PMI 用作奖励代理量，以确定一整类奖励—价值分解。为了解这一点，将（7）重新排列，可得到逐点分解

$$r_z(x \mid y) + V(x, y, z) = V(y, z) + \frac{1}{\alpha} i(x; z \mid y).$$

代入（2），对于每个上下文（y， z），可得：

$$J_z(\tilde{P}(x \mid y)) = V(y, z) + \sum_x \tilde{P}(x \mid y) \Big[\frac{1}{\alpha} i(x; z \mid y) - \frac{1}{\alpha} \log \frac{\tilde{P}(x \mid y)}{P(x \mid y)} \Big]. \tag{9}$$

因此，除仅依赖于上下文的项 $V(y,z)$（它不会影响优化器）外，在数值上相关的“类奖励”信号为 $\frac{1}{\alpha} i(x;z \mid y)$。也就是说，通过固定值 $i(x;z \mid y)$，我们便得到了一种构造方法，可由 $P(x|y)$ 构造后验 $P(x|y,z)$。

可容许 PMI 信号的约束。并非每一族函数 $i(\cdot;z \mid y)$ 都是可容许的。由于 $i(x;z \mid y)$ 必须与由单个环境联合分布 P 所诱导的后验对数比相一致，因此必须满足贝叶斯条件化所蕴含的归一化约束和相容性约束。具体而言，对于每个 (y,z)，

$$\log \sum_x P(x \mid y) \exp\{i(x;z \mid y)\} = 0,$$

从而 $P(x \mid y,z) = P(x \mid y) \exp\{i(x;z \mid y)\}$ 是一个有效的条件分布。此外，在不同的更新方向之间，顺序无关性 (A2) 还会通过交换性约束 (13) 对这些点互信息形式的交互项施加进一步的耦合。

7. 顺序无关性导出交换性约束

顺序无关性要求，后续过程只取决于已知的信息（信息事件），而不取决于其获知方式。在序贯／因果熵情境中，这是一项实质性的相干性要求，因为条件化结构和顺序都是建模的核心选择 [22, 23]。在当前的识别类中，点互信息形式的交互项由后验对数比确定，因此，要求在不同更新方向上使用同一个价值函数（假设 A2），会迫使两个方向的表示达成一致，从而导出可积性约束。

交换更新。我们不再假设变换 $P(x|y) \mapsto P(x|y,z)$，而是交换其方向，即 $P(z|y) \mapsto P(z|x,y)$。定义观察到 x 后优化 $\tilde{P}(z \mid y)$ 的式（2）对应形式：

$$J_x(\tilde{P}(z \mid y)) := \sum_z \tilde{P}(z \mid y)\left[r_x(z \mid y) - \frac{1}{\alpha} \log \frac{\tilde{P}(z \mid y)}{P(z \mid y)} + V(z,y,x)\right], \tag{10}$$

并对该更新同样施加后验识别：

$$P(z \mid y,x) := \arg\max_{\tilde{P}(z|y)} J_x(\tilde{P}(z \mid y)). \tag{11}$$

将 PMI 形恒等式应用于式（10），得到

$$\log \frac{P(z \mid y,x)}{P(z \mid y)} = \alpha\left[r_x(z \mid y) + V(z,y,x) - V(y,x)\right] = i(z;x \mid y). \tag{12}$$

由于 A2 施加了基于事件的一致性，$V(z,y,x) = V(x,y,z)$，且 $V(y,x) = V(x,y)$。此外，根据互信息的对称性，$i(z;x \mid y) = i(x;z \mid y)$。因此，两种更新顺序会得出相同的交互项（给定 y），从而得到交换性约束

$$r_z(x \mid y) - V(y,z) = r_x(z \mid y) - V(x,y). \tag{13}$$

约束（13）可视为一项可积性条件：一旦交互项固定（通过后验识别由点互信息确定），更新方向间的一致性便要求不同更新顺序下的奖励参数化彼此兼容。

8. 推广到变量的可数族

可数环境模型。令 $(X_i)_{i\in\mathbb{N}}$ 为一组可数个离散随机变量，其在乘积空间上的联合分布为 P。令 X、Y、Z 表示有限组坐标，并将相应的边缘分布记为$P(X,Y,Z)$。下文所有条件／边缘记号均由同一环境联合分布 P诱导。

最小适定性假设。固定一个有限的坐标组三元组 (X,Y,Z)，并在可数乘积空间上，使用环境联合分布的诱导边缘分布 $P(X,Y,Z)$。假设：

1. **正概率上下文**。对于每个我们写作 $P(\cdot \mid y)$ 或 $P(\cdot \mid y,z)$（或其交换形式）的上下文，假设 $P(y)>0$ 且 $P(y,z)>0$（相应地，$P(y,x)>0$），以使这些条件分布逐点有定义。

2. **有限对数归一化因子（指数可积性）**。对于每个感兴趣的 $(y,\ z)$，

$$Z(y,z):=\sum_{x}P(x\mid y)\exp\Big\{\alpha\big[r_z(x\mid y)+V(x,y,z)\big]\Big\}<\infty, \tag{14}$$

交换更新也类似。

备注（绝对连续性）。当所有条件分布都由同一个联合分布 P 诱导时，在正概率上下文上，绝对连续性是自动满足的：$P(x\mid y,z)>0\Rightarrow P\,x\mid y(>0)$，并且 $P(z\mid y,x)>0\Rightarrow P(z\mid y)>0$。我们不单独作此假设。

推论（软优化器的存在性与唯一性）。在 $Z(y,z)<\infty$ 的条件下，优化器存在且唯一，并且等于指数倾斜

$$\tilde{P}^{\star}(x\mid y;z)=\frac{P(x\mid y)\exp\{\alpha[r_z(x\mid y)+V(x,y,z)]\}}{Z(y,z)}.$$

事实上，对于任意 $q(\cdot\mid y)\ll P(\cdot\mid y)$，

$$J_z(q)=\frac{1}{\alpha}\log Z(y,z)-\frac{1}{\alpha}\mathrm{KL}\big(q(\cdot\mid y)\,\big\|\,\tilde{P}^{\star}(\cdot\mid y;z)\big),$$

因此 $J_z(q)\le\alpha^{-1}\log Z(y,z)$，当且仅当 $q=\tilde{P}^{\star}$ 时取等号。所有逐点恒等式（对数比、PMI 形式、规范不变性、可交换性）随后都适用于边缘三元组 (X,Y,Z)，与有限情形完全相同。

9. 讨论

我们的假设是有意设定得较强的，其回报是在概率条件分布与 KL 正则化优化原语之间建立清晰而精确的对应关系。

后验识别所确定的内容。在后验识别（A1）下，软优化器是同一联合分布 P 下真正的条件分布，因此后验对数比在每个点上都被确定。等价地，经由式（7），被识别的对象是交互项 $r_z(x\mid y)+V(x,y,z)-V(y,z)$。这是一个表示结果：点互信息并非假设而来，而是通过将优化器识别为同一环境联合分布下的贝叶斯后验而必然成立。

为何缺少校准时奖励存在歧义。由于只有交互项被确定，“奖励”与“软价值”的分解会继承仅依赖上下文的基线不变性（8）。这正是无法将 $r_z(x \mid y)$ 解读为后验更新的逐点奖励表示的确切障碍。

顺序鲁棒性作为相干性约束。 如果我们要求使用单一延续事件函数（A2）对不同更新方向上的后验进行识别，那么不同条件化顺序之间的相干性就会施加交换性约束(13)。这是一项可积性条件：一旦点互信息形式的交互项确定，按方向索引的奖励族就必须在不同更新顺序之间彼此协调。

含义。 心理学： 在识别类中，驱动新信息 z 引发行为变化的、依赖于 x的信号是条件对数证据比 $i(x;z \mid y)$ （按 $1/\alpha$ 缩放）。将其视为奖励代理时， $i(x;z \mid y)$ 量化了智能体在纳入超出基准上下文 y的 z 时，被“拉向”或推离结果 x 的程度。规范自由度（8）阐明了为何观测到的条件分布能确定相对 倾向的变化，却无法在未经校准的情况下确定绝对奖励基准。

经济学： 相较于萨维奇式主观期望效用，识别类并非一般性的偏好表征：精确的后验识别将与效用相关的交互项和后验对数似然比联系起来。因此，智能体的“仿佛收益”是信息性的，而非独立于信念的自由原始要素。KL 项体现了信息处理（或理性疏忽）成本，因此，除非将后果空间扩展到包含信息努力，否则对随机策略的偏好通常不会对混合呈线性。最后，在不同更新方向上施加单一的基于事件的延续价值，会导出路径无关约束（13）；这一相干性限制在经典主观期望效用中并不存在，并且恰恰会在顺序效应阻碍单一底层表示 [12, 18] 时遭到违反。

控制： 在同一环境 P下，后验识别为“控制即推断”划定了明确的边界：如果 KL 正则化更新确实 是贝叶斯条件分布，那么交互项就必须具有点互信息形式，而设计/学习自然应以已识别交互项为对象，而非绝对奖励。在 A2 下，模块化更新方向并非彼此独立的自由度；可交换性会将它们耦合起来。重要的是，这并不意味着所选更新方向上的设计自由度消失：可以针对实际采用的条件化顺序（如序贯决策）指定奖励/终端价值，但此时，后验识别加上 A2 会通过式（13）限定其他条件化顺序所兼容的表示。

局限性。 本分析依赖于在单一环境联合分布下进行精确的后验识别，并将延续价值视为已指定。放宽其中任一假设，都会回到类似后验（但并非与贝叶斯完全相同）的软更新，并重新引入可识别性缺口；本文仅在完全识别的情形下，明确分析精确的逐点约束。

10. 相关工作

KL 正则化控制与软价值。KL（相对熵）正则化会产生指数倾斜优化解和对数配分“软”价值，在线性可解控制 [19] 和控制—估计对偶性 [20]中，其结构尤为清晰。这些恒等式背后的变分基础是经典的 [4, 5, 7, 21]。

控制即推断与类后验更新。大量文献将（正则化）控制解释为概率推断，通常借助辅助图模型，其中最优策略对应于后验或变分后验 [8, 10, 11]。广义贝叶斯更新进一步阐明了损失倾斜后验分布何时源于优化 [1]。我们的设定限制更为严格：我们要求在单一环境联合分布 P 下实现精确的后验识别，并研究由此产生的逐点结果。

奖励歧义与可识别性。 众所周知，行为通常无法唯一确定奖励：奖励变换可以在保持策略不变的同时导出等价类 [14]。现代熵正则化逆向强化学习中的可识别性分析，刻画了何时能够恢复奖励 [2, 9, 24]。我们的“规范自由度”视角以识别后验为出发点：后验确定了一个基线不变交互项，而要指定逐点奖励，则需要在后验信息之外额外采用基线约定。

顺序与一致性。 序贯决策场景中的熵/KL正则化促使我们谨慎地明确条件化结构和信息流 [6, 13, 22, 23]。在此，我们通过要求使用单一的延续事件函数，确保不同更新方向间的一致性，从而得到交换性约束（13）。

11. 结论

我们施加了后验识别条件（优化器是单一联合分布 P 下的真实条件分布）和顺序无关性条件（不同方向共用单一延续事件函数）。后验识别迫使更新采用 PMI 形式（7）：后验对数比逐点确定了更新，因此也确定了 $i(x;z \mid y)$，并识别出交互项 $r_z(x \mid y) + V(x,y,z) - V(y,z)$。这揭示了基线规范自由度（8）：如果没有基线约定，后验就无法唯一确定奖励。最后，要求不同更新方向共用单一延续事件函数，会导出交换性约束（13），从而将按方向索引的奖励参数化联系起来，这是获得顺序稳健表示所必需的。

参考文献

[1] P. G. Bissiri、C. C. Holmes 和 S. G. Walker。更新信念分布的一般框架。皇家统计学会志：*B*辑（统计方法学），78(5):1103–1130，2016。[2] H. Cao、S. N. Cohen 和 Ł. Szpruch。逆强化学习中的可识别性。载于神经信息处理系统进展 *34*，2021。[3] T. M. Cover 和 J. A. Thomas。信息论基础。Wiley-Interscience，第2版，2006。[4] I. Csiszár。概率分布的 *I*-散度几何与极小化问题。概率年刊，3(1):146–158，1975。[5] M. D. Donsker 和 S. R. S. Varadhan。大时间下某些马尔可夫过程期望的渐近求值。II。纯粹与应用数学通讯，28(2):279–301，1975。[6] M. Geist、B. Scherrer 和 O. Pietquin。正则化马尔可夫决策过程理论。载于第*36*届国际机器学习会议论文集，第97卷机器学习研究论文集，第2160–2169页。PMLR，2019。[7] C. Hartmann、L. Richter、C. Schütte 和 W. Zhang。自由能的变分刻画：理论与算法。*Entropy*，19(11):626，2017。[8] H. J. Kappen、V. Gómez 和 M. Opper。将最优控制视为图模型推断问题。*CoRR*，abs/0901.0633，2009。[9] K. Kim、S. Garg、K. Shiragur 和 S. Ermon。逆强化学习中的奖励识别。载于M. Meila 和 T. Zhang（编），第*38*届国际机器学习会议论文集，第139卷机器学习研究论文集，第5496–5505页。PMLR，2021。[10] D. Koller 和 N. Friedman。概率图模型：原理与技术。麻省理工学院出版社，2009。[11] S. Levine。将强化学习与控制视为概率推断：教程与综述。*CoRR*，abs/1805.00909，2018。

[12] B. Maćkowiak、F. Matějka 和 M. Wiederholt。理性疏忽：综述。经济文献杂志，61(1):226–273，2023年3月。[13] G. Neu、A. Jonsson 和 V. Gómez。熵正则化马尔可夫决策过程的统一观点。*CoRR*，abs/1705.07798，2017。[14] A. Y. Ng、D. Harada 和 S. Russell。奖励变换下的策略不变性：理论及其在奖励塑造中的应用。载于第*16*届国际机器学习会议（*ICML*）论文集，第278–287页，1999。[15] P. A. Ortega 和 D. A. Braun。热力学：一种考虑信息处理成本的决策理论。英国皇家学会会刊*A*：数学、物理与工程科学，469(2153):20120683，2013。[16] P. A. Ortega、D. A. Braun、J. Dyer、K.-E. Kim 和 N. Tishby。信息论有界理性。2015。[17] C. E. Shannon。通信的数学理论。贝尔系统技术期刊，27:379–423，623–656，1948。[18] C. A. Sims。理性疏忽的影响。货币经济学杂志，50(3):665–690，2003。[19] E. Todorov。线性可解马尔可夫决策问题。载于神经信息处理系统进展 *19*，第1369–1376页，2006。[20] E. Todorov。最优控制与估计之间的一般对偶关系。载于第*47*届*IEEE*决策与控制会议（*CDC*）论文集，第4286–4292页，2008。[21] M. J. Wainwright 和 M. I. Jordan。图模型、指数族与变分推断。机器学习基础与趋势，1(1–2):1–305，2008。[22] B. D. Ziebart、J. A. Bagnell 和 A. K. Dey。通过最大因果熵原理对交互进行建模。载于第*27*届国际机器学习会议（*ICML*）论文集，第1255–1262页，2010。[23] B. D. Ziebart、J. A. Bagnell 和 A. K. Dey。最大因果熵原理在估计交互过程中的应用。*IEEE*信息论汇刊，59(4):1966–1980，2013。[24] B. D. Ziebart、A. L. Maas、J. A. Bagnell 和 A. K. Dey。最大熵逆强化学习。载于第*23*届*AAAI*人工智能会议（*AAAI*）论文集，第1433–1438页，2008。