arXiv:2301.01045v2 [cs.LG] 4 Jan 2023

奖励歧义下的风险规避马尔可夫决策过程

阮浩林香港城市大学数据科学学院，香港九龙塘
haolin.ruan@my.cityu.edu.hk

Zhi Chen香港城市大学商学院管理科学系，香港九龙塘 zhi.chen@cityu.edu.hk

ChinPang Ho香港城市大学数据科学学院，香港九龙塘
clint.ho@cityu.edu.hk

我们提出一种用于应对风险和奖励歧义的马尔可夫决策过程（MDP）分布鲁棒收益-风险模型。该模型优化均值表现与百分位表现的加权平均，并将分布鲁棒马尔可夫决策过程和分布鲁棒机会约束马尔可夫决策过程（两者均考虑奖励歧义）作为特例。考虑到未知奖励分布属于一个瓦瑟斯坦歧义集，我们推导出该模型的可处理重构形式。具体而言，我们证明，当只考虑确定性策略时，收益-风险模型也可以计入不确定转移核带来的风险；同时，基于百分位准则的分布鲁棒马尔可夫决策过程可以重构为风险水平经过调整后的名义对应形式。我们设计了一种可扩展的一阶算法来求解大规模问题，并通过数值实验展示了所提模型和算法的优势。

1. 引言

马尔可夫决策过程（马尔可夫决策过程）为随机动态环境中的序贯决策问题和强化学习提供了强大的建模框架（Puterman，2014年）。然而，在实践中，获取能够准确反映环境的马尔可夫决策过程模型参数始终是一项挑战，因为这些参数是根据有限且可能受到污染的数据估计得出的（Mannor 等，2007年）。此外，转移核和奖励函数等参数通常随时间变化，甚至具有不确定性，但在过度简化的设定中，它们却被近似为固定值（Mannor 等，2016年）。因此，马尔可夫决策过程输出的策略在实践中往往令人失望。

鲁棒马尔可夫决策过程通过允许转移核和奖励函数的未知值落在给定的歧义集中，来解决上述参数歧义问题（Behza- dian等人 2021年，Chen等人 2019年，Clement和Kroer 2021年a，Delgado等人 2016年）。随后，鲁棒马尔可夫决策过程寻求在所有转移核

和歧义集中的奖励函数上，使最坏情形期望回报最大化的策略。通过指定以高置信度包含未知转移核的歧义集，鲁棒马尔可夫决策过程的最优策略能够抵御参数歧义（Iyengar 2005年）。

本文聚焦于奖励函数存在歧义的情形，这种情形有时也称为模糊奖励马尔可夫决策过程（Alizadeh 等人，2015年；Regan 和 Boutilier，2010年、2011a,b、2012年）。这一特定问题也与模仿学习密切相关。模仿学习旨在训练智能体学习专家的某种行为，而可用的仅有专家的一些示范轨迹（Chen等人，2020年；何和Ermon，2016年；Osa 等人，2018年；Rashidinejad 等人，2021年）。采用逆向强化学习方法来学习能够完整表征专家偏好的奖励函数时（Brown 等人，2020年；Choi 和 Kim，2012年；Ng 等人，2000年），所得策略会受到奖励歧义的影响，因此在实践中可能表现不佳。

为处理奖励歧义，我们采用分布鲁棒优化（DRO）（Derman和Mannor，2020年）以及分布鲁棒机会约束规划（Chen等人，2007年；Postek等人，2018年）中的技术，并假设真实奖励分布属于某个歧义集。这种方法不要求精确指定奖励函数，而只需描述歧义集中的常见分布信息，例如支撑集、矩和形状；这些信息通常更容易获取或估计（Hanasusanto等人，2015年、2017年；Zymler等人，2013）。本文采用的分布鲁棒模型使用了Abdullah等人（2019年）、Calafiore和Ghaoui（2006年）以及Xie（2021）所述的瓦瑟斯坦歧义集。与可能包含过于极端的成员分布的φ-散度歧义集不同，瓦瑟斯坦集将支撑集内各点之间的接近程度纳入考量，因此其成员分布可能更加合理（Gao和Kleywegt，2022年）；另一方面，当样本数量太少，无法可靠估计矩时，瓦瑟斯坦集通常比基于矩的歧义集更合适（Yang，2020年）。基于这些原因，我们选择瓦瑟斯坦集，尽管文献中也考虑了其他类型的歧义集，例如嵌套歧义集（Xu和Mannor，2010年、2012年）以及基于Prohorov度量的歧义集（Erdo˘gan和Iyengar，2006年）。对于我们的分布鲁棒机会约束马尔可夫决策过程，我们还将证明，它与风险水平经过调整的名义对应问题等价。据我们所知，这是马尔可夫决策过程领域中首个确立分布歧义与风险之间相互转换关系的结果。

我们的收益-风险模型（RR）是一种风险规避型MDP模型，不仅考虑奖励歧义，还同时考虑回报的平均值和风险。将回报风险而非期望成本最小化的MDP称为风险感知型马尔可夫决策过程（也称风险敏感型或风险厌恶型MDP）（Ahmadi等（2021），B¨aauerle和Rieder 2017，Carpin等（2016），Haskell和Jain（2015），Huang和Haskell（2017））。在风险感知优化中，目标函数取为

一种风险度量，例如风险价值（VaR）（Delage 和 Mannor 2007年、2010年，Gilbert等人2017年）、条件风险价值（CVaR）（B¨auerle和Ott 2011，Chow等人2017，Huang和Guo 2016）以及其他谱风险度量（B¨auerle和Glauner 2021），还有期望效用的变体（Bernard等人 2022，Jaimungal等人 2022，Pflug和Wozabal 2007年）。

在这些风险度量中，VaR 和 CVaR 可以说是最受欢迎的两种，并且吸引了许多研究人员的关注（B¨auerle 和 Ott 2011，Chow 等，2017，Delage 和 Mannor 2007、2010，Gilbert 等，2017，Huang 和 Guo 2016）。使用 CVaR 的目的，是精确刻画（不确定奖励）分布的极端尾部；而 VaR 无法反映超出 VaR 的极端情形。众所周知，CVaR 是一种一致性风险度量，可以借助凸优化工具高效优化（Chen 和 Xie 2021）；相比之下，VaR 是一种更难处理的风险度量，因为它并非一致性风险度量。

VaR 的一个显著优势是估计稳定性（尤其是在奖励分布厚尾时（Sarykalin 等，2008）），这在数据驱动环境中尤为重要，因为此类环境下样本数量有限，决策者会依据模型的样本外表现来评估模型（Bertsimas 和 Thiele 2006，van de Berg 等，2022，Zheng 等，2016）。为作说明，我们举一个例子：考虑一个仅有 1 个状态 s 和 2 个动作 a_1 和 a_2 的单步马尔可夫决策过程（Sutton 和 Barto 2018）。在这个单步马尔可夫决策过程中，决策者在每个回合中只做出一次决策，并希望最大化这些回合奖励的 VaR/CVaR。我们考虑不确定奖励 $\tilde{r}_{s,\ a_1} \sim \mathbb{P}_{t\text{-dist}}$ 和 $\tilde{r}_{s,\ a_2} = \tilde{r}_{s,\ a_1} + \rho|s|$，其中 $\mathbb{P}_{t\text{-dist}}$ 服从 Student's t 分布，并改变其自由度 $\delta \in \{2,3,4\}$。我们设定平移比率 $\rho = \{0.05i\}_{i\in[5]}$；为了测试 VaR（相应地，CVaR）的估计准确性（其中风险阈值取 10%），我们将平移量 s 设为 $\mathbb{P}_{t\text{-dist}}$-VaR$_{0.1}[\tilde{r}_{s,a_1}]$（相应地，$\mathbb{P}_{t\text{-dist}}$-CVaR$_{0.1}[\tilde{r}_{s,a_1}]$），这两种风险度量都可以高效计算（详情见附录B）。我们将决策者的准确率评估为：在测试样本中，她选择了奖励 VaR/CVaR 较高的动作（即，动作 a_2）的样本所占比例；对于每一组准确率和平移比率，根据 Yamai 等（2002）的做法，决策者对每个状态-动作对都有 1000 个随机奖励样本，我们根据 10000 个测试样本来检验其准确率。

如图1所示，准确率会随移位比率 ρ 的增加而提高。随着 δ 减小， $\mathbb{F}$ 的重尾特征会更加明显，而风险价值的准确率显著高于条件风险价值，这表明对风险价值进行统计推断会比对条件风险价值进行统计推断更加准确。因此，样本集较小时，风险价值可能是更可取的选择。

我们的收益-风险模型受到软鲁棒准则/模型的启发。在优化文献中，该准则/模型对均值与鲁棒性能进行凸组合优化（Ben-Tal 等，2010）。近年来，软鲁棒性马尔可夫决策过程也颇受欢迎，决策者旨在

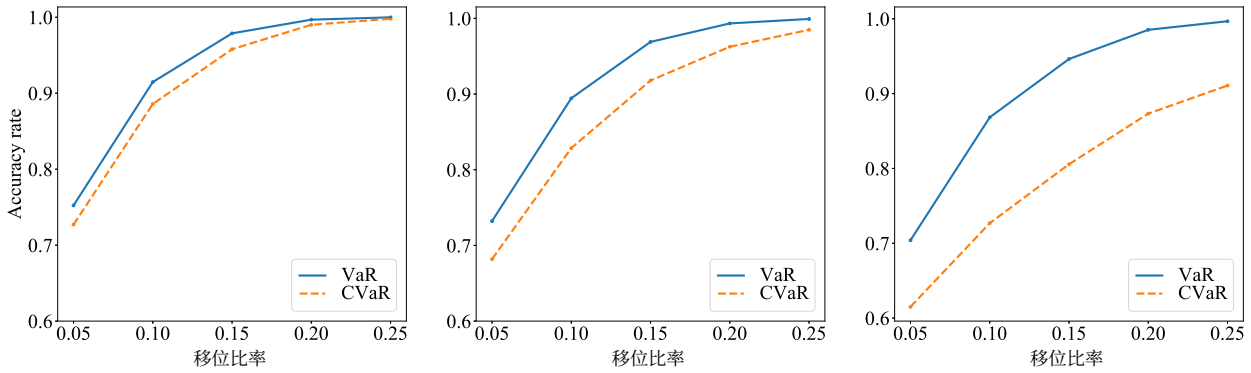


图1 决策者选择正确行动（从而使其奖励的VaR/CVaR最大化）的准确率： $\delta = 4$ （左）、 $\delta = 3$ （中）和 $\delta = 2$ （右）。

最大化均值与百分位表现的加权平均（Brown 等人，2020；Lobo 等，2020）。然而，与这些已有的软鲁棒马尔可夫决策过程不同，本文提出的收益-风险模型在两个方面存在根本区别：首先，这些已有的软鲁棒模型并未考虑奖励歧义，而我们利用分布鲁棒性来处理奖励歧义，从而能够针对奖励分布（歧义集内）中最具对抗性的实现进行对冲，因此我们的模型对奖励不确定性具有更强的鲁棒性（Chen等人，2019；Xu 和 Mannor，2010）；其次，我们选择风险价值作为风险度量，它与百分位表现有直接对应关系，而且如上所述，在数据驱动优化中往往更具优势。

我们的工作聚焦于基于模型的设定，所提出的模型受到经典的（对偶形式的）名义马尔可夫决策过程（Puterman 2014）和机会约束马尔可夫决策过程（Delage 和 Mannor 2010）的启发。值得注意的是，除了基于模型的设定外，关于鲁棒强化学习还有其他富有启发性且创新的研究，例如鲁棒TDC算法和鲁棒Q学习（Roy et al. 2017，Wang and Zou 2021）、鲁棒策略梯度（Wang and Zou 2022）、最小二乘策略迭代（Lagoudakis and Parr 2003）以及样本复杂度分析（Panaganti and Kalathil 2022）。需要指出的是，尽管无模型强化学习可用于为复杂环境学习令人满意的策略，但与环境进行大量交互的要求可能会使学习过程变慢（Kaiseretal. 2019）；而较高的样本效率是基于模型学习的一项显著优势（Sutton and Barto 2018）。我们还注意到，转移核存在歧义的马尔可夫决策过程是另一个活跃的研究方向，其中广泛采用分布鲁棒性（Clement and Kroer 2021b，Shapiro2016，2021，Xu and Mannor 2012）。

我们可以将我们的贡献总结如下（此外，我们还在附录 I 的表 2 中将我们的贡献与相关工作的贡献进行了比较）。

(i) 我们证明了，优化期望奖励的分布鲁棒模型可以重新表述为一个凸锥规划；该规划等价于在目标函数中加入凸正则化的名义马尔可夫决策过程。

*(ii)*对于分布鲁棒机会约束马尔可夫决策过程（DCC），我们证明其可以重新表述为在调整后的风险水平下的名义机会约束马尔可夫决策过程。这一发现弥合了风险与参数歧义之间的差距。

(iii) 结合*(i)* 和*(ii)*中提出的模型，我们提出了回报风险马尔可夫决策过程，该过程最大化奖励的期望与风险价值的加权平均（两者均针对奖励不确定性具有分布鲁棒性），具有较大的灵活性，并且在均值和百分位收益标准下均能表现良好。

(iv) 当仅考虑确定性策略时，我们证明了回报风险模型还可以纳入不确定转移核带来的风险，并将其等价地重新表述为混合整数二阶锥规划（MISOCP）。

(v) 为求解所提出的回报风险模型，我们设计了一种比MOSEK求解器更具可扩展性的一阶方法，因此在大规模问题上速度更快。

(vi) 在仿真和实证实验中，我们采用数据驱动设定，决策者旨在最大化随机奖励的期望和风险价值。我们比较了分布鲁棒马尔可夫决策过程（DRMDP）、分布鲁棒机会约束（DCC）、收益-风险马尔可夫决策过程（RR）、鲁棒马尔可夫决策过程（RMDP）（Delage 和 Mannor 2010）以及 BROIL（Brown 等人 2020）的表现。结果表明，第三种方法在期望和不同风险价值（风险阈值为 5%，10% 和 15%）下均表现最佳，展现了其优势以及对决策者在回报与风险之间不断变化的偏好的适应能力。

本文其余部分安排如下。第2节介绍背景。第3节和第4节分别研究分布鲁棒马尔可夫决策过程以及分布鲁棒机会约束模型，并推导其可处理的重构形式。结合这些提出的模型，我们在第5节提出收益-风险马尔可夫决策过程模型。第6节详细介绍为该收益-风险马尔可夫决策过程模型设计的一阶算法。我们比较分布鲁棒马尔可夫决策过程、分布鲁棒机会约束、收益-风险马尔可夫决策过程、鲁棒马尔可夫决策过程和 BROIL 的表现，并在第 7 节展示所提算法的优势。第8节给出结论。

2. 背景

我们考虑一个无限时域马尔可夫决策过程，其状态空间 $\mathcal{S}=\{1,\cdots,S\}$ 和动作空间 $\mathcal{A}=\{1,\cdots,A\}$ 均为有限集。令 $\boldsymbol{P}\in\mathbb{R}^{S\times A\times S}$ 为转移概率核，其中 $p_{s,\ a,\ s'}$ 表示在状态 $s\in\mathcal{S}$ 选择行动 $a\in\mathcal{A}$ 时转移至状态 $s'\in\mathcal{S}$ 的转移概率；因此，对于每个 $(s,\ a)\in\mathcal{S}\times\mathcal{A}$，$\boldsymbol{p}_{s,\ a}\in\Delta^S$ 表示转移概率分布。给定状态-动作对 $(s,\ a)$，智能体将获得期望奖励 $r_{s,\ a}\in\mathbb{R}$。为简化记号，我们将奖励函数表示为向量 $\boldsymbol{r}=\{r_{s,a}\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$。

我们寻求最优平稳随机策略 $\boldsymbol{\pi}=\{\boldsymbol{\pi}_s\}_{s\in\mathcal{S}}$，使得对所有 $s\in\mathcal{S}$ 均有 $\boldsymbol{\pi}_s\in\Delta^A$；在状态 $s\in\mathcal{S}$ 下，将以概率 $\pi_{s,\,a}$ 采取行动 $a\in\mathcal{A}$。最大化期望奖励的名义马尔可夫决策过程可表述为（Puterman，2014年）

$$\ell_{\mathrm{N}}=\max_{\boldsymbol{x}\in\mathcal{X}}\ \boldsymbol{r}^{\top}\boldsymbol{x}, \tag{1}$$

其中，可行集 $\mathcal{X}$ 由 $\mathcal{X}=\left\{\boldsymbol{x}\in\mathbb{R}_{+}^{SA}\ \middle|\ (\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})\boldsymbol{x}=\boldsymbol{p}_0\right\}$. 给出。这里，系数矩阵 $\boldsymbol{E}=\mathrm{diag}(\boldsymbol{e}^{\top},\cdots,\boldsymbol{e}^{\top})\in\mathbb{R}^{S\times SA}$，其中 S为全 1 向量，$\boldsymbol{e}\in\mathbb{R}^{A}$ 且 $\bar{\boldsymbol{P}}=(\bar{\boldsymbol{p}}_1,\cdots,\bar{\boldsymbol{p}}_S)^{\top}\in\mathbb{R}^{S\times SA}$，满足 $\bar{\boldsymbol{p}}_s=\{p_{s',a,s}\}_{(s',a)\in\mathcal{S}\times\mathcal{A}}$ 对所有 $s\in\mathcal{S}$均成立。对于每个 $(s,a)\in\mathcal{S}\times\mathcal{A}$，我们将 s^{th} 的子向量记为 $\boldsymbol{x}_s=\{x_i\}_{i\in\{(s-1)A+1,\cdots,sA\}}$；其 a^{th} 分量 $x_{s,a}$ 可解释为应用策略$\pi^{\star}_{s,a}=x^{\star}_{s,a}/(\sum_{a\in\mathcal{A}}x^{\star}_{s,a})\ \forall(s,a)\in\mathcal{S}\times\mathcal{A}$(Puterman 2014)[1]时，占据状态 s 并选择行动 a 的总折扣概率。折扣因子 $\gamma\in$属于 (0, 1)，初始状态的初始分布为 $\boldsymbol{p}_0\in\mathbb{R}_{++}^{S}$。问题 (1) 是一个线性规划，可通过单纯形法和内点法高效求解（Nocedal 和 Wright，2006年）。还可以通过对该问题对应的贝尔曼方程应用值迭代或策略迭代，高效计算最优策略（Bertsekas 和 Tsitsiklis，1995，Puterman，2014年）。

名义马尔可夫决策过程 (1) 未考虑奖励或转移核中的不确定性。为考虑奖励不确定性，Delage 和 Mannor（2010）假设随机奖励向量 $\tilde{\boldsymbol{r}}$ 服从已知的高斯分布 $\mathbb{P}$，并提出如下机会约束马尔可夫决策过程模型：

$$\ell_{\mathrm{CC}}(\varepsilon)=\begin{cases}\max\ y\\ \text{s.t.}\ \ \mathbb{P}[\tilde{\boldsymbol{r}}^{\top}\boldsymbol{x}\geq y]\geq 1-\varepsilon\\ \qquad\boldsymbol{x}\in\mathcal{X},\ y\in\mathbb{R}.\end{cases} \tag{2}$$

事实上，上述机会约束模型针对分布 $\mathbb{P}$ 最大化奖励的 VaR（风险水平为 $1-\varepsilon$）。由于假设 $\mathbb{P}$ 服从高斯分布，根据 Pr´ekopa（2013）的定理 10.4.1，可以将问题 (2) 重构为如下二阶锥规划：

$$\ell_{\mathrm{CC}}(\varepsilon)=\max_{\boldsymbol{x}\in\mathcal{X}}\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{r}}^{\top}\boldsymbol{x}]-\|\mathrm{F}^{-1}(1-\varepsilon)\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2,$$

其中，$\mathrm{F}^{-1}(\cdot)$ 是高斯分布 $\mathbb{P}$的累积分布函数的反函数，$\boldsymbol{\Sigma}$ 是 $\mathbb{P}$的协方差矩阵。借助 CPLEX、Gurobi 和 MOSEK 等先进的商业求解器，可以高效求解二阶锥规划（参见例如Ben-Tal 和 Nemirovski（2001年））。尽管机会约束马尔可夫决策过程 (2) 具有可处理性，但它需要将精确的底层奖励分布作为输入。此外，上述重构对于一般的分布 $\mathbb{P}$并不成立。

[1]根据 Puterman（2014），任何 $\boldsymbol{x}\in\mathcal{X}$ 都有这样的解释，因此我们可以通过这种方式求出本文所有所提模型的策略。

3. 分布鲁棒马尔可夫决策过程

在许多现实情形中，很难（甚至不可能）获得不确定奖励的真实分布。我们可能只能掌握一些确切信息，例如其矩和形状。作为处理这类情形最有效的方法之一，分布鲁棒优化方法将不确定性建模为一个随机变量，该变量服从某个未知的概率分布，而该分布属于一个歧义集。面对分布歧义，决策者寻求能够抵御歧义集中最具对抗性的分布的解。具体而言，在本文的背景下，我们假设不确定奖励的真实分布位于以某个参考分布 $\mathbb{P}$为中心、半径为 $\hat{}\theta \geq 0$ 的 Wasserstein 球内:

$$\mathcal{F}(\theta) = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^{SA}) \mid d_{\mathrm{W}}\big(\mathbb{P}, \hat{\mathbb{P}}\big) \leq \theta\}. \tag{3}$$

这里， $\mathcal{P}(\mathbb{R}^{SA})$ 表示定义在 $\mathbb{R}^{SA}$ 上的所有概率分布的集合；在 $\mathbb{R}^{SA}$ 中配备一般范数 $\|\cdot\|$ 时，两个分布 $\mathbb{P}_1$ 和 $\mathbb{P}_2$ 之间的瓦瑟斯坦距离为$d_{\mathrm{W}}(\mathbb{P}_1, \mathbb{P}_2) = $下确界$_{\mathbb{P}\in\mathcal{Q}(\mathbb{P}_1,\mathbb{P}_2)}$ $\mathbb{E}_{\mathbb{P}}[\|\tilde{\boldsymbol{r}}_1 - \tilde{\boldsymbol{r}}_2\|]$，其中，$\mathcal{Q}(\mathbb{P}_1, \mathbb{P}_2)$ 表示所有联合分布的集合，其边缘分布为 $\mathbb{P}_1$ 和 $\mathbb{P}_2$ ，分别支配 $\tilde{\boldsymbol{r}}_1$ 和 $\tilde{\boldsymbol{r}}_2$。

名义马尔可夫决策过程 (1) 中的随机参数是奖励期望，在实践中通常通过历史样本的均值进行估计。然而，当样本量较小时，这样的样本均值并不接近期望，反而被认为存在乐观偏差（参见例如，Smith 和 Winkler（2006年））。因此，基于样本的名义马尔可夫决策过程 (1) 可能产生令人不满意的策略，其样本外表现不佳。基于这一原因，一种可能的替代方案是最大化最坏情况期望奖励，如下述分布鲁棒马尔可夫决策过程中所示:

$$\ell_{\mathrm{DRMDP}}(\theta) = \max_{\boldsymbol{x}\in\mathcal{X}} \inf_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{r}}^{\top}\boldsymbol{x}]. \tag{4}$$

下面的命题给出了式 (4) 的等价锥规划。

PROPOSITION 1. 分布鲁棒马尔可夫决策过程 (4) 可以重构为一个锥规划

$$\ell_{\mathrm{DRMDP}}(\theta) = \max_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}}[\tilde{\boldsymbol{r}}^{\top}\boldsymbol{x}] - \theta\cdot\|\boldsymbol{x}\|_{*}.$$

不难发现，在参考分布 $\mathbb{P}$下，分布鲁棒马尔可夫决策过程可以视为名义马尔可夫决策过程（4）的一种凸正则化。具体而言，分布鲁棒马尔可夫决策过程中的凸正则化项为 $\theta\|\boldsymbol{x}\|_{*}$，其大小由 Wasserstein 半径 θ决定。有趣的是，我们还发现，（分布）乐观马尔可夫决策过程可以重构为带有（凹）正则化项 $-\theta\|\boldsymbol{x}\|_{*}$的反向锥规划。我们将这一结果放在附录 D 中。

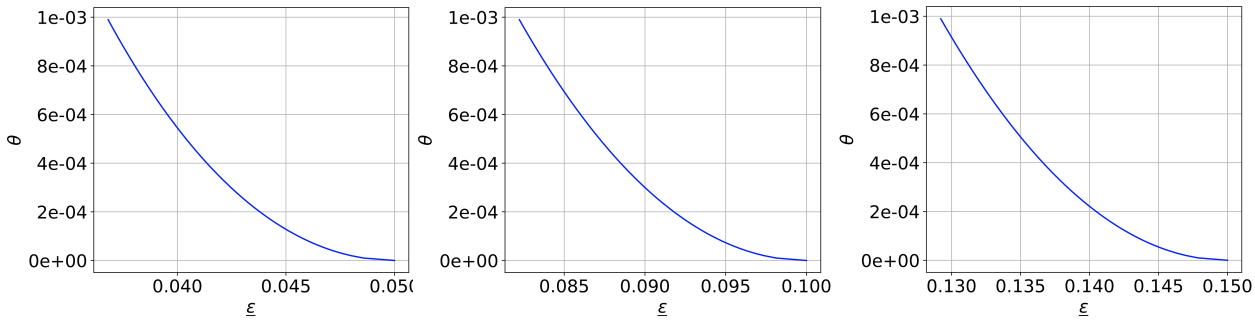


图 2 ε 相对于不同 θ的取值： $\varepsilon = 0.05$ （左）、 $\varepsilon = 0.1$ （中）和 $\varepsilon = 0.15$ （右）。

我们指出，问题（4）确实是 Jaimungal 等人（2022）所考虑的鲁棒优化问题的一个特例，其中我们采用了期望效用框架。与 Jaimungal 等人（2022）提出的策略梯度方法相比（其收敛性尚未得到证明），我们推导出了该问题的等价重构形式，即一个可处理的锥规划，可使用 Gurobi、MOSEK 和 CPLEX 等先进的商业求解器高效求解，并且还可无缝纳入我们在第5节提出的收益-风险模型的可处理重构形式中。

4. 分布鲁棒机会约束马尔可夫决策过程

本节中，我们将关注点从优化奖励期望转向其尾部表现，并通过研究机会约束马尔可夫决策过程来实现这一点。具体而言，我们仍然考虑 Wasserstein 歧义集 (3) ˆ 来处理分布歧义，同时在 Wasserstein 距离的定义中指定参考分布 $\mathbb{P}$ 和范数 $\|\cdot\|$。

对于前者，本节始终关注椭圆参考分布 ˆ $\mathbb{P} = \mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$[2] ，其概率密度分布为 $f(\boldsymbol{r}) = k \cdot g\left(\frac{1}{2}(\boldsymbol{r}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{r}-\boldsymbol{\mu})\right)$，其中 k 是正的归一化标量， $\boldsymbol{\mu}$ 是均值向量， $\boldsymbol{\Sigma}$ 是正定矩阵， gˆ 是生成函数。我们强调，对 $\mathbb{P}$ 作出这一假设并不严格，因为它仅是歧义集的中心。具体而言，我们提出的分布鲁棒机会约束马尔可夫决策过程可以涵盖所有类型的分布（只要它们位于歧义集内），并不局限于椭圆分布。正如我们将看到的，这些设定能够导出所提模型的可处理重构形式。关于椭圆分布的预备知识见附录C。

对于后者，我们采用与正定矩阵 $\boldsymbol{\Sigma}$相关联的马氏范数，由 $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}}$表示。注意，马氏范数 $\|\cdot\|_{\boldsymbol{\Sigma}}$的对偶范数是另一种马氏范数 $\|\cdot\|_{\boldsymbol{\Sigma}^{-1}}$，它由逆矩阵 $\boldsymbol{\Sigma}^{-1}$定义。

[2]请注意，第3节中的结果适用于一般的参考分布。

在分布鲁棒机会约束马尔可夫决策过程中，我们希望即使在最坏情况下，奖励也能以较高置信度不低于某个下界，并旨在通过求解以下问题来最大化这个下界

$$\ell_{\mathrm{DCC}}(\theta,\varepsilon)=\begin{cases}\max\ y\\ \text{s.t.}\ \inf\limits_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\varepsilon\\ \quad\ \boldsymbol{x}\in\mathcal{X},\ y\in\mathbb{R}.\end{cases}\tag{5}$$

值得特别指出的是，悲观机会约束马尔可夫决策过程 (5) 中的最坏情况机会约束，等价于具有更高风险水平的 (2) 中的名义机会约束。

Lemma 1. 假设在瓦瑟斯坦歧义集 (3) 中，参考分布为椭圆ˆ分布 $\mathbb{P}=\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$，且瓦瑟斯坦距离采用与正定矩阵 $\boldsymbol{\Sigma}$ 相关联的马氏范数。分布鲁棒机会约束

$$\forall\,\mathbb{P}\in\mathcal{F}(\theta):\mathbb{P}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\varepsilon\tag{6}$$

当且仅当 $\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\varepsilon$，其中 $\varepsilon=1-\Phi(\bar{\eta}^\star)\leq\varepsilon$，其 $\bar{\eta}^\star$ 可通过二分法计算，该方法搜索满足 $\eta\geq\Phi^{-1}(1-\varepsilon)$ 的最小值$\eta(\Phi(\eta)-(1-\varepsilon))-\int^{\eta^2/2}_{(\Phi^{-1}(1-\varepsilon))^2/2}kg(z)\mathrm{d}z\geq\theta.$

借助引理1，可以证明分布鲁棒机会约束马尔可夫决策过程（5）等价于风险水平更高的名义机会约束马尔可夫决策过程（2）。因此，分布鲁棒机会约束马尔可夫决策过程（5）可以重构为锥规划；更准确地说，由于我们选择了马氏范数，它可以重构为二阶锥规划。

Proposition 2. 假设在瓦瑟斯坦歧义集(3)中，参考分布为ˆ椭圆分布 $\mathbb{P}=\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$，且瓦瑟斯坦距离采用与正定矩阵 $\boldsymbol{\Sigma}$相关联的马氏范数。若风险阈值满足 $\varepsilon<0.5$，则分布鲁棒机会约束马尔可夫决策过程(5) 等价于以下二阶锥规划

$$\ell_{\mathrm{DCC}}(\theta,\varepsilon)=\max_{\boldsymbol{x}\in\mathcal{X}}\ \boldsymbol{\mu}^\top\boldsymbol{x}-\|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2,$$

其中 $\underline{\varepsilon}=1-\Phi(\bar{\eta}^\star)\leq\varepsilon$ 且 $\bar{\eta}^\star$ 为满足条件的最小 $\eta\geq\Phi^{-1}(1-\varepsilon)$ $\eta(\Phi(\eta)-(1-\varepsilon))-\int^{\eta^2/2}_{(\Phi^{-1}(1-\varepsilon))^2/2}kg(z)\mathrm{d}z\geq\theta.$

与第3节中的分布鲁棒马尔可夫决策过程类似，分布鲁棒机会约束马尔可夫决策过程也存在乐观对应形式，它等价于风险阈值更大的名义机会约束马尔可夫决策过程。我们将这一结果放在附录E中。

作为本节的总结，我们在图 2 中展示了 ε 与$\underline{\varepsilon}$之间的关系。事实上，对于任意固定的 ε，风险阈值 ε 与 Wasserstein 半径 θ之间存在一一对应关系。基于这一事实，对于数值实验（第 7 节）中的机会约束模型，我们只需校准风险阈值，而无需校准 Wasserstein 半径。

5. 回报风险马尔可夫决策过程

对于理性决策者而言，两类奖励是他们最为关注的：平均奖励和最坏情况奖励。然而，风险规避模型往往无法获得良好的平均回报，因为模型并未对此加以重视（Carpin 等（2016），Delage 和 Mannor（2010），Jiang 和 Powell（2018））。为兼顾这两方面的考量，我们以第3节和第4节中介绍的分布鲁棒马尔可夫决策过程和分布鲁棒机会约束模型为基础，提出回报风险马尔可夫决策过程，其目标是最大化最坏情况期望与奖励风险价值的加权平均，如下所示：

$$\ell_{\mathrm{RR}}(\alpha,\theta,\varepsilon)=\max_{\boldsymbol{x}\in\mathcal{X}}\ \alpha\inf_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{r}}^{\top}\boldsymbol{x}]+(1-\alpha)\inf_{\mathbb{P}\in\mathcal{F}'(\theta)}\mathbb{P}\text{-}\mathrm{VaR}_{\varepsilon}[\tilde{\boldsymbol{r}}^{\top}\boldsymbol{x}]. \tag{7}$$

这里，假设 Wasserstein 球 $\mathcal{F}(\theta)$ 配备了一般参考分布，并在 Wasserstein 距离的定义中采用L_2范数；而对于$\mathcal{F}'(\theta)$，则假设采用椭圆参考分布 $\hat{}\mathbb{P}=\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$ 以及与正定矩阵 $\boldsymbol{\Sigma}$ 相关的马氏范数。不难看出，通过改变 ε、 θ 和$\alpha\in\{0,1\}$，收益-风险马尔可夫决策过程 (7) 可将分布鲁棒马尔可夫决策过程 (4) 和分布鲁棒机会约束马尔可夫决策过程 (5) 作为特例。此外，通过选择分数型 α，收益-风险模型能够在风险与回报之间灵活权衡。下面的命题 3 给出了在这些假设下，回报风险马尔可夫决策过程 (7) 的等价二阶锥规划。

PROPOSITION 3. 假设在 (7) 中，*Wasserstein* 球 $\mathcal{F}(\theta)$ （分别地， $\mathcal{F}'(\theta)$）配备了一$\hat{}$般分布（分别地，椭圆参考分布 $\mathbb{P}=\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$），并且 $\mathcal{F}(\theta)$和 $\mathcal{F}'(\theta)$的 *Wasserstein* 距离定义中所用的范数分别为 L_2范数和与 $\boldsymbol{\Sigma}\succ 0$相关的马氏范数。假设风险阈值满足 $\varepsilon<0.5$，则回报风险马尔可夫决策过程 (7) 等价于一个二阶锥规划

$$\ell_{\mathrm{RR}}(\alpha,\theta,\varepsilon)=\max_{\boldsymbol{x}\in\mathcal{X}}\ \boldsymbol{\mu}^{\top}\boldsymbol{x}-\alpha\theta\cdot\|\boldsymbol{x}\|_2-(1-\alpha)\cdot\|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2, \tag{8}$$

其中 $\varepsilon_=1-\Phi(\bar{\eta}^{\star})\le\varepsilon$， $\bar{\eta}^{\star}$ 是满足 $\eta(\Phi(\eta)-(1-\varepsilon))-\int^{\eta^2/2}_{(\Phi^{-1}(1-\varepsilon))^2/2}kg(z)\mathrm{d}z\ge\theta$ 的最小 $\eta\ge\Phi^{-1}(1-\varepsilon)$，可通过二分法计算。

5.1. 不确定转移核的风险感知

采用 Lobo 等人（2020）的静态软鲁棒框架，也确实可以在我们的收益-风险模型中考虑转移核的不确定性。与 Lobo 等人（2020）一样，假设我们有由马尔可夫链蒙特卡罗生成的转移核有限样本 $\{\hat{}\boldsymbol{P}^i\}_{i\in[N]}$，其权重为 $\boldsymbol{w}\in\Delta^N:=\{\boldsymbol{w}\in\mathbb{R}^N_+\mid\boldsymbol{e}^{\top}\boldsymbol{w}=1\}$（参见例如 Kruschke（2010））。那么，我们提出的模型如下：

$$\max_{\boldsymbol{\pi}\in(\Delta^A)^S}\ \psi\cdot\mathbb{E}_{\hat{\mathbb{P}}}[g(\boldsymbol{\pi},\tilde{\boldsymbol{P}})]+(1-\psi)\cdot\hat{\mathbb{P}}\text{-}\mathrm{CVaR}_{\iota}[g(\boldsymbol{\pi},\tilde{\boldsymbol{P}})]. \tag{9}$$

$$
\begin{aligned}
&\text{最大值 } (1-\psi)(\eta - \frac{1}{1-\iota}\sum_{i\in[N]} y_i) + \psi\cdot\sum_{i\in[N]}(\boldsymbol{\mu}^\top \boldsymbol{x}^i - \alpha\theta\cdot\|\boldsymbol{x}^i\|_2 - (1-\alpha)\|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}^i\|_2) \\
&\text{s。t. } y_i - w_i\eta \geq \alpha\theta\cdot\|\boldsymbol{x}^i\|_2 + (1-\alpha)\cdot\|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}^i\|_2 - \boldsymbol{\mu}^\top\boldsymbol{x}^i && \forall i\in[N] \\
&\quad (\boldsymbol{E}-\gamma\cdot\boldsymbol{P}^i)\boldsymbol{x}^i = w_i\cdot\boldsymbol{p}_0 && \forall i\in[N] \\
&\quad \boldsymbol{x}^i \leq \tfrac{w_i}{1-\gamma}\boldsymbol{\pi} && \forall i\in[N] \\
&\quad x^i_{s,a} \geq \frac{w_i}{1-\gamma}(\pi_{s,\ a}-1) + \sum_{a'\in\mathcal{A}} x^i_{s,a'} && \forall (i,\ s,\ a)\in\mathcal{N}\times\mathcal{S}\times\mathcal{A} \\
&\quad \boldsymbol{\pi}\in(\Delta^A)^S\cap\{0,1\}^{SA}, \eta\in\mathbb{R}, \boldsymbol{x}^i\in\mathbb{R}^{SA}_+, \boldsymbol{y}\in\mathbb{R}^N_+ && \forall i\in[N].
\end{aligned}
$$

图 3 将 (9) 重新表述为混合整数二阶锥规划。

这里，（9）中的目标函数对于（转移核中的）不确定性同样是软鲁棒的，其中权重 $\psi\in[0,1]$ 作为鲁棒性的控制器，$\iota\in[0,1]$ 是（相对于不确定转移核的）风险阈值。加权经验分布 $\hat{\mathbb{P}}[\tilde{\boldsymbol{P}}=\hat{\boldsymbol{P}}^i]=w_i\ \forall i\in[N]$ 和函数

$$
\begin{aligned}
g(\boldsymbol{\pi},\boldsymbol{P}) = \max\ & \boldsymbol{\mu}^\top\boldsymbol{x} - \alpha\theta\cdot\|\boldsymbol{x}\|_2 - (1-\alpha)\cdot\|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2 \\
\text{s.t. } & x_{s,a} = \pi_{s,a}\cdot\sum_{a'\in\mathcal{A}} x_{s,a'} && \forall (s,a)\in\mathcal{S}\times\mathcal{A} \\
& (\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})\boldsymbol{x} = \boldsymbol{p}_0 \\
& \boldsymbol{x}\in\mathbb{R}^{SA}_+
\end{aligned}
$$

表示收益-风险模型的最优值，并附加约束：最优策略应为输入 $\boldsymbol{\pi}\in(\Delta^A)^S$，且 $\bar{\boldsymbol{P}}$ 是与输入转移核 $\boldsymbol{P}$ 对应的系数矩阵。

特别值得注意的是，若仅考虑确定性策略，则可以将 (9) 重新表述为混合整数二阶锥规划。

PROPOSITION 4. 若限制 $\boldsymbol{\pi}$ 为确定性策略（即 $\boldsymbol{\pi}\in(\Delta^A)^S\cap\{0,1\}^{SA}$），则问题 (9) 具有与图 *3* 所示等价的混合整数二阶锥规划重新表述。

我们指出，尽管与随机化策略相比，确定性策略似乎受到限制，但在某些情况下实际上更受青睐；例如，在某些医疗领域，由于实际和理念方面的原因，随机化策略无法实施，因此确定性策略可能是更合适的选择（Rosen 等，2006年）。此外，随机化策略部署后可能难以评估，且可复现性较差（Lobo 等，2020年）。

6. 一阶方法

本节介绍一种高效的一阶算法，用于求解收益-风险模型的等价形式 (8)。我们的算法基于交替方向线性化近端

乘子法（AD-LPMM）算法（Beck 2017年；Shefi 和 Teboulle 2014年）。该算法是交替方向乘子法（ADMM）算法的一种变体，Beck（2017年）也证明了其收敛速率为 $\mathcal{O}(1/N)$（其中 N 表示迭代次数）。所提出的分裂方法可高效更新 AD-LPMM 中的变量（其解可通过解析方法求得，或通过高效的二分法获得）。

对于ADMM算法的原始变量更新，需要求解目标函数中包含二次项的最小化问题；在AD-LPMM中，可以通过向目标函数中添加邻近项来线性化该二次项，从而使原始变量更新更加容易。为了实现我们的AD-LPMM算法，首先我们将引入辅助变量，并将（8）重写为如下最小化问题：

$$\begin{aligned}
&\min\ \alpha\theta\cdot\|\boldsymbol{x}\|_2+(1-\alpha)\cdot\|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{y}\|_2-\boldsymbol{\mu}^\top\boldsymbol{z}\\
&\text{s.t.}\ \ (\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})\boldsymbol{x}=\boldsymbol{p}_0\\
&\qquad\boldsymbol{x}=\boldsymbol{y}\\
&\qquad\boldsymbol{x}=\boldsymbol{z}\\
&\qquad\boldsymbol{x}\in\mathbb{R}^{SA},\boldsymbol{y}\in\mathbb{R}^{SA},\boldsymbol{z}\in\mathbb{R}^{SA}_+,
\end{aligned}\tag{10}$$

其中，按照AD-LPMM的思路，我们可以将决策变量分成两组并分别更新。（10）的增广拉格朗日函数为：

$$\begin{aligned}
&L(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z};\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\eta})\\
=\ &\alpha\theta\cdot\|\boldsymbol{x}\|_2+(1-\alpha)\Phi^{-1}(1-\underline{\varepsilon})\cdot\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{y}\|_2-\boldsymbol{\mu}^\top\boldsymbol{z}+\boldsymbol{\lambda}^\top((\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})\boldsymbol{x}-\boldsymbol{p}_0)+\boldsymbol{\xi}^\top(\boldsymbol{x}-\boldsymbol{y})\\
&+\boldsymbol{\eta}^\top(\boldsymbol{x}-\boldsymbol{z})+\tfrac{c}{2}\cdot\left\|\begin{array}{c}(\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})\boldsymbol{x}-\boldsymbol{p}_0\\ \boldsymbol{x}-\boldsymbol{y}\\ \boldsymbol{x}-\boldsymbol{z}\end{array}\right\|_2^2.
\end{aligned}$$

基于我们的分裂方法，我们将分别更新两组变量（$\boldsymbol{y}$，$\boldsymbol{z}$）和 $\boldsymbol{x}$。对于（$\boldsymbol{y}$，$\boldsymbol{z}$）的更新，我们定义两个原始变量更新算子

$$\mathfrak{P}_{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{\xi};c)=\arg\min_{\boldsymbol{y}}\,(1-\alpha)\Phi^{-1}(1-\underline{\varepsilon})\cdot\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{y}\|_2-\boldsymbol{\xi}^\top\boldsymbol{y}+\frac{c}{2}\cdot\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$$

并且 $\mathfrak{P}_{\boldsymbol{z}}(\boldsymbol{x},\ \boldsymbol{\eta};c)=\arg\min_{\boldsymbol{z}\geq\boldsymbol{0}}\ -\boldsymbol{z}^\top(\underline{\boldsymbol{\mu}}+\boldsymbol{\eta})+\frac{c}{2}\cdot\|\boldsymbol{x}-\boldsymbol{z}\|_2^2$；而对于 $\boldsymbol{x}$（即第二组变量）的更新，我们定义

$$\begin{aligned}
\mathfrak{P}_{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\eta};c,\nu,\hat{\boldsymbol{x}})=\ &\arg\min_{\boldsymbol{x}}\ \alpha\theta\cdot\|\boldsymbol{x}\|_2+\boldsymbol{x}^\top((\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})^\top\boldsymbol{\lambda}+\boldsymbol{\xi}+\boldsymbol{\eta})\\
&+\tfrac{c}{2}\cdot\left\|\begin{array}{c}(\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})\boldsymbol{x}-\boldsymbol{p}_0\\ \boldsymbol{x}-\boldsymbol{y}\\ \boldsymbol{x}-\boldsymbol{z}\end{array}\right\|_2^2+\tfrac{1}{2}\cdot\ell^2_{\boldsymbol{Q}(c,\nu)}(\boldsymbol{x}-\hat{\boldsymbol{x}}),
\end{aligned}$$

算法 1: 用于问题 (10) 的 AD-LPMM

输入: 弗罗贝尼乌斯范数 $\nu = \|(\boldsymbol{E} - \gamma \cdot \bar{\boldsymbol{P}})^\top (\boldsymbol{E} - \gamma \cdot \bar{\boldsymbol{P}}) + 2 \cdot \boldsymbol{I}\|_{\mathrm{F}}$，初始步长 $c_0 > 0$，步长增长率 $\beta > 0$，期望精度 δ，$\boldsymbol{x}^0$，$\boldsymbol{y}^0$，$\boldsymbol{z}^0$，$\boldsymbol{\lambda}^0$，$\boldsymbol{\xi}^0$，$\boldsymbol{\eta}^0$，$k \leftarrow 0$

当 $\left\| \begin{matrix} (\boldsymbol{E} - \gamma \cdot \bar{\boldsymbol{P}})\boldsymbol{x}^k - \boldsymbol{p}_0 \\ \boldsymbol{x}^k - \boldsymbol{y}^k \\ \boldsymbol{x}^k - \boldsymbol{z}^k \end{matrix} \right\|_\infty \geq \delta$ **执行**

// 原始变量更新

步骤 1: $\boldsymbol{y}^{k+1} \leftarrow \mathfrak{P}_{\boldsymbol{y}}(\boldsymbol{x}^k, \boldsymbol{\xi}^k; c_k)$;

步骤 2: $\boldsymbol{z}^{k+1} \leftarrow \mathfrak{P}_{\boldsymbol{z}}(\boldsymbol{x}^k, \boldsymbol{\eta}^k; c_k)$;

步骤 3: $\boldsymbol{x}^{k+1} \leftarrow \mathfrak{P}_{\boldsymbol{x}}(\boldsymbol{y}^{k+1}, \boldsymbol{z}^{k+1}, \boldsymbol{\lambda}^k, \boldsymbol{\xi}^k, \boldsymbol{\eta}^k; c_k, \nu, \boldsymbol{x}^k)$;

// 对偶更新

步骤 4: $\boldsymbol{\lambda}^{k+1} \leftarrow \boldsymbol{\lambda}^k + c_k \cdot ((\boldsymbol{E} - \gamma \cdot \bar{\boldsymbol{P}})\boldsymbol{x}^{k+1} - \boldsymbol{p}_0)$;

步骤 5: $\boldsymbol{\xi}^{k+1} \leftarrow \boldsymbol{\xi}^k + c_k \cdot (\boldsymbol{x}^{k+1} - \boldsymbol{y}^{k+1})$;

步骤 6: $\boldsymbol{\eta}^{k+1} \leftarrow \boldsymbol{\eta}^k + c_k \cdot (\boldsymbol{x}^{k+1} - \boldsymbol{z}^{k+1})$;

// 增大步长

步骤 7: $c_{k+1} \leftarrow c_k + \beta c_0$;

步骤 8: $k \leftarrow k + 1$;

end

输出: 解 $\boldsymbol{x}^k$

其中 $\boldsymbol{Q}(c, \nu) = c \cdot ((\nu - 2) \cdot \boldsymbol{I} - (\boldsymbol{E} - \gamma \cdot \bar{\boldsymbol{P}})^\top (\boldsymbol{E} - \gamma \cdot \bar{\boldsymbol{P}}))$，而 $\ell_{\boldsymbol{Q}}(\cdot)$（配备一个半正定矩阵 $\boldsymbol{Q}$）是加权向量范数，满足 $\ell_{\boldsymbol{Q}}(\boldsymbol{x}) = \sqrt{\boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x}}$。正如我们将在第6.3节中看到的，添加邻近项 $(1/2) \cdot \ell^2_{\boldsymbol{Q}(c,\ \nu)}(\boldsymbol{x} - \hat{\boldsymbol{x}})$后，$\boldsymbol{x}$ 的更新很快（可获得解析解）。注意，当 $\boldsymbol{Q}(c, \nu) \equiv \boldsymbol{0}$时，AD-LPMM中的更新便退化为ADMM中的更新。

我们现在在算法 1 中介绍 AD-LPMM。基本上，最耗时的计算集中在原始变量更新阶段，此时在其他变量固定为其上一次更新后的取值时，通过求解一个最小化问题来完成更新。我们很快将详细说明，由于采用了变量分裂法，原始变量更新也相当快速，并且可以得到解析解或通过二分法求得的解。这里我们选择一种在每次迭代中都递增的步长（增长率为 $\beta > 0$），这在实践中能够加快收敛。

6.1. 步骤 1 子问题：近端映射与投影

为求解 $\mathfrak{P}_y(\boldsymbol{x},\ \boldsymbol{\xi}; c)$，我们首先利用近端映射技术，并建立以下等价关系：

$$\begin{aligned}\mathfrak{P}_{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{\xi};c) &= \operatorname{Prox}_{\frac{(1-\alpha)\Phi^{-1}(1-\varepsilon)}{c}\cdot\|\cdot\|_{\boldsymbol{\Sigma}}}\left(\boldsymbol{x}+\tfrac{1}{c}\cdot\boldsymbol{\xi}\right)\\ &= \boldsymbol{x}+\tfrac{1}{c}\cdot\boldsymbol{\xi}-\tfrac{(1-\alpha)\Phi^{-1}(1-\varepsilon)}{c}\cdot\operatorname{Proj}_{\boldsymbol{B}_{\ell_{\boldsymbol{\Sigma}^{-1}}(\cdot)}}\left(\tfrac{1}{(1-\alpha)\Phi^{-1}(1-\varepsilon)}\cdot(c\cdot\boldsymbol{x}+\boldsymbol{\xi})\right),\end{aligned} \tag{11}$$

其中 $\operatorname{Prox}_{f(\cdot)}(\boldsymbol{x}) = \arg\min_{\boldsymbol{v}} f(\boldsymbol{v}) + \frac{1}{2}\cdot\|\boldsymbol{v}-\boldsymbol{x}\|_2^2$ 是近端映射算子，

$$\operatorname{Proj}_{\boldsymbol{B}_{\ell_{\boldsymbol{\Sigma}}(\cdot)}}(\boldsymbol{x}) = \underset{\boldsymbol{v}:\ell_{\boldsymbol{\Sigma}}(\boldsymbol{v})\le 1}{\arg\min}\ \frac{1}{2}\cdot\|\boldsymbol{v}-\boldsymbol{x}\|_2^2 \tag{12}$$

是到单位球上的投影算子 $\boldsymbol{B}_{\ell_{\boldsymbol{\Sigma}}(\cdot)} = \{\boldsymbol{x}\in\mathbb{R}^{SA} \mid \ell_{\boldsymbol{\Sigma}}(\boldsymbol{x}) \le 1\}$. 这里，式 (11) 中的第一个等号根据近端映射算子的定义成立，第二个等号则可参见 Beck（2017）中的例 6.4.7。例如，问题 (12) 可以通过二分法高效求解，以确定其最优对偶解 $\lambda^\star \ge 0$（随后即可立即求得最优原始解）；二分法的上界由附录 A.4 中的引理 2 给出。以下命题给出了求解过程 (11) 的时间复杂度以及二分法的伪代码。

PROPOSITION 5. 问题 $\mathfrak{P}_y(\boldsymbol{x},\boldsymbol{\xi};\ c)$ 可在时间 $\mathcal{O}(SA\log(1/\delta'))$内求解，其中 δ' 是二分法所需的精度。

6.2. 第2步中的子问题：逐分量更新

问题 $\mathfrak{P}_z(\boldsymbol{x},\ \boldsymbol{\eta}; c)$ 可分解为 SA 个单变量二次规划问题，每个问题都可解析求解。我们在下述命题中总结其时间复杂度及具体细节。

PROPOSITION 6. 问题 $\mathfrak{P}_z(\boldsymbol{x},\boldsymbol{\eta};\ c)$ 可在时间 $\mathcal{O}(SA)$内求解。

6.3. 第3步中的子问题：线性化与近端映射

与 ADMM 中的更新相比，在我们的 AD-LPMM 中，步骤 3 的更新目标函数中加入了一个邻近项 $(1/2)\cdot\ell^2_{\boldsymbol{Q}(c,\ \nu)}(\boldsymbol{x}-\hat{\boldsymbol{x}})$。按照第6节所述，通过选择 $\boldsymbol{Q}(\cdot,\ \cdot)$，我们可以将 $\mathfrak{P}_{\boldsymbol{x}}(\boldsymbol{y},\ \boldsymbol{z},\ \boldsymbol{\lambda},\ \boldsymbol{\xi},\ \boldsymbol{\eta};\ c,\ \nu,\ \hat{\boldsymbol{x}})$ 中的所有二次项线性化，从而利用近端映射技术解析地求得解（同时确保在算法 1 的每次迭代中 $\boldsymbol{Q}(c_k,\ \nu)$ 的半正定性）。下述命题给出了这一求解过程及其时间复杂度。

PROPOSITION 7. 问题 $\mathfrak{P}_{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\eta};\ c,\nu,\ \hat{\boldsymbol{x}})$ </style}可在时间 (SA)内求解。

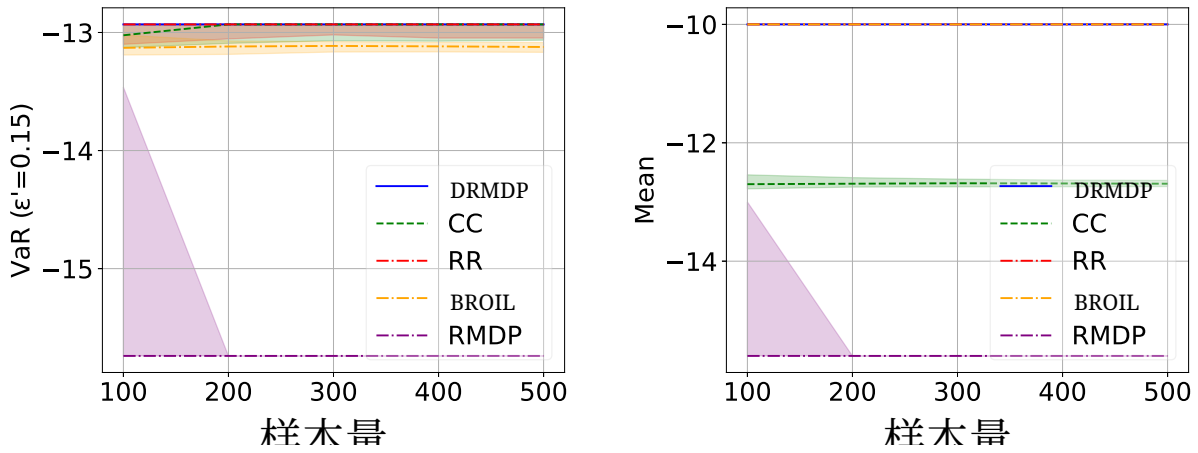


图4 实证研究。通过风险价值（风险阈值 $\varepsilon' = 15\%$）和奖励均值评估模型 DRMDP（4）、机会约束（2）、收益-风险马尔可夫决策过程（7）、RMDP 和 BROIL。阴影区域的上边缘和下边缘分别表示 100 次表现的第 95 和第 5 百分位数，实线表示中位数。

7. 数值实验

在本节中，我们进行两次数值实验，比较分布鲁棒马尔可夫决策过程 (4)、机会约束 (2)[3]，收益-风险马尔可夫决策过程 (7)、鲁棒马尔可夫决策过程（Delage 和 Mannor，2010年）和 BROIL（Brown 等人，2020年）的性能（关于后两个模型的更多详情，请参见附录 F 和 G）。在这两项实验中，我们使用不同的样本量（100、200、300、400、500）训练奖励函数。对于每个样本量，每个模型的性能均评估 100 次。每个模型的性能通过期望和风险价值进行评估，风险阈值为 $\varepsilon' \in \{5\%, 10\%, 15\%\}$。我们通过交叉验证进行参数选择（详情请参见附录 H.1）。

在第7.1节中，我们进行了一项仿真实验，其中的马尔可夫决策过程按照 Regan 和 Boutilier（2012）的方法随机生成；在第 7.2 节中，我们研究 Delage 和 Mannor（2010）提出的机器更换问题。正如证明所示，在本节中，分布鲁棒马尔可夫决策过程（4）的 Wasserstein歧义集采用一般参考分布，并使用 L_2-范数来计算 Wasserstein距离；对于 RR（7），我们在 Wasserstein歧义集 $\mathcal{F}(\theta)$ 的 Wasserstein距离定义中采用一般参考分布和 L_2-范数，而对于 $\mathcal{F}'(\theta)$，我们则采用椭圆参考分布 $\mathbb{P} = \mathbb{P}_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)}$，并使用与正定矩阵 $\boldsymbol{\Sigma}$ 相关的马氏范数来计算 Wasserstein距离。所有优化问题均由 MOSEK 在配备 32GB 内存的 2。3GHz 处理器上求解。

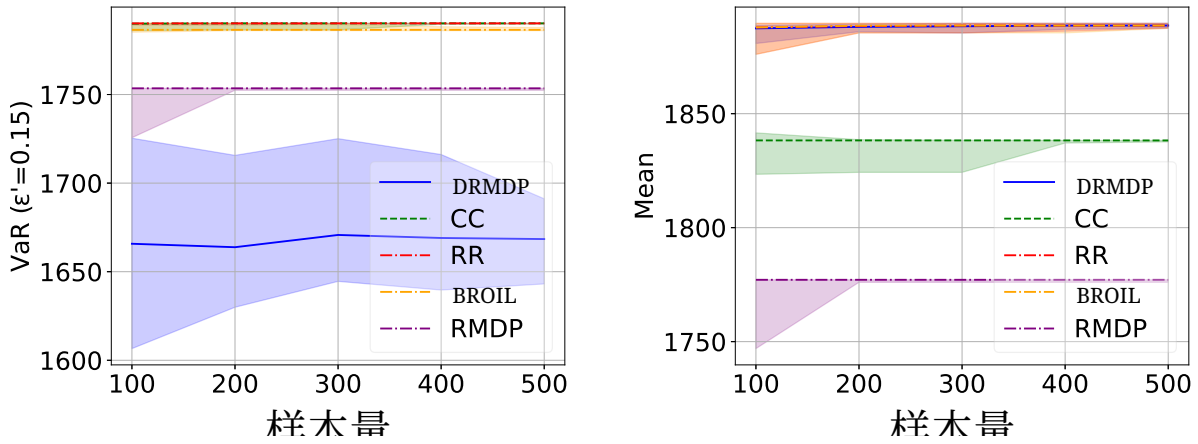


图5 仿真。采用风险价值（风险阈值$\varepsilon' = 15\%$）和奖励均值评估 DRMDP (4)、机会约束 (2)、收益-风险马尔可夫决策过程 (7)、RMDP 和 BROIL。阴影区域的上边缘和下边缘分别表示 100 次表现的第 95 和第 5 百分位数，实线表示中位数。

7.1. 仿真实验

在这项实验中，我们遵循 Regan 和 Boutilier（2012）中的实验设置，随机生成可达后继状态的数量和转移核（两者均为决策者所知）。有关实验设置的更多详情见附录 H.2。

如图 5 和图 7 所示（后者展示了风险阈值为 $\varepsilon' \in \{5\%, 10\%\}$ 时的风险价值，已移至附录 H.4），当决策者旨在优化尾部表现时，与分布鲁棒马尔可夫决策过程相比，机会约束是更优的选择；相反，当追求优化平均回报时，分布鲁棒马尔可夫决策过程的表现远胜于机会约束。值得注意的是，收益-风险马尔可夫决策过程将分布鲁棒马尔可夫决策过程和分布鲁棒机会约束模型作为特例，在所有评价标准下始终是最佳模型。特别是，收益-风险马尔可夫决策过程的百分位收益高于 BROIL（即不具备鲁棒性的模型），这表明分布鲁棒性的益处，以及风险价值这一风险度量在百分位表现优化方面的优势。正如预期，鲁棒马尔可夫决策过程最终产生了过于保守的策略，因此在大多数情形下，按所有评价标准衡量，其表现都很差。

7.2. 机器更换问题

在本实验中，我们沿用 Delage 和 Mannor（2010）的实验设置，并考虑这样一种情况：某工厂拥有大量机器，每台机器都遵循相同的底层马尔可夫决策过程（有关实验设置的更多详情，请参见附录 H.2）。我们的设置与 Delage 和 Mannor（2010）类似，但有以下不同：我们采用数据驱动设定

[3]正如我们在第 4 节所展示的，分布鲁棒机会约束等价于风险水平经过调整的名义机会约束，因此这里我们直接选择后者作为基准。

如上所述，并按照第7.1节中的做法，通过考察各种性能指标来评估我们的模型（及其策略）。

我们在图4和图8中报告了五种模型的整体表现（其中，VaR 对应 $\varepsilon' \in \{5\%, 10\%\}$ 的结果见附录H.5）。与之前的实验类似，RR 的表现始终优于或等同于机会约束和分布鲁棒马尔可夫决策过程中的最佳模型，并且在所有评价标准下均表现最佳。这再次体现了同时考虑期望性能、最坏情况性能和分布鲁棒性的优点。

7.3. 不同算法的计算时间

表1 MOSEK求解器和AD-LPMM算法的平均运行时间（秒），以及相对于MOSEK计算所得最优值的相对间隙（%）。

S=A	运行时间		相对间隙
	MOSEK	AD-LPMM	
40	**0.60**	2.79	< 0.1 %
70	5.58	**4.81**	< 0.1 %
100	25.50	**19.98**	0.2 %
130	93.54	**66.17**	< 0.1 %
160	444.06	**168.34**	0.4 %

在本节中，我们比较了 AD-LPMM 算法与当前最先进求解器 MOSEK 的计算时间。表 1 报告了 AD-LPMM 和 MOSEK 在不同问题规模下求解问题 (8) 的运行时间。结果表明，尽管问题规模较小时，我们的 AD-LPMM 算法比 MOSEK 求解器慢，但它展现出很强的可扩展性，在大规模问题上比 MOSEK 快得多（同时始终保持较高的解质量）；而且问题规模越大，这一优势越明显。

8. 结论

我们研究奖励函数存在歧义的风险感知型马尔可夫决策过程，并提出收益-风险模型。该模型用途广泛，可优化策略的平均表现与分位数表现的任意加权组合。此模型推广并结合了分布鲁棒马尔可夫决策过程和分布鲁棒机会约束马尔可夫决策过程的优点，因此在优化平均表现和百分位表现方面都十分有效。特别是，当输出策略为确定性策略时，收益-风险模型也能捕捉不确定转移核带来的风险。我们为所有提出的模型提供了易于处理的改写形式，并设计了一种 AD-LPMM算法，用于

收益-风险模型；该模型具有良好的可扩展性，在大规模问题上比 MOSEK求解器更快。实验结果展示了收益-风险模型的广泛适用性，以及该算法的可扩展性。

展望未来，我们认为，探索更高效的收益-风险马尔可夫决策过程（RR）求解方法十分重要，其中函数近似和策略梯度（Sutton 和 Barto，2018年）都是可行的选择。

参考文献

Abdullah、Mohammed Amin、Hang Ren、Haitham Bou Ammar、Vladimir Milenkovic、Rui Luo、Mingtian Zhang、Jun Wang。2019年。Wasserstein 鲁棒强化学习。*arXiv* 预印本 *arXiv:1907.13196*。

Ahmadi、Mohamadreza、Ugo Rosolia、Michel Ingham、Richard Murray、Aaron Ames。2021年。约束风险厌恶马尔可夫决策过程。第*35*届*AAAI*人工智能会议（*AAAI-21*）。Alizadeh、Pegah、Yann Chevaleyre、Jean-Daniel Zucker。2015年。马尔可夫决策过程中的基于近似遗憾的偏好引出。*2015*年*IEEE RIVF*计算与通信技术国际会议——研究、创新与未来愿景（*RIVF*）。IEEE，47–52。B¨aauerle、Nicole、Ulrich Rieder。2017年。部分可观测风险敏感马尔可夫决策过程。运筹学数学方法**42**(4) 1180–1196。B¨auerle、Nicole、Alexander Glauner。2021年。应用于马尔可夫决策过程的谱风险度量最小化。运筹学数学方法**94**(1) 35–69。B¨auerle、Nicole、Jonathan Ott。2011年。采用平均风险价值准则的马尔可夫决策过程。运筹学数学方法**74**(3) 361–379。Beck、Amir。2017年。优化中的一阶方法。SIAM。Behzadian、Bahram、Reazul Russel、Marek Petrik、Chin Pang Ho。2021年。使用鲁棒马尔可夫决策过程优化百分位准则。人工智能与统计国际会议1009–1017。Ben-Tal、Aharon、Dimitris Bertsi mas、David B Brown。2010年。歧义下优化的软鲁棒模型。《运筹学》**58**(4-part-2) 1220–1234。Ben-Tal、Aharon、Arkadi Nemirovski。2001年。现代凸优化讲义：分析、算法与工程应用。SIAM。Bernard、Carole、Silvana M Pesenti、Steven Vanduffel。2022年。鲁棒扭曲风险度量。*arXiv*预印本 *arXiv:2205.08850*。Bertsekas、Dimitri、John Tsitsiklis。1995。神经动态规划：概述。*1995*年第*34*届*IEEE*决策与控制会议论文集，第1卷。IEEE，560–564。Bertsimas、Dimitris、Aur´elie Thiele。2006年。鲁棒与数据驱动优化：不确定性下的现代决策。创新决策的模型、方法与应用。INFORMS，95–122。

Blanchet、Jose、Karthyek Murthy。2019年。通过最优传输量化分布模型风险。运筹学数学方法**44**(2) 565–600。Brown、Daniel、Scott Niekum、Marek Petrik。2020年。用于模仿学习的贝叶斯鲁棒优化。神经信息处理系统进展**33**2479–2491。Calafiore、Carlo、L El Ghaoui。2006年。关于分布鲁棒机会约束线性规划。优化理论与应用杂志**130**(1) 1–22。Carpin、Stefano、Yin-Lam Chow、Marco Pavone。2016年。在有限马尔可夫决策过程中使用总成本准则和平均风险价值进行风险厌恶决策。*2016*年*IEEE*国际机器人与自动化会议（*ICRA*）。IEEE，335–342。Chen、Xin、Melvyn Sim、Peng Sun。2007年。随机规划的鲁棒优化视角。《运筹学》**55**(6) 1058–1071。Chen、Xinyue、Zijian Zhou、Zheng Wang、Che Wang、Yanqiu Wu、Keith Ross。2020年。Bail：用于批量深度强化学习的最佳动作模仿学习。神经信息处理系统进展**33**18353–18363。Chen、Zhi、Daniel Kuhn、Wolfram Wiesemann。2018年。瓦瑟斯坦球上的数据驱动机会约束规划。*arXiv*预印本 *arXiv:1809.00210*。Chen、Zhi、Melvyn Sim、Huan Xu。2019年。具有无限约束歧义集的分布鲁棒优化。《运筹学》**67**(5) 1328–1344。Chen、Zhi、Weijun Xie。2021年。分布歧义下的风险价值共享。数理金融**31**(1) 531–559。Choi、Jaedeug、Kee-Eung Kim。2012年。针对多个奖励函数的非参数贝叶斯逆强化学习。神经信息处理系统进展**25**。Chow、Yinlam、Mohammad Ghavamzadeh、Lucas Janson、Marco Pavone。2017年。采用百分位风险准则的风险约束强化学习。机器学习研究杂志**18**(1) 6070–6120。Clement、Julien、Christian Kroer。2021年a。瓦瑟斯坦分布鲁棒马尔可夫决策过程的一阶方法。国际机器学习会议。PMLR，2010–2019。Clement、Julien Grand、Christian Kroer。2021年b。瓦瑟斯坦分布鲁棒马尔可夫决策过程的一阶方法。国际机器学习会议。PMLR，2010–2019。Delage、Erick、Shie Mannor。2007年。不确定马尔可夫决策过程中的百分位优化及其在高效探索中的应用。第*24*届国际机器学习会议论文集。PMLR，225–232。Delage、Erick、Shie Mannor。2010年。具有参数不确定性的马尔可夫决策过程的百分位优化。《运筹学》**58**(1) 203–213。Delgado、Karina、Leliane De Barros、Daniel Dias、Scott Sanner。2016年。具有不精确概率的马尔可夫决策过程的实时动态规划。人工智能**230** 192–223。

Derman, Esther, Shie Mannor. 2020. 强化学习中的分布鲁棒性与正则化。*arXiv*预印本 *arXiv:2003.02894* 。Erdo˘gan, Emre, Garud Iyengar. 2006. 模糊机会约束问题与鲁棒优化。《数学规划》**107**(1) 37–61。Gao, Rui, Anton Kleywegt. 2016. 基于Wasserstein距离的分布鲁棒随机优化。*arXiv*预印本 *arXiv:1604.02199* 。Gao, Rui, Anton Kleywegt. 2022. 基于Wasserstein距离的分布鲁棒随机优化。运筹学数学。Gilbert, Hugo, Paul Weng, Yan Xu. 2017. 优化偏好型马尔可夫决策过程中的分位数。*AAAI*人工智能会议论文集，第31卷。Hanasusanto, Grani, Vladimir Roitch, Daniel Kuhn, Wolfram Wiesemann. 2015. 不确定性量化与机会约束规划的分布鲁棒视角。《数学规划》**151**(1) 35–62。Hanasusanto, Grani, Vladimir Roitch, Daniel Kuhn, Wolfram Wiesemann. 2017. 均值与离散度信息下的模糊联合机会约束。《运筹学》**65**(3) 751–767。Haskell, William, Rahul Jain. 2015. 风险感知型马尔可夫决策过程的凸解析方法。《*SIAM*控制与优化杂志》**53**(3) 1569–1598。Ho, Jonathan, Stefano Ermon. 2016. 生成对抗模仿学习。神经信息处理系统进展**29**。Hogg, Robert V, Allen T Craig. 1995. 数理统计学导论。（第5版）。恩格尔伍德希尔斯，新泽西州。Huang, Wenjie, William Haskell. 2017. 马尔可夫决策过程的风险感知型Q学习。*2017*年*IEEE*第*56*届决策与控制年会（*CDC*）。IEEE，4928–4933。Huang, Yonghui, Xianping Guo. 2016. 连续时间有限时域半马尔可夫决策过程的最小平均风险价值。《*SIAM*优化杂志》**26**(1) 1–28。Iyengar, Garud. 2005. 鲁棒动态规划。运筹学数学**30**(2) 257–280。Jaimungal, Sebastian, Silvana M Pesenti, Ye Sheng Wang, Hariom Tatsat. 2022. 鲁棒风险感知型强化学习。《*SIAM*金融数学杂志》**13**(1) 213–226。Jiang, Daniel R, Warren B Powell. 2018. 基于分位数风险度量的风险厌恶型近似动态规划。运筹学数学**43**(2) 554–579。Kaiser, Lukasz, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. 2019. 面向Atari的基于模型的强化学习。*arXiv*预印本 *arXiv:1903.00374* 。

Kruschke, John K. 2010. 贝叶斯数据分析。*Wiley Interdisciplinary Reviews: Cognitive Science* **1**(5) 658–676。Lagoudakis, Michail G, Ronald Parr. 2003. 最小二乘策略迭代。机器学习研究杂志**41**107–1149。Lobo, Elita A, Mohammad Ghavamzadeh, Marek Petrik. 2020. 用于批量强化学习的软鲁棒算法。*arXiv*预印本 *arXiv:2011.14495* 。Mannor, Shie, Ofir Mebel, Huan Xu. 2016. 具有k-矩形不确定性的鲁棒马尔可夫决策过程。运筹学数学**41**(4) 1484–1509。Mannor, Shie, Duncan Simester, Peng Sun, John Tsitsiklis. 2007. 价值函数估计中的偏差与方差近似。《管理科学》**53**(2) 308–322。Ng, Andrew Y, Stuart J Russell, et al. 2000. 逆向强化学习算法。国际机器学习会议，第1卷。2。Nocedal, Jorge, Stephen Wright. 2006. 数值优化。施普林格科学与商业媒体。Osa, Takayuki, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters. 2018. 模仿学习的算法视角。*arXiv*预印本 *arXiv:1811.06711* 。Panaganti, Kishan, Dileep Kalathil. 2022. 生成模型下鲁棒强化学习的样本复杂度。人工智能与统计国际会议。PMLR，9582–9602。Petrik, Marek. 2010. 基于优化的近似动态规划。马萨诸塞大学阿默斯特分校。Petrik, Marek, Ronny Luss. 2016. 动态产品推荐中的可解释策略。不确定性人工智能会议。Pflug, Georg, David Wozabal. 2007. 投资组合选择中的模糊性。定量金融**7**(4) 435–442。Postek, Krzysztof, Aharon Ben-Tal, Dick Den Hertog, Bertrand Melenberg. 2018. 均值与离散度信息下带模糊随机约束的鲁棒优化。《运筹学》**66**(3) 814–833。Pr´ekopa, Andr´as. 2013. 随机规划，第324卷。施普林格科学与商业媒体。Puterman, Martin. 2014. 马尔可夫决策过程：离散随机动态规划。约翰·威立父子出版公司。Rashidinejad, Paria, Banghua Zhu, Cong Ma, Jiantao Jiao, Stuart Russell. 2021. 融合离线强化学习与模仿学习：悲观主义的故事。神经信息处理系统进展**34**。Regan, Kevin, Craig Boutilier. 2010. 使用非支配策略计算奖励不确定的马尔可夫决策过程中的鲁棒策略。*AAAI*人工智能会议论文集，第24卷。Regan, Kevin, Craig Boutilier. 2011a. 为马尔可夫决策过程诱导加性奖励函数。第二十二届国际人工智能联合会议。

Regan, Kevin，Craig Boutilier。2011b。奖励不确定的马尔可夫决策过程的稳健在线优化。第二十二届国际人工智能联合会议。Regan, Kevin，Craig Boutilier。2012。基于遗憾的马尔可夫决策过程奖励引出。*arXiv*预印本 *arXiv:1205.2619* 。Rosen, Laura，Orly Manor，Dan Engelhard，David Zucker。2006。为健康促进研究中的随机对照试验辩护。《美国公共卫生杂志》**96**(7) 1181–1186。Roy, Aurko，Huan Xu，Sebastian Pokutta。2017。模型失配下的强化学习。神经信息处理系统进展**30**。Sarykalin, Serge y，Gaia Serraino，Stan Uryasev。2008。风险管理与优化中的风险价值与条件风险价值比较。信息密集时代最先进的决策工具。Informs，270–294。Shapiro, Alexander。2016。概率测度的矩形集。《运筹学》**64**(2) 528–541。Shapiro, Alexander。2021。分布鲁棒最优控制与马尔可夫决策过程建模。《运筹学快报》**49**(5) 809–814。Shefi, Ron，Marc Teboulle。2014。基于乘子近端法的凸最小化分解方法的收敛速率分析。《*SIAM*优化杂志》**24**(1) 269–297。Smith, James，Robert L Winkler。2006。优化器诅咒：决策分析中的怀疑与决策后意外。《管理科学》**52**(3) 311–322。Sutton, Richard S，Andrew G Barto。2018。强化学习：导论。麻省理工学院出版社。van de Berg, Damien，Thomas Savage，Panagiotis Petsagkourakis，Dongda Zhang，Nilay Shah，Ehecatl Anto- nio del Rio-Chanona。2022。过程系统工程应用中的数据驱动优化。《化学工程科学》**248**117135。Wang, Yue，Shaofeng Zou。2021。具有模型不确定性的在线鲁棒强化学习。神经信息处理系统进展**34**7193–7206。Wang, Yue，Shaofeng Zou。2022。鲁棒强化学习的策略梯度法。*arXiv*预印本 *arXiv:2205.07344* 。Xie, Weijun。2021。关于具有Wasserstein距离的分布鲁棒机会约束规划。《数学规划》**186**(1) 115–155。Xu, Huan，Shie Mannor。2010。分布鲁棒马尔可夫决策过程。神经信息处理系统进展**23**2505–2513。Xu, Huan，Shie Mannor。2012。分布鲁棒马尔可夫决策过程。运筹学数学**37**(2)。

Yamai, Yasuhiro，Toshinao Yoshiba，等。2002。预期损失与风险价值的比较分析：其估计误差、分解与优化。货币与经济研究**20**(1) 87–121。Yang, Insoon。2020。Wasserstein分布鲁棒随机控制：一种数据驱动方法。*IEEE*自动控制汇刊**66**(8) 3863–3870。Yu, Pengqian，Huan Xu。2015。马尔可夫决策过程中的分布鲁棒对应形式。*IEEE*自动控制汇刊**61**(9) 2538–2543。Zheng, Kan，Zhe Yang，Kuan Zhang，Periklis Chatzimisios，Kan Yang，Wei Xiang。2016。面向5G的移动网络大数据驱动优化。*IEEE*网络**30**(1) 44–51。Zymler, Steve，Daniel Kuhn，Ber¸c Rustem。2013。具有二阶矩信息的分布鲁棒联合机会约束。《数学规划》**137**(1) 167–198。

A. 结果的证明

A.1. 第3节结果的证明

命题 *1* 的证明。只需将式 (4) 的目标函数改写如下:

$$\begin{aligned}\inf_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] &= -\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{E}_{\mathbb{P}}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] \\ &= -\min_{\lambda\geq 0}\left\{\lambda\theta - \int_{\mathbb{R}^{SA}} \inf_{\boldsymbol{\xi}\in\mathbb{R}^{SA}} (\lambda\|\boldsymbol{\xi}-\boldsymbol{r}\| + \boldsymbol{\xi}^\top \boldsymbol{x})\, \mathrm{d}\hat{\mathbb{P}}\boldsymbol{r}\right\} \\ &= -\min_{\lambda\geq\|\boldsymbol{x}\|_*}\left\{\lambda\theta - \int_{\mathbb{R}^{SA}} \boldsymbol{r}^\top \boldsymbol{x}\, \mathrm{d}\hat{\mathbb{P}}\boldsymbol{r}\right\} \\ &= \mathbb{E}_{\hat{\mathbb{P}}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] - \theta\|\boldsymbol{x}\|_*,\end{aligned}$$

其中，第二个等式由 Gao 和 Kleywegt（2016年）中的定理1得出，第三个等式则由强锥对偶性得出

$$\inf_{\boldsymbol{\xi}\in\mathbb{R}^K} (\lambda\|\boldsymbol{\xi}-\boldsymbol{r}\| + \boldsymbol{\xi}^\top \boldsymbol{x}) = \begin{cases} \boldsymbol{r}^\top \boldsymbol{x} & \lambda \geq \|\boldsymbol{x}\|_* \\ -\infty & \lambda \in [0, \|\boldsymbol{x}\|_*). \end{cases}$$

将上述改写代入后，即可完成证明。 Q.E.D.

A.2. 第4节结果的证明

引理*1*的证明。 注意到，(6) 等价于

$$\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}\left[\tilde{\boldsymbol{r}}^\top \boldsymbol{x} < y\right] \leq \varepsilon \Longleftrightarrow \sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}\left[\tilde{\boldsymbol{r}}^\top \boldsymbol{x} \leq y\right] \leq \varepsilon,$$

其中，将左侧的严格不等式替换为右侧的弱不等式后，两者等价；参见 Gao 和 Kleywegt（2016年）的命题3。考察风险价值的定义可知,

$$\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}\left[\tilde{\boldsymbol{r}}^\top \boldsymbol{x} \leq y\right] \leq \varepsilon \Longleftrightarrow \sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}\text{-}\mathrm{VaR}_{1-\varepsilon}\left[y - \tilde{\boldsymbol{r}}^\top \boldsymbol{x}\right] \leq 0.$$

根据 Chen 和 Xie（2021年）的推论4.9以及马氏范数的假设，有

$$\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}\text{-}\mathrm{VaR}_{1-\varepsilon}\left[y - \tilde{\boldsymbol{r}}^\top \boldsymbol{x}\right] = \mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}\text{-}\mathrm{VaR}_{1-\underline{\varepsilon}}\left[y - \tilde{\boldsymbol{r}}^\top \boldsymbol{x}\right].$$

换言之，椭圆分布 $\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$ 在风险阈值ε 下的最坏情况风险价值，等于采用较小风险阈值 $\underline{\varepsilon} \leq \varepsilon$ 时的名义椭圆分布风险价值（这将对应于更高的风险水平）。因此，我们得到

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}\text{-}\mathrm{VaR}_{1-\varepsilon}\left[y - \tilde{\boldsymbol{r}}^\top \boldsymbol{x}\right] \leq 0 &\Longleftrightarrow \mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}\text{-}\mathrm{VaR}_{1-\underline{\varepsilon}}[y - \tilde{\boldsymbol{r}}^\top \boldsymbol{x}] \leq 0 \\ &\Longleftrightarrow \mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}\left[\tilde{\boldsymbol{r}}^\top \boldsymbol{x} \leq y\right] \leq \underline{\varepsilon} \\ &\Longleftrightarrow \mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}\left[\tilde{\boldsymbol{r}}^\top \boldsymbol{x} \geq y\right] \geq 1 - \underline{\varepsilon},\end{aligned}$$

其中，最后一个等价关系源于 $\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$ 是连续分布。 Q.E.D.

命题 *2* 的证明。根据引理1，(5) 中的第一个约束等价于

$$\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}\left[\tilde{\boldsymbol{r}}^\top \boldsymbol{x} \geq y\right] \geq 1-\underline{\varepsilon},$$

其中 $\underline{\varepsilon}=1-\Phi(\bar{\eta}^\star) \leq \varepsilon$ 且 $\bar{\eta}^\star$ 是满足以下条件的最小 $\eta \geq \Phi^{-1}(1-\varepsilon)$

$$\eta(\Phi(\eta)-(1-\varepsilon))-\int_{\left(\Phi^{-1}(1-\varepsilon)\right)^2/2}^{\eta^2/2} kg(z)\mathrm{d}z \geq \theta.$$

该约束可进一步写为

$$\begin{aligned}\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x} \geq y] \geq 1-\underline{\varepsilon} &\iff \Phi((\boldsymbol{\mu}^\top \boldsymbol{x}-y)/\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}}) \geq 1-\underline{\varepsilon}\\ &\iff \boldsymbol{\mu}^\top \boldsymbol{x}-y \geq \Phi^{-1}(1-\underline{\varepsilon})\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma}\boldsymbol{x}}\\ &\iff \boldsymbol{\mu}^\top \boldsymbol{x}-y \geq \|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2,\end{aligned}$$

其中，第一个等价关系由椭圆分布的线性性得出，第二个等价关系是因为 $\Phi(\cdot)$ 单调不减，最后一个等价关系则是因为 $1-\varepsilon \geq 0.5$（由$\varepsilon \leq \varepsilon < 0.5$可得）。注意，最优值在 $y^\star=\boldsymbol{\mu}^\top \boldsymbol{x}-\|\Phi^{-1}(1-\varepsilon)\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2$ 处取得，将其代入问题 (5) 的目标函数，即可完成证明。

Q.E.D.

A.3. 第 5 节结果的证明

命题 *3* 的证明。根据命题 1 和命题 2，我们有

$$\inf_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] = -\theta\|\boldsymbol{x}\|_2 + \mathbb{E}_{\hat{\mathbb{P}}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x}]$$

and

$$\inf_{\mathbb{P}\in\mathcal{F}'(\theta)} \mathbb{P}\text{-VaR}_{1-\varepsilon}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] = \boldsymbol{\mu}^\top \boldsymbol{x} - \|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2$$

其中 $\underline{\varepsilon}$ 如所述。将上述两个方程代入 (7) 并重新整理各项，即可完成证明。

Q.E.D.

ˆ命题 *4* 的证明。根据 $\mathbb{T}$ 的定义，问题 (9) 可改写为:

$$\max_{\boldsymbol{\pi}\in(\Delta^A)^S} \psi \sum_{i\in[N]} w_i \cdot g(\boldsymbol{\pi},\hat{\boldsymbol{P}}^i) + (1-\psi)\max_{\eta\in\mathbb{R}}\left\{\eta - \frac{1}{1-\iota}\sum_{i\in[N]} w_i(\eta - g(\boldsymbol{\pi},\hat{\boldsymbol{P}}^i))^+\right\}.$$

通过引入辅助决策变量 $\boldsymbol{y}\in\mathbb{R}^N$，可以进一步将其改写为:

$$\begin{aligned}\max\ & \psi \sum_{i\in[N]} w_i \cdot g(\boldsymbol{\pi},\hat{\boldsymbol{P}}^i) + (1-\psi)\left(\eta - \frac{1}{1-\iota}\sum_{i\in[N]} y_i\right)\\ \text{s.t.}\ & y_i \geq w_i(\eta - g(\boldsymbol{\pi},\hat{\boldsymbol{P}}^i)) \qquad \forall i\in[N]\\ & \boldsymbol{\pi}\in(\Delta^A)^S, \boldsymbol{y}\in\mathbb{R}^N_+, \eta\in\mathbb{R}.\end{aligned} \tag{13}$$

在这里，我们可以表示

$$\begin{aligned} w_i \cdot g(\boldsymbol{\pi}, \boldsymbol{P}) = \max\ & \boldsymbol{\mu}^\top \boldsymbol{x} - \alpha\theta \cdot \|\boldsymbol{x}\|_2 - (1-\alpha) \cdot \|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2 \\ \text{s.t.}\ & x_{s,a} = \pi_{s,a} \cdot \sum_{a' \in \mathcal{A}} x_{s,a'} && \forall (s,a) \in \mathcal{S} \times \mathcal{A} \\ & (\boldsymbol{E} - \gamma \cdot \bar{\boldsymbol{P}})\boldsymbol{x} = w_i \cdot \boldsymbol{p}_0 \\ & \boldsymbol{x} \in \mathbb{R}_+^{SA} \end{aligned} \tag{14}$$

如 Lobo 等人所述 （2020年）。然后，通过结合（13）和（14），我们可以将问题（9）重新表述为 ：

$$\begin{aligned} \max\ & \psi \sum_{i \in [N]} (\boldsymbol{\mu}^\top \boldsymbol{x}^i - \alpha\theta \cdot \|\boldsymbol{x}^i\|_2 - (1-\alpha) \cdot \|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}^i\|_2) + (1-\psi)(\eta - \frac{1}{1-\iota} \sum_{i \in [N]} y_i) \\ \text{s.t.}\ & y_i - w_i \eta \geq \alpha\theta \cdot \|\boldsymbol{x}^i\|_2 + (1-\alpha) \cdot \|\Phi^{-1}(1-\underline{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}^i\|_2 - \boldsymbol{\mu}^\top \boldsymbol{x}^i && \forall i \in [N] \\ & x^i_{s,a} = \pi_{s,a} \cdot \sum_{a' \in \mathcal{A}} x^i_{s,a'} && \forall i \in [N], (s,a) \in \mathcal{S} \times \mathcal{A} \\ & (\boldsymbol{E} - \gamma \cdot \bar{\boldsymbol{P}}^i)\boldsymbol{x}^i = w_i \cdot \boldsymbol{p}_0 && \forall i \in [N] \\ & \boldsymbol{\pi} \in (\Delta^A)^S, \eta \in \mathbb{R}, \boldsymbol{x}^i \in \mathbb{R}_+^{SA}, \boldsymbol{y} \in \mathbb{R}_+^N && \forall i \in [N]. \end{aligned}$$

现在只需关注第二组约束

$$x^i_{s,a} = \pi_{s,a} \cdot \sum_{a' \in \mathcal{A}} x^i_{s,a'}\ \forall i \in [N], (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{15}$$

由于我们只考虑确定性策略 $\boldsymbol{\pi} \in \{0,1\}^{SA}$和 $\sum_{a \in \mathcal{A}} x^i_{s,\ a} \in [0, w_i/(1-\gamma)]$ （参见 Petrik（2010）中的引理 C.10），因此问题（15）的 McCormick松弛（参见 Petrik和Luss（2016））为：

$$\begin{cases} x^i_{s,a} \leq \sum_{a' \in \mathcal{A}} x^i_{s,a'} \\ x^i_{s,a} \leq \dfrac{w_i}{1-\gamma} \pi_{s,a} \\ x^i_{s,a} \geq 0 \\ x^i_{s,a} \geq \dfrac{w_i}{1-\gamma}(\pi_{s,a} - 1) + \sum_{a' \in \mathcal{A}} x^i_{s,a'} \end{cases}$$

对于所有 $i \in [N]$, $(s,\ a) \in \mathcal{S} \times \mathcal{A}$。我们的结论由以下事实得出：当 $\pi \in \{0,1\}$ （即，区间 $[0,1]$ 的端点值）时，McCormick 松弛是精确的。 Q.E.D.

A.4. 第6节结果的证明

命题 *5* 的证明。 根据（11），只需着重求解 $\text{Proj}^{\boldsymbol{B}_{\ell_{\boldsymbol{\Sigma}}(\cdot)}}(\boldsymbol{x})$。通过特征值分解，我们有 $\boldsymbol{\Sigma} = \boldsymbol{G}^\top \boldsymbol{D} \boldsymbol{G}$[4]，其中 $\boldsymbol{D} = \text{diag}(d_1,\ \cdots,\ d_{SA})$，因此我们有：

$$\begin{aligned} \text{Proj}_{\boldsymbol{B}_{\ell_{\boldsymbol{\Sigma}}(\cdot)}}(\boldsymbol{x}) = \ & \arg\min\ \tfrac{1}{2} \cdot \|\boldsymbol{v} - \boldsymbol{x}\|_2^2 \\ & \text{s.t.}\quad \boldsymbol{v}^\top \boldsymbol{G}^\top \boldsymbol{D} \boldsymbol{G} \boldsymbol{v} \leq 1 \\ & \boldsymbol{v} \in \mathbb{R}^{SA}. \end{aligned}$$

[4]这里的特征值分解不计入二分法（或 AD-LPMM算法）的时间复杂度，因为该过程用于计算（求解（8）之前的） $\boldsymbol{\Sigma}^{1/2}$ in (8)。

通过变量替换 $\boldsymbol{u}=\boldsymbol{G}\boldsymbol{v}$ 并令 $\boldsymbol{b}=\boldsymbol{G}\boldsymbol{x}$，只需关注以下等价问题：

$$\begin{aligned} \arg\min\ & \tfrac{1}{2}\cdot\|\boldsymbol{u}-\boldsymbol{b}\|_2^2 \\ \text{s.t.}\quad & \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{u}\le 1 \\ & \boldsymbol{u}\in\mathbb{R}^{SA}, \end{aligned} \tag{16}$$

由此可得 $\boldsymbol{v}^\star=\boldsymbol{G}^\top\boldsymbol{u}^\star$。(16) 的拉格朗日函数（引入对偶变量 $\zeta\in\mathbb{R}_+$ 后）为

$$L(\boldsymbol{u};\zeta)=\frac{1}{2}\cdot\|\boldsymbol{u}-\boldsymbol{b}\|_2^2+\zeta(\boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{u}-1).$$

由于 (16) 是一个凸优化问题，KKT 条件是原始解和对偶解满足最优性的充分条件：

$$\begin{cases} \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{u}\le 1 \\ \zeta\ge 0 \\ \zeta(\boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{u}-1)=0 \\ \nabla_{\boldsymbol{u}}L(\boldsymbol{u};\zeta)=\boldsymbol{u}-\boldsymbol{b}+2\zeta\cdot\boldsymbol{D}\boldsymbol{u}=0, \end{cases}$$

其中，对于 $\zeta=0$，有

$$\begin{cases} \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{u}\le 1 \\ \boldsymbol{u}-\boldsymbol{b}=0; \end{cases}$$

而当 $\zeta>0$ 时，有

$$\begin{cases} \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{u}=1 \\ (\boldsymbol{I}+2\zeta\cdot\boldsymbol{D})\boldsymbol{u}-\boldsymbol{b}=0. \end{cases}$$

因此，若 $\boldsymbol{b}^\top\boldsymbol{D}\boldsymbol{b}\le 1$，则有 $\boldsymbol{u}^\star=\boldsymbol{b}$；若 $\boldsymbol{b}^\top\boldsymbol{D}\boldsymbol{b}>1$，则只需解方程 $g(\zeta)=1$，其中

$$g(\zeta)=\sum_{i\in[SA]}\frac{d_i b_i^2}{(1+2\zeta d_i)^2}.$$

函数 g 在 $[0,+\infty)$ 上单调递减，且 $\lim_{\zeta\to+\infty}g(\zeta)=0$，因此我们可以在区间 $[0,\zeta]$ 上（其中 ζ：$g(\zeta)\le 1$ 是我们在引理 2 中给出的搜索上界）应用二分法，以找到 $\zeta^\star$ 并求得 $u_i^\star=b_i/(1+2\zeta^\star d_i)\ \forall i\in[SA]$。伪代码见算法 2。

求解 $\mathfrak{P}_{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{\xi};c)$ 的时间复杂度主要由二分法决定，其时间复杂度为 $\mathcal{O}(\log(1/\delta'))$。我们的结论基于以下事实：每次迭代中的计算耗时为 $\mathcal{O}(SA)$。

Q.E.D.

Lemma 2. 对于所有 $\zeta\ge(1/(2d_{i''}))(b_{i'}\sqrt{SAd_{i'}}-1)$，不等式 $g(\zeta)\le 1$ 均成立，其中 $i'\in\operatorname{argmax}_{i\in[SA]}d_ib_i^2$ 和 $i''\in\operatorname{argmin}_{i\in[SA]}d_i$

算法 2： 问题 (16) 的二分法

输入： 期望精度 δ'，初始下界 $\zeta \leftarrow 0$ 和上界 $\zeta \geq$ $\quad \bar{} \quad 0$

如果 $g(0) \leq 1$ **则**
  $\boldsymbol{u} \leftarrow \boldsymbol{b}$;
end
else
  当 $|\zeta - \bar{\zeta}| \geq \delta'$ **时**
    $\zeta \leftarrow 0.5(\bar{\zeta} + \underline{\zeta})$;
    如果 $g(\zeta) >= 1$ **那么**
      $\underline{\zeta} \leftarrow \zeta$;
    end
    else
      $\bar{\zeta} \leftarrow \zeta$;
    end
  end
  对于 $i = 1, \ldots, SA$ **执行**
    $u_i = b_i/(1 + 2\zeta d_i)$;
  end
end

输出： 解 $\boldsymbol{u}$

证明。注意，

$$\begin{aligned} g(\zeta) &\leq \sum_{i \in [SA]} \frac{d_{i'} b_{i'}^2}{(1 + 2\zeta d_i)^2} \\ &\leq \frac{SA d_{i'} b_{i'}^2}{(1 + 2\zeta d_{i''})^2}, \end{aligned}$$

由此可得

$$\frac{SA d_{i'} b_{i'}^2}{(1 + 2\zeta d_{i''})^2} \leq 1 \Rightarrow g(\zeta) \leq 1.$$

我们的结论　因此，只需重新排列左侧不等式中的各项即可得出　　.

Q.E.D.

根据引理 2，对于算法 2，可以选择 $\zeta = (1/(2d_{i''}))(b_{i'}\sqrt{SAd_{i'}} - 1)$，其中 $i' \in \arg\max_{i \in [SA]} d_i b_i^2$，且 $i'' \in \arg\min_{i \in [SA]} d_i$。

命题 6 的证明。注意，只需解决 i^{th} 子问题：

$$\arg\min_{z \geq 0} \frac{c}{2} z^2 - (cx_i + \mu_i + \eta_i) z = \max\left\{0, \frac{1}{c}(cx_i + \mu_i + \eta_i)\right\}$$

对所有 $i \in [SA]$ 均成立，因此得出我们的结论。　　Q.E.D.

命题 7 的证明。根据 $\boldsymbol{Q}(\cdot,\cdot)$ 的定义，我们有

$$
\begin{aligned}
&\mathfrak{P}_{\boldsymbol{x}}(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{\xi},\boldsymbol{\eta};c,\nu,\hat{\boldsymbol{x}})\\
&= \arg\min_{\boldsymbol{x}}\ \alpha\theta\cdot\|\boldsymbol{x}\|_2+\boldsymbol{x}^\top((\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})^\top\boldsymbol{\lambda}+\boldsymbol{\xi}+\boldsymbol{\eta})+\frac{c}{2}\cdot\left\|\begin{array}{c}(\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})(\boldsymbol{x}-\hat{\boldsymbol{x}})+(\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})\hat{\boldsymbol{x}}-\boldsymbol{p}_0\\ \boldsymbol{x}-\hat{\boldsymbol{x}}+\hat{\boldsymbol{x}}-\boldsymbol{y}\\ \boldsymbol{x}-\hat{\boldsymbol{x}}+\hat{\boldsymbol{x}}-\boldsymbol{z}\end{array}\right\|_2^2\\
&\quad+\tfrac{1}{2}\cdot\ell^2_{\boldsymbol{Q}(c,\nu)}(\boldsymbol{x}-\hat{\boldsymbol{x}})\\
&= \arg\min_{\boldsymbol{x}}\ \alpha\theta\cdot\|\boldsymbol{x}\|_2+\boldsymbol{x}^\top((\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})^\top\boldsymbol{\lambda}+\boldsymbol{\xi}+\boldsymbol{\eta})+\frac{c}{2}\cdot\left\|\begin{array}{c}(\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})(\boldsymbol{x}-\hat{\boldsymbol{x}})\\ \boldsymbol{x}-\hat{\boldsymbol{x}}\\ \boldsymbol{x}-\hat{\boldsymbol{x}}\end{array}\right\|_2^2\\
&\quad+c\cdot\boldsymbol{x}^\top\left((\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})^\top\left((\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})\hat{\boldsymbol{x}}-\boldsymbol{p}_0\right)+2\cdot\hat{\boldsymbol{x}}-\boldsymbol{y}-\boldsymbol{z}\right)+\tfrac{1}{2}\cdot\ell^2_{\boldsymbol{Q}(c,\nu)}(\boldsymbol{x}-\hat{\boldsymbol{x}})\\
&= \arg\min_{\boldsymbol{x}}\ \frac{\alpha\theta}{c\nu}\cdot\|\boldsymbol{x}\|_2+\boldsymbol{x}^\top\boldsymbol{w}+\frac{1}{2}\cdot\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2\\
&= \arg\min_{\boldsymbol{x}}\ \frac{\alpha\theta}{c\nu}\cdot\|\boldsymbol{x}\|_2+\frac{1}{2}\cdot\|\boldsymbol{x}-(\hat{\boldsymbol{x}}-\boldsymbol{w})\|_2^2\\
&= \left(1-\frac{\frac{\alpha\theta}{c\nu}}{\max\{\|\boldsymbol{w}\|_2,\frac{\alpha\theta}{c\nu}\}}\right)\cdot(\hat{\boldsymbol{x}}-\boldsymbol{w})
\end{aligned}
$$

其中记 $\boldsymbol{w} = \frac{1}{c\nu}\cdot\left((\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})^\top\boldsymbol{\lambda}+\boldsymbol{\xi}+\boldsymbol{\eta}\right)+\frac{1}{\nu}\cdot\hat{}\hat{}$，

$\left((\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})^\top\left((\boldsymbol{E}-\gamma\cdot\bar{\boldsymbol{P}})\boldsymbol{x}-\boldsymbol{p}_0\right)+2\cdot\boldsymbol{x}-\overline{\boldsymbol{y}}-\boldsymbol{z}\right)$

最后一个等式依据，例如，Beck（2017年）中的例6.1.9成立。

计算时间主要取决于计算 $\|\boldsymbol{w}\|_2$，其为 $\mathcal{O}(SA)$。 Q.E.D.

B. Student's t 分布的 VaR 和 CVaR 的评估

自由度为 t 的 Student 分布在阈值 ε 下的风险价值实际上是其概率密度函数（PDF）的下侧 ε 百分位数，可以查阅表格获得，例如，Hogg 和 Craig （1995）（其中列出了一些常用的 $\varepsilon<0.5$ 值）。我们按如下方式计算条件风险价值（假设已知自由度为 $\delta>1$ 、 $v:=\mathbb{P}_{t\text{-dist}}$- 风险价值 $_\varepsilon(\tilde{r}$ ））：

$$
\begin{aligned}
\mathbb{P}_{t\text{-dist}}\text{-CVaR}_\varepsilon(\tilde{r}) &= \tfrac{1}{\varepsilon}\cdot\frac{\Gamma(\frac{\delta+1}{2})}{(\pi\delta)^{\frac{1}{2}}\Gamma(\frac{\delta}{2})}\int_{-\infty}^{v}\frac{r}{(1+\frac{r^2}{\delta})^{\frac{\delta+1}{2}}}\mathrm{d}r\\
&= \tfrac{1}{\varepsilon}\cdot\frac{\delta^{\frac{1}{2}}\cdot\Gamma(\frac{\delta+1}{2})}{2\pi^{\frac{1}{2}}\Gamma(\frac{\delta}{2})}\int_{-\infty}^{1+\frac{v^2}{\delta}}u^{-\frac{k+1}{2}}\mathrm{d}u\\
&= -\frac{\delta^{\frac{1}{2}}\cdot\Gamma(\frac{\delta+1}{2})}{\varepsilon\pi^{\frac{1}{2}}(\delta-1)\Gamma(\frac{\delta}{2})}\cdot\left(1+\tfrac{v^2}{\delta}\right)^{-\frac{k-1}{2}},
\end{aligned}
$$

其中，第一个等号来自条件风险价值的定义以及本文中 t 分布的概率密度函数，第二个等号则由换元积分法得出。

C. 椭圆分布的预备知识

椭圆参考分布 $\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$ 的概率密度分布为

$$
f(\boldsymbol{r})=k\cdot g\left(\frac{1}{2}(\boldsymbol{r}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{r}-\boldsymbol{\mu})\right),
$$

其中，k 是正的归一化标量，$\boldsymbol{\mu}$ 是均值向量，$\boldsymbol{\Sigma}$ 是正定矩阵，g 是生成函数。椭圆分布是一个广泛的分布族，包含多元正态分布、多元 t 分布和多元逻辑斯蒂分布等特例。椭圆分布的一个显著性质是线性性：椭圆分布随机变量的任意线性组合仍服从椭圆分布。也就是说，对于任意随机向量 $\tilde{\boldsymbol{r}} \sim \mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$，有 $\tilde{\boldsymbol{r}}^\top \boldsymbol{x} \sim \mathbb{P}(\mu_{\boldsymbol{x}},\sigma_{\boldsymbol{x}}^2,g)$，其中 $\mu_{\boldsymbol{x}} = \boldsymbol{\mu}^\top \boldsymbol{x}$ 且 $\sigma_{\boldsymbol{x}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}$。实际上，我们可以将该组合表示为 $\tilde{\boldsymbol{r}}^\top \boldsymbol{x} = \mu_{\boldsymbol{x}} + \sigma_{\boldsymbol{x}} \tilde{z}$，其中 $\tilde{z} \sim \mathbb{P}_{(0,1,g)}$ 是一个标准椭圆分布随机变量，其概率密度函数和累积分布函数分别为 $\phi(z) = k \cdot g(z^2/2)$ 和 $\Phi(x) = \int_{-\infty}^{x} k \cdot g(z^2/2)\mathrm{d}z$，。举一个具体例子，我们来仔细考察标准正态分布，其归一化标量和生成函数分别为 $k = 1/\sqrt{2\pi}$ 和 $g(x) = \exp(-x)$。

———

—

D. 分布鲁棒乐观马尔可夫决策过程

与鲁棒模型相比，如果决策者希望进一步了解马尔可夫决策过程，有时会更倾向于探索而非利用。因此，我们可以改用乐观对应形式，关注最优情形，从而引出以下分布鲁棒乐观马尔可夫决策过程：

$$\ell_{\mathrm{O}}(\theta) = \max_{\boldsymbol{x} \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{F}(\theta)} \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x}]. \tag{17}$$

与鲁棒情形不同，在这里，我们的决策取决于可能达到的最佳（期望）结果，这恰好体现了乐观主义。我们将式 (17) 的重构总结如下。

PROPOSITION 8. 分布鲁棒乐观马尔可夫决策过程(17) 等价于一个优化问题

$$\ell_{\mathrm{O}}(\theta) = \max_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] + \theta \|\boldsymbol{x}\|_*.$$

证明。 只需将式 (17) 的目标函数改写如下：

$$\sup_{\mathbb{P} \in \mathcal{F}(\theta)} \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] = - \inf_{\mathbb{P} \in \mathcal{F}(\theta)} \mathbb{E}_{\mathbb{P}}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] = -(\mathbb{E}_{\hat{\mathbb{P}}}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] - \theta \|\boldsymbol{x}\|_*) = \mathbb{E}_{\hat{\mathbb{P}}}[\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] + \theta \|\boldsymbol{x}\|_*,$$

其中，第二个等式的推导与命题 1 的证明类似。 Q.E.D.

命题 8 中的等价改写是一个反向锥规划，通常是非凸的。不过，只要 $\|\cdot\|_*$ 是常用的 L_1-范数或 L_∞-范数，就可以将其重写为混合整数线性规划。此类混合整数线性规划可以用最先进的方法求解。

E. 分布乐观的机会约束模型

在分布乐观型机会约束马尔可夫决策过程模型中，我们关注的是最优情形：以较高概率，奖励不低于某个我们要最大化的下界。形式化地，分布乐观型机会约束马尔可夫决策过程模型表述如下:

$$\ell_{\mathrm{DOCC}}(\theta,\varepsilon)=\left\{\begin{array}{ll}\max & y\\ \text{s.t.} & \sup\limits_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\varepsilon\\ & \boldsymbol{x}\in\mathcal{X},\ y\in\mathbb{R}.\end{array}\right. \tag{18}$$

乐观机会约束模型 (18) 也等价于一个名义机会约束模型，但其风险水平较低。在正式证明这一论点之前，先引入以下两个引理。

LEMMA 3. 随机向量 $\tilde{\boldsymbol{r}}$ 在集合 $\mathcal{R}$中取值的最大概率，

$$\sup_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}[\tilde{\boldsymbol{r}}\in\mathcal{R}], \tag{19}$$

等价于

$$\min_{\lambda\geq 0}\left\{\lambda\theta+\int_{\boldsymbol{r}\in\mathbb{R}^{SA}}(\lambda\cdot\mathbf{dist}(\boldsymbol{r},\mathcal{R})-1)^-\mathrm{d}\hat{\mathbb{P}}\boldsymbol{r}\right\}.$$

这里，我们用 $\mathbf{距离}(\boldsymbol{r},\mathcal{R})=$ 下确界$\{\|\boldsymbol{r}-\hat{}\boldsymbol{r}\|\mid\hat{}\boldsymbol{r}\in\mathcal{R}\}$ 表示向量 $\boldsymbol{r}\in\mathbb{R}^{SA}$ 到集合 $\mathcal{R}\subseteq\mathbb{R}^{SA}$的距离。

证明。 根据 Gao 和 Kleywegt（2016年）的定理1或 Blanchet 和 Murthy（2019年）的定理1，不确定性量化问题（19）等价于

$$\min_{\lambda\geq 0}\left\{\lambda\theta-\int_{\boldsymbol{r}\in\mathbb{R}^{SA}}\inf_{\boldsymbol{w}\in\mathbb{R}^{SA}}\{\lambda\|\boldsymbol{w}-\boldsymbol{r}\|-\mathbb{I}[\boldsymbol{w}\in\mathcal{R}]\}\mathrm{d}\hat{\mathbb{P}}\boldsymbol{r}\right\}, \tag{20}$$

其中 $\mathbb{I}$ 是 0-1 示性函数。考虑上述最小化问题目标函数中的第二项，可得

$$\inf_{\boldsymbol{w}\in\mathbb{R}^{SA}}\{\lambda\|\boldsymbol{w}-\boldsymbol{r}\|-\mathbb{I}[\boldsymbol{w}\in\mathcal{R}]\}=-(\lambda\cdot\mathbf{dist}(\boldsymbol{r},\mathcal{R})-1)^-. \tag{21}$$

事实上，如果 $\boldsymbol{r}\in\mathcal{R}$（其中，$\mathbf{距离}(\boldsymbol{r},\ \mathcal{R})=0$），那么通过选择 $\boldsymbol{w}=\boldsymbol{v}$，可得

$$\inf_{\boldsymbol{w}\in\mathbb{R}^{SA}}\{\lambda\|\boldsymbol{w}-\boldsymbol{r}\|-\mathbb{I}[\boldsymbol{w}\in\mathcal{R}]\}=-1=-(\lambda\cdot\mathbf{dist}(\boldsymbol{r},\mathcal{R})-1);$$

而如果 $\boldsymbol{r}\notin\mathcal{R}$，则可得

$$\begin{aligned}\inf_{\boldsymbol{w}\in\mathbb{R}^{SA}}\{\lambda\|\boldsymbol{w}-\boldsymbol{r}\|-\mathbb{I}[\boldsymbol{w}\in\mathcal{R}]\}&=\min\left\{\inf_{\boldsymbol{w}\in\mathcal{R}}\{\lambda\|\boldsymbol{w}-\boldsymbol{r}\|-1\},\inf_{\boldsymbol{w}\notin\mathcal{R}}\lambda\|\boldsymbol{w}-\boldsymbol{r}\|\right\}\\&=\min\left\{\inf_{\boldsymbol{w}\in\mathcal{R}}\{\lambda\|\boldsymbol{w}-\boldsymbol{r}\|-1\},0\right\}\\&=-(\lambda\cdot\mathbf{dist}(\boldsymbol{r},\mathcal{R})-1)^-.\end{aligned}$$

将表达式 (21) 代入问题 (20)，即可得到所需结果。根据 Gao 和 Kleywegt（2016年）的命题 3，无论 $\mathcal{R}$ 是开集还是闭集，该结果均成立。 Q.E.D.

LEMMA 4. 分布鲁棒乐观机会约束

$$\inf_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}[\tilde{\boldsymbol{r}} \in \mathcal{R}] \leq \varepsilon \tag{22}$$

风险阈值为 $\varepsilon \in (0,1)$ 时，当且仅当

$$\mathbb{P}\text{-CVaR}_{\varepsilon}[-\mathbf{dist}(\tilde{\boldsymbol{r}}, \bar{\mathcal{R}})] \geq -\frac{\theta}{1-\varepsilon},$$

其中， $\bar{\mathcal{R}} = \mathbb{R}^{SA} \setminus \mathcal{R}$ 是集合 $\mathcal{R}$ 的补集， $\mathcal{R}$表示不期望事件。

证明。 我们首先将 (22) 重新表述为

$$\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}[\tilde{\boldsymbol{r}} \in \bar{\mathcal{R}}] \geq 1-\varepsilon.$$

利用引理 3，上述约束等价于

$$\min_{\lambda\geq 0} \left\{ \lambda\theta + \int_{\boldsymbol{r}\in\mathbb{R}^{SA}} (\lambda \cdot \mathbf{dist}(\boldsymbol{r}, \bar{\mathcal{R}}) - 1)^{-} \mathrm{d}\hat{\mathbb{P}}\boldsymbol{r} \right\} \geq 1-\varepsilon. \tag{23}$$

左侧问题可表示为

$$\min\left\{ \min_{\lambda>0} \left\{ \lambda\theta + \int_{\boldsymbol{r}\in\mathbb{R}^{SA}} (\lambda \cdot \mathbf{dist}(\boldsymbol{r}, \bar{\mathcal{R}}) - 1)^{-} \mathrm{d}\hat{\mathbb{P}}\boldsymbol{r} \right\}, 1 \right\}.$$

由于 $1 \geq 1-\varepsilon$，上述改写意味着约束（23）等价于

$$\min_{\lambda>0} \left\{ \lambda\theta + \int_{\boldsymbol{r}\in\mathbb{R}^{SA}} (\lambda \cdot \mathbf{dist}(\boldsymbol{r}, \bar{\mathcal{R}}) - 1)^{-} \mathrm{d}\hat{\mathbb{P}}\boldsymbol{r} \right\} \geq 1-\varepsilon.$$

将两边乘以（λ（$1-\varepsilon$））$^{-1} > 0$，得到

$$\min_{\tau<0} \left\{ \frac{1}{1-\varepsilon} \int_{\boldsymbol{r}\in\mathbb{R}^{SA}} (-\mathbf{dist}(\boldsymbol{r}, \bar{\mathcal{R}}) - \tau)^{+} \mathrm{d}\hat{\mathbb{P}}\boldsymbol{r} + \tau \right\} \geq -\frac{\theta}{1-\varepsilon},$$

结合这一事实

$$\min_{\tau\geq 0} \left\{ \frac{1}{1-\varepsilon} \int_{\boldsymbol{r}\in\mathbb{R}^{SA}} (-\mathbf{dist}(\boldsymbol{r}, \bar{\mathcal{R}}) - \tau)^{+} \mathrm{d}\hat{\mathbb{P}}\boldsymbol{r} + \tau \right\} \geq 0 \geq -\frac{\theta}{1-\varepsilon},$$

等价于

$$\min_{\tau\in\mathbb{R}} \left\{ \frac{1}{1-\varepsilon} \int_{\boldsymbol{r}\in\mathbb{R}^{SA}} (-\mathbf{dist}(\boldsymbol{r}, \bar{\mathcal{R}}) - \tau)^{+} \mathrm{d}\hat{\mathbb{P}}\boldsymbol{r} + \tau \right\} \geq -\frac{\theta}{1-\varepsilon},$$

ˆ⁻其中左侧本质上是 $\mathbb{P}$-条件风险价值$_{\varepsilon}$[−**分布**（$\tilde{\boldsymbol{r}}$， $\mathcal{R}$）]。 Q.E.D.

现在，我们可以证明机会约束模型与其乐观对应形式（采用调整后的风险阈值）之间的等价性。

LEMMA 5. 假设在 *Wasserstein*歧义集 (3) 中，参考分布为椭圆分布 $\mathbb{P} = \mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$，且 *Wasserstein*距离采用与正定矩阵 $\boldsymbol{\Sigma}$ 相关联的马氏范数。分布乐观的鲁棒机会约束

$$\exists\, \mathbb{P} \in \mathcal{F}(\theta) : \mathbb{P}[\tilde{\boldsymbol{r}}^{\top}\boldsymbol{x} \geq y] \geq 1-\varepsilon$$

当且仅当 $\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}[\tilde{\boldsymbol{r}}^{\top}\boldsymbol{x} \geq y] \geq 1-\bar{\varepsilon}$，其中 $\bar{\varepsilon} = 1-\Phi(\eta^{\star}) \geq \varepsilon$ 且 $\eta^{\star}$ 是满足 $\eta(\Phi(\eta) - (1-\varepsilon))$ $+ \int_{\eta^2/2}^{(\Phi^{-1}(1-\varepsilon))^2/2} kg(z)\mathrm{d}z \leq \theta$. 的最小值 $\eta \leq \Phi^{-1}(1-\varepsilon)$

证明。 我们首先考察单个分布乐观鲁棒机会约束

$$\exists\,\mathbb{P}\in\mathcal{F}(\theta):\mathbb{P}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\varepsilon$$

对于某个一般的系数向量 $\boldsymbol{x}\in\mathbb{R}^{SA}$。上述机会约束等价于

$$\sup_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\varepsilon\iff\sup_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}> y]\geq 1-\varepsilon\iff\inf_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\leq y]\leq\varepsilon,$$

对于第一个等价关系，根据 Gao 和 Kleywegt（2016年）的命题3，将严格不等式替换为弱不等式不会改变结果。考察风险价值的定义，我们注意到

$$\inf_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\leq y]\leq\varepsilon\Longleftrightarrow\inf_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}\text{-VaR}_{1-\varepsilon}[y-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}]\leq 0.$$

因此，利用风险价值的平移不变性，只需证明

$$\inf_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}\text{-VaR}_{1-\varepsilon}[-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}]\triangleq\inf_{v\in\mathbb{R}}\left\{v\mid\inf_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}[-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}>v]\leq\varepsilon\right\}.\tag{24}$$

根据引理 4和马氏范数的假设，我们有

$$\begin{aligned}\inf_{\mathbb{P}\in\mathcal{F}(\theta)}\mathbb{P}\left[-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}>v\right]\leq\varepsilon&\iff\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}\text{-CVaR}_\varepsilon[-\mathbf{dist}(\tilde{\boldsymbol{r}},\bar{\mathcal{R}})]\geq-\frac{\theta}{1-\varepsilon}\\&\iff-\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}\text{-CVaR}_\varepsilon[-(-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}-v)^+]\leq\frac{\theta\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}^{-1}}}{1-\varepsilon},\end{aligned}$$

其中 $\bar{\mathcal{R}}=\left\{\boldsymbol{r}\mid-\boldsymbol{r}^\top\boldsymbol{x}\leq v\right\}$，并且我们利用闭式解

$$\mathbf{dist}(\tilde{\boldsymbol{r}},\bar{\mathcal{R}})=\left(-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}-v\right)^+/\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}^{-1}};$$

参见 Chen等人（2018年）中的引理 2，例如。

为简单起见，令 $\mathbb{P}_{\mathrm{S}}=\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$。根据椭圆分布的性质，对于 $\tilde{\boldsymbol{r}}\sim\mathbb{P}_{\mathrm{S}}$ 和任意实向量 $\boldsymbol{x}$，有 $-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\sim\mathbb{P}_{(\mu_{\mathrm{S}},\sigma^{2\mathrm{S}}\,g)}=\mathbb{P}_{(-\boldsymbol{\mu}^\top\boldsymbol{x},\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x},g)}$。我们将其概率密度函数记为

$$h(z)=\frac{k}{\sigma_{\mathrm{S}}}\cdot g\left(\frac{(z-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}\right).$$

约束的左侧还可以进一步变换为

$$\begin{aligned}&-\mathbb{P}_{\mathrm{S}}\text{-CVaR}_\varepsilon[-(-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}-v)^+]\\=\;&-\mathbb{E}_{\mathbb{P}_{\mathrm{S}}}[-(-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}-v)^+\mid-(-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}-v)^+\geq\mathbb{P}_{\mathrm{S}}\text{-VaR}_\varepsilon[-(-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}-v)^+]]\\=\;&-\frac{1}{1-\varepsilon}\int_{-\infty}^{\sup\{z\mid-(z-v)^+\geq\mathbb{P}_{\mathrm{S}}\text{-VaR}_\varepsilon[-(-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}-v)^+]\}}-(z-v)^+h(z)\mathrm{d}z\\=\;&\frac{1}{1-\varepsilon}\int_{v}^{\sup\{z\mid-(z-v)^+\geq\mathbb{P}_{\mathrm{S}}\text{-VaR}_\varepsilon[-(-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}-v)^+]\}}(z-v)h(z)\mathrm{d}z\\=\;&\frac{1}{1-\varepsilon}\int_{v}^{\mathbb{P}_{\mathrm{S}}\text{-VaR}_{1-\varepsilon}[-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}]}(z-v)h(z)\mathrm{d}z,\end{aligned}$$

其中，最后一个等式由.....得出

$$\begin{aligned}
&\sup\{z \mid -(z-v)^+ \geq \mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{\varepsilon}[-(-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}-v)^+]\} \\
=\ &\sup\{z \mid \min\{v-z,0\} \geq \mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{\varepsilon}[\min\{v+\tilde{\boldsymbol{r}}^\top \boldsymbol{x},0\}]\} \\
=\ &\sup\{z \mid \min\{-z,-v\} \geq \mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{\varepsilon}[\min\{\tilde{\boldsymbol{r}}^\top \boldsymbol{x},-v\}]\} \\
=\ &\sup\{z \mid -z \geq \mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{\varepsilon}[\min\{\tilde{\boldsymbol{r}}^\top \boldsymbol{x},-v\}]\} \\
=\ &\sup\{z \mid z \leq \mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{1-\varepsilon}[\max\{-\tilde{\boldsymbol{r}}^\top \boldsymbol{x},v\}]\} \\
=\ &\sup\{z \mid z \leq \mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{1-\varepsilon}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}]\} \\
=\ &\mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{1-\varepsilon}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}].
\end{aligned}$$

这里，第二个等式源于风险价值的平移不变性，第三个等式由$-v \geq \mathbb{P}_{\mathrm{S}}$-风险价值$_{\varepsilon}[\min\{\tilde{\boldsymbol{r}}^\top \boldsymbol{x},-v\}]$得出，第五个等式则是因为对于任意 $\varepsilon \in (0,1)$，分布乐观鲁棒风险价值满足

$$v = \inf_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}\text{-}\mathrm{VaR}_{1-\varepsilon}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}] \leq \mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{1-\varepsilon}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}], \tag{25}$$

因此，$-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}$和 $\max\{-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}, v\}$的 $1-\varepsilon$分位数相同。

令 $q_{1-\varepsilon} = \mathbb{P}_{\mathrm{S}}$-风险价值$_{1-\varepsilon}[-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}]$，根据其定义，它满足

$$\frac{q_{1-\varepsilon}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}} = \mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{1-\varepsilon}\left[\frac{-\tilde{\boldsymbol{r}}^\top \boldsymbol{x}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right] = \mathbb{P}^0_{(0,1,g)}\text{-}\mathrm{VaR}_{1-\varepsilon}[\tilde{z}] = \Phi^{-1}(1-\varepsilon),$$

这里，第一个等式成立是因为风险价值具有平移不变性和正齐次性，而最后一个等式则根据标准椭圆分布$\mathbb{P}_{(0,1,g)}$下风险价值的定义得出。

根据约束的最后一种重新表述，我们还可以得到

$$\frac{1}{1-\varepsilon}\int_v^{q_{1-\varepsilon}} (z-v)h(z)\mathrm{d}z = \frac{1}{1-\varepsilon}\int_v^{q_{1-\varepsilon}} z\cdot\frac{k}{\sigma_{\mathrm{S}}}\cdot g\left(\frac{(z-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}\right)\mathrm{d}z - \frac{v}{1-\varepsilon}\int_v^{q_{1-\varepsilon}} \frac{k}{\sigma_{\mathrm{S}}}\cdot g\left(\frac{(z-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}\right)\mathrm{d}z.$$

对于其第一个分量，我们有

$$\begin{aligned}
&\frac{1}{1-\varepsilon}\int_v^{q_{1-\varepsilon}} z\cdot\frac{k}{\sigma_{\mathrm{S}}}\cdot g\left(\frac{(z-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}\right)\mathrm{d}z\\
=\;&\frac{1}{1-\varepsilon}\int_v^{q_{1-\varepsilon}} \frac{z-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\cdot k\cdot g\left(\frac{(z-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}\right)\mathrm{d}z+\frac{1}{1-\varepsilon}\int_v^{q_{1-\varepsilon}}\frac{\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\cdot k\cdot g\left(\frac{(z-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}\right)\mathrm{d}z\\
=\;&\frac{\sigma_{\mathrm{S}}}{1-\varepsilon}\int_v^{q_{1-\varepsilon}} \frac{z-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\cdot k\cdot g\left(\frac{(z-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}\right)\mathrm{d}\left(\frac{z-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)+\\
&\frac{\mu_{\mathrm{S}}}{1-\varepsilon}\left(\Phi\left(\frac{q_{1-\varepsilon}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)-\Phi\left(\frac{v-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)\right)\\
=\;&\frac{\sigma_{\mathrm{S}}}{1-\varepsilon}\int_{\frac{v-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}}^{\frac{q_{1-\varepsilon}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}} t\cdot k\cdot g\left(\frac{t^2}{2}\right)\mathrm{d}\left(\frac{z-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)+\frac{\mu_{\mathrm{S}}}{1-\varepsilon}\left(\Phi\left(\frac{q_{1-\varepsilon}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)-\Phi\left(\frac{v-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)\right)\\
=\;&\frac{\sigma_{\mathrm{S}}}{1-\varepsilon}\int_{\frac{(v-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}}^{\frac{(q_{1-\varepsilon}-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}} k\cdot g(z)\mathrm{d}z+\frac{\mu_{\mathrm{S}}}{1-\varepsilon}\left(\Phi\left(\frac{q_{1-\varepsilon}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)-\Phi\left(\frac{v-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)\right),
\end{aligned}$$

而对于第二个分量，有

$$\begin{aligned}
\frac{v}{1-\varepsilon}\int_v^{q_{1-\varepsilon}}\frac{k}{\sigma_S}\cdot g\left(\frac{(z-\mu_S)^2}{2\sigma_S^2}\right)\mathrm{d}z\;&=\;\frac{v}{1-\varepsilon}\int_{\frac{v-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}}^{\frac{q_{1-\varepsilon}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}} k\cdot g\left(\frac{z^2}{2}\right)\mathrm{d}z\\
&=\;\frac{v}{1-\varepsilon}\left(\Phi\left(\frac{q_{1-\varepsilon}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)-\Phi\left(\frac{v-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)\right).
\end{aligned}$$

因此，将该约束与式（25）结合，可得问题（24）的以下等价表达式:

$$\begin{aligned}
&\inf\; v\\
&\text{s.t.}\int_{\frac{(v-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}}^{\frac{(q_{1-\varepsilon}-\mu_{\mathrm{S}})^2}{2\sigma_{\mathrm{S}}^2}} k\cdot g(z)\mathrm{d}z+\frac{\mu_{\mathrm{S}}-v}{\sigma_{\mathrm{S}}}\left(\Phi\left(\frac{q_{1-\varepsilon}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)-\Phi\left(\frac{v-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)\right)\le\frac{\theta\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}^{-1}}}{\sigma_{\mathrm{S}}}=\theta\\
&\quad v\le\mathbb{P}_{\mathrm{S}}\text{-}\mathrm{VaR}_{1-\varepsilon}[-\tilde{\boldsymbol{r}}^\top\boldsymbol{x}]\\
&\quad v\in\mathbb{R},
\end{aligned}$$

其中，等号由马氏范数的定义可得。令 $\eta=(v-\mu_{\mathrm{S}})/\sigma_{\mathrm{S}}$，此时最优情形下的风险价值变为

$$\begin{aligned}
&\inf\;\mu_{\mathrm{S}}+\sigma_{\mathrm{S}}\eta\\
&\text{s.t.}\int_{\eta^2/2}^{(\Phi^{-1}(1-\varepsilon))^2/2} k\cdot g(z)\mathrm{d}z-\eta\cdot(1-\varepsilon-\Phi(\eta))\le\theta\\
&\quad\eta\le\Phi^{-1}(1-\varepsilon)\\
&\quad\eta\in\mathbb{R}.
\end{aligned}\tag{26}$$

该函数

$$V(\eta)\triangleq\int_{\eta^2/2}^{(\Phi^{-1}(1-\varepsilon))^2/2} k\cdot g(z)\mathrm{d}z-\eta\cdot(1-\varepsilon-\Phi(\eta))$$

在（$-\infty$, $\Phi^{-1}(1-\varepsilon)$）上单调递减，因为对于任意 $\eta<\Phi^{-1}(1-\varepsilon)$，都有

$$V'(\eta)=-\eta\cdot k\cdot g\left(\frac{\eta^2}{2}\right)-(1-\varepsilon)+\Phi(\eta)+\eta\phi(\eta)=\Phi(\eta)-(1-\varepsilon)<0.$$

因此，问题 (26) 可以通过二分算法高效求解，并可得到所声称的最优 $\eta^\star$。最后，结果如下：

$$\begin{aligned}\exists\,\mathbb{P}\in\mathcal{F}(\theta):\mathbb{P}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\varepsilon &\iff -y\geq\sigma_{\mathrm{S}}\underline{\eta}^\star+\mu_{\mathrm{S}}\\ &\iff\frac{-y-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\geq\underline{\eta}^\star\\ &\iff\Phi\left(\frac{-y-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right)\geq\Phi(\underline{\eta}^\star)\\ &\iff\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}\left[\frac{\tilde{\boldsymbol{r}}^\top\boldsymbol{x}-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\geq\frac{y-\mu_{\mathrm{S}}}{\sigma_{\mathrm{S}}}\right]\geq 1-\bar{\varepsilon}\\ &\iff\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\bar{\varepsilon}.\end{aligned}$$

Q.E.D.

利用引理5中的 $\bar{\varepsilon}$，现在我们可以推导分布乐观型机会约束模型 (18) 的二阶锥重构。

PROPOSITION 9. 假设在 *Wasserstein* 歧义集 (3) 中，参考分布是一个椭圆分布 $\mathbb{P}=\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}$，且 *Wasserstein* 距离采用与正定矩阵 $\boldsymbol{\Sigma}$ 相关联的马氏范数。如果风险阈值满足 $\varepsilon\leq\bar{\varepsilon}<0.5$，则分布乐观型机会约束马尔可夫决策过程 (18) 等价于以下二阶锥规划

$$\ell_{\mathrm{DOCC}}(\theta,\varepsilon)=\max_{\boldsymbol{x}\in\mathcal{X}}\ \boldsymbol{\mu}^\top\boldsymbol{x}-\|\Phi^{-1}(1-\bar{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2,$$

其中 $\bar{\varepsilon}=1-\Phi(\underline{\eta}^\star)\geq\varepsilon$，其中 $\underline{\eta}^\star$ 是满足以下条件的最小 $\eta\leq\Phi^{-1}(1-\varepsilon)$：

$$\eta(\Phi(\eta)-(1-\varepsilon))+\int_{\eta^2/2}^{\left(\Phi^{-1}(1-\varepsilon)\right)^2/2}kg(z)\mathrm{d}z\leq\theta.$$

证明。根据引理5，式（18）中的第一个约束等价于

$$\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\bar{\varepsilon},$$

其中 $\bar{\varepsilon}=1-\Phi(\underline{\eta}^\star)\geq\varepsilon$ 满足条件的最小 $\eta\leq\Phi^{-1}(1-\varepsilon)$ 为 $\eta^\star$

$$\eta(\Phi(\eta)-(1-\varepsilon))+\int_{\eta^2/2}^{\left(\Phi^{-1}(1-\varepsilon)\right)^2/2}kg(z)\mathrm{d}z\leq\theta,$$

可进一步变换如下：

$$\begin{aligned}\mathbb{P}_{(\boldsymbol{\mu},\boldsymbol{\Sigma},g)}[\tilde{\boldsymbol{r}}^\top\boldsymbol{x}\geq y]\geq 1-\bar{\varepsilon}&\iff\Phi((\boldsymbol{\mu}^\top\boldsymbol{x}-y)/\sqrt{\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}})\geq 1-\bar{\varepsilon}\\ &\iff\boldsymbol{\mu}^\top\boldsymbol{x}-y\geq\Phi^{-1}(1-\bar{\varepsilon})\sqrt{\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}}\\ &\iff\boldsymbol{\mu}^\top\boldsymbol{x}-y\geq\|\Phi^{-1}(1-\bar{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2,\end{aligned}$$

其中，第一个等价关系由椭圆分布的线性性得出，第二个等价关系是因为累积分布函数 $\Phi(\cdot)$ 单调不减，第三个等价关系则因为 $\bar{\varepsilon} < 0.5$。由于最优值在 $y = \boldsymbol{\mu}^\top \boldsymbol{x} - \|\Phi^{-1}(1-\bar{\varepsilon})\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2$ 时取得，将此等式代入目标函数（18），即可完成证明。

Q.E.D.

F. 关于鲁棒马尔可夫决策过程的更多细节

正如 Delage 和 Mannor（2010）所引入的，鲁棒马尔可夫决策过程通过考虑预定义歧义集内不确定参数的最坏情况实现，来最大化总期望回报：

$$\max_{\boldsymbol{\pi}\in\Pi} \min_{r^0\in\mathcal{R}, r^1\in\mathcal{R},\cdots} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^t(s_t) \mid s_0 \propto \boldsymbol{p}_0, \boldsymbol{\pi}\right], \tag{27}$$

其中，Π 是所有平稳随机策略的集合，r^t 和 s_t 分别表示时间阶段 t的奖励和状态。与 Delage 和 Mannor（2010）一样，我们将 $\mathcal{R}$ 设为随机奖励向量的 99% 置信椭球，作为不确定集。

G. 关于 BROIL 的更多细节

与我们的收益-风险模型类似，BROIL（Brown 等人，2020年）也旨在寻找一种策略，使均值表现和百分位表现的加权平均值最大化：

$$\max_{\boldsymbol{\pi}\in\Pi} \lambda \cdot \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^t(s_t) \mid s_0 \propto \boldsymbol{p}_0, \boldsymbol{\pi}\right] + (1-\lambda)\cdot \mathrm{CVaR}_{\varepsilon}\left[\sum_{t=0}^{\infty} \gamma^t r^t(s_t) \mid s_0 \propto \boldsymbol{p}_0.\boldsymbol{\pi}\right], \tag{28}$$

其中 $\lambda \in [0,1]$ 是权重。给定作为（n）奖励样本矩阵的 $\boldsymbol{R} \in \mathbb{R}^{SA\times n}$，BROIL 可表示为如下线性规划：

$$\max_{\boldsymbol{x}\in\mathcal{X}, y\in\mathbb{R}} \lambda \cdot \frac{1}{n}\boldsymbol{e}^\top \boldsymbol{R}^\top \boldsymbol{x} + (1-\lambda)\cdot\left(y - \frac{1}{\varepsilon}\cdot\frac{1}{n}\boldsymbol{e}^\top(y\cdot\boldsymbol{e} - \boldsymbol{R}^\top\boldsymbol{x})\right).$$

请注意，BROIL 与我们的收益-风险模型有两个主要区别：首先，BROIL 使用条件风险价值作为其风险度量，而我们的收益-风险模型采用风险价值；其次，尽管我们的目标函数考虑了（回报的均值和风险价值的）分布鲁棒性，BROIL 只计算回报的名义均值和条件风险价值。

H. 实验的补充细节与结果 H.1. 参数选择的补充细节

在仿真和实证研究中，我们均使用交叉验证进行参数选择。对于 DRMDP (4)，θ的候选集合为 $\{0, 2, \cdots, 18\}$；对于 CC (2)，ε的候选集合为 $\{i\varepsilon'/5\}_{i\in[5]}$；对于 RR (7)，我们选择 θ，使得 ε 在 $\{i\varepsilon'/5\}_{i\in[5]}$之间变化，并选择 $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$；对于 BROIL (28)，我们选择 $\lambda \times \varepsilon \in \{0, 0.25, 0.5, 0.75, 1\} \times \{0.05, 0.1, 0.15\}$；对于 RMDPs (27)，与 Delage 和 Mannor（2010）一样，我们将 $\mathcal{R}$ 设为随机奖励向量的 99% 置信椭球，作为不确定集。

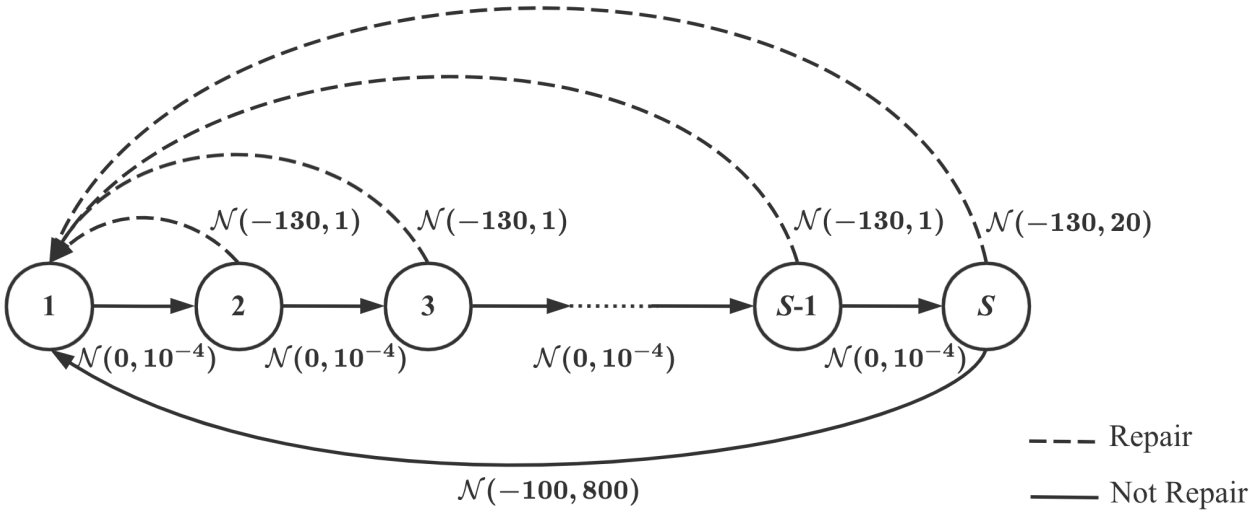


图 6 具有固定高斯奖励的机器更换问题。

H.2. 仿真实验的补充细节

我们考虑 $S=10$ 个状态、 $A=10$ 个动作、均匀初始状态分布，以及折扣因子 $\gamma=0.95$。对于每个状态 $s\in[S]$，可达后继状态的数量为 $\lceil\log S\rceil$。我们从多元正态分布 $\mathcal{N}(\boldsymbol{\mu}',\ \boldsymbol{\Sigma}')$ 中采样真实奖励，其中对于每个 $k\in[SA]$， μ'_k按如下方式生成：首先，我们从$\{0,1\}$的离散均匀分布中采样一个数（0 或 1）。如果结果为 0，则从正态分布 $\mathcal{N}(50,\ 100)$ 中生成 μ'_k；否则，从 $\mathcal{N}(90,\ 100)$ 中生成。奖励的标准差以同样的方式，使用另外两个正态分布 $\mathcal{N}(3,\ 9)$ 和 $\mathcal{N}(18,\ 9)$ 生成。完成上述过程后，将标准差和均值均截断为非负值。奖励的相关矩阵按如下方式生成：首先，我们采样一个矩阵 $\boldsymbol{R}\in\mathbb{R}^{SA\times SA}$，其所有元素均独立地从 $[0.25,1]$ 中均匀采样；然后得到相关矩阵 $\mathrm{diag}(\boldsymbol{d})\boldsymbol{V}\mathrm{diag}(\boldsymbol{d})$，其中 $\boldsymbol{V}=\boldsymbol{R}^\top\boldsymbol{R}$ 和 $\boldsymbol{d}=\{d_i\}_{i\in[SA]}=\{1/\sqrt{V_{ii}}\}_{i\in[SA]}$。

—

H.3. 实证研究的补充细节

在本实验中，每台机器都遵循相同的底层马尔可夫决策过程，其状态集为 $\mathcal{S}=[S]$，包含 $S=50$，动作集则仅包含两个动作：维修机器或不维修。状态转移是确定性的，折扣因子为 0.8。奖励取决于当前状态和行动，且所有奖励均相互独立并服从正态分布。图 6 展示了生成随机奖励的真实底层分布。

H.4. 仿真实验的补充结果 H.5. 实证研究的补充结果

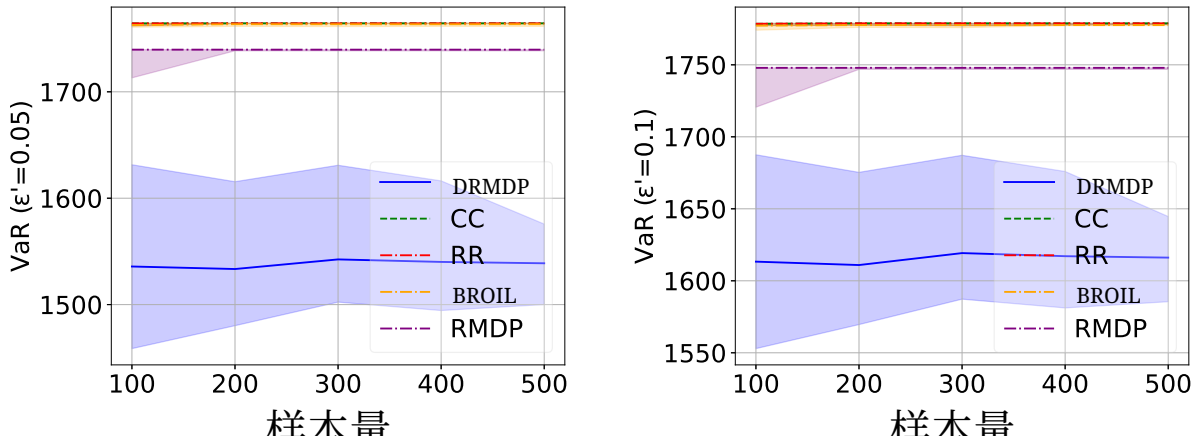


图 7 仿真。采用风险价值（风险阈值 $\varepsilon' \in \{5\%, 10\%\}$）评估模型 DRMDP (4)、机会约束 (2)、收益-风险马尔可夫决策过程 (7)、RMDP 和 BROIL。阴影区域的上边缘和下边缘分别表示 100 次性能结果的第 95 和第 5 百分位数，实线表示中位数。

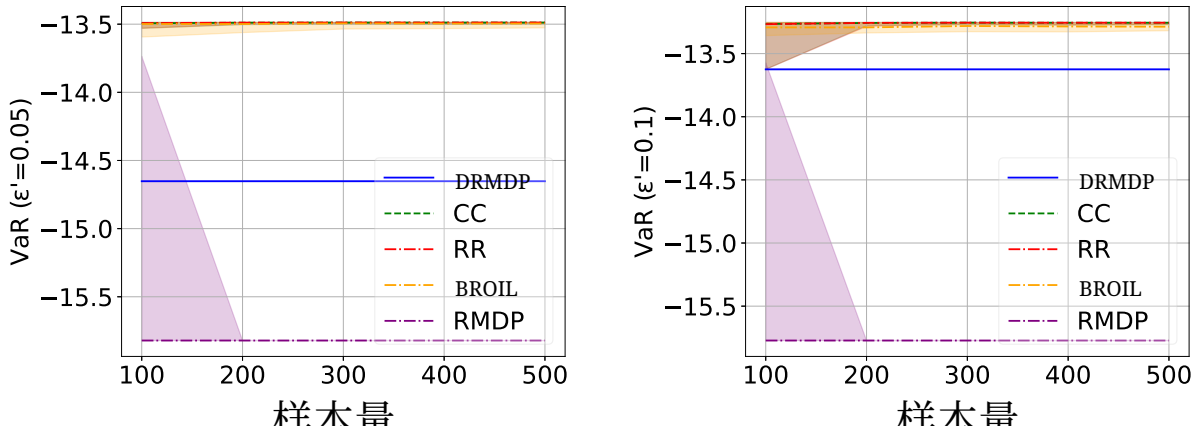


图 8 实证结果。采用风险价值评估的模型：DRMDP (4)、机会约束 (2)、收益-风险马尔可夫决策过程 (7)、RMDP 和 BROIL（风险阈值 $\varepsilon' \in \{5\%, 10\%\}$）。阴影区域的上边缘和下边缘分别表示 100 次表现的第 95 和第 5 百分位数，实线表示中位数。

I. 相关工作

表 2 总结了与我们工作相关的文献。我们指出，与表 2 中的相关工作相比，我们的收益-风险模型是唯一考虑风险歧义的模型；此外，我们还设计了一种快速的一阶算法来求解该模型，从而提升了模型在大规模问题中的实用性。

表 2 相关工作。

论文	不确定性	鲁棒性	歧义集	风险度量	软鲁棒性
Delage 和 Mannor（2010年）	奖励 and 转移核	-	-	VaR	No
Xu 和 Mannor（2010年）	奖励 and 转移核	DRO	嵌套	-	No
Yu 和 Xu（2015）	奖励 and 转移核	DRO	（一般）嵌套	-	No
Brown 等人（2020）	奖励	-	-	CVaR	Yes
Gilbert 等人（2017）	奖励	-	-	VaR	No
Lobo 等人（2020）	转移核	-	-	CVaR	Yes
Yang（2020）	转移核	DRO	瓦瑟斯坦	-	No
本文	奖励	DRO	瓦瑟斯坦	VaR	Yes