Identity Concealment Games: How I Learned to Stop Revealing and Love the Coincidences [⋆]

Mustafa O. Karabag [a], Melkior Ornik [b], Ufuk Topcu [c]

[a] *Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, TX, USA*

[b] *Department of Aerospace Engineering, University of Illinois at Urbana–Champaign, IL, USA*

[c] *Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, TX, USA*

Abstract

In an adversarial environment, a hostile player performing a task may behave like a non-hostile one in order not to reveal its identity to an opponent. To model such a scenario, we define identity concealment games: zero-sum stochastic reachability games with a zero-sum objective of identity concealment. To measure the identity concealment of the player, we introduce the notion of an average player. The average player's policy represents the expected behavior of a non-hostile player. We show that there exists an equilibrium policy pair for every identity concealment game and give the optimality equations to synthesize an equilibrium policy pair. If the player's opponent follows a non-equilibrium policy, the player can hide its identity better. For this reason, we study how the hostile player may learn the opponent's policy. Since learning via exploration policies would quickly reveal the hostile player's identity to the opponent, we consider the problem of learning a near-optimal policy for the hostile player using the game runs collected under the average player's policy. Consequently, we propose an algorithm that provably learns a near-optimal policy and give an upper bound on the number of sample runs to be collected.

Key words: Identity concealment; Deception; Game theory; Markov models; Offline learning.

1 Introduction

In an adversarial environment, an agent interacts with a non-cooperative opponent. For a hostile agent, it may be important not to expose its identity since the opponent might attempt to hinder the agent's operation knowing that the agent is hostile. For instance, intelligence services often instruct the agents who are under surveillance to *dry-clean*, that is, to evade surveillance in a way that looks accidental, not intentional, since intentional evasions cause suspicion (Macintyre 2018). This behavior motivated video games such as Spy Party (Hecker 2018) and Garry's Mod Guess Who (Newman 2015) where the goal is to complete tasks behaving like a non-playable character, i.e., a bot. While identity concealment is a significant behavior in reality, it has not been studied in the literature of zero-sum games, which is a common formalism of adversarial settings (Kardes & Hall 2005).

We formalize the above notion of *identity concealment* in two-player zero-sum reachability games. We consider a graph as the state space of the game. The goal of a *hostile player* is to reach a set of target states, but in a way that its behavior looks similar to the behavior of non-hostile players, i.e., the hostile player aims to make its win look coincidental. The goal of the opponent is to distinguish between hostile and non-hostile players. As a reference point, we introduce an abstract notion of an *average player* to measure the identity concealment of a hostile player. The average player's policy represents the expected behavior of non-hostile players. For example, in the cyber interaction scenario shown in Figure 1, hostile players are attackers who perform a denial-of-service attack against a server, and average players are real clients interacting with the server. The attackers' goal is to overwhelm the server and make it fail to provide service to real clients while not being identified. The server is the opponent that aims to distinguish the attackers from real clients. We measure identity concealment by the cumulative Kullback-Leibler (KL) divergence between the action distribution of a hostile player and the action distribution of an average player over a game run. As the KL objective function increases, the opponent can dis-

[⋆] This paper was not presented at any IFAC meeting. Corresponding author M. O. Karabag.

Email addresses: karabag@utexas.edu (Mustafa O. Karabag), mornik@illinois.edu (Melkior Ornik), utopcu@utexas.edu (Ufuk Topcu).

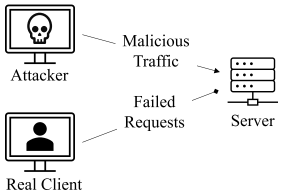


Figure 1. A cyber interaction scenario as a two-player game. The game is played between a client and the server. The server does not know the identity of the client, whether the client is an attacker or a real client. The states of the game represent the remaining times for the client's processed requests if there are any. At every time, the client can disconnect, make a request or wait. The server can accept or reject the client's potential request.

tinguish a hostile player from a non-hostile player more easily. For example, in the cyber interaction scenario, a game run can represent the history of the client's requests and the server's responses. In particular, possibly complex or time-varying behavior (such as repeated requests for unrelated computationally heavy resources) may indicate that the client is hostile, i.e., an attacker.

We define the *identity concealment game* as the two-player zero-sum reachability game with the cumulative KL divergence objective function. We first identify the conditions for which the value of the KL objective function is finite, i.e., the opponent is never sure the player is hostile. Then, we show that there exists an equilibrium policy pair for a hostile player and the opponent, which can be synthesized using value iteration.

The hostile player can achieve a lower value than the equilibrium value if the opponent follows a non-equilibrium policy. In this case, an equilibrium policy is not necessarily optimal for the hostile player against an imperfect opponent. The hostile player needs to learn and respond to the opponent's suboptimal policy to achieve the optimal value. However, the player's ability to learn in the described setting is limited in that an active learner would quickly reveal its identity during exploration. We consider the question of whether it is possible to learn a near-optimal policy offline by solely using the game runs collected under the average player's policy. The output policy needs to be near-optimal in that the KL objective function is ε-optimal, and the probability of winning is at least $1 - \lambda$ where ε and λ are the input parameters of the algorithm.

We provide an algorithm that solely uses a finite number of runs collected under the average player's policy to learn a near-optimal policy. To show the near-optimality in the KL objective, we utilize and improve some of the probably approximately correct Markov decision processes (PAC-MDP) learning results (Fiechter 1994, Kearns & Singh 2002, Strehl & Littman 2008). To show the near-optimality in the probability of winning, we show that under the output policy, the hostile player can lose the game only if an unknown state, i.e., a state with a low number of samples, is visited. Then, we show that the unknown states cannot be visited with a high probability if the number of sample runs is high enough.

We give the proofs of some technical results in (Karabag et al. 2021*b*) due to lack of space.

2 Related Work

The KL objective function is used for different purposes including *deception in supervisory control* (Karabag et al. 2021*a*), *game balancing* (Grau-Moya et al. 2018), *inverse reinforcement learning* (Boularias et al. 2011), and *reinforcement learning* (Fox et al. 2016, Peters et al. 2010). Karabag et al. (2021*a*) utilized Sanov's theorem (Cover & Thomas 2012) and the KL divergence of the path distributions in MDPs for deception in supervisory control. In that paper, the supervisor designs a reference policy to an agent, which is supposed to follow this policy, but it deviates from the reference policy to achieve a malicious task. The goal of the supervisor is to design a reference policy that minimizes deviations. While we use the objective function for the same purpose, this paper differs from (Karabag et al. 2021*a*) in that the opponent (analogue of the supervisor) does not design the average player's policy (analogue of the reference policy). Instead, the opponent designs a policy that determines the observability of the player (analogue of the agent). Grau-Moya et al. (2018) used the KL divergence objective for game balancing in two-player stochastic games. Aside from the contextual differences, the objective function in (Grau-Moya et al. 2018) has a discount factor. We, on the other hand, do not have a discount factor that significantly differs the proof for the existence of an equilibrium. Goal and plan obfuscation (Kulkarni et al. 2019, Keren et al. 2016) are similar to the concept of identity concealment. We consider a measure based on statistical hypothesis testing, whereas the cited works consider measures based on the distance of the observation sequences generated by a game run.

The learning algorithm provided in this paper is related to PAC-MDP algorithms (Kearns & Singh 2002, Strehl & Littman 2008). While these algorithms guarantee near-optimality after a finite number of suboptimal actions, there are no guarantees on the suboptimality of the transient learning period due to exploration. In the adversarial setting described in this paper, the use of PAC-MDP algorithms would reveal the identity of the player during the learning period. The algorithm provided in this paper uses a fixed policy, the average player's policy, to learn, whereas PAC-MDP algorithms learn in an exploratory manner. The learning algorithm provided in this paper is also related to off-policy evaluation and optimization (Farajtabar et al. 2018, Precup et al. 2000, Yu et al. 2020, Kidambi et al. 2020, Levine et al. 2020) as we collect offline samples using a behavior policy that is not the target policy to be evaluated or optimized. In detail, our algorithm is similar to model-based off-policy optimization (Yu et al. 2020, Kidambi et al. 2020). The existing offline learning literature considers the single-objective discounted infinite horizon (Puterman 2014) setting. On the other hand, we consider a multi-objective infinite horizon setting, where one of the objectives, probability of winning, is undiscounted, and the other objective is the KL divergence. Similar to the existing model-based offline learning methods, we show the near-optimality in the KL divergence by showing that the learned model is close to the actual model. To show the near-optimality in the probability of winning, we use a new approach that utilizes the near-optimality of the KL objective function.

The sample complexity of offline policy optimization is dependent on the distributional shift between the behavior and optimal policies (Levine et al. 2020). However, quantifying the distributional shift is challenging since it requires knowing the statistical properties of the processes induced by the policies. To ensure that the learned policy does not suffer from distributional shift, existing learning methods use KL divergence as a regularizer (Schulman et al. 2015). We, on the other hand, use the equilibrium value of KL divergence to reason about the maximum distributional shift that a near-optimal policy can have and guarantee near-optimality in the probability of winning: Since the maximum distributional shift of the near-optimal policy is bounded, we can bound the probability of losing for the hostile player. This approach allows us to give an explicit bound on the number of samples required to synthesize a near-optimal policy. Unlike the existing model-based off-policy optimization works that provide sample complexity bounds with agnostic dependencies on the distributional shift (Ross & Bagnell 2012, Uehara & Sun 2021), we give a bound that has a known dependency on the distributional shift thanks to the known equilibrium value of the game.

3 Preliminaries

In this section, we give preliminaries on two-player stochastic games and objective functions for the games.

3.1 Two-Player Stochastic Games and Markov Decision Processes

A *two-player stochastic game* is a tuple $\mathcal{G} = (\mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, \mathcal{P}, s_0)$ where $\mathcal{S}$ is a finite set of states, $\mathcal{A}^1$ is a finite set of actions for Player 1, $\mathcal{A}^2$ is a finite set of actions for Player 2, $\mathcal{P} : \mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2 \times \mathcal{S} \to [0, 1]$ is the transition probability function, and $s_0 \in \mathcal{S}$ is the initial state. We note that the available actions can be state-dependent. We use fixed sets, $\mathcal{A}^1$ and $\mathcal{A}^2$, for the available actions of players for clarity of presentation. For every state $s \in \mathcal{S}$, $\sum_{q \in \mathcal{S}} \mathcal{P}(s, a^1, a^2, q) = 1$ for all $(a^1, a^2) \in \mathcal{A}^1 \times \mathcal{A}^2$. We use S to denote the cardinality of $\mathcal{S}$, and A to denote the maximum of the cardinalities of $\mathcal{A}^1$ and $\mathcal{A}^2$. The *successor states* of state s is denoted by $Succ(s)$ where state q is in $Succ(s)$ if and only if there exist actions $(a^1, a^2) \in \mathcal{A}^1 \times \mathcal{A}^2$ such that $\mathcal{P}(s, a^1, a^2, q) > 0$. State s is *absorbing* if and only if $Succ(s) = \{s\}$, and there is only one available action for each player. The set of all absorbing states is $\mathcal{S}^{abs}$.

The game has infinite steps. At every time step t, Players 1 and 2 choose their actions, a^1_t and a^2_t, simultaneously and transition to state s_{t+1} from s_t with probability $\mathcal{P}(s_t, a^1_t, a^2_t, s_{t+1})$. The history $h_t = s_0 a^1_0 a^2_0 \dots s_{t-1} a^1_{t-1} a^2_{t-1} s_t$ at time t is the sequence of all previous states and actions. The set of all possible histories at time t is denoted by $\mathcal{H}_t$.

A *(history-dependent) policy* for Player i is a sequence $\pi^i = \mu^i_0 \mu^i_1 \dots$ where each $\mu^i_t : \mathcal{H}_t \times \mathcal{A}^i \to [0, 1]$ is a decision function such that $\sum_{a^i \in \mathcal{A}^i} \mu^i_t(h_t, a^i) = 1$ for all $h_t \in \mathcal{H}_t$. Given the history h_t, we use $\mu^i_t(h_t)$ to denote the action distribution under Player i's policy π^i at time t and state s_t. A *stationary policy* for Player i is a sequence $\pi^i = \mu^i \mu^i \dots$ such that $\mu^i : \mathcal{S} \times \mathcal{A}^i \to [0, 1]$ and $\sum_{a^i \in \mathcal{A}^i} \mu^i(s, a^i) = 1$ for all $s \in \mathcal{S}$. The set of all policies for Player i is denoted by Π^i. The set of all stationary policies Player i is denoted by $\Pi^{i,St}$. For state s, we use $\pi^i(s)$ to denote the action distribution under Player i's stationary policy π^i. A *run* $\gamma = s_0 a^1_0 a^2_0 s_1 a^1_1 a^2_1 \dots$ is an infinite sequence states and actions under policies π^1 and π^2 such that $\mathcal{P}(s_t, a^1, a^2, s_{t+1})\ \mu^1_t(h_t, a^1)\ \mu^2_t(h_t, a^2) > 0$ for all $t \geq 0$, i.e., all transitions are feasible. The probability distribution of runs under π^1 and π^2 is denoted by Γ^{π^1,π^2}. The probability distribution of histories at time t is denoted by Γ^{π^1,π^2}_t. The *(undiscounted) occupancy measure* of state s is the expected number of times state s is visited and is equal to $\sum_{t=0}^{\infty} \Pr^{\pi^1,\pi^2}(s_t = s | s_0)$.

A *Markov decision process* (MDP) is a tuple $\mathcal{M} = (\mathcal{S}', \mathcal{A}', \mathcal{P}', s'_0)$ where $\mathcal{S}'$ is a finite set of states, $\mathcal{A}'$ is a finite set of actions, $\mathcal{P}' : \mathcal{S}' \times \mathcal{A}' \times \mathcal{S}' \to [0, 1]$ is the transition probability function, and s'_0 is the initial state. A two-player stochastic game where one of the players uses a known, fixed policy is an MDP.

3.1.1 Zero-Sum Objective and Equilibrium Policies

A payoff function $c : \mathcal{S} \times \Delta_{|\mathcal{A}^1|} \times \Delta_{|\mathcal{A}^2|} \to \mathbb{R}$ maps the state and action distributions of the players to a payoff value where Δ_k is the k-dimensional probability simplex. At time t, Players 1 and 2 with policies $\pi^1 = \mu^1_0 \mu^1_1 \dots$ and $\pi^2 = \mu^2_0 \mu^2_1 \dots$ receive a payoff of $c(s_t, \mu^1_t(h_t), \mu^2_t(h_t))$.

Let $X_t = (s_t, a^1_t, a^2_t)$ be a random variable consisting of the random state and actions of the players at time t. For the random process $\{X_t\}$, the *hitting time* τ to the set $\mathcal{S}^{abs}$ of absorbing states is a random variable defined by $\tau = \min\{t \geq 0 | s_t \in \mathcal{S}^{abs}\}$ taking values in $\mathbb{N} \cup \{\infty\}$. Using the hitting time, the zero-sum objective function

$$C(\pi^1, \pi^2) = \mathbb{E}\left[\sum_{t=0}^{\tau} c(s_t, \mu_t^1(h_t), \mu_t^2(h_t))\right]$$

is the expected cumulative payoff until the random stopping time τ where the expectation is over the randomness of policies, π^1 and π^2, and the dynamics $\mathcal{P}$ of the game. Player 1 is the *minimizer*, and Player 2 is the *maximizer* of the zero-sum objective.

Let P^1 and P^2 denote the fixed sets of feasible policies for Players 1 and 2, respectively. A pair $(\pi^{1,*}, \pi^{2,*}) \in P^1 \times P^2$ of policies is an *equilibrium policy pair* if and only if

$$\sup_{\pi^2 \in P^2} C(\pi^1, \pi^{2,*}) \leq C(\pi^{1,*}, \pi^{2,*}) \leq \inf_{\pi^1 \in P^1} C(\pi^1, \pi^{2,*}).$$

If such an equilibrium policy pair $(\pi^{1,*}, \pi^{2,*})$ exists, $v^* = C(\pi^{1,*}, \pi^{2,*})$ is the *equilibrium value* of the game.

3.1.2 Reachability Objective

The event of eventually reaching set D is denoted by $\Diamond D$. Under policies π^1 and π^2, the probability of reaching set D from state s is denoted by $\Pr^{\pi^1,\pi^2}(\Diamond D | s)$. The probability of reaching set D from state s in L time steps is denoted by $\Pr^{\pi^1,\pi^2}(\Diamond_{\leq L} D | s)$.

$\mathcal{S}^R$ denotes the set of winning states for Player 1 for the reachability objective. Player 1 *wins* if and only if the game run $\gamma = s_0 a_0^1 a_0^2 s_1 a_1^1 a_1^2 \ldots$ satisfies $s_t \in \mathcal{S}^R$ for some $t \geq 0$, i.e., γ satisfies $\Diamond \mathcal{S}^R$. A policy π^1 for Player 1 is *winning* if $\min_{\pi^2 \in \Pi^2} \Pr^{\pi^1,\pi^2}(\Diamond \mathcal{S}^R | s_0) = 1$. We denote the set of winning policies for Player 1 by $\Pi^{1,win}$, and we denote the set of winning stationary policies for Player 1 by $\Pi^{1,St,win} = \Pi^{1,St} \cap \Pi^{1,win}$. For simplicity of presentation, we assume that all winning states are absorbing.

3.2 Kullback–Leibler (KL) Divergence

The support of a discrete probability distribution Q is denoted by $Supp(Q)$. For discrete probability distributions Q_1 and Q_2 where $Supp(Q_1) = \mathcal{X}$, the *Kullback–Leibler (KL) divergence* between Q_1 and Q_2 is $KL(Q_1 || Q_2) = \sum_{x \in \mathcal{X}} Q_1(x) \log(Q_1(x)/Q_2(x))$ where log denotes the natural logarithm. We define $0 \log(0/0) = 0$. Data processing inequality (Cover & Thomas 2012) states that any transformation $T : \mathcal{X} \to \mathcal{Y}$ satisfies $KL(Q_1 || Q_2) \geq KL(T(Q_1) || T(Q_2))$.

Let $\pi^{1'}$ be a policy for Player 1. Note that

$$KL\big(\Gamma_t^{\pi^1,\pi^2} || \Gamma_t^{\pi^{1'},\pi^2}\big) \leq KL\big(\Gamma_{t+1}^{\pi^1,\pi^2} || \Gamma_{t+1}^{\pi^{1'},\pi^2}\big)$$

due to the data processing inequality. We define

$$KL\big(\Gamma^{\pi^1,\pi^2} || \Gamma^{\pi^{1'},\pi^2}\big) = \lim_{t \to \infty} KL\big(\Gamma_t^{\pi^1,\pi^2} || \Gamma_t^{\pi^{1'},\pi^2}\big).$$

The limit either converges or diverges to ∞ due to monotonicity of $KL\big(\Gamma_t^{\pi^1,\pi^2} || \Gamma_t^{\pi^{1'},\pi^2}\big)$. We denote $KL\big(\Gamma^{\pi^1,\pi^2} || \Gamma^{\pi^{1'},\pi^2}\big)$ with $KL(\pi^1, \pi^2 || \pi^{1'}, \pi^2)$ for notational simplicity.

4 Problem Statement

We consider a two-player stochastic game $\mathcal{G} = (\mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, \mathcal{P}, s_0)$. In this game, Player 1 aims to satisfy the reachability objective $\Diamond \mathcal{S}^R$, i.e., win the game. Player 1 may be *hostile*, i.e., it may aim to reach $\mathcal{S}^R$ for malicious purposes. For example, in the cyber interaction scenario shown in Figure 1, Player 1 may be a client performing a denial-of-service attack. Player 2 is not aware of the identity of Player 1. In the cyber interaction scenario, Player 2 represents the server and does not know whether the client is an attacker. When Player 1 is hostile, its goal is not to expose its identity while winning the game. Player 2, on the other hand, aims to detect the identity of Player 1, i.e., determine whether Player 1 is hostile, in addition to making Player 1 lose the game. In the cyber interaction scenario, the goal of Player 1, i.e., a hostile client, is to perform an attack while not being detected by the server, and the goal of Player 2, i.e., the server, is to provide service to well-meaning clients while identifying hostile clients. In this setting, we assume that both players have full information on the current state and full information on each other's previous actions.

We consider an average player as the reference point to measure identity concealment. The average player's policy encodes the expected behavior of a non-hostile player interacting with Player 2, and can be used to measure how much Player 1's policy π^1 exposes its identity and how well Player 2's policy π^2 distinguishes hostile agents. The average player's policy π^{Av} is not necessarily designed to win the game against Player 2, but the average player can accidentally win the game due to the stochasticity of the environment or its policy. For example, in the cyber interaction scenario, π^{Av} may represent the average behavior of non-hostile clients, i.e., real users, interacting with the server. These clients may cause a denial-of-service, but their goal is not necessarily to cause a breakdown. We assume that the average player's policy π^{Av} is common knowledge. We also have the following assumption which ensures the computational tractability of the problems to be proposed.

Assumption 1. *The average player's policy π^{Av} is stationary on the state space $\mathcal{S}$, i.e., $\pi^{\mathsf{Av}} \in \Pi^{1,St}$.*

Because an average player can win the game with a positive probability, a win in the game does not immediately identify Player 1 as hostile. Therefore, Player 1 aims to make its win look accidental and indistinguishable from an average player's win. On the flip side, Player 2 aims to design its policy in a way that the identity concealment of Player 1 is minimized, i.e., an average player and a hostile Player 1 produce different game runs.

We define the identity exposure payoff at time t as the KL divergence between the action distribution $\mu_t^1(h_t)$ under Player 1's policy $\pi^1 = \mu_0^1\mu_1^1\ldots$ and the action distribution $\pi^{\mathsf{Av}}(s_t)$ under the average player's policy π^{Av}. Formally, the payoff of Players 1 and 2 is

$$c(s_t, \mu_t^1(h_t), \mu_t^2(h_t)) := KL(\mu_t^1(h_t)||\pi^{\mathsf{Av}}(h_t)) \qquad (1)$$

at time t. For clarity of presentation, we restrict the feasible policy spaces of Players 1 and 2 to stationary policies, i.e., $\pi^1 \in \Pi^{1,St}$ and $\pi^2 \in \Pi^{2,St}$. In this case, the payoff of Players 1 and 2 is

$$\begin{aligned} c(s_t, \pi^1(s_t), \pi^2(s_t)) &= KL(\pi^1(s_t)||\pi^{\mathsf{Av}}(s_t)) \\ &= \sum_{a^1\in\mathcal{A}^1} \pi^1(s_t, a^1)\log\left(\frac{\pi^1(s_t, a^1)}{\pi^{\mathsf{Av}}(s_t, a^1)}\right) \end{aligned}$$

at time t. This payoff decreases when $\pi^1(s_t)$ gets more similar to $\pi^{\mathsf{Av}}(s_t)$. The payoff is 0 when the action distributions of Player 1 and the average player is the same, i.e., $\pi^1(s_t) = \pi^{\mathsf{Av}}(s_t)$.

Using the hitting time τ that is the first hitting time to the set $\mathcal{S}^{abs}$ of absorbing states, the *zero-sum identity concealment objective function* is

$$C(\pi^1, \pi^2) := \mathbb{E}\left[\sum_{t=0}^{\tau} KL(\pi^1(s_t)||\pi^{\mathsf{Av}}(s_t))\right]$$

where the expectation is over the randomness of policies, π^1 and π^2, and the dynamics $\mathcal{P}$ of the game. Player 1 is the *minimizer* and Player 2 is the *maximizer* of the zero-sum objective: Hostile Player 1 aims to behave similar to the average players and Player 2 aims to distinguish hostile Player 1 from the average players. The zero-sum objective accounts for the total identity exposure of Player 1 until the effective end of the game, i.e., reaching an absorbing state. We discuss the relationship of this objective function with hypothesis testing in Section 4.1.

We define an *identity concealment game* $\mathcal{IC} = (\mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, \mathcal{P}, s_0, \mathcal{S}^R, \pi^{\mathsf{Av}})$ as a two-player stochastic game with objective functions

$$C(\pi^1, \pi^2) = \mathbb{E}\left[\sum_{t=0}^{\tau} KL(\pi^1(s_t)||\pi^{\mathsf{Av}}(s_t))\right]$$

and

$$1 - \Pr^{\pi^1,\pi^2}(\Diamond\mathcal{S}^R|s_0),$$

where Player 1 is the minimizer and Player 2 is the maximizer for both functions. We remark that the game is not well-defined due to multiple objective functions, and one needs combine these objective functions to find optimal policies. Chen et al. (2013) showed that when multiple objective functions are combined with a conjunction where each predicate is a threshold for an objective function, it is, in general, computationally hard to compute optimal policies. To bypass computational issues, we pose the following problem that constrains the value of the objective function $1 - \Pr^{\pi^1,\pi^2}(\Diamond\mathcal{S}^R|s_0)$ to 0, i.e., Player 1 must use a winning policy.

Problem 2. For a given identity concealment game $\mathcal{IC}$, determine whether there exists an equilibrium pair $(\pi^{1,*}, \pi^{2,*}) \in \Pi^{1,St,win} \times \Pi^{2,St}$ of policies such that

$$\sup_{\pi^2\in\Pi^{2,St}} C(\pi^1, \pi^{2,*}) \leq C(\pi^{1,*}, \pi^{2,*})$$

and

$$C(\pi^{1,*}, \pi^{2,*}) \leq \inf_{\pi^1\in\Pi^{1,St,win}} C(\pi^1, \pi^{2,*}).$$

Remark 3. We state that Player 1 aims to win with probability 1 and assume that there exists such a winning policy. If there is not such a policy, one can use the weighted zero-sum objective function $C(\pi^1, \pi^2) + \alpha(1 - \Pr^{\pi^1,\pi^2}(\Diamond\mathcal{S}^R|s_0))$ where $\alpha \in \mathbb{R}\cup\{\infty\}$ is the weight of the winning objective. When winning in the game is a hard constraint for Player 1, i.e., when $\alpha = \infty$, the weighted zero-sum objective function recovers Problem 2.

Player 1 can achieve a lower value than the equilibrium value of the game if Player 2 uses a suboptimal non-equilibrium policy. In this case, the optimal policy of Player 1 is not necessarily the equilibrium policy, and the hostile Player 1 may need to learn Player 2's policy to synthesize the optimal policy. While it is possible to learn Player 2's policy with a high amount of exploration, learning in this way is undesirable in the described adversarial setting since naive exploration would quickly reveal the identity of the hostile Player 1. Furthermore, Player 1 may not be able to collect game runs by directly interacting with Player 2 and may only observe Player 2's interactions with average players. Hence, the hostile Player 1's goal is to learn Player 2's policy from runs collected under the average player's policy and compute the optimal policy. We propose the following problem.

Problem 4. For an identity concealment game $\mathcal{IC}$, let $\pi^{2,\circ}$ be Player 2's policy that is unknown a priori to Player 1 and $\pi^{1,\circ} = \arg\min_{\pi\in\Pi^{1,win}} C(\pi^1, \pi^{2,\circ})$. Given $\varepsilon > 0$, $\lambda \in [0, 1]$, and $\delta \in [0, 1]$, find an algorithm that uses a finite number of runs that are collected only using the average player's policy so that the (potentially history-dependent) output policy π^1 of the algorithm satisfies

$$C(\pi^1, \pi^{2,\circ}) \leq C(\pi^{1,\circ}, \pi^{2,\circ}) + \varepsilon$$

and

$$\Pr^{\pi^1, \pi^{2,\circ}}(\Diamond \mathcal{S}^R | s_0) \geq 1 - \lambda,$$

with probability at least $1 - \delta$.

4.1 KL Divergence Payoffs and Hypothesis Testing

The identity concealment game has a KL objective function $C(\pi^1, \pi^2) = \mathbb{E}\left[\sum_{t=0}^{\tau} KL(\pi^1(s_t)||\pi^{\mathsf{Av}}(s_t))\right]$ motivated by statistical hypothesis testing. The sum of KL stage payoffs is equal to the KL divergence between the probability distribution of runs under Player 1's policy π^1 and Player 2's policy π^2, and the probability distribution of runs under the average player's policy π^{Av} and Player 2's policy π^2. Formally, as we explain later in the proof of Lemma 6, we have

$$\mathbb{E}\left[\sum_{t=0}^{\tau} KL(\pi^1(s_t)||\pi^{\mathsf{Av}}(s_t))\right] = KL\left(\pi^1, \pi^2 || \pi^{\mathsf{Av}}, \pi^2\right).$$

By Sanov's theorem, $\exp\left(-nKL\left(\pi^1, \pi^2||\pi^{\mathsf{Av}}, \pi^2\right)\right)$ measures the probability that n random game runs with a hostile Player 1 occur under the average player's policy. Consequently, as the number n of game runs or $KL\left(\pi^1, \pi^2||\pi^{\mathsf{Av}}, \pi^2\right)$ increases Player 2 is more likely to identify a hostile player. More formally, as $nKL\left(\pi^1, \pi^2||\pi^{\mathsf{Av}}, \pi^2\right)$ increases, the accuracy of the likelihood-ratio test (Hogg et al. 1977) increases. The goal of Player 1 is thus to minimize $KL\left(\pi^1, \pi^2||\pi^{\mathsf{Av}}, \pi^2\right)$, while the goal of Player 2 is to maximize this value.

5 Equilibrium Policies for Identity Concealment Games

In this section, we prove the existence of an equilibrium for the identity concealment game and provide the optimality equations to compute it.

If there exists an equilibrium, and the equilibrium value for value for $C(\pi^1, \pi^2)$ is infinite in Problem 2, and all winning stationary policies are equally good for Player 1. We mainly focus on the more interesting case that there exists a winning stationary policy $\pi^1 \in \Pi^{1,St,win}$ such that $\max_{\pi^2 \in \Pi^{2,St}} C(\pi^1, \pi^2) < \infty$.

We define that action a^1 is *permissible* for Player 1 at state s if $\pi^{\mathsf{Av}}(s, a^1) > 0$. For example, all actions are permissible at every state for Player 1 for the identity concealment game given in Figure 2. Note that if Player 1 takes an impermissible action with a positive probability, then $C(\pi^1, \pi^2)$ is infinite, i.e., with a positive probability Player 2 is certain that Player 1 is not an average player, since an event happens with a positive probability under Player 1's policy and zero probability under the average player's policy. Note that removing the impermissible actions does not change the equilibrium value if the equilibrium value is finite.

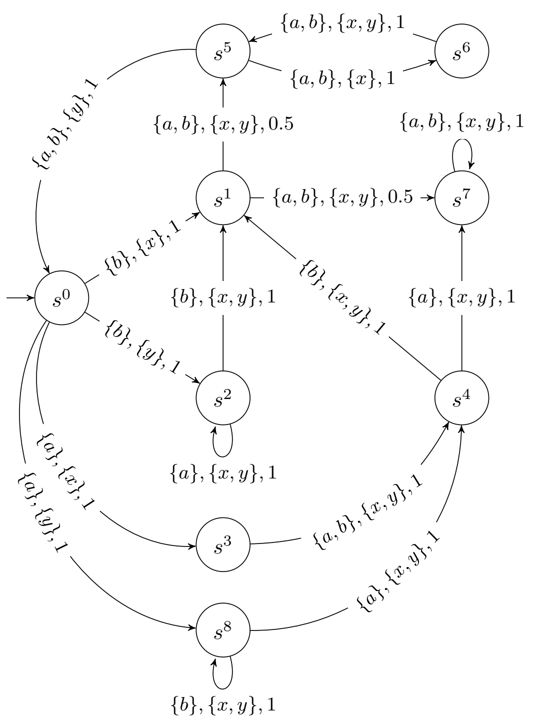


Figure 2. A identity concealment game where the actions of Player 1 are a and b, and the actions of Player 2 are x and y. Nodes are the states of the game. The initial state is s^0 and s^7 is the only winning state for Player 1. The average player's policy π^{Av} takes actions a and b uniformly randomly at every state. A label D^1, D^2, p of a directed edge from s to q means $\mathcal{P}(s, a^1, a^2, q) = p$ for every $a^1 \in D^1$ and $a^2 \in D^2$. A stationary trapping policy for Player 2 takes action x at every state.

Given that the equilibrium value is finite, without loss of generality, we assume that all impermissible actions are removed from all states.

Assumption 5. *Every available action is permissible for Player* 1.

To find equilibrium policies for Problem 2, we first identify the states at which Player 2 can win the game with a positive probability. These states can be found with a procedure similar to solving reachability games on graphs (Chatterjee et al. 2008).

State s is a *trap* state for Player 1 if there exists a policy $\pi^2 \in \Pi^2$ that satisfies $\Pr^{\pi^{\mathsf{Av}}, \pi^2}(\Diamond \mathcal{S}^R | s) = 0$. For example states s^5 and s^6 are trap states in Figure 2. The set of trap states is denoted by $\mathcal{S}^{trap}$. The set of trap states is easy to find: Since π^{Av} is stationary, it induces an

MDP for Player 2 given the game. On this MDP, to find $\pi^{2,trap}$, we solve a reach-avoid problem where the objective is to avoid the winning states $\mathcal{S}^R$ for Player 2. There exists a stationary policy $\pi^{2,trap} \in \Pi^{2,St}$ that minimizes $\Pr^{\pi^{\mathsf{Av}},\pi^2}(\Diamond \mathcal{S}^R | s)$ for every $s \in \mathcal{S}$ (Baier & Katoen 2008). Since the trapping policy minimizes the winning probability of an average player for every state, we have $s \in \mathcal{S}^{trap}$ if and only if $\Pr^{\pi^{\mathsf{Av}},\pi^{2,trap}}(\Diamond \mathcal{S}^R | s) = 0$.

Under Assumption 5, we have $\Pr^{\pi^1,\pi^{2,trap}}(\Diamond \mathcal{S}^R | s) = 0$ for every $s \in \mathcal{S}^{trap}$ and $\pi^1 \in \Pi^1$. To observe this, we consider two directed graphs. The policy pair $(\pi^{\mathsf{Av}}, \pi^{2,trap})$ induces a Markov chain. Let $G^{(\pi^{\mathsf{Av}},\pi^{2,trap})}$ be a directed graph that represents the feasible transitions on this Markov chain. In this directed graph the states $\mathcal{S}^{trap}$ are not connected to $\mathcal{S}^R$ since $\Pr^{\pi^{\mathsf{Av}},\pi^{2,trap}}(\Diamond \mathcal{S}^R | s) = 0$ for every $s \in \mathcal{S}^{trap}$. Similarly, the policy pair $(\pi^1, \pi^{2,trap})$ induces another Markov chain. Let $G^{(\pi^1,\pi^{2,trap})}$ be a directed graph that represents the feasible transitions on this Markov chain. Under Assumption 5, $(\pi^1, \pi^{2,trap})$ must be a subgraph of $G^{(\pi^{\mathsf{Av}},\pi^{2,trap})}$. This is because π^{Av} takes every available action for Player 1 with a positive probability. Since $G^{(\pi^1,\pi^{2,trap})}$ is a subgraph of $G^{(\pi^{\mathsf{Av}},\pi^{2,trap})}$, the states $\mathcal{S}^{trap}$ are not connected to $\mathcal{S}^R$ in $G^{(\pi^1,\pi^{2,trap})}$. Hence, $\Pr^{\pi^1,\pi^{2,trap}}(\Diamond \mathcal{S}^R | s) = 0$ for every $s \in \mathcal{S}^{trap}$ and $\pi^1 \in \Pi^1$.

A stationary winning policy $\pi^1 \in \Pi^{1,St,win}$ never visits a trap state regardless of Player 2's policy under Assumption 5. We show this by a contradiction. Consider policies $\pi^1 \in \Pi^{1,St,win}$ and $\pi^2 \in \pi^{2,St}$ that reach a trap state with a positive probability from the initial state, i.e., $\Pr^{\pi^1,\pi^2}(\Diamond \mathcal{S}^{trap} | s_0) > 0$. Consider a policy $\pi^{2'}$ that is the same as π^2 for all states in $\mathcal{S} \setminus \mathcal{S}^{trap}$ and is the same as $\pi^{2,trap}$ for all states in $\mathcal{S}^{trap}$. We have $\Pr^{\pi^1,\pi^{2'}}(\Diamond \mathcal{S}^{trap} | s_0) = \Pr^{\pi^1,\pi^2}(\Diamond \mathcal{S}^{trap} | s_0) > 0$ since $\pi^{2'}$ is the same as π^2 for all states in $\mathcal{S} \setminus \mathcal{S}^{trap}$. We also have $\Pr^{\pi^1,\pi^{2'}}(\Diamond \mathcal{S}^R | s) = 0$ for all $s \in \mathcal{S}^{trap}$ since a policy $\pi^{2'}$ is the same as $\pi^{2,trap}$ for all states in $\mathcal{S}^{trap}$. Hence, we have $\Pr^{\pi^1,\pi^{2'}}(\Diamond \mathcal{S}^R | s_0) < 1$ which contradicts with the fact that π^1 is a winning policy. Without Assumption 5, there could be a $\pi^1 \in \Pi^{1,St,win}$ that visits a trap state. All such policies take an impermissible action with a positive probability and yield an infinite objective value.

We find the set $\mathcal{S}^+$ of potentially winning states for which there exists a policy π^1 for Player 1 that reaches $\mathcal{S}^R$ with probability 1 for all $\pi^2 \in \Pi^2$ and avoids $\mathcal{S}^{trap}$. For example, s^0, s^3, s^4, s^7, and s^8 are the potentially winning states in Figure 2. We remark that there might be some states from which Player 2 can avoid the trap states with probability 1, but never reach the winning states. For example, s^2 is such a state in Figure 2. The set $\mathcal{S}^+$ of potentially winning states can be found by iteratively expanding $\mathcal{S}^R$ as in the attractor computation for two-player reachability games (Chatterjee et al. 2008). We note that stationary policies for Player 1 suffice to achieve maximal $\mathcal{S}^+$ against all possible policies of Player 2 since the game has the Markov property. If a pair of equilibrium policies exist, then only the states in $\mathcal{S}^+$ are visited with a positive probability since from all states in $\mathcal{S} \setminus \mathcal{S}^+$, there exists a policy for Player 2 such that $\mathcal{S}^R$ is reached with a probability strictly less than 1. We define that at state $s \in \mathcal{S}^+$ action a^1 is *safe* for Player 1 if and only if all potential successor states are in $\mathcal{S}^+$, i.e., $\mathcal{P}(s, a^1, a^2, q) = 0$ for all $a^2 \in \mathcal{A}^2$ and $q \in \mathcal{S} \setminus \mathcal{S}^+$. Note that for every state in $\mathcal{S}^+$, there exists a safe action due to the construction of $\mathcal{S}^+$. For example, a and b are safe actions for states s^3,s^7, and s^8, and a is the only safe action for states s^0 and s^4.

Having identified the set of states from which Player 1 can win the game with probability 1, we now focus on the existence of equilibrium policies. We note that the stage payoff $\sum_{a^1 \in \mathcal{A}^1} \pi^1(s_t, a^1) \log\left(\frac{\pi^1(s_t,a^1)}{\pi^{\mathsf{Av}}(s_t,a^1)}\right)$ is a convex function of the policy parameters of the minimizer, i.e., Player 1 and a concave function of the policy parameters of the maximizer, i.e., Player 2. Zero-sum stochastic games with such payoffs have an equilibrium when the payoffs are discounted (Başar & Olsder 1998). However, the game that we consider does not have a discount. To show the existence of an equilibrium, we need to prove additional properties of the identity concealment game.

Lemma 6 shows that if the initial state is in the potentially winning states, then there exists a (stationary) policy that makes the KL objective function finite. Furthermore, the occupancy measure at all states, but the states in $\mathcal{S}^{abs}$ must be finite in order to have a finite value for the KL objective function. To show this, we consider the visitation distributions for an arbitrary state since the KL divergence between the visitation distributions is a lower bound on the KL objective function. Proof of Lemma 6 shows that the policy pairs that induce infinite occupancy measure for a state that is not in $\mathcal{S}^{abs}$ lead to a visitation distribution that has infinite KL divergence from the distribution under the average player's policy. Formally, a pair (π^1, π^2) of (history-dependent) policies is *prolonging* if $\sum_{t=0}^{\infty} \Pr^{\pi^1,\pi^2}(s_t = s | s_0) = \infty$ for some $s \in \mathcal{S}^+ \setminus \mathcal{S}^{abs}$. All prolonging pairs (π^1, π^2) of policies satisfy $C(\pi^1, \pi^2) = \infty$. For example, consider state s^8 of the identity concealment game shown in Figure 2. There exists history-dependent policies for Player 1 that induces $\sum_{t=0}^{\infty} \Pr^{\pi^1,\pi^2}(s_t = s | s^0) = \infty$ if s^8 is reached with a positive probability. All such policy pairs have $C(\pi^1, \pi^2) = \infty$. We use properties given in Lemma 6 to show the existence of an equilibrium.

Lemma 6. *If $s_0 \in \mathcal{S}^+$, then there exists a winning policy $\pi^{1,fin} \in \Pi^{1,win}$ that satisfies $C(\pi^{1,fin}, \pi^2) < \infty$, and $\sum_{t=0}^{\infty} \Pr^{\pi^{1,fin},\pi^2}(s_t = s | s_0) < \infty$ for all $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$ and $\pi^2 \in \Pi^2$.*

If $\sum_{t=0}^{\infty} \Pr^{\pi^{1,inf},\pi^2}(s_t = s|s_0) = \infty$ *for some* $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$ *and* $\pi^{1,inf} \in \Pi^1$*, then* $C(\pi^{1,inf}, \pi^2) = \infty$.

We use the additional Lemma 7 to prove Lemma 6. The proof of Lemma 7 follows from that $\sum_{n \in C'} \mathcal{D}^1(n)n = \infty$ where $n \in C'$ if and only if $\mathcal{D}^1(n) > c_1 \exp(-nc_2/2)$.

Lemma 7. *Let* $\mathcal{D}^1$ *and* $\mathcal{D}^2$ *be discrete probability distributions such that* $Supp(\mathcal{D}^1), Supp(\mathcal{D}^2) \subseteq \mathbb{N}$*. If* $\sum_{n=0}^{\infty} \mathcal{D}^1(n)n = \infty$ *and there exist* $c_1, c_2 \in (0, \infty)$ *such that* $\mathcal{D}^2(n) \leq c_1 \exp(-c_2 n)$*, then* $KL(\mathcal{D}^1||\mathcal{D}^2) = \infty$.

Proof of Lemma 6. We prove the first part of the lemma by constructing a policy using safe actions and the definition of potentially winning states. For the second part, we lower bound the KL objective using the data processing inequality, and the time distributions at the states in $\mathcal{S}^+ \setminus \mathcal{S}^R$. We show that the lower bound and, consequently, the objective function are infinite if the Player 1's policy has infinite occupancy measure at $\mathcal{S}^+ \setminus \mathcal{S}^R$.

We show the existence of a stationary $\pi^{1,fin}$ by construction. At states in $\mathcal{S}^+$, $\pi^{1,fin}$ takes all permissible, safe actions uniformly randomly. To show $C(\pi^{1,fin}, \pi^2) < \infty$, we first note that by definition of conditional expectation, $C(\pi^{1,fin}, \pi^2) = \sum_{s \setminus \mathcal{S}^{abs}} \sum_{t=0}^{\infty} \Pr^{\pi^{1,fin},\pi^2}(s_t = s|s_0) KL(\pi^1(s)||\pi^{\mathsf{Av}}(s))$.

For every $s \in \mathcal{S}^+$, we have $KL(\pi^1(s)||\pi^{\mathsf{Av}}(s)) < \infty$ since $\pi^{1,fin}$ takes only permissible actions. Let $\bar{c} = \max_{s \in \mathcal{S}^+} KL(\pi^1(s)||\pi^{\mathsf{Av}}(s))$.

By definition of $\mathcal{S}^+$, there must exist a state $s_t \in \mathcal{S}^+$ and such that $\Pr^{\pi^{1,fin},\pi^2}(s_{t+1} \in \mathcal{S}^R|s_t) > 0$ for all π^2. Similarly, $\Pr^{\pi^{1,fin},\pi^2}(\Diamond_{\leq S} \mathcal{S}^R|h_t) > 0$ for all $h_t \in \mathcal{H}_t$. Since the game ends in every S steps with a positive probability, we have $\sum_{s \in \mathcal{S} \setminus \mathcal{S}^{abs}} \sum_{t=0}^{\infty} \Pr^{\pi^{1,fin},\pi^2}(s_t = s|s_0) \leq \bar{t} < \infty$. Therefore, we have $C(\pi^{1,fin}, \pi^2) \leq \bar{c}\bar{t} < \infty$.

We now prove that if $\sum_{t=0}^{\infty} \Pr^{\pi^{1,inf},\pi^2}(s_t = s) = \infty$ for some $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$, then $C(\pi^{1,inf}, \pi^2)$ is infinite. We first represent the objective function $C(\pi^1, \pi^2)$ as the KL divergence between the probability distribution of runs under Player 1's policy π^1 and Player 2's policy π^2, and the probability distribution of runs under the average player's policy π^{Av} and Player 2's policy π^2. In detail,

$$C(\pi^1, \pi^2) = \mathbb{E}\left[\sum_{t=0}^{\tau} \sum_{a^1 \in \mathcal{A}^1} \mu_t^1(h_t, a^1) \log\left(\frac{\mu_t^1(h_t, a^1)}{\pi^{\mathsf{Av}}(s_t, a^1)}\right)\right]$$

$$= \mathbb{E}\left[\sum_{t=0}^{\infty} \sum_{a^1 \in \mathcal{A}^1} \mu_t^1(h_t, a^1) \log\left(\frac{\mu_t^1(h_t, a^1)}{\pi^{\mathsf{Av}}(s_t, a^1)}\right)\right]$$

$$= \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \sum_{\substack{s_{t+1} \in \mathcal{S} \\ a^1 \in \mathcal{A}^1 \\ a^2 \in \mathcal{A}^2}} \mu_t^1(h_t, a^1)\mu_t^2(h_t, a^2)\mathcal{P}(s_t, a^1, a^2, s_{t+1})$$

$$\log\left(\frac{\mu_t^1(h_t, a^1)\mu_t^2(h_t, a^2)\mathcal{P}(s_t, a^1, a^2, s_{t+1})}{\pi^{\mathsf{Av}}(s_t, a^1)\mu_t^2(h_t, a^2)\mathcal{P}(s_t, a^1, a^2, s_{t+1})}\right)\Bigg]$$

$$= \lim_{t \to \infty} \sum_{\gamma_t \in Supp(\Gamma_t^{\pi^1,\pi^2})} \Pr^{\pi^1,\pi^2}(\gamma_t) \log\left(\frac{\Pr^{\pi^1,\pi^2}(\gamma_t)}{\Pr^{\pi^{\mathsf{Av}},\pi^2}(\gamma_t)}\right)$$

where the first equality is by definition, the second equality is because for $t \geq \tau$ $s_t \in \mathcal{S}^{abs}$ and every state $s \in \mathcal{S}^{abs}$ has a single action inducing 0 KL divergence cost, and the last inequality is due to the chain rule of KL divergence and Markovianity of the game.

Let $\Diamond \bigcirc D$ denote the event of eventually reaching set D starting from the next time step. The game run $\gamma = s_0 a_0^1 a_0^2 s_1 a_1^1 a_1^2 \ldots$ satisfies $\Diamond \bigcirc D$ if and only if there exists $s_t \in D$ for some $t \geq 1$. We first claim that for all $\pi^2 \in \Pi^2$, $t \geq 0$, and $h_t \in \mathcal{H}_t$ such that $s_t \in \mathcal{S}^+ \setminus \mathcal{S}^R$, we have $\Pr^{\pi^{\mathsf{Av}},\pi^2}(\Diamond \bigcirc \{s_t\}|h_t) < 1$. In words, a state in $\mathcal{S}^+ \setminus \mathcal{S}^R$ will not be visited back with a positive probability. Note that stationary policies for Player 2 suffice to maximize the reachability probability to a state in the MDP induced by π^{Av}. If $\Pr^{\pi^{\mathsf{Av}},\pi^2}(\Diamond \bigcirc \{s_t\}|h_t) = 1$ for some $\pi^2 \in \Pi^{2,St}$, then we have $\Pr^{\pi^{\mathsf{Av}},\pi^2}(\Diamond \mathcal{S}^R|s_t) = 0$ and s_t must be a trap state. This yields a contradiction since $s \in \mathcal{S}^+$ and $\mathcal{S}^+ \cap \mathcal{S}^{trap} = \emptyset$. Thus, for all $\pi^2 \in \Pi^2$, $t \geq 0$, and $h_t \in \mathcal{H}_t$ such that $s_t \in \mathcal{S}^+ \setminus \mathcal{S}^R$, we have $\Pr^{\pi^{\mathsf{Av}},\pi^2}(\Diamond \bigcirc \{s_t\}|h_t) < 1$. Let $N_s^{\mathsf{Av},2}$ be a random variable denoting the number of times that $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$ appears in a random run under π^{Av} and π^2. Similarly, let $N_s^{1,2}$ be a random variable denoting the number of times that $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$ appears in a random run under π^1 and π^2. Since $\Pr^{\pi^{\mathsf{Av}},\pi^2}(\Diamond \bigcirc \{s\}|h_t, s_t = s) < 1$, there exists a $c \in [0, 1)$ such that $\Pr(N_s^{\mathsf{Av},2} = k) \leq c^{k-1}$ for all $k \geq 1$ and $\Pr(N_s^{\mathsf{Av},2} = 0) \leq 1$. If $\sum_{t=0}^{\infty} \Pr^{\pi^{1,inf},\pi^2}(s_t = s) = \infty$ for some $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$, then by Lemma 7, we have $KL(N_s^{1,2}||N_s^{\mathsf{Av},2}) = \infty$ for some $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$. By the data processing inequality, for all $s \in S$ we have $KL(N_s^{1,2}||N_s^{\mathsf{Av},2}) \leq KL(\pi^{1,inf}, \pi^2||\pi^{\mathsf{Av},inf}, \pi^2)$. Thus, $KL(\pi^{1,inf}, \pi^2||\pi^{\mathsf{Av}}, \pi^2) = C(\pi^{1,inf}, \pi^2) = \infty$. □

We note three facts: 1) All prolonging pairs $(\pi^{inf,1}, \pi^2)$ of policies for which $\sum_{t=0}^{\tau} \Pr^{\pi^{inf,1},\pi^2}(s_t = s) = \infty$ for some $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$, satisfy $C(\pi^{1,inf}, \pi^2) = \infty$, 2) There exists a winning policy $\pi^{1,fin}$ for Player 1 that has $C(\pi^{1,fin}, \pi^2) < \infty$, and 3) The payoff for each time step is a convex, continuous function of Player 1's action distribution and a concave, continuous function of Player 2's action distribution. Due to these facts it is sufficient to consider only the stationary policies to find an equilibrium policy pair (Patek & Bertsekas 1999). When these conditions hold, Bellman's optimality equation leads to a unique fixed point, and there exists a stationary policy for a player that induces the optimal set of occupancy measures and achieves Bellman optimality when the other player's policy is fixed [1] .

Proposition 8. *For an identity concealment game $\mathcal{IC}$, if there exists a winning policy π^1 for Player 1 for which $\Pr^{\pi^1,\pi^2}(\Diamond\mathcal{S}^R|s_0) = 1$ for all $\pi^2 \in \Pi^2$, then there exists an equilibrium pair $(\pi^{1,*},\pi^{2,*}) \in \Pi^{1,St,win} \times \Pi^{2,St}$ of policies such that*

$$\sup_{\pi^2\in\Pi^{2,St}} C(\pi^1,\pi^{2,*}) \leq C(\pi^{1,*},\pi^{2,*})$$

and

$$C(\pi^{1,*},\pi^{2,*}) \leq \inf_{\pi^1\in\Pi^{1,St,win}} C(\pi^1,\pi^{2,*}).$$

Proof of Proposition 8. We consider two cases, the equilibrium value for the KL objective is finite and infinite.

We first consider the finite case. The existence of an equilibrium follows from the conditions given in (Patek & Bertsekas 1999): 1) All prolonging pairs $(\pi^{inf,1},\pi^2)$ of policies for which $\sum_{t=0}^{\tau}\Pr^{\pi^{inf,1},\pi^2}(s_t = s) = \infty$ for some $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$, satisfy $C(\pi^{1,inf},\pi^2) = \infty$, 2) There exists a winning policy $\pi^{1,fin}$ for Player 1 that has $C(\pi^{1,fin},\pi^2) < \infty$, and 3) The payoff for each time step is a convex function of Player 1's action distribution and a concave function of Player 2's action distribution. We show that these conditions hold for the identity concealment games and prove the existence of an equilibrium.

Without loss of generality, assume that Player 1 only takes actions that are safe. Let Player 1's actions $\mathcal{A}^1(s)$ be enumerated from 1 to $|\mathcal{A}^1|$ and Player 2's actions $\mathcal{A}^2$ be enumerated from 1 to $|\mathcal{A}^2|$ for all $s \in \mathcal{S}$. Denote the n-dimensional probability simplex by Δ^n.

Define a zero-sum two-player stochastic game $\hat{\mathcal{G}} = (\mathcal{S},\hat{\mathcal{A}}^1,\hat{\mathcal{A}}^2,\hat{\mathcal{P}},s_0,\mathcal{S}^R)$ with compact action spaces where $\hat{\mathcal{A}}^1$ and $\hat{\mathcal{A}}^2$ are metric set of actions for Players 1 and 2, respectively. Player 1 and 2's policies are $\hat{\pi}^1$ and $\hat{\pi}^2$, respectively. At time t, Player 1's decision function is $\hat{\mu}_t^1 : \mathcal{H}_t \to \Delta^{|\mathcal{A}^1|}$ and Player 2's decision function is $\hat{\mu}_t^2 : \mathcal{H}_t \to \Delta^{|\mathcal{A}^2|}$. Player 1 and 2's feasible policies are $\hat{\Pi}^1$ and $\hat{\Pi}^2$, respectively. Let $\hat{\Pi}^{1,win}$ be the set of winning policies for Player 1 such that $\hat{\Pi}^{1,win} = \left\{\pi^1 | \min_{\hat{\pi}^2\in\hat{\Pi}^2}\Pr^{\hat{\pi}^1,\hat{\pi}^2}(\Diamond\mathcal{S}^R|s_0) = 1\right\}$. We define $\hat{\mathcal{A}}^1$, $\hat{\mathcal{A}}^2$, and $\hat{\mathcal{P}}$ such that the following is satisfied:

$$\hat{\mathcal{P}}(s,\hat{a}^1,\hat{a}^2,q) = \sum_{i=1}^{|\mathcal{A}^1|}\sum_{j=1}^{|\mathcal{A}^2|}\hat{a}^1(i)\hat{a}^2(j)\mathcal{P}(s,i,j,q)$$

for all $s \in \mathcal{S}$, $\hat{a}^1 \in \hat{\mathcal{A}}^1 = \Delta^{|\mathcal{A}^1|}$, $\hat{a}^2 \in \hat{\mathcal{A}}^2 = \Delta^{|\mathcal{A}^2|}$, and $q \in \mathcal{S}$. Define payoff function $\hat{c}(s,\hat{a}^1,\hat{a}^2) = \sum_{a^1\in\mathcal{A}^1}\hat{a}^1(i)\log\left(\hat{a}^1(i)/\pi^{\mathsf{Av}}(s,a^1)\right)$ for all $s \in \mathcal{S}^+ \setminus \mathcal{S}^R$, $\hat{c}(s,\hat{a}^1,\hat{a}^2) = 0$ for all $s \in \mathcal{S}^R$, and $\hat{c}(s,\hat{a}^1,\hat{a}^2) = \infty$ for all $s \in \mathcal{S} \setminus \mathcal{S}^+$. We consider $\hat{\mathcal{G}}$ with the objective function $\hat{C}(\hat{\pi}^1,\hat{\pi}^2) = \mathbb{E}^{\hat{\pi}^1,\hat{\pi}^2}\left[\sum_{t=0}^{\tau}\hat{c}(s,\hat{a}_t^1,\hat{a}_t^2)\right]$ where Player 1 is the minimizer and Player 2 is the maximizer. Note that the payoff function is a convex function of $\hat{a}^1$ and a concave function of $\hat{a}^2$. We also note that by definition $\hat{C}(\hat{\pi}^1,\hat{\pi}^2) = \mathbb{E}^{\hat{\pi}^1,\hat{\pi}^2}\left[\sum_{t=0}^{\tau}\hat{c}(s,\hat{a}_t^1,\hat{a}_t^2)\right]$ is equal to the value of $C(\pi^1,\pi^2) = \mathbb{E}\left[\sum_{t=0}^{\tau}KL(\mu_t^1(s_t)||\pi^{\mathsf{Av}}(s_t))\right]$ if for all $s \in \mathcal{S}$, $t \geq 0$, we have $\hat{\mu}_t^1 = [\mu_t^1(s,1),\ldots,\mu_t^1(s,|\mathcal{A}^1|)]$ and $\hat{\mu}_t^2 = [\mu_t^2(s,1),\ldots,\mu_t^2(s,|\mathcal{A}^2|)]$.

Due to Lemma 6 and the above equivalence between the objective functions of $\mathcal{IC}$ and $\hat{\mathcal{G}}$, all prolonging policy pairs $(\hat{\pi}^{1,inf},\hat{\pi}^2)$ has an infinite objective value for Player 1. Similarly, due to Lemma 6, there exists a policy $\hat{\pi}^{1,fin}$ that incurs a finite objective value for Player 1 for all policies of Player 2. Note that by construction $\hat{\mathcal{G}}$ and $\hat{c}$, every policy pair $(\hat{\pi}^1,\hat{\pi}^2)$ with $\hat{C}(\hat{\pi}^1,\hat{\pi}^2) < \infty$ reaches $\mathcal{S}^R$ with probability 1. Also note that there exists a $\hat{\pi}^2$ for every $\hat{\pi}^1 \in \hat{\Pi}^1 \setminus \hat{\Pi}^{1,win}$ that makes $\hat{C}(\hat{\pi}^1,\hat{\pi}^2) = \infty$. Hence, we limit the feasible policies of Player 1 to $\hat{\Pi}^{1,win}$.

Since all prolonging policy pairs incur an infinite objective value for Player 1 and there exists a policy $\hat{\pi}^{1,fin}$ that incurs a finite objective value for Player 1, by Proposition 4.6 of (Patek & Bertsekas 1999), the equilibrium value is unique and there exists an equilibrium policy pair for $\hat{\mathcal{G}}$. Furthermore, there exists stationary pair $(\hat{\pi}^{1,*},\hat{\pi}^{2,*}) \in \hat{\Pi}^{1,St,win} \times \hat{\Pi}^{2,St}$ of policies which achieve an equilibrium, i.e.,

$$\sup_{\hat{\pi}^2\in\hat{\Pi}^2}\hat{C}(\hat{\pi}^1,\hat{\pi}^2) \leq \hat{C}(\hat{\pi}^{1,*},\hat{\pi}^{2,*}) \leq \inf_{\hat{\pi}^1\in\hat{\Pi}^{1,win}}\hat{C}(\hat{\pi}^1,\hat{\pi}^{2,*}).$$

We also note that the convexity of $\hat{c}(s_t,\hat{a}_t^1,\hat{a}_t^2)$ in $\hat{a}^1$ and the concavity in $\hat{a}^2$ implies that the deterministic policies suffice for Player 1 and Player 2 in $\hat{\mathcal{G}}$. Since there is a one-to-one mapping between the deterministic policies of $\hat{\mathcal{G}}$ and all policies of $\mathcal{IC}$, there also exists an equilibrium stationary policy pair for $\mathcal{IC}$, i.e., there exists

[1] The game model in (Patek & Bertsekas 1999) has metric action spaces and the convex-concave property of the payoff function is not needed. We consider a game model with a finite set of actions and need the convex-concave property to show that the deterministic policies are sufficient for the equivalent game model with the metric action spaces. Consequent the randomized policies are sufficient for the game model that we consider. We discuss this implication in the proof of Proposition 8.

$(\pi^{1,*}, \pi^{2,*}) \in \Pi^{1,St,win} \times \Pi^{2,St}$ such that

$$\sup_{\pi^2 \in \Pi^2} C(\pi^1, \pi^{2,*}) \leq C(\pi^{1,*}, \pi^{2,*}) \leq \inf_{\pi^1 \in \Pi^{1,win}} C(\pi^1, \pi^{2,*}).$$

Restricting the policy spaces to stationary policies yields the desired result. Note that $\mathcal{S}^R$ is eventually reached with probability 1 under this equilibrium policy pair since the occupancy measures at $\mathcal{S} \setminus \mathcal{S}^R$ are finite.

Finally, consider the infinite case. Since the KL objective function is infinite, we must have $s_0 \notin \mathcal{S}^+$ due to Lemma 6. Since there exists a stationary winning policy π^1, but $s_0 \notin \mathcal{S}^+$, it implies that all winning policies take an impermissible action with a positive probability. Let stationary policy π^2 for Player 2 be equal to $\pi^{2,trap}$ for the states in $\mathcal{S} \setminus \mathcal{S}^+$ and take actions uniformly randomly for the states in $\mathcal{S}^+$. Every $\pi^1 \in \Pi^{1,win}$ has $C(\pi^1, \pi^2) = \infty$ and $\Pr^{\pi^1,\pi^2}(\Diamond \mathcal{S}^R | s_0) = 1$. Hence, (π^1, π^2) is an equilibrium policy pair for every $\pi^1 \in \Pi^{1,win}$. □

Remark 9. For clarity of presentation, we restrict the feasible policy spaces of the players to $\Pi^{1,St,win}$ and $\Pi^{2,St}$. The equilibrium pair $(\pi^{1,*}, \pi^{2,*}) \in \Pi^{1,St,win} \times \Pi^{2,St}$ of policies from Proposition 8 also satisfy

$$\sup_{\pi^2 \in \Pi^2} C(\pi^1, \pi^{2,*}) \leq C(\pi^{1,*}, \pi^{2,*}) \leq \inf_{\pi^1 \in \Pi^{1,win}} C(\pi^1, \pi^{2,*}).$$

Knowing that the stationary policies suffice to find an equilibrium policy pair, we can represent the payoff of each step as a function of Player 1's policy and find a set of equilibrium policies via value iteration. Let $\pi^1(s)$ and $\pi^{\mathsf{Av}}(s)$ denote the vector of Player 1's and average player's policies at state s, respectively. Also, let $v(s)$ denote the payoff-to-go at state s such that $v(s) = 0$ for all $s \in \mathcal{S}^R$ and $v(s) = \infty$ for all $s \in \mathcal{S} \setminus \mathcal{S}^+$. By the first-order optimality conditions, for all $s \in \mathcal{S} \setminus \mathcal{S}^+$, we have

$$v(s) = \min_{\pi^1(s)} \Bigg(KL\left(\pi^1(s) || \pi^{\mathsf{Av}}(s)\right) + \max_{\pi^2(s)} \sum_{q \in \mathcal{S}} \sum_{\substack{a^1 \in \mathcal{A}^1(s) \\ a^2 \in \mathcal{A}^2(s)}} \mathcal{P}(s, a^1, a^2, q) \pi^1(s, a^1) \pi^2(s, a^2) v(q) \Bigg)$$

where the arguments of the minimum are the equilibrium policies for Player 1. Similarly, by the first-order optimality conditions, we can show that for all $s \in \mathcal{S} \setminus \mathcal{S}^+$, the equilibrium policies for Player 2 satisfy

$$\pi^2(s) = \arg\max_{\pi^{2'}(s)} \sum_{a^1 \in \mathcal{A}^1(s)} \pi^{\mathsf{Av}}(s, a^1) \exp\Bigg(\sum_{\substack{q \in \mathcal{S} \\ a^2 \in \mathcal{A}^2(s)}} \mathcal{P}(s, a^1, a^2, q) \pi^{2'}(s, a^2) v(q) \Bigg)^{-1}.$$

6 Offline Learning of Player 2's Policy

In this section, we give an algorithm to learn Player 2's policy and synthesize a near-optimal policy for Player 1.

Algorithm 1 takes the game model $\mathcal{IC}$ and n sample runs (with infinite lengths [2]) collected under the average player's policy (Line 1). A potentially winning state is *known* if there are a total of at least m sample transitions from that state in the sample runs (Line 6). Otherwise, the state is *unknown*. Let (i, j) denote the label of the transition in the i-th sample run at time j. For every known state s, we create an ascending order of the sample transitions from s where the index of the sample runs has a higher priority. An example ordering is $(1, 0), (1, 3), (2, 1)$. For every known state, Algorithm 1 estimates Player 2's policy using the first m action samples from that state as $\pi^2(s)$(Line 7). In Algorithm 1, we consider a modified game $\mathcal{IC}'$ (Lines 8-9) where the unknown states are also in the winning states. After constructing $\mathcal{IC}'$, we solve for the optimal winning policy $\pi^{1,'}$ when Player 2's policy is the estimated policy π^2 (Line 10). The output policy π^1 for the original game $\mathcal{IC}$ is history-dependent and uses one-bit of extra memory compared to a stationary policy. The memory bit tracks whether an unknown state has been visited yet. The output policy π^1 uses the optimal policy $\pi^{1,'}$ (synthesized in Line 10) against Player 2's estimated policy until reaching an unknown state. If an unknown state has been visited in the history, the output policy uses the average player's policy.

Algorithm 1 considers a modified game where the unknown states are also in the winning states. This game construction ensures that the optimal value for the modified game is lower than that of the original game when the respective sets of winning policies are considered. After reaching an unknown state, Player 1 follows the average player's policy and induces zero KL divergence payoff. The policy construction ensures that Player 1's policy has the same objective value for both the modified and the original games and, consequently, is near-optimal for the original game. On the other hand, after reaching an unknown state, Player 1 may not win the original game since it does not necessarily take safe actions. While the learned policy may not be a winning policy, we later show that reaching an unknown state and losing the game happens with a low probability.

We define some notation before discussing the properties of the algorithm. The equilibrium value $C(\pi^{1,*}, \pi^{2,*})$ of the game is denoted by v^*. Note that there is a unique value v^* due to Proposition 8. Player 2's true policy is $\pi^{2,\circ}$. For a state s, the total number of collected sample transitions from s is $\hat{m}_s$,

[2] Since Algorithm 1 utilizes only m samples per state, in practice, we need to store at most mS transitions.

Algorithm 1 Offline learning of Player 2's policy and policy optimization for Player 1

```
Input: An identity concealment game $\mathcal{IC}$, $n$ independent sample runs under $(\pi^{\mathsf{Av}},\pi^{2,\circ})$.
Output: A policy $\pi^1$ for Player 1.
$\mathcal{S}^K := \emptyset$.
for $s \in \mathcal{S}^+$ do
    if $\hat{m}_s \geq m$ then
        $\mathcal{S}^K := \mathcal{S}^K \cup \{s\}$ .
        Set $\pi^2(s)$ as the empirical distribution of first $m$ actions of Player 2 at state $s$.
$\mathcal{S}^U := \mathcal{S}^+ \setminus \mathcal{S}^K$, $\mathcal{S}^{end} := \mathcal{S}^U \cup \mathcal{S}^R$.
Construct a modified identity concealment game $\mathcal{IC}'$ that is the same as $\mathcal{IC}$ except that all states in $\mathcal{S}^{end}$ are absorbing, and $\mathcal{S}^{end}$ is the set of winning states.
For $\mathcal{IC}'$, synthesize the optimal stationary policy $\pi^{1,'}$ using the estimated policy $\pi^2(s)$ for all $s \in \mathcal{S}^K$.
for $t = 0, \ldots$ do
    For every $h_t = s_0 a_0^1 a_0^2 \ldots s_t$, define $\mu_t^1(h_t)$ such that $\mu_t^1(h_t) := \pi^{1,'}(s_t)$ if $s_i \in \mathcal{S}^K$ for all $0 \leq i \leq t$, and $\mu_t^1(h_t) := \pi^{\mathsf{Av}}(s_t)$ otherwise.
For $\mathcal{IC}$, define the policy $\pi^1 := \mu_0^1 \mu_1^1 \ldots$.
```

and the empirical action frequencies for Player 2 using only the first m samples drawn from $\pi^{2,\circ}$ is π^2. Player 1's optimal winning policy against $\pi^{2,\circ}$ is $\pi^{1,\circ}$, i.e., $\pi^{1,\circ} = \arg\min_{\pi^1 \in \Pi^{1,win}} C(\pi^1, \pi^{2,\circ}) = \arg\min_{\pi^1 \in \Pi^{1,win}} \mathbb{E}\left[\sum_{t=0}^{\tau} KL(\mu_t^1(h_t) || \pi^{\mathsf{Av}}(s_t)) | \pi^1, \pi^{2,\circ}\right]$. For $\mathcal{S}^+$, c_{max} denotes the maximum KL divergence between the safe action distributions for Player 1 and the action distribution for π^{Av}, i.e., $c_{max} = \max_{s \in \mathcal{S}^+, a^1 \in \mathcal{A}^1} (\log(\pi^{\mathsf{Av}}(s))^{-1}$ subject to a^1 is safe.

We have the following assumption on Player 2's policy. Assumption 10 ensures tractability of estimation for the transition probabilities.

Assumption 10. *$\pi^{2,\circ}$ is stationary on $\mathcal{S}$.*

Algorithm 1 satisfies the requirements given in Problem 4 in two steps: 1) The objective value incurred by π^1 is close to the optimal value for the known states since Player 2's policy will be estimated accurately for these states. For the unknown states, π^1 will incur 0 payoff since π^1 is the same with π^{Av} for these states. Overall, the KL objective value will be close to the optimal value under π^1. 2) If the number of sample runs is large enough, unknown states are reached with low probability under π^{Av}. If the unknown states are visited with high probability under π^1, then the objective value would be large since the deviation of π^1 from π^{Av} would be large. However, since the KL objective function is near optimal, the unknown states are visited with low probability under π^1, and the probability of losing is low for Player 1.

We define that a stationary policy pair (π^1, π^2) has an (L, β')-*contraction*, if $\min_{s \in \mathcal{S}^+} \Pr\left(\Diamond_{\leq L} \mathcal{S}^R \cup \mathcal{S}^U | s_0 = s\right) \geq 1 - \beta'$. To show the near optimality of the output policy, we make the following assumption, which ensures the finiteness of the expected length of a game run.

Assumption 11. *The policy pair (π^1, π^2) has an $\left(L, \beta - \frac{\varepsilon(1-\beta)^2}{c_{max} L}\right)$-contraction where β is a constant strictly lower than 1.*

The validity of Assumption 11 can be checked after the termination of the algorithm since both π^1 and π^2 are known. If the assumption is violated, one can increase β and rerun the algorithm. We remark that $\beta - (\varepsilon(1-\beta)^2)/(c_{max} L)$ is an increasing function of β, and the policy pair (π^1, π^2) has to have a $(S, \beta - (\varepsilon(1-\beta)^2)/(c_{max} S))$-contraction for some $\beta < 1$ since otherwise π^1 has to incur infinite payoff. Therefore, there always exists $\beta < 1$ that satisfies the assumption. We note that having a contraction, e.g., a discount factor, is a common assumption in reinforcement learning to ensure the boundedness of the objective function (Sutton & Barto 2018).

The following proposition shows that Algorithm 1 indeed results in a near-optimal policy using only the game runs collected under the average player's policy.

Proposition 12. *Let $w = (v^* + \log(2) + \varepsilon)/\lambda$. Under Assumptions 1, 10, and 11, if*

$$m \geq \frac{4c_{max}^2 L^4 \left(2\log(2)A + \log\left(2S/\delta\right)\right)}{(1-\beta)^4 \varepsilon^2}$$

and

$$n \geq e^{2w} \log\left(4/\delta\right)/2 + 2Se^{w}m$$

in Algorithm 1, then the output policy π^1 satisfies

$$C(\pi^1, \pi^{2,\circ}) - C(\pi^{1,\circ}, \pi^{2,\circ}) \leq \varepsilon$$

and

$$\Pr^{\pi^1, \pi^{2,\circ}}(\Diamond \mathcal{S}^R | s_0) \geq 1 - \lambda,$$

with probability at least $1 - \delta$.

We use a series of lemmas to prove Proposition 12. Lemma 13 shows that with high probability, the estimated action distribution π^2 and the actual action distribution $\pi^{2,\circ}$ are close for all known states[3]. The proof follows Sanov's theorem and Pinsker's inequality combined with the union bound (Weissman et al. 2003).

[3] One can use all available sample transitions instead of the first m samples. While this approach yields a concentration bound of the same order (see Lemma 3 of (Karabag & Topcu 2018)) and may improve the estimates for some states, we use only the first m samples for every state since the performance bound given in Lemma 14 requires uniform coverage.

Lemma 13. *For any $\delta_K \in (0, 1]$, given m independent transitions sampled from $\pi^{2,\circ}(s)$ for each $s \in \mathcal{S}^K$, with probability at least $1 - \delta_K$,*

$$\left\|\pi^{2,\circ}(s) - \pi^2(s)\right\|_1 \leq \sqrt{\frac{2(\log(2)A + \log(S/\delta_K))}{m}}$$

for all $s \in \mathcal{S}^K$.

Lemma 14 shows that if the estimated and actual transition probability distributions are close, and the policy pair $(\pi^1, \pi^{2,\circ})$ has an L-step contraction, then the values of the objective function are close for $(\pi^1, \pi^{2,\circ})$ and (π^1, π^2). The paper (Strehl & Littman 2008) showed this property for $(1, \beta)$-contractions. We extend this result to (L, β)-contractions and show that difference between the value functions is bounded by representing the KL objective as a sum of payoffs per time step. Since (π^1, π^2) has L-step contraction lower than or equal to $\beta - \varepsilon(1-\beta)^2/(c_{max}L)$ and $\|\pi^{2,\circ}(s) - \pi^2(s)\|_1 \leq \varepsilon(1-\beta)^2/(c_{max}L^2)$ for all $s \in \mathcal{S}^K$, then $(\pi^1, \pi^{2,\circ})$ has (L, β)-contraction. Since $(\pi^1, \pi^{2,\circ})$ has (L, β)-contraction, the KL objective value is bounded by $\sum_{i=0}^{\infty} Lc_{max}\beta^i = Lc_{max}/(1-\beta)$ from every initial state in $\mathcal{S}^+ \setminus \mathcal{S}^R$. Because the estimated and true transition probabilities are close as in Lemma 14, and $(\pi^1, \pi^{2,\circ})$ has (L, β)-contraction, the $\|\Gamma^{\pi^1,\pi^2} - \Gamma^{\pi^1,\pi^{2,\circ}}\|_1$ is bounded by $\varepsilon(1-\beta)/(2Lc_{max})$. Since the KL objective value is bounded from every initial state and the distributions of game runs induced by (π^1, π^2) and $(\pi^1, \pi^{2,\circ})$ are close to each other, the KL objective differs by at most $\varepsilon/2$.

Lemma 14. *If $\|\pi^{2,\circ}(s) - \pi^2(s)\|_1 \leq \frac{\varepsilon(1-\beta)^2}{c_{max}L^2}$ for all $s \in \mathcal{S}^K$,*

$$|C(\pi^1, \pi^2) - C(\pi^1, \pi^{2,\circ})| \leq \frac{\varepsilon}{2}.$$

The following lemmas show that the probability of losing is low if the number of sample trajectories is high. Lemma 15 shows that if a state is unknown, then the probability of reaching that state is low. The proof is an application of the Chernoff-Hoeffding inequality. We use the fact that number of collected action samples from a state is higher than the number of sample runs that visit the state. Since the unknown states does not have enough sample transitions, it implies that these states are visited with a low probability.

Lemma 15. *Let $\hat{m}_D$ denote the number of transitions from set D of states using n runs independently sampled under policies $(\pi^{\mathsf{Av}}, \pi^{2,\circ})$. For $m' \geq \hat{m}_D$ and $1/2 \geq \sigma > m'/n$, with probability at least $1 - 2\exp(-2n(\sigma - m'/n)^2)$, we have $\Pr^{\pi^{\mathsf{Av}},\pi^{2,\circ}}(\lozenge D|s_0) \leq \sigma$.*

Lemma 16 shows that if a state is reached with high probability under π^1 and with low probability under π^{Av}, then the value of the objective function is high. The proof follows from that the KL divergence between the distributions of game runs is lower bounded by the KL divergence between the reachability probability to set D by the data processing inequality. Since the unknown states are visited with low probability under π^{Av}, visiting these states with high probability causes deviations from the average player and incurs a high value for the KL objective function.

Lemma 16. *Let $D \subseteq \mathcal{S}$. If $\Pr^{\pi^1,\pi^{2,\circ}}(\lozenge D|s_0) > \frac{-(v^*+\log(2)+\varepsilon)}{\log\left(\Pr^{\pi^{\mathsf{Av}},\pi^{2,\circ}}(\lozenge D|s_0)\right)}$, then $C(\pi^1, \pi^{2,\circ}) > v^* + \varepsilon$.*

The proof of Proposition 12 consists of two parts. We first show that the learned policy π^1 is near optimal. At the unknown states, π^1 is the same with the average player's policy π^{Av} and incurs 0 payoff. Consider a modified identity concealment game where the unknown states are included in the winning states. The equilibrium value of the modified game is less compared to the original game. For the known states, Player 2's estimated policy will be close to the true policy due to Lemma 13. Since the estimated policy is accurate, π^1 is near optimal for the modified game. Thus, π^1 is near optimal due to Lemma 14. To show that the probability of losing is small, we use Lemmas 15 and 16. Since the ratio between the numbers of sample paths and sample transitions is high enough, Lemma 15 implies that the probability of reaching an unknown state is bounded under π^{Av}. Since π^1 is near optimal, the probability of reaching an unknown state is also bounded under the learned policy due to Lemma 16.

Proof of Proposition 12. At time t define $\pi^{1,\triangleleft}$ such that $\pi^{1,\triangleleft}(s) := \pi^{1,\circ}(s)$ if $s_i \in \mathcal{S}^K$ for all $0 \leq i < t$, and $\pi^{1,\triangleleft}(s) := \pi^{\mathsf{Av}}(s)$ otherwise. For notational convenience, define $w := (v^* + \log(2) + \varepsilon)/\lambda$.

By Lemma 13, if $m \geq \frac{4c_{max}^2 L^4\left(2\log(2)|\mathcal{S}^2| + \log\left(\frac{2}{\delta}\right)\right)}{(1-\beta)^4\varepsilon^2}$ in Algorithm 1, then with probability at least $1 - \delta/2$, we have $\left\|\pi^{2,\circ}(s) - \pi^2(s)\right\|_1 \leq \varepsilon(1-\beta)^2/(c_{max}L^2)$ for all $s \in \mathcal{S}^K$ [4]. Then by Lemma 14, we have $|C(\pi^{1,\triangleleft}, \pi^{2,\circ}) - C(\pi^{1,\triangleleft}, \pi^2)| \leq \varepsilon/2$ and $|C(\pi^1, \pi^2) - C(\pi^1, \pi^{2,\circ})| \leq \varepsilon/2$ with probability at least $1 - \delta/2$. Since $C(\pi^1, \pi^2) \leq C(\pi^{1,\triangleleft}, \pi^2)$ due to the optimality of π^1 against π^2, we have $|C(\pi^{1,\triangleleft}, \pi^{2,\circ}) - C(\pi^1, \pi^{2,\circ})| \leq \varepsilon$ with probability at least $1 - \delta/2$. We also have that $C(\pi^{1,\triangleleft}, \pi^{2,\circ}) \leq$

[4] The concentration bound given in in the lemma requires independent transitions. However, the transition in the sample runs may not be independent in general. Despite the dependent samples, the concentration bound can still be used for our analysis. A detailed discussion on the dependence of sample transitions and the use of this bound is given in (Strehl & Littman 2008). Alternatively, one can use a concentration bound that can handle dependent transitions and random stopping times (e.g., Lemma 3 of (Karabag & Topcu 2018)) and get the same order of convergence.

$C(\pi^{1,\circ}, \pi^{2,\circ})$ since $\pi^{1,\circ}$ and $\pi^{1,\triangleleft}$ induce the same payoff until reaching an unknown state, and $\pi^{1,\circ}$ induces a nonnegative payoff after reaching an unknown state whereas $\pi^{1,\triangleleft}$ induces 0 payoff after reaching an unknown state. Consequently, with probability at least $1-\delta/2$, we have

$$C(\pi^1, \pi^{2,\circ}) \leq C(\pi^{1,\circ}, \pi^{2,\circ}) + \varepsilon.$$

We now show that if $n/m \geq e^{2w} \log\left(4/\delta\right)/2m + 2Se^w$ and m is as above, $\Pr^{\pi^1,\pi^2}(\Diamond \mathcal{S}^R | s_0) \geq 1-\lambda$ with probability at least $1-\delta/2$. Let $\mathcal{S}^U$ be the set of unknown states. Define $y := e^w \log\left(4/\delta\right)/(2mS)$ and $c := y+2$. We have $n/m \geq Sce^w$. Also define $c' := \left(y + \sqrt{y}\sqrt{y+4}+2\right)/2$. Note that $y \geq 0$ and $c \geq c' \geq 1$. The number of sample transitions from $\mathcal{S}^U$ is lower than mS by the definition. By Lemma 6, with probability at least $1 - 2\exp\left(-2n(1-1/c)^2 e^{-2w}\right)$, we have $\Pr^{\pi^{\mathsf{Av}},\pi^{2,\circ}}(\Diamond \mathcal{S}^U | s_0) \leq e^{-w}$. Since $c \geq c' \geq 1$, we have $2\exp\left(-2n(\frac{c-1}{c})^2 e^{-2w}\right) \leq 2\exp\left(-2mSc'(\frac{c'-1}{c'})^2 e^{-w}\right) = \delta/2$. Thus, we have $\Pr^{\pi^{\mathsf{Av}},\pi^{2,\circ}}(\Diamond \mathcal{S}^U | s_0) \leq e^{-w}$ with probability at least $1-\delta/2$.

If $C(\pi^1, \pi^{2,\circ}) \leq C(\pi^{1,\circ}, \pi^{2,\circ}) + \varepsilon$, we have $C(\pi^1, \pi^{2,\circ}) \leq v^* + \varepsilon$ since $C(\pi^{1,\circ}, \pi^{2,\circ}) \leq v^*$. By Lemma 7, the probability $\Pr^{\pi^1,\pi^{2,\circ}}(\Diamond \mathcal{S}^U | s_0) \leq \lambda$ with probability at least $1-\delta$ since $\Pr^{\pi^{\mathsf{Av}},\pi^{2,\circ}}(\Diamond \mathcal{S}^U | s_0) \leq e^{-w}$ with probability at least $1-\delta/2$. Since Player 1 can lose the game only by reaching an unknown state, the probability of losing is at most λ with probability at least $1-\delta/2$.

Combining the near-optimality result for the objective function and the result for the probability of losing, we conclude that π^1 satisfies

$$C(\pi^1, \pi^{2,\circ}) \leq C(\pi^{1,\circ}, \pi^{2,\circ}) + \varepsilon.$$

and

$$\Pr^{\pi^1,\pi^2}(\Diamond \mathcal{S}^R | s_0) \geq 1-\lambda$$

with probability at least $1-\delta$. □

7 Numerical Examples

In this section, we give numerical examples of the equilibrium policies for identity concealment games and offline policy optimization for Player 1.

7.1 Detection of Hostile Clients in Cyber Interactions

We show the effect of identity concealment on the detection of hostile clients in the cyber interaction scenario shown in Figure 1. The game is played between a client (Player 1) and the server (Player 2), and the states represent the remaining times for the client's processed requests, if there are any. At every time, the client can disconnect, make a request, or wait. The server can accept or reject the client's potential request. However, the server cannot reject the request again if the client has been rejected previously for that request. If the request is accepted, it takes a certain number of steps to process the request. The attacker's goal is to cause a denial of service by overwhelming the server, and it wins the game if and only if the server concurrently processes multiple requests of the client. At every state, the real clients' policy, i.e, the average player's policy, is randomized, and is more likely to make a request if there are no requests being processed or there is a rejected request. The details of the setting are given in (Karabag et al. 2021*b*).

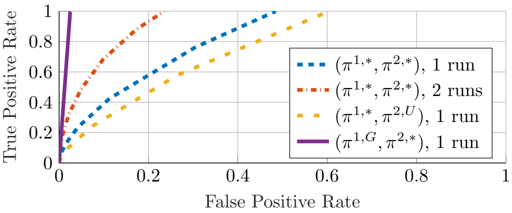


Figure 3. Receiver operating characteristic curve of the likelihood ratio classifier that identifies hostile clients. True positive rate is the ratio of detected attackers to all attackers. False positive rate is the ratio of real clients identified as an attacker to all clients.

In Figure 3, we observe that when the server uses its equilibrium policy $\pi^{2,*}$, hostile clients are identified with high accuracy compared to policy $\pi^{2,U}$ that accepts or rejects the requests with equal probabilities. This is because, unlike $\pi^{2,U}$, the equilibrium policy $\pi^{2,*}$ is state-dependent, and using $\pi^{2,*}$ the server can drive the game into a state where the hostile client's behavior is different from the real clients' behaviors. Similarly, a hostile client is less likely to be detected when it uses its equilibrium policy $\pi^{1,*}$ compared to the greedy policy $\pi^{1,G}$ that makes a request at every time step. We also observe that an additional interaction, i.e., a game run, improves the accuracy of classification as explained in Section 4.

7.2 Equilibrium Policies for a Pursuit-Evasion Game

We show the behavior for hostile Player 1 in a pursuit-evasion game. Player 1 is an evader and Player 2 is a pursuer. The environment is a two-dimensional grid where each node represents an intersection. At each time step, every intersection is occupied with probability 0.5. If the pursuer's intersection is clear, it can move in $+x$, $-x$, $+y$, $-y$ directions by 1 or stay at its current intersection. If the intersection is occupied, the pursuer stays. Regardless of the state of its intersection, the evader can move in all directions by 1 or 2 blocks, or stay at the current intersection. We encode the states of the game using the relative distances on x and y coordinates between the players, the states of the evader's intersection, and the state of the pursuer's intersection. If the block distance between the players is 6 or more the evader wins

Table 1
The policies of the average and hostile evaders. O denotes an occupied intersection and C denotes a clear intersection.

Evader Type	State	Probability of Action								
		Stay	x	2x	y	2y	-y	-2y	-x	-2x
Average	Intersection Occupied	0.80	0.10	0	0.05	0	0.05	0	0	0
	Intersection Clear	0	0.40	0.10	0.20	0.05	0.20	0.05	0	0
Hostile	(2,0,C,C)	0	0.41	0.23	0.12	0.06	0.12	0.06	0	0
	(1,0,C,C)	0	0.31	0.33	0.09	0.09	0.09	0.09	0	0
	(1,0,O,C)	0	0.62	0	0.19	0	0.19	0	0	0
	(1,0,O,O)	0.74	0.16	0	0.05	0	0.05	0	0	0

the game, and if the block distance is 0, the pursuer wins the game, i.e., traps the average evader.

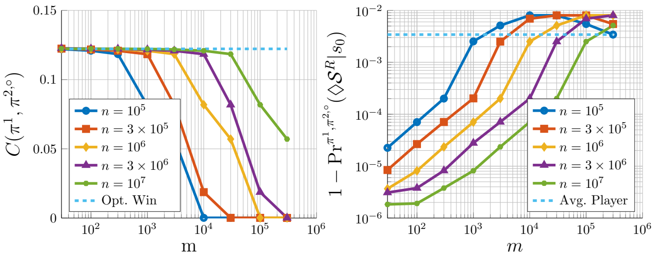
The average player's policy only depends on whether its intersection is occupied and is given in Table 1. If the intersection is occupied, with high probability the average player waits the intersection to be cleared and with low probability the average player "illegally" crosses the intersection. If its intersection is clear, with high probability the average player moves with a normal speed, otherwise it "overspeeds", i.e., moves by 2 blocks.

The equilibrium policy of the hostile evader is given in Table 1 for some states. If the evader is close to the pursuer, for instance at state $(1, 0, C, C)$, the evader tends behave more suspiciously by "overspeeding". When the distance is higher, for instance at state $(2, 0, C, C)$, the evader does not flee and behaves more closely to the average player. Similarly, the hostile evader chooses to move at state $(1, 0, O, C)$ since otherwise the evader may get caught. On the other hand, the hostile evader chooses to stay at state $(1, 0, O, O)$ with high probability since the pursuer has to wait. Overall, the hostile evader behaves dissimilar to the average player only if it is in a dangerous position, which matches the intuition of concealing the identity by behaving in an unsuspicious way.

7.3 *Offline Learning of the Pursuer's Policy*

In this example, we show the empirical performance of the proposed offline learning algorithm for different number of sample runs n and number of estimation samples m per state. Note that we do not give optimality guarantees for the demonstrated values of m and n. We use the same environment with the previous example where the initial state s_0 is $(1, 0, O, O.)$. The pursuer's policy $\pi^{2,\circ}$ is defined as follows. At each time step the pursuer stops tracking the evader, and the evader wins with probability 0.2. If the pursuer does not stop, it takes allowed actions with uniform probabilities.

In Figure 4a, we observe that the evader is able to learn the pursuer's suboptimal policy and lower the objective function compared to the equilibrium value of the game.



(a) (b)

Figure 4. The value of the objective function and the probability of losing for different values of m and n. The dashed line in (a) marks the value of the objective function for the optimal winning policy. The dashed line in (b) marks the probability of losing under the average player's policy.

For lower values of n/m, the value of the objective function is lower than the value of the objective function under the optimal safe policy. If n/m is lower, then fewer states become known and the hostile evader reaches unknown states with higher probabilities. Resultingly, the evader follows the average player's policy and incurs 0 payoff, which lowers the value. When $m = 3 \times 10^5$ and $n = 10^5$, all states are unknown, and the output policy is equal to the average player's policy. In Figure 4b, if n/m is low, then the probability of losing is high for the hostile evader since it follows the average player's policy with high probability. In fact, for some values of n/m the probability of losing is higher than the probability that the average player loses the game. This result matches the intuition behind Lemma 16 and the n/m ratio given in Proposition 12: The learned policy may reach unknown states with higher probability compared to the average evader's policy, and to ensure that the probability of losing is low, the n/m should be sufficiently high.

8 Conclusion

We formalized the notion of identity concealment zero-sum games and defined identity concealment games. We showed that there exists a stationary equilibrium policy pair for identity concealment games. We then showed that a hostile player can learn a near optimal policy if the opponent is not following an equilibrium policy. In detail, we presented an algorithm that solely uses a finite number of game runs collected under the average player's policy. The output of the algorithm is a policy for the player that guarantees near optimality in the identity concealment objective and the probability of winning.

Acknowledgements

This work was supported in part by the Air Force Research Laboratory, USA under award number FA9550-19-1-0169, the Army Research Office, USA under award number W911NF-23-1-0317, the Defense Advanced Research Projects Agency under award number

D19AP00004, and the Office of Naval Research under award number N00014-23-1-2651.

References

Baier, C. & Katoen, J.-P. (2008), *Principles of Model Checking*, MIT Press.

Başar, T. & Olsder, G. J. (1998), *Dynamic noncooperative game theory*, SIAM.

Boularias, A., Kober, J. & Peters, J. (2011), Relative entropy inverse reinforcement learning, *in* 'Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics', pp. 182–189.

Chatterjee, K., Henzinger, T. A. & Piterman, N. (2008), 'Algorithms for Büchi games', *arXiv preprint arXiv:0805.2620* .

Chen, T., Forejt, V., Kwiatkowska, M., Simaitis, A. & Wiltsche, C. (2013), On stochastic games with multiple objectives, *in* 'International Symposium on Mathematical Foundations of Computer Science', Springer, pp. 266–277.

Cover, T. M. & Thomas, J. A. (2012), *Elements of Information Theory*, John Wiley & Sons.

Farajtabar, M., Chow, Y. & Ghavamzadeh, M. (2018), More robust doubly robust off-policy evaluation, *in* 'International Conference on Machine Learning', PMLR, pp. 1447–1456.

Fiechter, C.-N. (1994), Efficient reinforcement learning, *in* 'Proceedings of the Seventh Annual Conference on Computational Learning Theory', ACM, pp. 88–97.

Fox, R., Pakman, A. & Tishby, N. (2016), Taming the noise in reinforcement learning via soft updates, *in* 'Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence', AUAI Press, pp. 202–211.

Grau-Moya, J., Leibfried, F. & Bou-Ammar, H. (2018), Balancing two-player stochastic games with soft Q-learning, *in* 'Proceedings of the 27th International Joint Conference on Artificial Intelligence', AAAI Press, pp. 268–274.

Hecker, C. (2018), 'SpyParty', http://www.spyparty.com/. Online.

Hogg, R. V., Tanis, E. A. & Zimmerman, D. L. (1977), *Probability and statistical inference*, Vol. 993, Macmillan New York.

Karabag, M. O., Ornik, M. & Topcu, U. (2021*a*), 'Deception in supervisory control', *IEEE Transactions on Automatic Control* .

Karabag, M. O., Ornik, M. & Topcu, U. (2021*b*), 'Identity concealment games: How I learned to stop revealing and love the coincidences', *arXiv preprint arXiv:2105.05377* .

Karabag, M. O. & Topcu, U. (2018), On the sample complexity of vanilla model-based offline reinforcement learning with dependent samples, *in* 'Thirty-Seventh AAAI Conference on Artificial Intelligence', AAAI Press, pp. 8195–8202.

Kardes, E. & Hall, R. (2005), 'Survey of literature on strategic decision making in the presence of adversaries'.

Kearns, M. & Singh, S. (2002), 'Near-optimal reinforcement learning in polynomial time', *Machine learning* **49**(2-3), 209–232.

Keren, S., Gal, A. & Karpas, E. (2016), Privacy preserving plans in partially observable environments., *in* 'Proceedings of the 25th International Joint Conference on Artificial Intelligence', pp. 3170–3176.

Kidambi, R., Rajeswaran, A., Netrapalli, P. & Joachims, T. (2020), 'Morel: Model-based offline reinforcement learning', *Advances in neural information processing systems* **33**, 21810–21823.

Kulkarni, A., Srivastava, S. & Kambhampati, S. (2019), A unified framework for planning in adversarial and cooperative environments, *in* 'Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence', pp. 2479–2487.

Levine, S., Kumar, A., Tucker, G. & Fu, J. (2020), 'Offline reinforcement learning: Tutorial, review, and perspectives on open problems', *arXiv preprint arXiv:2005.01643* .

Macintyre, B. (2018), *The spy and the traitor: the greatest espionage story of the Cold War*, Viking.

Newman, G. (2015), 'Garry's Mod Guess Who', https://gmod.facepunch.com/. Online.

Patek, S. D. & Bertsekas, D. P. (1999), 'Stochastic shortest path games', *SIAM Journal on Control and Optimization* **37**(3), 804–824.

Peters, J., Mulling, K. & Altun, Y. (2010), Relative entropy policy search, *in* 'Twenty-Fourth AAAI Conference on Artificial Intelligence'.

Precup, D., Sutton, R. S. & Singh, S. P. (2000), Eligibility traces for off-policy policy evaluation, *in* 'Proceedings of the Seventeenth International Conference on Machine Learning', pp. 759–766.

Puterman, M. L. (2014), *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons.

Ross, S. & Bagnell, D. (2012), Agnostic system identification for model-based reinforcement learning, *in* 'Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012', icml.cc / Omnipress. **URL:** *http://icml.cc/2012/papers/833.pdf*

Schulman, J., Levine, S., Abbeel, P., Jordan, M. & Moritz, P. (2015), Trust region policy optimization, *in* 'International conference on machine learning', PMLR, pp. 1889–1897.

Strehl, A. L., Li, L. & Littman, M. L. (2009), 'Reinforcement learning in finite MDPs: PAC analysis', *Journal of Machine Learning Research* **10**(Nov), 2413–2444.

Strehl, A. L. & Littman, M. L. (2008), 'An analysis of model-based interval estimation for Markov decision processes', *Journal of Computer and System Sciences* **74**(8), 1309–1331.

Sutton, R. S. & Barto, A. G. (2018), *Reinforcement learning: An introduction*, MIT press.

Uehara, M. & Sun, W. (2021), 'Pessimistic model-based offline reinforcement learning under partial coverage', *arXiv preprint arXiv:2107.06226* .

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S. & Weinberger, M. J. (2003), 'Inequalities for the l1 deviation of the empirical distribution', *Hewlett-Packard Labs, Tech. Rep* .

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C. & Ma, T. (2020), 'Mopo: Model-based offline policy optimization', *Advances in Neural Information Processing Systems* **33**, 14129–14142.

A The Proofs for Technical Results

Complete versions of the proof sketches are available at (Karabag et al. 2021*b*) due to the lack of space.

We use technical Lemmas 17 and 18 to prove Lemma 7.

Lemma 17. *Let $\mathcal{D}$ be a discrete probability distribution such that $\mathcal{D}(n) \geq 0$ if $n \in \mathbb{N}$ and $\mathcal{D}(n) = 0$ otherwise, and let $c_1, c_2 \in (0, \infty)$ be arbitrary constants. Define set D such that $n \in D$ if and only if $\mathcal{D}(n) > c_1 \exp(-nc_2)$. If $\sum_{n=0}^{\infty} \mathcal{D}(n)n = \infty$, we have*

$$\sum_{n \in D} \mathcal{D}(n) \log \left(\frac{\mathcal{D}(n)}{c_1 \exp(-nc_2)} \right) = \infty.$$

Lemma 18. *For all $n \geq 0$, the optimal value of $\min_{x,y \in \mathbb{R}^n} KL(x||y)$ subject to $0 \leq x_i \leq y_i$ for all $i \in [n]$ and $\sum_{i=1}^{n} y_i \leq c$, is $-c \exp(-1)$.*

Proof of Lemma 7. We partition $\mathbb{N}$ into three disjoint sets D_1, D_2, and D_3 where $n \in D_1$ if $\mathcal{D}^1(n) \leq \mathcal{D}^2(n) \leq c_1 \exp(-c_2 n)$, $n \in D_2$ if $\mathcal{D}^2(n) < \mathcal{D}^1(n) \leq c_1 \exp(-c_2 n)$, and $n \in D_3$ if $\mathcal{D}^2(n) \leq c_1 \exp(-c_2 n) < \mathcal{D}^1(n)$.

We first lower bound the KL divergence on subsets D_1 and D_2. For subset D_1 we have

$$\sum_{n \in D_1} \mathcal{D}^2(n) \leq \sum_{n=0}^{\infty} \mathcal{D}^2(n) \leq \sum_{n=0}^{\infty} \frac{c_1}{\exp(c_2 n)} = \frac{c_1}{\exp(c_2) - 1}.$$

By Lemma 18, we have

$$\sum_{n \in D_1} \mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{\mathcal{D}^2(n)} \right) \geq -\frac{c_1 \exp(-1)}{\exp(c_2) - 1} \qquad \text{(A.1)}$$

since $\mathcal{D}^1(n) \leq \mathcal{D}^2(n)$ for all $n \in D_1$ and $\sum_{n \in D_1} \mathcal{D}^2(n) \leq \frac{c_1}{\exp(c_2)-1}$. For subset D_2 we have

$$\sum_{n \in D_2} \mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{\mathcal{D}^2(n)} \right) \geq 0 \qquad \text{(A.2)}$$

since $\mathcal{D}^2(n) < \mathcal{D}^1(n)$ and consequently $\mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{\mathcal{D}^2(n)} \right) > 0$ for all $n \in D_2$.

Therefore, $KL(\mathcal{D}^1 || \mathcal{D}^2)$ is equal to

$$\begin{aligned}
&\sum_{n=0}^{\infty} \mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{\mathcal{D}^2(n)} \right) && \text{(A.3a)} \\
&= \sum_{n \in D_1} \mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{\mathcal{D}^2(n)} \right) + \sum_{n \in D_2} \mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{\mathcal{D}^2(n)} \right) \\
&+ \sum_{n \in D_3} \mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{\mathcal{D}^2(n)} \right) && \text{(A.3b)} \\
&\geq -\frac{c_1 \exp(-1)}{\exp(c_2) - 1} + \sum_{n \in D_3} \mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{\mathcal{D}^2(n)} \right) && \text{(A.3c)} \\
&\geq -\frac{c_1 \exp(-1)}{\exp(c_2) - 1} + \sum_{n \in D_3} \mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{c_1 \exp(-c_2 n)} \right) && \text{(A.3d)}
\end{aligned}$$

where (A.3c) is due to (A.1) and (A.2), and (A.3d) is due to $\mathcal{D}^2(n) \leq c_1 \exp(-c_2 n)$.

We note that $n \in D_3$ if and only if $\mathcal{D}^1(n) > c_1 \exp(-c_2 n)$, and $\sum_{n=0}^{\infty} \mathcal{D}^1(n) n = \infty$. By Lemma 17, we have $\sum_{n \in D_3} \mathcal{D}^1(n) \log \left(\frac{\mathcal{D}^1(n)}{c_1 \exp(-c_2 n)} \right) = \infty$. Therefore, $KL(\mathcal{D}^1 || \mathcal{D}^2) = \infty$. □

Proof sketch for Lemma 13. By Lemma 14 of (Strehl et al. 2009), with probability at least $1 - \delta_k / S$, we have $\|\pi^2(s) - \pi^{2,*}(s)\|_1 \leq \sqrt{\frac{2(\log(2^A - 2) + \log(S/\delta_k))}{m}}$. Combining this with a union bound over $\mathcal{S}^K$, we get the desired result. □

Proof sketch for Lemma 14. We first establish that if the policy pair (π^1, π^2) has $\left(L, \beta - \frac{\varepsilon(1-\beta)^2}{c_{max} L}\right)$-contraction, and $\|\pi^2(s) - \pi^{2,\circ}(s)\|_1 \leq \frac{\varepsilon(1-\beta)^2}{c_{max} L^2}$ for all $s \in \mathcal{S}^K$, then $(\pi^1, \pi^{2,\circ})$ has (L, β)-contraction. We show this property by induction, noting that the flow difference under these two policies is at most $\frac{\varepsilon(1-\beta)^2}{c_{max} L}$ at every L-steps. Next, we show $|C(\pi^1, \pi^2) - C(\pi^1, \pi^{2,\circ})| \leq \varepsilon/2$. Since $\|\pi^2(s) - \pi^{2,\circ}(s)\|_1 \leq \frac{\varepsilon(1-\beta)^2}{c_{max} L^2}$, the flow difference under policy pairs (π^1, π^2) and $(\pi^1, \pi^{2,\circ})$ is bounded by $\frac{\varepsilon(1-\beta)}{2 c_{max} L}$. Since $(\pi^1, \pi^{2,\circ})$ has (L, β)-contraction, the different flow eventually reaches an end state and incurs 0 payoff. Due to (L, β)-contraction and the bounded payoff c_{max}, this flow difference incurs at most $\varepsilon/2$ difference in the value functions. □

Proof of Lemma 15. Let $\hat{m}_D^{unq}$ denote the number of sample runs that contain a transition from D. By Chernoff's inequality, we have

$$\begin{aligned}&\Pr\left(\left|\Pr^{\pi^{\mathsf{Av}},\pi^{2,*}}(\Diamond D|s_0) - \hat{m}_D^{unq}/n\right| \geq \sigma - \hat{m}_D^{unq}/n\right)\\ &\quad\leq 2\exp\left(-n\left(\sigma - \hat{m}_D^{unq}/n\right)^2/2\right)\end{aligned}$$

where the outer probability is over the randomness of sample paths.

Note that $\hat{m}_D^{unq} \leq \hat{m}_D \leq m'$ since $\hat{m}_D^{unq}$ is the number of paths with a transition from D and $\hat{m}_D$ is the total number of transitions from D. Therefore, we have

$$\begin{aligned}&\Pr\left(\Pr^{\pi^{\mathsf{Av}},\pi^{2,*}}(\Diamond D|s_0) \geq \sigma\right)\\ &\quad\leq \Pr\left(\left|\Pr^{\pi^{\mathsf{Av}},\pi^{2,*}}(\Diamond D|s_0) - \hat{m}_D^{unq}/n\right| \geq \sigma - \hat{m}_D^{unq}/n\right)\\ &\quad\leq 2\exp\left(-n\left(\sigma - \hat{m}_D^{unq}/n\right)^2/2\right)\\ &\quad\leq 2\exp\left(-n\left(\sigma - m'/n\right)^2/2\right)\end{aligned}$$

which yields the desired result. □

Proof of Lemma 16. Let $\rho^1 = \Pr^{\pi^1,\pi^{2,\circ}}(\Diamond D|s_0)$ and $\rho^{\mathsf{Av}} = \Pr^{\pi^{\mathsf{Av}},\pi^{2,\circ}}(\Diamond D|s_0)$. Note that $C(\pi^1,\pi^{2,\circ}) = KL\left(\pi^1,\pi^{2,\circ}||\pi^{\mathsf{Av}},\pi^{2,\circ}\right) \geq KL\left(Ber\left(\rho^1\right)||Ber\left(\rho^{\mathsf{Av}}\right)\right)$ due to the data processing inequality. Therefore, it suffices to show that if $\rho^1 > \rho^{\mathsf{Av}}$ and $\rho^1 > \frac{v^*+\log(2)+\varepsilon}{-\log(\rho^{\mathsf{Av}})}$, then $KL\left(Ber\left(\rho^1\right)||Ber\left(\rho^{\mathsf{Av}}\right)\right) > v^* + \varepsilon$.

We have

$$\begin{aligned}&KL\left(Ber\left(\rho^1\right)||Ber\left(\rho^{\mathsf{Av}}\right)\right) &\text{(A.6a)}\\ &= \rho^1\log\left(\frac{\rho^1}{\rho^{\mathsf{Av}}}\right) + \left(1-\rho^1\right)\log\left(\frac{1-\rho^1}{1-\rho^{\mathsf{Av}}}\right) &\text{(A.6b)}\\ &\geq \rho^1\log\left(\frac{\rho^1}{\rho^{\mathsf{Av}}}\right) + \left(1-\rho^1\right)\log\left(1-\rho^1\right) &\text{(A.6c)}\\ &\geq -\rho^1\log\left(\rho^{\mathsf{Av}}\right) - \log(2) &\text{(A.6d)}\end{aligned}$$

where (A.6d) is because of that $\min x\log(x) + (1-x)\log(1-x)$ subject to $x \in [0,1]$ is $-\log(2)$. We get the desired result by rearranging the terms in (A.6d). □

Proof of Lemma 17. Fix arbitrary constants $c_1, c_2 \in (0,\infty)$. We partition $\mathbb{N}$ into three disjoint subsets D_1, D_2, and D_3 such that $n \in D_1$ if and only if $\mathcal{D}(n) \leq c_1\exp(-nc_2)$, $n \in D_2$ if and only if $c_1\exp(-nc_2) \leq \mathcal{D}(n) < c_1\exp(-nc_2/2)$, and $n \in D_3$ otherwise. Also define $D := D_2 \cup D_3$.

$\sum_{n\in D_1\cup D_2}\mathcal{D}(n)n \leq \sum_{n\in D_1\cup D_2} c_1\exp(-nc_2/2)n \leq \sum_{n=0}^{\infty} c_1\exp(-nc_2/2)n = \frac{c_1\exp(c_2/2)}{(\exp(c_2/2)-1)^2} < \infty$ since $c_1, c_2 \in (0,\infty)$.

Since $\mathcal{D}(n)n \geq 0$ for all $n \in \mathbb{N}$, we have

$$\sum_{n=0}^{\infty}\mathcal{D}(n)n = \sum_{n\in D_1\cup D_2}\mathcal{D}(n)n + \sum_{n\in D_3}\mathcal{D}(n)n.$$

Since $\sum_{n=0}^{\infty}\mathcal{D}(n)n$ diverges and $\sum_{n\in D_1\cup D_2}\mathcal{D}(n)n$ converges, we must have $\sum_{n\in D_3}\mathcal{D}(n)n = \infty$.

We have

$$\begin{aligned}&\sum_{n\in D}\mathcal{D}(n)\log\left(\frac{\mathcal{D}(n)}{c_1\exp(-nc_2)}\right) &\text{(A.7a)}\\ &\quad= \sum_{n\in D_2\cup D_3}\mathcal{D}(n)\log\left(\frac{\mathcal{D}(n)}{c_1\exp(-nc_2)}\right) &\text{(A.7b)}\\ &\quad\geq \sum_{n\in D_3}\mathcal{D}(n)\log\left(\frac{\mathcal{D}(n)}{c_1\exp(-nc_2)}\right) &\text{(A.7c)}\\ &\quad\geq \sum_{n\in D_3}\mathcal{D}(n)\log\left(\frac{c_1\exp(-nc_2/2)}{c_1\exp(-nc_2)}\right) &\text{(A.7d)}\\ &\quad= \infty &\text{(A.7e)}\end{aligned}$$

where (A.7c) is due to $\mathcal{D}(n) \geq c_1\exp(-nc_2)$ for all $n \in D_2$, (A.7d) is due to $\geq c_1\exp(-nc_2/2)$ for all $n \in D_3$, and (A.7e) is due to $\sum_{n\in D_3}\mathcal{D}(n)n = \infty$ and $c_2 > 0$. □

Proof sketch for Lemma 18. The proof follows from the convexity of KL divergence and the first-order optimality condition. □