概率攻击图中主动防御的最优诱饵资源分配

Haoxiang Ma 佛罗里达大学 盖恩斯维尔，美国 hma2@ufl.edu

Shuo Han 伊利诺伊大学芝加哥分校 芝加哥，美国 hanshuo@uic.edu

Nandi Leslie 雷神技术公司 阿灵顿县，美国 n andi.o.leslie@raytheon.com

Charles Kamhoua 美国陆军研究实验室 盖恩斯维尔，美国 charles.a.ka mhoua.civ@mail.mil

Jie Fu 佛罗里达大学 盖恩斯维尔，美国 fujie@ufl.edu

arXiv:2301.01336v1 [cs.MA] 3 Jan 2023

摘要

本文研究主动防御系统的综合问题。在这类系统中，防御者可以部署欺骗性目标，并改变攻击者采取行动的成本；攻击者的目标是攻陷系统中的安全资产。我们使用一种形式化安全模型——概率攻击图，来建模攻击者与系统之间的交互。防御者通过部署虚假目标／诱饵，试图分散攻击者的注意力，使其不再攻陷真实目标；通过提高某些攻击动作的成本，试图阻止攻击者采取特定策略，从而增强防御能力。为在诱饵资源有限且受操作约束的情况下优化防御，我们将综合问题表述为一个双层优化问题：防御者设计系统时，会预先考虑攻击者在因欺骗而掌握系统虚假信息的情况下所作出的最优响应。尽管一般形式的双层优化问题是 NP 难的，我们证明，在某些假设条件下，该问题可以转化为约束优化问题。我们提出了一种算法，利用新颖的激励设计方法进行投影梯度上升，从而近似求解该约束优化问题。通过大量数值实验，我们验证了所提方法的有效性。

关键词

攻击图，欺骗，马尔可夫决策过程

ACM 引用格式：

Haoxiang Ma、Shuo Han、Nandi Leslie、Charles Kamhoua 和 Jie Fu。2023年。概率攻击图中的主动防御最优诱饵资源分配。载于第*22*届自主智能体与多智能体系统国际会议论文集（*AAMAS 2023*），伦敦，英国，*2023*年*5*月*29*日至*6*月*2*日，IFAAMAS，共8页。

1 引言

主动防御是指一类防御机制，防御者可借此检测正在进行的攻击、利用欺骗分散攻击者的注意力，或通过随机化增加攻击系统的难度。本文提出了一种数学框架和求解方法，用于综合具备欺骗能力的主动防御系统。

我们首先使用概率攻击图来描述攻击规划问题。概率攻击图可视为一种马尔可夫决策过程（MDP），其中包含一组攻击目标状态。攻击图（AG）[7] 可用于对计算机网络建模。它们广泛应用于网络安全领域，用于识别能够防止所有已知攻击所需的最小漏洞/传感器子集[13, 16]。概率攻击图引入了攻击动作结果的不确定性，以刻画随机环境中的动作失败。例如，[5, 6], 攻击图中的概率转移反映了由基于网络的随机化所产生的不确定性。在概率攻击图建模框架下，我们研究如何将诱饵资源分配为虚假目标，以分散攻击者的注意力，使其转而攻击虚假目标；以及如何调整攻击动作成本，以阻止攻击者到达真实目标。

诱饵资源分配与动作成本调整的联合设计可建模为一个双层优化问题，该问题通常是NP难的 [2]。在预先定义潜在诱饵状态、且防御者只需为诱饵分配资源/奖励的假设下，我们证明，双层优化问题可以等价地表示为一个约束优化问题。为了通过投影梯度上升高效地求解该约束优化问题，我们建立了两个重要联系：首先，我们证明，防御者期望的攻击策略投影到可实现攻击策略空间的步骤可以通过逆强化学习（IRL）[21] 完成。从本质上说，逆强化学习通过塑造攻击者的感知奖励，使理性攻击者模仿防御者选定的策略。其次，梯度上升步骤可以通过策略改进来完成；策略改进是策略迭代中用于最大化防御者总奖励的一个子程序。投影梯度上升能够保证收敛到该非凸约束优化问题的一个（局部）最优解。

相关工作。 本文研究的主动防御策略合成与斯塔克尔伯格安全博弈（SSG）（综述见 [18]）及其通过双层优化求解的方法密切相关。在 SSG 中，防御者需要利用有限资源保护一组目标，而攻击者则根据其掌握的防御者策略信息选择最优攻击策略。在 [12], 中，作者研究了

第*22*届自主智能体与多智能体系统国际会议论文集，*A. Ricci*、*W. Yeoh*、*N. Agmon*、*B. An*（编），*2023*年*5*月*29*日至*6*月*2*日，伦敦，英国。© 2023年国际自主智能体与多智能体系统基金会（www.ifaamas.org）。保留所有权利。

安全对策分配，并利用攻击图根据已分配的资源评估网络安全性。然而，SSG 未考虑欺骗所引入的非对称信息。在 [20], 中，作者引入奖励塑造，以激励智能体遵循目标策略。然而，在我们的设定中，目标策略可能不可行，因为防御者旨在诱使攻击者到达虚假目标，而攻击者可能不会刻意避开真实目标。

欺骗会向攻击者提供错误或不完整的信息。在 [19], 中，作者提出了一种安全博弈，用于分配有限的诱饵资源，以向网络攻击者掩饰网络配置。这种基于诱饵的欺骗会操纵对手对收益矩阵的认知。在 [1], 中，作者研究了确定性攻击图中的蜜罐分配，并利用极小极大定理确定最优分配策略。在 [10],中，作者研究了可由防御者利用欺骗性和防护性资源进行修改的有向无环攻击图。他们提出一种基于混合整数线性规划（MILP）的算法，以确定图中欺骗性和防护性资源的分配。在 [3], 中，他们利用蜜罐加固网络，使攻击者无法区分真实目标和虚假目标。在 [11], 中，作者在攻击图中设置虚假边以阻止攻击者，并采用 MILP 求解最优方案。

与现有工作相比，我们的工作做出了以下贡献：首先，我们不对攻击图的结构作任何假设，而是研究概率攻击图，而非确定性攻击图。由于攻击者可以在概率攻击图中采用随机化策略，因此无法构建收益矩阵并应用极小极大定理来分配诱饵资源。其次，我们同时考虑有限诱饵资源的分配和攻击动作成本的调整，并分析攻击者在欺骗所造成的虚假信息影响下的最优响应。第三，我们提出了一种受激励设计启发的高效算法，用于综合防御策略。在假设攻击者是理性的，且无法区分诱饵与真实目标的前提下，我们通过调整动作奖励并合理分配诱饵资源，证明可以塑造攻击者的行为，使被误导的攻击者受到激励，采取能够最大化防御者奖励的攻击策略。最后，我们在不同规模的问题上测试了该方法的可扩展性。

2 预备知识与问题形式化

记号。 令$\mathbf{R}$ 表示实数集，$\mathbf{R}^n$ 表示实数 n向量集。令$\mathbf{R}^n_{>0}$（分别为$\mathbf{R}^n_{<0}$）表示正（分别为负）实数 n向量集。我们用$\mathbf{1}$ 表示全 1 向量。给定向量 $z \in \mathbf{R}^n$，令 z_i 表示其第 i个分量。给定有限集 Z，其上的概率分布集合记为 $\mathrm{Dist}(Z)$。给定 $d \in \mathrm{Dist}(Z)$，d的支撑集记为 $\mathrm{Supp}(d) = \{z \in Z \mid d(z) > 0\}$。令 I_B 为示性函数，即若 $x \in B$，则$I_B(x) = 1$；否则 $I_B(x) = 0$。

我们研究配备主动防御（将在后文给出正式定义）的系统中，防御者（玩家 1，代词为 she/her）与攻击者（玩家 2，代词为 he/him/his）之间的对抗交互，

首先，我们介绍一种称为概率攻击图的形式化模型，用以描述攻击者如何计划实现攻击目标。接着，我们介绍带有欺骗的主动防御对策。

攻击规划问题。 攻击规划问题建模为一个概率攻击图，

$$M = (S, A, P, \nu, \gamma, F, R_2),$$

其中 S 是状态集（攻击图中的节点），A 是攻击动作集，$P: S \times A \to \mathrm{Dist}(S)$是概率转移函数，使得 $P(s'|s,a)$表示在状态 s采取动作 a后到达状态 s'的概率；$\nu \in \mathrm{Dist}(S)$是初始状态分布，$\gamma \in (0,1]$ 是折扣因子。攻击者的目标由目标状态集 F和目标奖励函数 R_2：$F \times A \to \mathrm{R}_{\geq 0}$描述，该函数为每个状态-动作对（$s,a$）（其中 $s \in F$且 $a \in A$）赋予一个非负值，表示攻击者到达该目标所获得的奖励。通过定义 $R_2(s,a) = 0$（对于任意 $s \in S \setminus F$，$a \in A$），可以将奖励函数扩展到整个状态空间。为描述攻击的终止，我们引入唯一的汇状态$s\mathrm{sink} \in S \setminus F$，使得对于所有 $a \in A$，$P(s\mathrm{sink}|s\mathrm{sink}\ a) = 1$；并且对于任意目标 $s \in F$和 $a \in A$，$P(s\mathrm{sink}|s,a) = 1$。

概率攻击图刻画了网络安全中遇到的目标导向攻击 [8, 14], , 在这类攻击中，攻击者通过到达目标状态来攻陷某些关键网络主机。概率攻击图 [10, 17] 利用概率转移函数描述攻击动作的不确定结果，并推广了确定性攻击图 [7]。

攻击者从初始状态 $S_0 \sim \nu$出发，旨在最大化其折扣总奖励。随机有限记忆攻击策略是一个函数 $\pi: S^* \to \mathrm{Dist}(A)$，它将有限运行$\rho \in S^*$ 映射为行动上的分布 $\pi(\rho)$。如果策略仅取决于最近的状态，即 $\pi: S \to \mathrm{Dist}(A)$，则称其为马尔可夫策略。我们仅考虑马尔可夫策略，因为只需在马尔可夫策略中搜索即可找到最优攻击策略。

设(Ω, F)为$(S_0, A_0, (S_t, A_t)_{t>1})$的典范样本空间，其配备 Borel σ-代数$\mathrm{F} = \mathrm{B}(\Omega)$和 $\Omega = S \times A \times \prod_{t=1}^{\infty}(S \times A)$。由马尔可夫策略$\pi$诱导的$(\Omega, \mathrm{F})$上的概率测度$\mathrm{Pr}^\pi$满足：$\mathrm{Pr}^\pi(S_0 = s) = \mu_0(s)$，$\mathrm{Pr}^\pi(A_0 = a \mid S_0 = s) = \pi(s,a)$，$\mathrm{Pr}^\pi(S_t = s' \mid (S_k, A_k)_{k<t}) = P(s' \mid S_k, A_k)$，以及$\mathrm{Pr}^\pi(A_t = a \mid (S_k, A_k)_{k<t}, S_t) = \pi(S_t, a)$。

给定马尔可夫策略 π: $S \to \mathrm{Dist}(A)$，我们将攻击者的价值函数 V_2^π: $S \to \mathbf{R}$ 定义为

$$V_2^\pi(s) = \mathbf{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_2(S_k, A_k) | S_0 = s],$$

其中，$\mathbf{E}_\pi$ 表示在概率测度 $\mathbf{Pr}^\pi$下的期望。

欺骗式主动防御。 我们假设防御者知道攻击者的目标，该目标由元组 F, R_2给出，即目标状态和目标奖励函数。防御者可采用以下主动防御机制：

- 通过欺骗进行防御：防御者采用一种称为“揭露伪装”的欺骗方法。具体而言，防御者在马尔可夫决策过程 M 中有一组状态 $D \subset S \setminus F$，可以将其设为虚假目标状态，并赋予虚假目标奖励 $\vec{y} \in \mathbf{R}^{|D|}$。

攻击者无法区分真实目标 F 和虚假目标 D。

- 通过状态-动作奖励修改进行防御：防御者拥有一组 $W \subset (S\backslash(F\cup D))\times A$ 马尔可夫决策过程（MDP）M中的状态-动作对，其奖励可以被修改。一旦状态-动作对（s, a）的奖励被修改，攻击者的感知奖励 $R_2(s, a) < 0$，即状态 s 下攻击动作 a的代价为 $-R_2(s, a)$。

防御者可以决定如何分配其诱饵资源，以及有限的状态-动作奖励修改能力。

定义 1（约束下的诱饵分配）。 防御者的诱饵分配设计是一个非负实值向量 $\vec{y} \in \mathbf{R}_{\geq 0}^{|S|}$，满足 $\vec{y}(s) = 0$对任意 $s \in S\backslash D$均成立，并受约束$\mathbf{1}^{\mathsf{T}}\vec{y} \leq h$，其中 $h \geq 0$为某个值。给定诱饵分配 $\vec{y}$，攻击者的感知奖励函数定义为

$$R_2^{\vec{y}}(s,a) = \begin{cases} \vec{y}(s) & \text{if } \vec{y}(s) > 0, \\ R_2(s,a) & \text{if } \vec{y}(s) = 0. \end{cases}$$

定义 2（动作奖励修改）。 给定集合 $W \subset (S\setminus(F\cup D)) \times A$, 防御者的动作奖励修改是一个非正的奖励值向量 $\vec{x} \in \mathbf{R}_{\leq 0}^{|S\times A|}$，满足 $\vec{x}(s, a) = 0$，对任意 $(s, a) \notin W$均成立。给定一个动作奖励修改 $\vec{x}$，攻击者的感知奖励函数定义如下

$$R_2^{\vec{x}}(s,a) = \begin{cases} \vec{x}(s,a) & \text{if } \vec{x}(s,a) < 0, \\ R_2(s,a) & \text{if } \vec{x}(s,a) = 0. \end{cases}$$

请注意，防御者不会考虑修改（虚假或真实）目标状态 $F \cup D$ 的状态-动作奖励，因为一旦到达 $F \cup D$ 中的某个状态，攻击就会终止。

定义 3. 防御者的主动防御策略是一个包含动作奖励修改 $\vec{x}$ 和诱饵分配设计 $\vec{y}$的元组 $(\vec{x}, \vec{y})$。

由于动作奖励修改独立于诱饵分配设计，因此给定防御者策略 $(\vec{x}, \vec{y})$ 时的奖励函数是 $R_2^{\vec{x}}$和 $R_2^{\vec{y}}$的组合，故予以省略。

假设 1. 攻击过程在以下两种情况下终止：攻击成功，即攻击者到达目标$s \in F$；或者攻击被拦截，即攻击者到达分配了诱饵的状态。

我们的问题可以非正式地表述如下。

问题 1. 在上述攻击规划场景中，确定防御者的策略，分配诱饵资源并修改动作奖励，从而在攻击者采取最优响应的情况下，最大化攻击者到达虚假目标的概率。

3 主要结果

在本节中，我们首先定义固定动作奖励修改和诱饵资源分配（$\vec{x}$, $\vec{y}$）下攻击者的感知规划问题。然后，我们说明主动防御的设计可以表述为一个双层优化问题。我们研究所构建的双层优化问题的特殊性质，以开发一种基于优化的方法来综合主动防御策略。

3.1 A双层优化公式

防御者的策略会以下列方式改变攻击者对攻击规划问题的感知:

定义 4（带有修改后奖励和诱饵的感知攻击规划问题）。 设动作奖励修改为 $\vec{x}$，诱饵分配为 $\vec{y}$，攻击者原始规划问题为 $M = (S, A, P, v, \gamma, F, R_2)$，则攻击者的感知规划问题由以下具有终止状态的马尔可夫决策过程定义:

$$M(\vec{x}, \vec{y}) = (S, A, P^{\vec{y}}, v, \gamma, F \cup D^{\vec{y}}, R_2^{\vec{x},\vec{y}}),$$

其中， S, A, v, γ与 M中的相同， $D^{\vec{y}} = \{s \in D \mid \vec{y}(s) \neq 0\}$为诱饵目标状态，且为吸收状态。转移函数 $P^{\vec{y}}$ 由原始转移函数 P 得到，仅将 $D^{\vec{y}}$中的所有状态设为吸收状态。奖励 $R^{\vec{x}}{}_{,\ 2}{}^{\vec{y}}$根据定义 1 和定义 2 定义。

攻击者的感知价值为

$$V_2^{\pi}(v; \vec{x}, \vec{y}) = \mathbf{E}_{\pi}\Big[\sum_{k=0}^{\infty} \gamma^k R_2^{\vec{x},\vec{y}}(S_k, A_k) \mid S_0 \sim v\Big],$$

其中，$\mathbf{E}_{\pi}$是由 π从马尔可夫决策过程 $M(\vec{x}, \vec{y})$诱导的概率测度$\mathbf{Pr}^{\pi}$下的期望。

防御者的欺骗目标由奖励函数$R_1^{y} : S \to \mathbf{R}$给出，定义如下:

$$R_1^{\vec{y}}(s) = \begin{cases} 1 & \text{if } s \in D^{\vec{y}}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

给定概率测度$\mathbf{Pr}^{\pi}$，将防御者的价值记为

$$V_1^{\pi}(v; \vec{y}) = \mathbf{E}_{\pi}\Big[\sum_{k=0}^{\infty} \gamma^k R_1(S_k) \mid S_0 \sim v\Big].$$

根据此奖励定义，价值 $V_1^{\pi}(v; \vec{y})$表示攻击者到达 $D^{\vec{y}}$中的虚假目标的概率。

为形式化欺骗目标，我们引入如下防御者优选攻击策略的概念。

定义 5（防御者的优选攻击策略）。 给定攻击者的感知规划问题 $M(\vec{x}, \vec{y})$，其中（$\vec{x}, \vec{y}$）是固定的主动防御策略。设 π和 π'是两种对攻击者而言价值相同的攻击策略，即 $V_2^{\pi}($ v $=$ $V_2^{\pi'}$ v $\vec{x}$ $\vec{y}$ π π $; \vec{x}, \vec{y})$ $(;,)$。当且仅当以下条件成立时，防御者严格偏好策略 而非，

$$V_1^{\pi}(v; \vec{y}) > V_1^{\pi'}(v; \vec{y}).$$

换言之，如果两种策略对攻击者同样有利，那么防御者会偏好到达虚假目标的概率更高的策略。

那么，综合最优主动防御策略（$\vec{x}$, $\vec{y}$）的问题可数学化表述为

问题 2。

$$\begin{aligned} \max_{\vec{x}\in X, \vec{y}\in Y} \quad & V_1^{\pi^*}(v; \vec{y}) \\ \text{s.t.} \quad & \pi^* \in \operatorname*{argmax}_{\pi} V_2^{\pi}(v; \vec{x}, \vec{y}). \end{aligned}$$

其中，$X = \mathbf{R}_{\leq 0}^{|W|}$ 和 $Y = \{\vec{y} \mid \forall s \in S \setminus D,\ \vec{y}(s) = 0$ 和 $\mathbf{1}^\top \vec{y} \leq h\}$ 分别是变量 $\vec{x}$ 和 $\vec{y}$ 的取值范围。

换言之，防御者选择 $(\vec{x}, \vec{y})$，使得攻击者在其感知攻击规划问题中的最优响应成为防御者最偏好的攻击策略，因为该策略可最大化防御者的价值。

3.2 转化为约束优化问题

众所周知，双层优化问题是强NP难的[4]。然而，在某些条件下，可以证明双层优化问题等价于一个约束优化问题。

令 $\Pi(\vec{x}, \vec{y}) = \{\pi \mid V_2^\pi(v; \vec{x}, \vec{y}) = \max_\pi V_2^\pi(v; \vec{x}, \vec{y})\}$，该集合是在给定变量 $\vec{x}$ 和 $\vec{y}$ 的情况下，攻击者的感知规划问题中的最优策略集合。于是，双层优化问题等价地写成以下约束优化问题：

$$\begin{aligned} \max_{\pi^*, \vec{x} \in X, \vec{y} \in Y} \quad & V_1^{\pi^*}(v; \vec{y}) \\ \text{s.t.} \quad & \pi^* \in \Pi(\vec{x}, \vec{y}). \end{aligned} \tag{2}$$

这反过来等价于

$$\begin{aligned} \max_{\pi^*} \quad & V_1^{\pi^*}(v; \vec{y}) \\ \text{s.t.} \quad & \pi^* \in \bigcup_{\vec{x} \in X, \vec{y} \in Y} \Pi(\vec{x}, \vec{y}). \end{aligned} \tag{3}$$

这里，约束意味着，在给定 $\vec{x}$、$\vec{y}$ 的所有可能取值时，攻击者的响应 π^* 可以从最优攻击策略集合中选取。

根据防御者价值函数的定义，可以注意到，$V_1^\pi(v; \vec{y})$ 不取决于 $\vec{y}$ 的确切取值，而只取决于对于每个状态 $s \in D$，$\vec{y}(s) > 0$ 是否成立。形式化地，

引理1。对于任意 $\vec{y}_1$、$\vec{y}_2 \in Y$，若 $\vec{y}_1(s) = 0 \implies \vec{y}_2(s) = 0$，反之亦然，则 $V_1^\pi(v; \vec{y}_1) = V_1^\pi(v; \vec{y}_2)$。

证明。给定两个不同的向量 $\vec{y}_1$ 和 $\vec{y}_2$，我们可以分别构造两个马尔可夫决策过程：$M_1 := M(\vec{x}, \vec{y}_1) = (S, A, P^{\vec{y}_1}, v, \gamma, F, R_1)$ 和 $M_2 := M(\vec{x}, \vec{y}_2) = (S, A, P^{\vec{y}_2}, v, \gamma, F, R_1)$。

如果 $\vec{y}_1(s) = 0$ 当且仅当 $\vec{y}_2(s) = 0$，那么 M_1 的转移函数 $P^{\vec{y}_1}$ 与 M_2 的转移函数 $P^{\vec{y}_2}$ 相同（见定义4）。此外，防御者奖励函数 $R_{\vec{y}_1}$ 也等于 $R_{\vec{y}_2}$（见（1）），由于转移动态和奖励相同，因此我们有 $V_1^\pi(v; \vec{y}_1) = V_1^\pi(v; \vec{y}_2)$。□

接下来，为了消除 $V_1^\pi(v; \vec{y})$ 对 $\vec{y}$ 的依赖，我们作出如下假设：

假设2。给定诱饵分配所在的状态集合 $D^{\vec{y}} = \{s \in D \mid \vec{y}(s) \neq 0\}$。

在这一假设下，我们只需假设给定集合 D 中的所有状态都必须分配非零的诱饵资源。也就是说 $D^{\vec{y}} = D$。

这一假设进一步将防御者的综合问题转化为一个约束优化问题。

$$\begin{aligned} \max_{\pi^*} \quad & V_1^{\pi^*}(v) \\ \text{s.t.} \quad & \pi^* \in \overline{\Pi} \triangleq \bigcup_{\vec{y} \in Y, \vec{x} \in X} \Pi(\vec{x}, \vec{y}), \\ & \vec{y}(s) > 0, \forall s \in D. \end{aligned} \tag{4}$$

由于上述问题是一个标准的约束优化问题，因此可以使用投影梯度法获得局部最优解：

$$\pi^{k+1} = \mathrm{proj}_{\overline{\Pi}}(\pi^k + \eta \nabla V_1^{\pi^k}(v)).$$

其中，$\mathrm{proj}_\Pi(\pi)$ 表示将策略 π 投影到策略空间 Π 上，而 η 是步长。

3.3 将逆强化学习与 ProjectGradientAscent 联系起来

执行投影梯度上升的一个关键步骤是，对任意策略 $\hat{}\pi$，计算其投影 $\mathrm{proj}_\Pi(\hat{}\pi)$。然而，这并非易事，因为集合 $\bar{\Pi}$ 包含一组攻击策略，其中每个策略都对应于向量 $(\vec{x}, \vec{y})$ 的一种取值。因此，$\bar{\Pi}$ 没有紧凑的表示形式。接下来，我们提出一种用于计算投影的新算法。

首先，根据投影的定义，可以注意到，这一投影步骤等价于求解以下优化问题：

$$\begin{aligned} \min_{\pi} \quad & \mathbf{D}(\pi, \hat{\pi}) \\ \text{s.t.} \quad & \pi \in \overline{\Pi}, \\ & \vec{y}(s) > 0; \forall s \in D. \end{aligned} \tag{5}$$

其中，$\mathbf{D}(\pi, \hat{}\pi)$ 表示两个策略 $\pi, \hat{}\pi$ 之间的距离。

距离函数$\mathbf{D}$可选为策略诱导的马尔可夫链之间的Kullback–Leibler（KL）散度，定义如下。

定义 6. 给定一个马尔可夫决策过程 $M = (S, A, P, v)$和两个马尔可夫策略 π_1, π_2。设 $M_{\pi_1} = (S, P_1, v)$和 $M_{\pi_2} = (S, P_2, v)$分别为在 M下由 π_1和 π_2诱导的两个马尔可夫链。KL散度$\mathbf{D}_{\mathrm{KL}}(M_{\pi_1} \| M_{\pi_2})$（从 M_{π_2}到 M_{π_1}的相对熵）定义为

$$\mathbf{D}_{\mathrm{KL}}(M_{\pi_1} \| M_{\pi_2}) = \sum_{\rho \in S^*} \mathbf{Pr}_1(\rho) \log \frac{\mathbf{Pr}_1(\rho)}{\mathbf{Pr}_2(\rho)},$$

其中，$\mathbf{Pr}_i(\rho)$是对于 $i = 1, 2$，马尔可夫链 M_{π_i}中路径 ρ的概率。

式（5）中的 KL 散度可表示为

$$\begin{aligned} \mathbf{D}_{\mathrm{KL}}(M_\pi(\vec{x}, \vec{y}) \| M_{\hat{\pi}}(\vec{x}, \vec{y})) &= \sum_\rho \widehat{\mathbf{Pr}}(\rho) \log \frac{\widehat{\mathbf{Pr}}(\rho)}{\mathbf{Pr}(\rho | \vec{x}, \vec{y})} \\ &= \sum_\rho \widehat{\mathbf{Pr}}(\rho) \log \widehat{\mathbf{Pr}}(\rho) - \sum_\rho \widehat{\mathbf{Pr}}(\rho) \log \mathbf{Pr}(\rho | \vec{x}, \vec{y}), \end{aligned} \tag{6}$$

其中，$\widehat{\mathbf{Pr}}(\rho)$ 是马尔可夫链$M_{\hat{\pi}}(\vec{x}, \vec{y})$ 中路径 ρ 的概率，而 $\mathbf{Pr}(\rho | \vec{y})$ 是由策略 π 诱导的马尔可夫链 $M_\pi(\vec{x}, \vec{y})$ 中路径 ρ 的概率。

由于在 $\hat{}\pi$ 固定时，式 (6) 求和中的第一项是常数，因此KL散度最小化问题等价于以下最大化问题：

$$\max_{\vec{x}\in X,\vec{y}\in Y} \quad \sum_{\rho} \widehat{\mathbf{Pr}}(\rho) \log \mathbf{Pr}(\rho|\vec{x},\vec{y}) \tag{7}$$

$$\text{s.t.} \quad \vec{y}(s) > 0; \forall s \in D, \tag{8}$$

$$\mathbf{1}^{\mathsf{T}}\vec{y} \leq h. \tag{9}$$

问题 (7) 可通过扩展最大熵（MAXENT）逆强化学习算法 [21], 求解，该算法最初是在没有约束的情况下开发的。众所周知，逆强化学习旨在根据专家演示推断奖励函数，使生成这些演示的专家策略达到最优。使用逆强化学习执行投影，其直观理解如下：目标是计算一对向量 $(\vec{x},\ \vec{y})$，以改变攻击者的感知奖励函数，使得攻击者在给定 $(\vec{x},\ \vec{y})$ 时的最优策略接近“专家策略” $\hat{}\pi$，同时满足 $\vec{y}$ 的约束。

为处理诱饵资源约束 (9)，我们使用对数障碍函数近似该约束，并通过基于梯度的数值优化计算最优解 $\vec{y}^*$。

考虑约束 $\mathbf{1}\tau\vec{y} \leq h$，我们实现障碍函数以近似不等式约束，并将优化问题改写为：

$$\max_{\vec{x},\vec{y}} \sum_{\rho} \widehat{\mathbf{Pr}}(\rho) \log \mathbf{Pr}(\rho|\vec{x},\vec{y}) + \frac{1}{t}\log(h - \mathbf{1}^{\mathsf{T}}\vec{y})$$

$$\text{subject to: } \vec{y}(s) = 0, \quad \forall s \in S \setminus D.$$

其中，t 是对数障碍函数的权重参数。在我们的实验中，t 固定为1000。

由于约束 $\vec{y}(s) = 0, \forall s \in S\backslash D$ 可以纳入决策变量 $\vec{y}$ 的定义域，因此我们可以使用梯度上升来获得使目标函数达到最大值的最优 $\vec{x}^*, \vec{y}^*$。具体而言，$\vec{x}$和 $\vec{y}$可以通过 $\vec{x}^{k+1} = \text{projx}(\vec{x}^k + \eta_x \nabla L(\vec{x},\vec{y}))$、$\vec{y}^{k+1} = \text{projy}(\vec{y}^k + \eta_y \nabla L(\vec{x},\vec{y}))$进行更新。

3.4 梯度上升步的策略改进

经过投影步骤得到策略 π^k 及其对应的向量（$\vec{x},\ \vec{y}$）后，我们旨在计算一次梯度上升，以提高目标函数的值

$$V_1^{k+1}(v) = V_1^k(v) + \nabla V_1^k(v),$$

其中，V_1^k（v）表示在第 k次迭代中，给定攻击策略 π^k 时评估得到的防御者的价值。

在这一步中，我们针对防御者奖励函数 R_1^y 执行策略改进步骤，该步骤相对于_进行；此时，该函数不再依赖于 $\vec{y}$，因为集合 $D^{\vec{y}}$ 被固定为常数集合 D。文献 [9, 15] 证明，策略改进是优化价值函数的一步牛顿更新。

具体而言，策略改进是计算

$$\tilde{\pi}^{k+1}(s,a) = \frac{\exp\left((R_1(s,a) + \gamma V_1^k(s'))/\tau\right)}{\sum_{a\in A}\exp\left((R_1(s,a) + \gamma V_1^k(s'))/\tau\right)},$$

迭代 $k+1$ 时的策略通过执行投影步骤 ((5)) 获得，其中 $\hat{}\pi \triangleq \tilde{\pi}_{k+1}$。

当 $|V_1^{k+1}$（v）$- V_1^k$（v）$| \leq \epsilon$时，迭代停止，其中 ϵ是人工设定的阈值。输出为一个元组（$\vec{x}^*,\ \vec{y}^*$），即（局部）最优主动防御策略。由于转换后的约束优化问题具有非凸约束集，我们在此只能得到一个局部最优主动防御策略。不过，我们可以从不同的初始策略开始，并选取其中最优的一个。此外，假设防御者在不考虑攻击者目标的情况下求解自身问题，则可以得到防御者目标的上界。我们可以选择目标函数最接近该上界的解。

备注1。 在我们的问题中，我们假设集合 D 已给定。如果集合 D 未给定，那么该问题就变成了组合问题。如果集合D 未给定，而是需要从候选状态集中确定，那么一种简单的方法是枚举所有可能的组合，计算每个子集对应的防御者的价值，并选出使防御者的价值最高的子集。研究该组合问题是否具有次模性或超模性会很有意思，但这超出了本文的研究范围。

总之，所提出的算法从初始策略 $\tilde{\pi}^0$开始，使用逆强化学习求出投影 π^0 及其对应的向量（$\vec{x}^0,\ \vec{y}^0$），这些向量共同构成攻击者的感知奖励函数，使得 π^0 为最优策略。随后，通过策略改进将 π^0 更新为 $\tilde{\pi}^1$。通过交替进行投影和策略改进，该过程持续进行，直到满足停止准则 $|V_1^{k+1}$（v）$- V_1^k$（v）$| \leq \epsilon$。

4 实验

我们通过两组示例展示所提出的方法：一组是概率攻击图，另一组是在随机网格世界中形式化的攻击规划问题。所有案例研究均使用配备 Intel i7-11700K 处理器和 32GB内存的工作站。

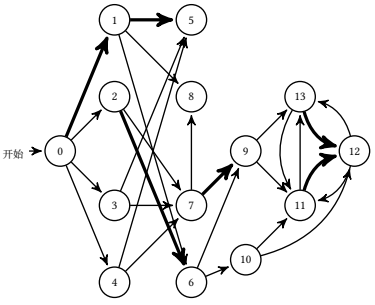


图1：概率攻击图。

图1展示了一个概率攻击图，其目标集为 $F = \{10\}$，动作集为 $\{a,\ b,\ c,\ d\}$。为清晰起见，图中仅展示动作 a 对应的转移，其中粗箭头（相应地，细箭头）表示较高（相应地，较低）的转移概率。例如，$P(0,\ a) = \{1:0.7,\ 2:0.1,\ 3:0.1,\ 4:0.1\}$ [1]。

考虑诱饵状态集合$\mathrm{D} = \{11, 13\}$。回顾一下，防御者奖励函数为 $R_1(s) = 1$，对所有 $s \in D$ 均成立。假设没有资源被

[1]确切的转移函数见补充文件。

分配给 D，且 D 中的所有状态都是汇状态，那么攻击者从初始状态 0 到达目标集 F 的概率为 60。33%。与此同时，防御者的期望值为 0。149。也就是说，攻击者有 14。9% 的概率到达 D 中的一个诱饵状态，从而使攻击终止。

给定有限资源 $1\tau\vec{y} \leq 3$，诱饵资源分配结果为 $\vec{y}(11) = \vec{y}(13) = 1.313$。根据给定的诱饵资源分配，攻击者到达目标集 F的概率为 8.63%，且在初始状态 0 下，防御者的期望奖励为 0.653。因此，通过分配资源设置诱饵来吸引攻击者，防御者显著降低了攻击者到达目标状态的概率（降低 85%），并使防御者的价值提高了 3.38 倍。

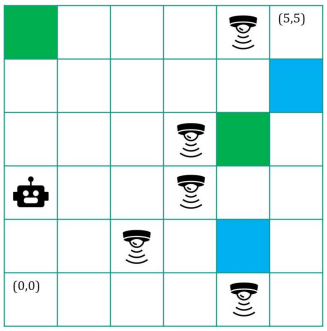


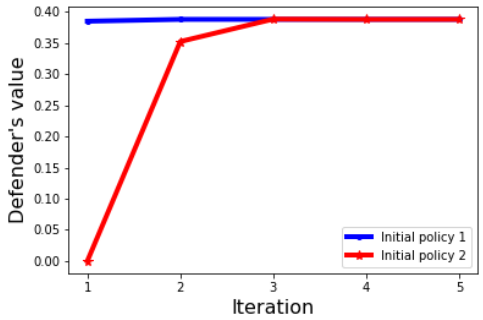
图 2： 6×6 **网格世界示例。**

接下来，我们考虑图 2 所示的随机 6×6 网格世界中的机器人运动规划问题。攻击者/机器人旨在到达一组目标状态，同时避免被防御者检测到。攻击者可以向四个罗盘方向移动。给定一个动作，例如“N”，攻击者以 $1-2\alpha$的概率进入预期单元格，并以 α 的概率进入相邻单元格，即西侧和东侧的单元格。在我们的实验中，α取值为 0.1。状态（i, j）表示第 i 行、第 j 列的单元格。

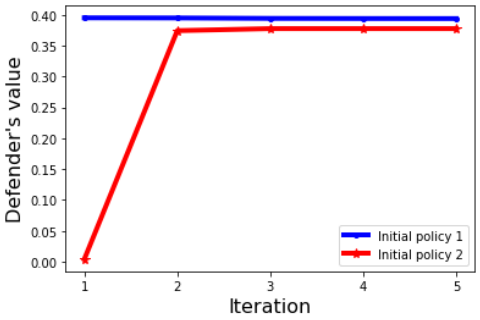
防御者部署了图 2 所示的传感器，用于检测攻击者是否出现。因此，一旦攻击者进入传感器状态，其任务就会失败。诱饵集 D 由蓝色单元格表示，目标集 F 由绿色单元格表示。

初始状态位于 (2，0)，图中用机器人表示。我们测试以下三种场景：不分配诱饵资源、仅分配诱饵资源，以及同时分配诱饵资源并修改奖励。结果如表 1 所示。如果不将资源分配给诱饵，攻击者有 98。98% 的概率在避开传感器状态的同时到达目标集 F，而防御者的期望值为 3.56×10^{-6}。当防御者可以将总预算为 4 的资源分配给诱饵时，诱饵资源分配结果为 $\vec{y}((1，4)) = 2.016$，$\vec{y}((4，5)) = 1.826$。由于使用对数障碍函数来施加约束，当资源接近上界时，对数障碍函数会在梯度上升中产生很大的惩罚，因此防御者不会用尽所有诱饵资源。

在给定的资源分配下，攻击者到达目标集 F 的概率为 9.9%，防御者在初始状态的期望值为 0.3877。在诱饵资源分配中



(a) 使用不同的初始策略进行收敛。



(b) 通过诱饵资源分配和动作奖励修改实现收敛。

图3：给定 $D = \{(1,4), (4,5)\}$ **时，** 6×6 **网格世界示例中防御者价值的收敛趋势。**

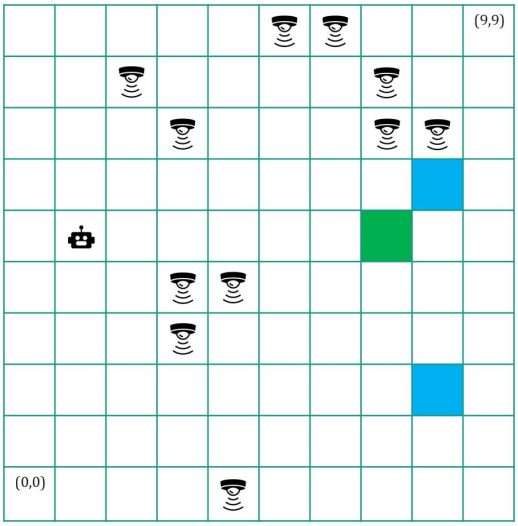


图4： 10×10 **网格世界示例。**

在动作奖励修改实验中，允许防御者修改状态 (4，4) 下的所有动作奖励，以及状态 (4，0)、(4，1) 和 (4，2) 下动作‘N’的奖励。结果表明，防御者向诱饵 (1，4) 分配了 1.938，向诱饵 (2，5) 分配了 1.734。同时，状态 (4，0) 下动作‘N’的奖励被修改为 -1，而状态 (4，1) 下相同动作的奖励被

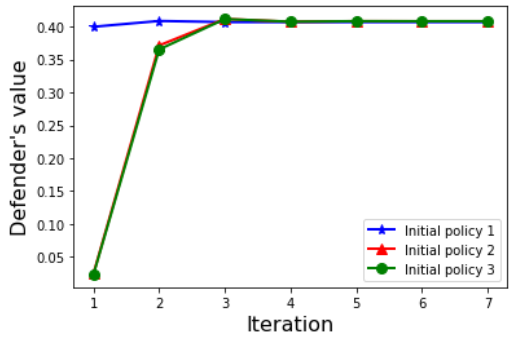


图5： 10×10 **网格世界中防御者的价值收敛趋势。**

修改为 -0.94，状态 (4，2) 下动作“N”的奖励被修改为 -0.904；防御者还会将状态 (4，4) 下动作“W”“S”“N”的奖励修改为 -1。将诱饵资源分配结果与诱饵资源分配和动作奖励修改结果进行比较，可以发现，允许修改动作奖励后，防御者降低了攻击者到达目标的概率（降低 13.13%）。与此同时，防御者的期望值提高了 1.62%。

值得注意的是，由于优化问题存在非线性，在不同初始条件下，算法会收敛到不同的解，如图 3a 和图 3a 所示。图中，初始策略 1 的生成方式是假设攻击者到达诱饵时获得 1 的奖励，到达目标状态时获得 0 的奖励。这对防御者目标而言是理想的，但对该优化问题来说不可行，因为在攻击者的感知规划问题中，无论分配给诱饵多少资源，到达真实目标始终会获得 1 的奖励。初始策略 2 是随机生成的。在本次实验中，不同初始策略对应的目标函数值相近。

为了测试诱饵集 D 对结果的影响，我们将诱饵的位置重新分配到 $\{(0,2),(5,3)\}$。结果如表 2 所示。基于新的配置，如果我们不分配诱饵资源，攻击者以 98.97% 的概率到达目标集，且初始状态下防御者的价值为 7.61×10^{-8}。如果防御者可以向诱饵分配资源，我们的方法得到 $\vec{y}((0,2)) = 1.141$ 和 $\vec{y}((5,3)) = 1.0$。攻击者到达目标集的概率为 3.99%，防御者的期望值为 0.6991。如果允许防御者修改与前一个示例中相同的一组状态-动作奖励，我们的算法得到 $\vec{y}((0,2)) = 0.985$ 和 $\vec{y}((5,3)) = 1.068$。$(4,0)$ 处的动作“N”的奖励修改为 -1，$(4,1)$ 处的相同动作修改为 -0.85，$(4,2)$ 处动作“N”的奖励修改为 -0.081；防御者还会将 $(4,4)$ 处所有动作的奖励修改为 -1。在此配置下，攻击者到达目标集的概率为 0.286%（与仅分配诱饵资源相比降低了 93%），防御者的期望值为 0.7301（与仅分配诱饵资源相比提高了 4.4%）。通过更改集合 D 的配置，我们展示了集合 D 的配置会影响攻击者到达目标集的概率以及防御者的期望值：第二个集合 $D = \{(0,2),(5,3)\}$ 似乎优于第一个集合 $D = \{(1,4),(4,5)\}$。

表 1：给定 $D = \{(1,4),(4,5)\}$ **时，** 6×6 **网格世界中的实验结果。**

	无诱饵	仅诱饵	诱饵和动作奖励
攻击者的价值	98.98%	9.9%	8.6%
防御者的价值	3.56×10^{-6}	0.3877	0.394

表 2： 6×6 **网格世界在** $D = \{(0,2),(5,3)\}$**中的实验结果。**

	无诱饵	仅有诱饵	诱饵和动作奖励
攻击者的价值	98.97%	3.99%	0.286%
防御者的价值	7.61×10^{-8}	0.6991	0.7301

接下来，为了测试可扩展性，我们将网格世界的规模扩大到 10×10，如图4所示。在大型网格世界示例中，我们只进行诱饵资源分配。传感器、诱饵集和目标集采用与 6×6 网格世界相同的记号表示。防御者的奖励函数仍为 $R_1(s) = 1$，适用于所有 $s \in D$。假设初始状态位于(5,1)。当防御者不分配诱饵资源时，攻击者到达目标的概率为82.43%，且防御者在初始状态下的期望值为0.0024。当允许防御者为诱饵分配资源时，我们的算法得出 $\vec{y}((2,8)) = 1.350, \vec{y}((6,8)) = 1.235$。在给定诱饵资源的情况下，攻击者到达目标的概率降至5.1%（降低94%），而防御者在初始状态下的期望值升至0.4034。我们还使用不同的初始策略测试防御者的收敛趋势，如图5所示。初始策略1的生成方式与 6×6 示例中的初始策略1类似。初始策略2和3是随机生成的策略。从图5可以看出，不同的初始策略会使目标函数收敛到相近的值。考虑到算法的可扩展性，10×10 网格世界示例的计算时间为185.94秒，而 6×6 示例的计算时间为22.51秒。运行时间表明，我们的算法可以扩展到中等规模的问题。需要指出的是，影响运行时间的不仅是状态空间的大小，还有诱饵的选择和数量。

5 结论与未来工作 K

我们提出了一种用于主动防御系统中诱饵分配和奖励修改的数学框架与算法。我们的技术方法可应用于许多安全关键系统，在这些系统中，概率攻击图是根据系统中的已知漏洞构建的。该公式及其求解方法可扩展至广泛的对抗性交互场景，在这些场景中可以部署欺骗式主动防御。未来，我们将考虑更复杂的攻击与防御目标，并研究在攻击者目标或能力存在不确定性时如何分配诱饵。除了本文研究的“揭露伪装”之外，我们还将研究如何通过操纵攻击者攻陷真实目标所获得的感知奖励来“隐藏真相”。

参考文献

[1] A. H. Anwar、C. Kamhoua 和 N. Leslie。2020年。网络欺骗博弈中的攻击图蜜罐配置。载于*2020*年国际计算、网络与通信会议 *(ICNC)*。502–506。[2] Stephan Dempe 和 Alain Zemkoho。2020年。双层优化。施普林格。[3] Karel Durkota、Viliam Lis`y、Branislav Bošansk`y 和 Christopher Kiekintveld。2015年。使用攻击图博弈进行最优网络安全加固。载于第二十四届国际人工智能联合会议。[4] Pierre Hansen、Brigitte Jaumard 和 Gilles Savard。1992年。线性双层规划的新分支定界规则。《*SIAM*科学与统计计算杂志》13，5（1992年），1194–1217。[5] Jin Hong 和 Dong-Seong Kim。2012年。HARMs：用于网络安全分析的分层攻击表示模型。载于澳大利亚信息安全管理会议。SRI安全研究所，埃迪斯科文大学，珀斯，西澳大利亚州，9。[6] Jin B. Hong 和 Dong Seong Kim。2016年。使用安全模型评估移动目标防御的有效性。《*IEEE*可信与安全计算汇刊》13，2（2016年3月），163–177。[7] S. Jha、O. Sheyner 和 J. Wing。2002年。攻击图的两种形式化分析。载于第*15*届*IEEE*计算机安全基础研讨会论文集。*CSFW-15*。49–63。[8] Harjinder Singh Lallie、Kurt Debattista 和 Jay Bal。2020年。网络安全中攻击图与攻击树可视化语法综述。《计算机科学评论》35 2020 100219 10 1016 2019 100219（2月），。[9] https://doi.org/ 。/j.cosrev. 。Omid Madani。 。论策略迭代作为牛顿法及多项式策略迭代算法。载于第十八届全国人工智能会议（加拿大艾伯塔省埃德蒙顿）。美国人工智能协会，美国，273–278。[10] Stephanie Milani、Weiran Shen、Kevin S. Chan、Sridhar Venkatesan、Nandi O. Leslie、Charles Kamhoua 和 Fei Fang。2020年。运用欺骗之力的攻击图安全博弈。载于安全决策与博弈论（《计算机科学讲义》），Quanyan Zhu、John S. Baras、Radha Poovendran 和 Juntao Chen（编）。施普林格国际出版，Cham，147–167。https://doi.org/10.1007/978-3-030-64793-3_8[11] Stephanie Milani、Weiran Shen、Kevin S Chan、Sridhar Venkatesan、Nandi O Leslie、Charles Kamhoua 和 Fei Fang。2020年。运用欺骗之力基于攻击图的安全博弈。载于安全决策与博弈论国际会议。施普林格，147–167。[12] Thanh H. Nguyen、Mason Wright、Michael P. Wellman 和 Satinder Singh。2018年。多阶段攻击图安全博弈：启发式策略及实证博弈论分析。《安全与通信网络》2018年（2018年12月），1–28。[13] Steven Noel 和 Sushil Jajodia。2008年。利用攻击图优化入侵检测系统传感器部署和警报优先级排序。《网络与系统管理杂志》16卷，3期（2008年），259–275。[14] Steven Noel、Sushil Jajodia、Lingyu Wang 和 Anoop Singhal。2010年。利用攻击图测量网络安全风险。《下一代计算国际期刊》1卷，1期（2010年），135–147。[15] Martin L Puterman。2014年。马尔可夫决策过程：离散随机动态规划。约翰·威立父子公司。[16] Oleg Sheyner、Joshua Haines、Somesh Jha、Richard Lippmann 和 Jeannette M Wing。2002年。攻击图的自动生成与分析。载于*2002*年*IEEE*安全与隐私研讨会论文集。IEEE，273–284。[17] Anoop Singhal 和 Xinming Ou。2017年。利用概率攻击图对企业网络进行安全风险分析。载于网络安全度量。施普林格，53–73。[18] Arunesh SINHA、Fei FANG、Bo AN、Christopher KIEKINTVELD 和 Milind TAMBE。2018年。斯塔克尔伯格安全博弈：回顾十余年的成功。第二十七届国际人工智能联合会议论文集（*IJCAI-18*），瑞典斯德哥尔摩，*7*月*13*日至*19*日（2018年7月），5494–5501。[19] Omkar Thakoor、Milind Tambe、Phebe Vayanos、Haifeng Xu、Christopher Kiek- intveld 和 Fei Fang。2019年。用于策略性欺骗的网络伪装博弈。载于安全决策与博弈论（《计算机科学讲义》），Tansu Alpcan、Yevgeniy Vorobeychik、John S. Baras 和 György Dán（编）。施普林格国际出版，Cham，525–541。https://doi.org/10.1007/978-3-030- 32430-8 31_[20] Guojun Wu、Yanhua Li、Zhenming Liu、Jie Bao、Yu Zheng、Jieping Ye 和 Jun Luo。2019年。奖励推进：最大因果熵原理下的策略变换。*arXiv*预印本 *arXiv:1907.05390*（2019年）。[21] Brian D Ziebart、Andrew L Maas、J Andrew Bagnell、Anind K Dey 等。2008年。最大熵逆强化学习。载于*AAAI*，第8卷。美国伊利诺伊州芝加哥，1433–1438。