Quantification of Secrecy in Partially Observed Stochastic Discrete Event Systems

Jun Chen*, Member, IEEE, Mariam Ibrahim* and Ratnesh Kumar, Fellow, IEEE

Abstract—While cryptography is used to protect the content of information (e.g., a message) by making it undecipherable, behaviors (as opposed to information) may not be encrypted, and may only be protected by partially or fully hiding through creation of ambiguity by providing covers that generate indistinguishable observations from secrets. Having a cover together with partial observability does cause ambiguity about the system behaviors to be kept secret, yet some information about secrets may still be leaked due to statistical difference between the occurrence probabilities of the secrets and their covers. In this paper, we propose a Jensen-Shannon divergence (JSD) based measure to quantify secrecy loss in systems modeled as partially-observed stochastic discrete event systems (stochastic PODES), which quantifies the statistical difference between two distributions, one over the observations generated by secret and the other over those generated by cover. We further show that the proposed JSD measure for secrecy loss is equivalent to the mutual information between the distribution over possible observations and that over possible system status (secret versus cover). Since an adversary is likely to discriminate more if he/she observes for a longer period, our goal is to evaluate the worst-case loss of secrecy as obtained in limit over longer and longer observations. Computation for the proposed measure is also presented. Illustrative examples, including one with side channel attack, are provided to demonstrate the proposed computation approach.

Note to Practitioners — Secrecy is the ability to hide private information. For communicated information, this can be done through encryption or access control. But the same is not doable for system behaviors, and in contrast, cover is introduced for providing ambiguity. Quantifying the ability to hide secrets is a challenge. This paper provides a means to quantify this in terms of a type of distance measure between a secret and its cover. A computation of the same is also provided for partially-observed stochastic discrete event systems, and illustrated through a cache’s side-channel secrecy loss example.

Keywords—Discrete event systems (DES), stochastic systems, partial observability, Jensen-Shannon divergence (JSD), secrecy quantification.

I. INTRODUCTION

The rapid progress in information and communication technology has made it possible for an adversary to eavesdrop and/or attack confidential or private communication. While cryptography is used to protect the content of information (e.g., a message) by making it undecipherable, the same technique may not be used to hide behaviors which may not be encrypted. In such cases, secrecy can instead be attained through creation of ambiguity, caused for example by partial observation that ambiguates secrets from covers, where the secrets are system behaviors desired to be kept confidential, whereas the covers are the complementary system behaviors that generate the same observations as the secrets, creating ambiguity. Researchers in the field of security and privacy have explored many techniques for hiding secrets based on ambiguity schemes such as, Steganography and Watermarking [1], [2], Network level Anonymization [3], and Software Obfuscation [4].

Various notions of information secrecy have been explored in literature. References [5]–[7] defined the non-interference for input-output systems as a property in which the outputs that are observable to an adversary should not depend on any secret input so that the adversary does not deduce anything about the secret input by observing the output. Non-interference is a logical notion that is either satisfied or violated, and as such it does not allow the quantification of the degree to which a system may violate the property. Accordingly, the notion is enriched for probabilistic systems for which the degree of interference can be quantified in terms of the amount of information leaked by a system to an observer. The amount of information leakage, in turn, is measured by the loss of uncertainty about the inputs due to the observation of the outputs, i.e., the difference between the prior and posterior entropies of the inputs, namely the mutual information between inputs and outputs [5]. While such a quantification of information leakage is satisfactory for long periods of system operation (since entropy measures uncertainty in an average sense), it is of limited use for systems in which an adversary makes a single observation. To address this situation, the average case measure of entropy was replaced by its best case measure corresponding to minimum uncertainty, namely min-entropy, in the definition of mutual information [7].

In general, a secret can be a property of a sequence of executions, and not just a single execution, and this general situation has also been examined in the literature. For example in the setting of discrete event systems (DESs), the definition of secrecy examined in [8] requires that the execution of behaviors constituting a secret must be masked to an ob-
server through indistinguishable behaviors that are non-secret (i.e., cover). This is indeed analogous to the notion of non-interference, which by virtue of being logical has the same limitation that it cannot quantify the degree to which a system is interfering (or leaks information).

For probabilistic DESs, where each discrete transition is associated with a certain occurrence probability, more powerful notions of secrecy can be defined. For example, [9] used Jensen-Shannon divergence between the distributions of a secret versus its cover as a way to quantify the secrecy, which is measured as the divergence of two distributions over the set of feasible observations, one being the probabilities of secret behaviors and the other being those of cover behaviors. An approximation algorithm for computing an upper bound of JSD was also provided in [9]. Another attempt to generalize secrecy from logical to stochastic DESs is provided in [10], where, unlike the setting of mutual information based characterization of information leakage, the authors consider the difference between the prior and posterior distributions (before and after any observations) of the secret states, and require it to be upper bounded. The corresponding verification problem turns out to be undecidable. In another paper [11], the same authors proposed the notion of Step-Based Almost Current-State Opacity requiring the probability of revealing the secret must be upper bounded at each time step. This notion is decidable, but stringent since it is defined for each individual step. In contrast, the definition of $S_r$-secrecy proposed by [12] bounds the probability of revealing the secret over the set of all behaviors, as opposed to for each step. It is shown that $S_r$-Security can be viewed as a generalization of the logical secrecy defined in [8], and that it is a variant of the divergence used in [9]. Related work on K-step and infinite step opacity are explored in [13], [14]. The above mentioned secrecy notions (also referred to opacity in literatures), along with related articles have been reviewed in a recent survey [15].

In this paper, we propose a JSD based quantification to measure the secrecy loss in partially-observed stochastic discrete event systems (stochastic PODES), which are Markovian generators of arbitrary long sequences with transitions being partially-observed. (Stochastic PODES are equivalent to partially observed labeled Markov chains; see below.) The proposed JSD based quantification for secrecy loss is shown to be equivalent to the mutual information between the distribution over possible observations and that over possible status of system execution (secret versus cover). A recursive method for JSD computation is also presented: Given the distribution with respect to length-$(n-1)$ sequences, and the length-1 dynamics of the underlying partially-observed model, it computes the JSD of length-$n$ sequences. Under certain conditions, this recursion reaches a fixed point and provides “limiting” JSD measure, quantifying the worst case statistical difference that is defined over arbitrary long sequences. Since JSD is always bounded between 1 and 0, this worst case value is also bounded, providing an upper bound to the quantification of the amount of information leaked about secrets due to statistical difference between the observations of secrets versus covers. We derive the above recursion, and next construct an observer model of the given stochastic PODES, which is then used to develop a state-based computation of the fixed point JSD measure.

The computation of JSD for a PODES is challenging since a finite-state DES under partial observations is potentially infinite-state (with the state-space being conditional state distributions following observations). However, a finite-state observer representation is possible, which we construct and employ for divergence computation. This observer model is not a Markov chain model since the transition probabilities are no longer scalars, rather matrices (not necessarily square). Secrecy has also been studied in the context of partially observed labeled Markov chain (POLMC). See for example [15], [16] and the reference therein. While there is slight difference between PODES and POLMC (in a POLMC, transitions are not labeled and the states are partially observed whereas in PODES, transitions are labeled and it is the transitions that are partially observed), this does not affect the expressibility, as partial observability of states can be expressed as partial observability of transitions, and vice-versa. Note that [9] considers JSD over all finite length of observations for a terminating PODES, while this work formulates the limiting JSD for non-terminating PODES, where the limiting JSD provides an upper bound to the level of loss of secrecy that is achieved when an observer is able to wait for arbitrary long observations.

The rest of this paper is organized as follows. Section II presents notations and preliminaries. Divergence based quantification of secrecy loss is provided in Section III, whereas Section IV presents an observer based computation of worst-case JSD resulting from arbitrary long observations. Section V illustrates the proposed approach through practical examples, while the paper is concluded in Section VI. The appendices include proofs of lemmas and a review on related information theoretic notions utilized in this paper.

II. NOTATIONS AND PRELIMINARIES

A. Stochastic PODESs

For an event set $\Sigma$, define $\Sigma := \Sigma \cup \{\epsilon\}$, where $\epsilon$ denotes “no-event”. The set of all finite length event sequences over $\Sigma$, including $\epsilon$ is denoted as $\Sigma^*$, and $\Sigma^+ := \Sigma^* - \{\epsilon\}$. A trace is a member of $\Sigma^*$ and a language is a subset of $\Sigma^*$. We use $s \leq t$ to denote if $s \in \Sigma^*$ is a prefix of $t \in \Sigma^*$, and $|s|$ to denote the length of $s$ or the number of events in $s$. For $L \subseteq \Sigma^*$, its prefix-closure is defined as $pr(L) := \{s \in \Sigma^*: \exists t \in \Sigma^*: st \in L\}$ and $L$ is said to be prefix-closed (or simply closed) if $pr(L) = L$, i.e., whenever $L$ contains a trace, it also contains all the prefixes of that trace. For $s \in \Sigma^*$ and $L \subseteq \Sigma^*$, $L \setminus s := \{t \in \Sigma^*: st \in L\}$ denotes the set of traces in $L$ after $s$.

A stochastic PODES can be modeled by a stochastic automaton $G = (X, \Sigma, \alpha, x_0)$, where $X$ is the set of states, $\Sigma$ is the finite set of events, $x_0 \in X$ is the initial state, and $\alpha : X \times \Sigma \times X \rightarrow [0,1]$ is the probability transition function [17], and $\forall x \in X, \sum_{\sigma \in \Sigma} \sum_{x' \in X} \alpha(x, \sigma, x') = 1$. A non-stochastic PODES can be modeled as the same 4-tuple, but by replacing the transition function with $\alpha : X \times \Sigma \times
$X \to \{0, 1\}$, and a non-stochastic PODES is deterministic if
\[ \forall x \in X, \sigma \in \Sigma, \sum_{x' \in \mathcal{X}} \alpha(x, \sigma, x') \in \{0, 1\}. \]
The transition probability function $\alpha$ can be generalized to $\alpha : X \times \Sigma^* \times X$ in a natural way. Define the language generated by $G$ as $L(G) := \{ s \in \Sigma^* \mid \exists x \in X, \alpha(x_0, s, x) > 0 \}$. For a given $G$, a component $C = (X_C, \alpha_C)$ of $G$ is a “subgraph” of $G$, i.e., $X_C \subseteq X$ and $\forall x, x' \in X_C$ and $\sigma \in \Sigma$, $\alpha_C(x, \sigma, x') = \alpha(x, \sigma, x')$ whenever the latter is positive, and $\alpha_C(x, \sigma, x') = 0$ otherwise. $C$ is said to be a strongly connected component (SCC) or irreducible if $\forall x, x' \in X_C, \exists s \in \Sigma^*$ such that $\alpha_C(x, s, x') > 0$. A SCC $C$ is said to be closed if for each $x \in X_C$, $\sum_{\sigma \in \Sigma} \sum_{x' \in X_C} \alpha_C(x, \sigma, x') = 1$. The states which belong to a closed SCC are recurrent states and the remaining states (that do not belong to any closed SCC) are transient states. Another way to identify recurrent versus transient states is to consider the steady-state state distribution $\pi^*$ as the fixed-point of $\pi^* = \pi^* \Omega$, where $\pi^*$ is a row-vector with the same size as $X$, and $\Omega$ is the transition matrix with $i$th entry being the transition probability $\sum_{\sigma \in \Sigma} \alpha(i, \sigma, j)$. In case $\Omega$ is periodic with period $d \neq 1$, we consider the set of fixed-points of $\pi^* = \pi^* \Omega^d$. Then any state $i$ is recurrent if and only if there exists a reachable fixed point $\pi^*$ such that the $i$th entry of $\pi^*$ is nonzero. Identifying the set of recurrent states can be done polynomially, by the algorithm presented in [18].

The events executed by a PODES can be partially observed by an observer (i.e., an adversary). Such limited observation capability of an observer can be represented as an observation mask, $M : \Sigma \to \Sigma$, where $\Delta$ is the set of observed symbols and $M(\epsilon) = \epsilon$. An event $\sigma$ is unobservable if $M(\sigma) = \epsilon$. The set of unobservable events is denoted as $\sum_u$ and the set of observable events is then given by $\Sigma = \Sigma_u$. The observation mask can be generalized in natural way to $\Sigma^*$ with $M(\epsilon) = \epsilon$ and $\forall s \in \Sigma^*, \sigma \in \Sigma, M(s \sigma) = M(s) M(\sigma)$.

### B. Secret/non-secret behaviors and refined plant

Suppose $K \subseteq \Sigma^*$ models the secret behaviors (traces), whereas the remaining traces in $L - K$ can be viewed as its cover. Let the stochastic automaton $G = (X, \Sigma, \alpha, x_0)$ with generated language $L(G) = L$ be the system model, and the deterministic automaton $R = (Y, \Sigma, \beta, y_0)$, which specifies the secret behaviors $K$, be such that $L(R) = K$. Then a refinement of $G$ with respect to $R$, denoted $G^R$, can be used to capture the property-satisfying/violating traces in form of the reachability of certain non-secret states (the state has $D$ in its second coordinate), and is given by $G^R := (X \setminus Y, \Sigma, \gamma, (x_0, y_0))$, where $Y = Y \cup \{D\}$, and $\forall (x, \bar{y}), (x', \bar{y}') \in X \times Y$, $\sigma \in \Sigma, \gamma((x, \bar{y}), \sigma, (x', \bar{y}')) = \alpha(x, \sigma, x')$ if the following holds:

\[ (\bar{y}, \bar{y'}) \in Y \land \beta(\bar{y}, \sigma, \bar{y'}) > 0 \]

\[ \forall (\bar{y} = \bar{y'}) = D \lor \left( \bar{y}' = D \land \sum_{y \in Y} \beta(\bar{y}, \sigma, y) = 0 \right), \]

and otherwise $\gamma((x, \bar{y}), \sigma, (x', \bar{y}')) = 0$. Note that here $D$ is an added state to capture the traces in $L - K$. Then it can be seen that the refined plant $G^R$ has the following properties: (1) $L(G^R) = L(G)$; (2) any property-satisfying trace $s \in L(G)$ but not in $L(R)$ transitions the refinement $G^R$ to a non-secret state; (3) for each $s \in L(G) = L(G^R)$, $\sum_{x \in X} \alpha(x_0, s, x) = \sum_{(x, \bar{y}) \in X \times Y} \gamma((x_0, y_0), s, (x', \bar{y}))$, i.e., the occurrence probability of each trace in $G^R$ is the same as that in $G$. For $(x, \bar{y}), (x', \bar{y}') \in X \times Y$, and $\delta \in \Delta$, define the set of traces originating at $(x, \bar{y})$, terminating at $(x', \bar{y}')$ and executing a sequence of unobservable events followed by a single observable event with observation $\delta$ as $L_{G^R}((x, \bar{y}), \delta, (x', \bar{y}')) := \{ s \in \Sigma^* | s = u \sigma, M(u) = \epsilon, M(\sigma) = \delta, \gamma((x, \bar{y}), s, (x', \bar{y}')) > 0 \}$. Define $\alpha_L((x, \bar{y}), (x', \bar{y}')) := \sum_{s \in L_{G^R}((x, \bar{y}), \delta, (x', \bar{y}'))} \gamma((x, \bar{y}), s, (x', \bar{y}'))$, and denote it as $\theta((x, \bar{y}), \delta, (x', \bar{y}'))$. Therefore $\theta((x, \bar{y}), \delta, (x', \bar{y}'))$ is the probability of all traces originating at $(x, \bar{y})$, terminating at $(x', \bar{y}')$ and executing a sequence of unobservable events followed by a single observable event with observation $\delta$. Also for $i = (x, \bar{y}), j = (x', \bar{y}')$, define $\lambda_{ij} = \sum_{\sigma \in \Sigma_u} \gamma(i, \sigma, j)$ as the probability of transitioning from $(x, \bar{y})$ to $(x', \bar{y}')$ while executing a single unobservable event. Then \[ \theta((x, \bar{y}), \delta, (x', \bar{y}')) = \sum_{\sigma \in \Sigma_u} \lambda_{i \delta} \cdot \theta((x, \bar{y}), \delta, (x, \bar{y}')), \]

where the first term on the right hand side (RHS) corresponds to transitioning in at least two steps (i.e., intermediate k unobservably, and k to j with a single observable event at the end), whereas the second term on RHS corresponds to transitioning in exactly one step [19]. Thus, for each $\delta \in \Delta$, all the probabilities $\theta((i, \delta), j, x \times Y)$ can be found by solving the following matrix equation [20]:

\[ \Theta(\delta) = \Lambda \Theta(\delta) + \Gamma(\delta), \]

where $\Theta(\delta), \Lambda$ and $\Gamma(\delta)$ are all $|X \times Y| \times |X \times Y|$ square matrices whose $ij$th elements are given by $\theta(i, \delta, j)$, $\lambda_{ij}$ and $\sum_{\sigma \in \Sigma_u} \gamma(i, \bar{y}, j, (x, \bar{y}'))$, respectively.

**Remark 1:** To find $\Theta(\delta)$ using Equation (1), we need to solve $\Theta(\delta) = (I - \Lambda)^{-1} \Gamma(\delta)$. The complexity of matrix inverse is $O(|X|^3 \times |Y|^3)$ and the complexity of matrix multiplication is $O(|X|^3 \times |Y|^3)$, and so overall complexity is $O(|X|^3 \times |Y|^3)$. Since the number of secret states and number of non-secret states are both upper bounded by the number of states in $G^R$, which is $O(|X| \times |Y|)$, the complexity of finding $\Theta(\delta)$ for all $\delta \in \Delta$ using Equation (1) is bounded by $O(|\Delta| \times |X|^3 \times |Y|^3)$.

Note also $G^R$ has $O(|X| \times |Y|)$ states and $O(|\Sigma| \times |X|)$ transitions per state since only the $G$ part is non-deterministic, whereas the complexity for identifying all the non-secret recurrent states in $G^R$ is cubic in the number of states in $G^R$ and linearly in the number of transitions in $G^R$, respectively [21]. So the overall complexity for finding all $\Theta(\delta)$ using Equation (1) is $O(\Delta \times |X|^3 \times |Y|^3 + |\Sigma| \times |X|^2 \times |Y|^2)$. ■

**Example 1:** Fig. 1(a) is an example of a stochastic automaton $G$. The set of states is $X = \{0, 1, 2\}$ with initial state $x_0 = 0$, event set $\Sigma = \{a, b, c\}$. A state is depicted as a node, whereas a transition is depicted as an edge between its origin and termination states, with its event name and probability value labeled on the edge. The observation mask $M$ is such that $\Theta(\epsilon) = \epsilon$ and for all other events $\sigma \in \{a, b, c\}$, $M(\sigma) = \sigma$. Suppose $R$ is given in Fig. 1(b), i.e.,
As is standard in information theory, a base–2 logarithm has been used. For more detail please refer to [22].

Given a probability distribution \( p \) over discrete set \( A \), the entropy of \( p \) is defined as

\[
H(p) = -\sum_{a \in A} p(a) \log p(a).
\]

Given two probability distributions \( p \) and \( q \) over \( A \), the Kullback-Leibler (KL) divergences between \( p \) and \( q \), denoted as \( D_{KL}(p,q) \), is defined as

\[
D_{KL}(p,q) = \sum_{a \in A} p(a) \log \frac{p(a)}{q(a)}.
\]

For given \( \lambda_1 > 0 \) and \( \lambda_2 > 0 \) satisfying \( \lambda_1 + \lambda_2 = 1 \), the Jensen-Shannon Divergence (JSD) between \( p \) and \( q \), denoted as \( D(p,q) \), is defined as

\[
D(p,q) = \lambda_1 D_{KL}(p,\lambda_1 p + \lambda_2 q) + \lambda_2 D_{KL}(q,\lambda_1 p + \lambda_2 q),
\]

which is equivalent to

\[
D(p,q) = H(\lambda_1 p + \lambda_2 q) - \lambda_1 H(p) - \lambda_2 H(q).
\]

Given two probability distributions \( p \) over \( A \) and \( q \) over \( B \), the mutual information between \( p \) and \( q \) is defined as

\[
I(p,q) = \sum_{a \in A} \sum_{b \in B} p(a,b) \log \frac{Pr(a,b)}{Pr(a)Pr(b)}.
\]

Mutual information can also be equivalently defined as

\[
I(p,q) = H(p) - H(p|q),
\]

where the condition entropy \( H(p|q) \) is given as

\[
H(p|q) = -\sum_{a \in A} \sum_{b \in B} p(a)Pr(b|a) \log Pr(b|a).
\]

### III. Divergence-based Secrecy Quantification

For any \( n \in \mathbb{N} \), and a length-\( n \) observation \( o \in \Delta^n \), let \( p_n(o) \) denote the probability of observation \( o \). Since the occurrences of observations of length \( n \) are mutually disjoint, \( \sum_{o \in \Delta^n} p_n(o) = 1 \), i.e., \( p_n \) is a probability distribution over \( \Delta^n \). Then its entropy is given as:

\[
H(p_n) = -\sum_{o \in \Delta^n} p_n(o) \log p_n(o).
\]

**Lemma 1:** The entropy \( p_n \) as defined above for length-\( n \) observation can be recursively computed as follows:

\[
H(p_n) = H(p_{n-1}) - \sum_{o \in \Delta^{n-1}} p_{n-1}(o) \sum_{\delta \in \Delta} Pr(\delta|o) \log Pr(\delta|o).
\]

**Proof:** See appendix.

Observations in \( \Delta^n \) can be generated by secrets (behaviors in \( K \)) or by covers (behaviors in \( L - K \)), and so define two more probability distributions over \( \Delta^n \): probability that an observation \( o \in \Delta^n \) is generated by some secret in \( K \), denoted \( p_n^s(o) \), versus that is generated by some cover in \( L - K \), denoted \( p_n^c(o) \):

\[
p_n^s(o) := \frac{Pr(s \in K \cap M^{-1}(o))}{Pr(s \in K \cap M^{-1}(\Delta^n))},
\]

\[
p_n^c(o) := \frac{Pr(s \in (L - K) \cap M^{-1}(o))}{Pr(s \in (L - K) \cap M^{-1}(\Delta^n))}.
\]

Further define \( \lambda_n^s := Pr(s \in K \cap M^{-1}(\Delta^n)) \) to be the probability of secrets and \( \lambda_n^c := Pr(s \in (L - K) \cap M^{-1}(\Delta^n)) \) to be the probability of covers, respectively, generating length-\( n \) observation. Then it is easy to show that \( \lambda_n^s + \lambda_n^c = 1 \) for
all \( n \in \mathbb{N} \). The entropy of \( p_n^s \) and \( p_n^c \) are given, respectively, by:

\[
H(p_n^s) = - \sum_{o \in \Delta^n} p_n^s(o) \log p_n^s(o) \quad (2)
\]

\[
H(p_n^c) = - \sum_{o \in \Delta^n} p_n^c(o) \log p_n^c(o). \quad (3)
\]

The ability of an intruder to identify secret versus cover behaviors based on observations of length \( n \), depends on the disparity between the two distributions \( p_n^s \) versus \( p_n^c \). If \( p_n^s \) and \( p_n^c \) are identical, i.e., with “zero disparity”, there is no way to statistically tell apart secrets from covers, and in that case there is perfect secrecy. However, when \( p_n^s \) and \( p_n^c \) are different, then one could characterize the ability of an intruder to discriminate secrets from covers based on length-\( n \) observations, using the JSD between \( p_n^s \) and \( p_n^c \), denoted as \( D(p_n^s, p_n^c) \). This JSD is given by the following weighted sum of a pair of Kullback-Leibler (KL) divergences between, respectively, \( p_n^s \) and \( p_n^c \), and their weighted sum:

\[
D(p_n^s, p_n^c) = \lambda_n^s D_{KL}(p_n^s,\lambda_n^s p_n^s + \lambda_n^c p_n^c) + \lambda_n^c D_{KL}(p_n^c,\lambda_n^c p_n^c + \lambda_n^s p_n^s) = H(\lambda_n^s p_n^s + \lambda_n^c p_n^c) - \lambda_n^s H(p_n^s) - \lambda_n^c H(p_n^c) = H(p_n) - \lambda_n^s H(p_n^s) - \lambda_n^c H(p_n^c), \quad (4)
\]

where \( D_{KL} \) represents the KL-divergence. Note that JSD is symmetric in its arguments and bounded by 0 and 1.

\textbf{Remark 2:} The work reported in [23] also uses JSD measure for determining statistical difference in Markovian models of genetic sequences from phylogenetically proximal organisms, which however is not related to secrecy as no information hiding through partial observation is involved. The computation of such JSD is necessarily for \textit{fully-observed} and over finite length genetic sequences, which can be easily done numerically.

We first show that the JSD measure as considered in this paper is indeed a useful measure of information revealed, by formally establishing in the following theorem that it equals the mutual information between the observations \( p_n \) and the status (secret vs. cover) of system executions. (Note mutual-information is a well accepted measure of the information revealed about one random variable form the observations of another.) This status can be captured by a bi-valued random variable \( \Lambda_n \), defined for each \( n \in \mathbb{N} \), such that \( Pr(\Lambda_n = s) = \lambda_n^s \) and \( Pr(\Lambda_n = c) = \lambda_n^c \). With a slight abuse of notation, also denote its distribution as \( \Lambda_n \), which corresponds to the distribution of the system executing secret versus cover.

\textbf{Theorem 1:} The JSD as defined in (4) is equivalent to the mutual information of \( \Lambda_n \) and \( p_n \), i.e.,

\[
D(p_n^s, p_n^c) = I(\Lambda_n, p_n).
\]

\textbf{Proof:} According to the definition of mutual information, we have

\[
I(\Lambda_n, p_n) = H(p_n) - H(p_n|\Lambda_n).
\]

The conditional entropy \( H(p_n|\Lambda_n) \) can be expressed as follows:

\[
H(p_n|\Lambda_n) = - \lambda_n^s \sum_{o \in \Delta^n} Pr(o|\Lambda_n = s) \log Pr(o|\Lambda_n = s) - \lambda_n^c \sum_{o \in \Delta^n} Pr(o|\Lambda_n = c) \log Pr(o|\Lambda_n = c)
\]

\[
- \lambda_n^s \sum_{o \in \Delta^n} p_n^s(o) \log p_n^s(o) - \lambda_n^c \sum_{o \in \Delta^n} p_n^c(o) \log p_n^c(o)
\]

\[
= \lambda_n^s H(p_n^s) + \lambda_n^c H(p_n^c),
\]

where we utilize the fact that \( Pr(o|\Lambda_n = s) = p_n^s(o) \) and \( Pr(o|\Lambda_n = c) = p_n^c(o) \). Substituting \( H(p_n|\Lambda_n) \) into the definition of mutual information \( I(\Lambda_n, p_n) \), and considering relationship in (4), we have

\[
I(\Lambda_n, p_n) = H(p_n) - \lambda_n^s H(p_n^s) - \lambda_n^c H(p_n^c) = D(p_n^s, p_n^c)
\]

Thus the proof is completed.

\textbf{Remark 3:} Theorem 1 establishes the equivalence of the JSD in (4) and the mutual information of \( \Lambda_n \) and \( p_n \), the latter of which measures the mutual dependence between length-\( n \) observations and status of system execution (secret versus cover). When \( D(p_n^s, p_n^c) = I(\Lambda_n, p_n) = 0 \), length-\( n \) observations are independent of system execution status, and thus no secret information can be leaked through length-\( n \) observations. On the other hand, when \( D(p_n^s, p_n^c) = I(\Lambda_n, p_n) > 0 \), the dependence of length-\( n \) observations and system status can be measured by the JSD, \( D(p_n^s, p_n^c) \), which in turn quantifies the extent to which system secrecy can be leaked by length-\( n \) observations.

\textbf{A. Recursive Characterization}

An intruder is likely to discriminate more if he/she observes for a longer period, and accordingly, our goal is to evaluate the worst-case loss of secrecy, as obtained in the limit: \( \lim_{n \to \infty} D(p_n^s, p_n^c) \). This worst-case JSD provides an upper bound to quantify the amount of information leaked about secrets.

To compute the worst-case loss of secrecy, we first develop a recursive computation for \( D(p_{n+1}^s, p_{n+1}^c) \), relating it to distributions of length-(\( n+1 \)) observations and divergence of length-\( n \) distributions. For \( o \in \Delta^n \) and \( \delta \in \Delta \), define the distributions of secret versus cover upon a single observation \( \delta \) following a history of observation \( o \):

\[
p^{s|o}(\delta) := \frac{Pr(s \in K \cap M^{-1}(o\{\delta\}))}{Pr(s \in K \cap M^{-1}(o\{\Delta\}))}
\]

\[
p^{c|o}(\delta) := \frac{Pr(s \in (L-K) \cap M^{-1}(o\{\delta\}))}{Pr(s \in (L-K) \cap M^{-1}(o\{\Delta\}))}
\]

Further define \( \lambda^{s|o} := Pr(s \in K \cap M^{-1}(o\{\Delta\})) / Pr(o) \) and \( \lambda^{c|o} := Pr(s \in (L-K) \cap M^{-1}(o\{\Delta\})) / Pr(o) \). Then again we have \( \lambda^{s|o} + \lambda^{c|o} = 1 \). Following the definition of JSD, we have
The following notations:

The conditional state distribution is \( \pi_{n}^{o} \).

Lemma 2: Given observation \( o \), the length-1 JSD between \( p_{n}^{o} \) and \( p_{n}^{0} \) can be computed as:

\[
D(p_{n}^{o}, p_{n}^{0}) = H(\lambda_{n}^{o} | p_{n}^{o}) + \lambda_{n}^{o} H(p_{n}^{o}) - \lambda_{n}^{o} H(p_{n}^{0}) - \lambda_{n}^{0} H(p_{n}^{0}).
\]  

(5)

The following lemma characterizes the computation of the length-1 JSD given observation \( o \).

Lemma 3: Using the above lemma, we can next provide the following recursive computation for JSD.

\[
D(p_{n}^{o}, p_{n}^{0}) = H(\lambda_{n}^{o} | p_{n}^{o}) + \lambda_{n}^{o} H(p_{n}^{o}) - \lambda_{n}^{o} H(p_{n}^{0}) - \lambda_{n}^{0} H(p_{n}^{0}).
\]  

(6)

Proof: See appendix.

Using the above lemma, we can next provide the following recursive computation for JSD.

```
Lemma 4:
```

 IV. OBSERVER-BASED COMPUTATION OF WORST-CASE SECRECY LOSS

The state-based characterization of \( \lim_{n \to \infty} D(p_{n}^{o}, p_{n}^{0}) \) requires the computation of \( \lim_{n \to \infty} P_{n-1}(\pi) \) which can be accomplished with the help of an observer that we introduce next. An observer tracks the possible system states following each observation, and also allows the computation of the corresponding state distribution. We let \( Obs \) denote an observer automaton with state set as the power set of states of refined plant, namely, \( Z \subseteq 2^{X \times Y} \), so that each node \( z \in Z \) of the observer is a subset of system states, i.e., \( z \subseteq X \times Y \), and we use \( |z| \) to denote the number of system states in \( z \). \( Obs \) is initialized at node \( z_{0} = \{(x_{0},y_{0})\} \), and there is a transition labeled with \( \delta \in \Delta \) from node \( z \) to \( z' \) if and only if every element of \( z' \) is reachable from some elements of \( z \) along a trace that ends in the only observation \( \delta \), i.e., \( z' = \{(x',y') \in X \times Y : \exists (x,y) \in z, L_{G,R}((x,y),\delta, (x',y')) \neq \emptyset \} \). Associated with this transition is the transition probability matrix \( \Theta_{z,\delta,z'} \) of size \( |z| \times |z'| \) (a submatrix of \( \Theta(\delta) \) matrix introduced earlier), whose \( i,j \)th element \( \theta_{i,j} \) is given by the transition probability from \( i \)th element \( (x,y) \) of \( z \) to \( j \)th element \( (x',y') \) of \( z' \) while producing the observation \( \delta \), and equals \( \alpha(L_{G,R}((x,y),\delta, (x',y'))) \).

Example 2: Consider the models of Fig. 2, where \( M(u) = \epsilon, M(a) = a \) and \( M(b) = b \). Then the observer \( Ohs \) is given as Fig. 3.

Associated with each observation \( o \in \Delta^{*} \), there is a reachable state distribution \( \pi(o) \) as discussed earlier. Let the state \( z \) be reached in \( Obs \) following observation \( o \). Then obviously the number of positive elements of \( \pi(o) \) is the same as the number of elements in \( z \). Then with a slight abuse of notation, we also use \( \pi(o) \) to denote the row-vector containing only positive elements, and of same size as the number of elements in the node reached by \( o \) in \( Obs \). Then \( \pi(o) \) can also be recursively computed as follows: for any \( o \in \Delta^{*}, \delta \in \Delta: \pi(\epsilon) = 1 \) and \( \pi(o\delta) = \frac{\pi(o) \Theta_{z_{a},\delta} z_{a}}{\pi(o) \Theta_{z_{a},\delta} z_{a}} \), where \( z_{a} \) and \( z_{a\delta} \) are the nodes reached in \( Obs \) following \( o \) and \( o\delta \) respectively.
Then it can be seen that along any cycle in \textit{Obs}, the distribution upon completing the cycle is a function of the distribution upon entering the cycle, through a sequence of transition matrix-multiplications and their normalization. In case of steady-state, those two distributions will be the same, namely, a fixed point of that function.

Given the \textit{Obs} with state space \( Z \) for system \( G \) and specification \( R \), let \( \Theta \) be a \((\sum_{z} |z|) \times (\sum_{z} |z|)\) square matrix, whose \( ij \)th block is the \( |z_i| \times |z_j| \) matrix \( \sum_{\delta} \Theta_{z_i, \delta, z_j} \). The fix point distribution associated with \( \Theta \) can be obtained by solving \( \pi^* = \pi^* \Theta \), where \( \pi^* \) is a row vector of size \( \sum_{z} |z| \). For each \( z_i \in Z \), let \( p(z_i) \) be the summation of the \( i \)th block of \( \pi^* \), then \( z_i \) is said to be \textit{recurrent} if \( p(z_i) > 0 \). Also note that for each \( z \in Z \), exists a sufficiently large \( N \) such that

\[
p(z) = \sum_{o \in A^N} p_r(o) \rightarrow p_N(o).
\]

In other words, \( p(z) \) is the probability of all sufficiently long observations that reach the observer state \( z \). With a slight abuse of notations, define \( \lambda^* \) as the summation of the elements of \( \pi^* \) corresponding to secret states, i.e., \( \lambda^* := \pi^* \mathcal{I}^*, \) and \( \lambda^c := 1 - \lambda^* \).

\textbf{Example 3:} For the observer in Example 2, \( \sum_{z} |z| = 8 \) and so \( \Theta \) is a \( 8 \times 8 \) matrix given as:

\[
\tilde{\Theta} = \begin{bmatrix}
0 & 0.1 & 0.1 & 0.7 & 0.1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.3 & 0.7 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.7 & 0 & 0 & 0.3 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{bmatrix},
\]

and

\[
\pi^* = \begin{bmatrix}
0 & 0 & 0 & 0.75 & 0 & 0.07 & 0.05 & 0.13 \\
\end{bmatrix}.
\]

Therefore \( p(z_0) = p(z_1) = 0, p(z_2) = 0.75, p(z_3) = 0.12 \) and \( p(z_4) = 0.13 \).

For a set of recurrent nodes \( \{z_1, z_2, \ldots, z_n\} \) that forms a SCC in \textit{Obs}, define a set of distributions \( \{\pi_{z_1}^*, \pi_{z_2}^*, \ldots, \pi_{z_n}^*\} \) to be a set of steady-state distributions for a given set of recurrent nodes, denoted as \( \{\{\pi_{z_1, k}, \ldots, \pi_{z_n, k}\}, k \in N\} \). Then if steady-state always exists, for any sufficiently long observation that reaches a recurrent node \( z \), there exists \( k \in N \) such that \( \pi(o) = \pi_{z, k}^* \). Denote \( p(z, k) := Pr\{o \mid o \text{ reaches } z \text{ and } \pi(o) = \pi_{z, k}^*\} \).

Note that when the set of steady state distributions is a singleton, and hence unique, \( p(z, k) = p(z) \).

\textbf{Example 4:} Let us revisit Example 2. It can be seen that \( z_2, z_3 \) and \( z_4 \) are recurrent nodes, and each of them forms a SCC. We have \( \pi_{z_2}^* = [1 \ 0] \), \( \pi_{z_4}^* = [1 \ 0] \), and while there are multiple solutions to the equation set \( \pi_{z_3}^* = \frac{\pi_{z_2}^* \Theta_{z_2, u, z_3}}{\pi_{z_2}^* \Theta_{z_2, u, z_3}} \) and \( \pi_{z_3}^* = \frac{\pi_{z_4}^* \Theta_{z_4, k, z_3}}{\pi_{z_4}^* \Theta_{z_4, k, z_3}} \), only \( \pi_{z_3}^* = [0.5833 \ 0.4167] \) is reachable. Thus each set of recurrent nodes is a singleton set, and each with a unique fixed-point distribution. Therefore, for each recurrent node \( z, p(z, k) = p(z) \).
Let $\mathcal{T}^s_x$ and $\mathcal{T}^c_x$ be indicator column vectors with binary entries of size $|z|^2$ for identifying, within $z'$, the secret and cover states, respectively. For each steady-state distribution $\pi^*_z$ of each node $z$, define:

\[
\lambda^s|\pi^*_z,k := \sum_{\delta \in \Delta} \pi^*_z,\delta|z,\delta'|\mathcal{T}^s_x, \\
\lambda^c|\pi^*_z,k := \sum_{\delta \in \Delta} \pi^*_z,\delta|z,\delta'|\mathcal{T}^c_x, \\
p^s|\pi^*_z,k(\delta) := \frac{\pi^*_z,\delta|z,\delta'|\mathcal{T}^s_x}{\lambda^s|\pi^*_z,k}, \\
p^c|\pi^*_z,k(\delta) := \frac{\pi^*_z,\delta|z,\delta'|\mathcal{T}^c_x}{\lambda^c|\pi^*_z,k}.
\]

**Example 5:** Then for Example 2, $\mathcal{T}^s_2 = [0 1]^T$, $\mathcal{T}^c_2 = [1 0]^T$, $\mathcal{T}^s_3 = [0 1]^T$, $\mathcal{T}^c_3 = [1 0]^T$, $\mathcal{T}^s_4 = [1]^T$ and $\mathcal{T}^c_4 = [0]^T$. For $z_2$ and $\pi^*_2$,

\[
\lambda^s|\pi^*_2 = 0, \\
\lambda^c|\pi^*_2 = 1, \\
p^s|\pi^*_2(a) = p^c|\pi^*_2(a) = p^s|\pi^*_2(b) = 0.
\]

For $z_3$ and $\pi^*_3$,

\[
\lambda^s|\pi^*_3 = 0.5833, \\
\lambda^c|\pi^*_3 = 0.4167, \\
p^s|\pi^*_3(a) = \frac{\pi^*_3,\Theta_{a,2,2}|\mathcal{T}^s_3}{\lambda^s|\pi^*_3} = 0.3, \\
p^s|\pi^*_3(b) = \frac{\pi^*_3,\Theta_{b,2,2}|\mathcal{T}^s_3}{\lambda^s|\pi^*_3} = 0.7, \\
p^c|\pi^*_3(a) = \frac{\pi^*_3,\Theta_{a,2,2}|\mathcal{T}^c_3}{\lambda^c|\pi^*_3} = 0.3, \\
p^c|\pi^*_3(b) = \frac{\pi^*_3,\Theta_{b,2,2}|\mathcal{T}^c_3}{\lambda^c|\pi^*_3} = 0.7.
\]

For $z_4$ and $\pi^*_4$,

\[
\lambda^s|\pi^*_4 = 1, \\
\lambda^c|\pi^*_4 = 0, \\
p^s|\pi^*_4(a) = \frac{\pi^*_4,\Theta_{a,4,4}|\mathcal{T}^s_4}{\lambda^s|\pi^*_4} = 0.3, \\
p^s|\pi^*_4(b) = \frac{\pi^*_4,\Theta_{b,4,4}|\mathcal{T}^s_4}{\lambda^s|\pi^*_4} = 0.7, \\
p^c|\pi^*_4(a) = p^c|\pi^*_4(b) = 0.
\]

**Theorem 2:** Consider a system $G$ and specification $R$. Then under Assumption 1, the worst case secrecy loss, i.e., JSD between $p^*_n$ and $p^c_n$ when $n \to \infty$, is given by:

\[
\lim_{n \to \infty} D(p^*_n, p^c_n) = H(\ell^s) + \sum_{z,k} \sum_{z \in \text{recurrent}, k \in \mathbb{N}} \left( -H(\{\lambda^s|\pi^*_z,k, \lambda^c|\pi^*_z,k\}) + D(p^s|\pi^*_z,k, p^c|\pi^*_z,k) \right).
\]

The next assumption assumes that for each set of recurrent nodes in $\text{Obs}$, there only exists one set of steady-state distributions.

**Assumption 2:** For each set of recurrent nodes in $\text{Obs}$, $k = 1$, i.e., the set of steady-state distributions is unique, so that $p(z, k) = p(z)$.

**Theorem 3:** Consider a system $G$ and specification $R$. Then under Assumptions 1 and 2, the worst case secrecy loss, i.e., JSD between $p^*_n$ and $p^c_n$ when $n \to \infty$, is given by:

\[
\lim_{n \to \infty} D(p^*_n, p^c_n) = 0.7219 - 0.1176 = 0.6043.
\]

Thus, for the system in Fig. 2, the worst case secrecy loss, as measured by the limiting JSD, is 0.6043.

**V. CACHE SIDE-CHANNEL ATTACK EXAMPLE**

In this section, a modified version of cache side-channel attack example adopted from [25] is considered. When a host program executes on the system, its memory accesses contain information that might help an attacker determine the secret of whether or not host is accessing the cache memory. Suppose the attacker executes a program on the same processor, and shares the same cache as the host program (see Fig. 4). If the host holds its own data in cache, its cache access results in a hit ($H_{hit}$), but if the attacker evicts the host’s data in the cache lines by requesting cache access, it would result a miss ($H_{miss}$). Similarly, when the host requires cache data, it evicts the cache lines which hold the attacker’s data, which make the attacker’s future cache access “miss” ($A_{miss}$). On the other hand, $A_{hit}$ occurs when attacker access the data while it is held in cache. These behaviors may give the attacker information to infer the host’s cache accesses. The system is described as in Fig. 4. Note that as long as the host requests cache access, the attacker will for sure witness an $A_{miss}$, and at that point it would occupy the cache, thereby fully knowing the cache status.

Now suppose, in order to prevent any information leak, the system (i.e., the processor) practices attack protection

\[
\begin{align*}
\text{Assumption 1:} & \quad \text{Assume that for any sufficiently long observations } o_1 \leq o_2, \text{ if } \text{Obs} \text{ reaches the same node following } o_1 \text{ and } o_2, \text{ then } \pi(o_1) = \pi(o_2). \\
\text{Then following the above definitions and Lemma 4, next theorem provides computation of } & \lim_{n \to \infty} D(p^*_n, p^c_n), \text{ under Assumption 1.}
\end{align*}
\]
systems, where information about system secrets may be revealed through the side-channel of observable inputs/outputs. Statistical difference, in the form of the Jensen-Shannon Divergence measure between the influence of secrets versus covers on the observations, is employed to quantify the loss of secrecy. We showed that this JSD measure is equivalent to the mutual information between the distribution over possible observations and that over possible status of system execution (secret versus cover), and proposed the computation of the “limiting” JSD as a measure of worst-case secrecy loss, resulting from their longer and longer observations. The computation of limiting JSD required developing a recursion relating JSD over length-\( n \) sequences to distributions over length-(\( n-1 \)) sequences through the 1-step dynamics of underlying system model. We also presented an observer-based approach for computing the fixed-point of recursion, and also the limiting JSD. Illustrative examples, including one based on side channel attack, are provided to demonstrate the proposed notions and associated computation. For terminating DESs, one can simply add unobservable self-loops at terminating states with probability 1, and proceed as in this paper. Future work can consider generalizing the results, by computing the JSD for systems that may not satisfy the two assumptions about convergence and uniqueness, respectively, made in the paper.

**APPENDIX**

*Proof for Lemma 1:*

\[
H(p_n) = - \sum_{o \in \Delta^n} p_n(o) \log p_n(o) \\
= - \sum_{o \in \Delta^n} \sum_{\delta \in \Delta} p_n(o\delta) \log p_n(o\delta) \\
= - \sum_{o \in \Delta^n} \sum_{\delta \in \Delta} p_{n-1}(o) Pr(\delta|o) \log(p_{n-1}(o)Pr(\delta|o)) \\
= - \sum_{o \in \Delta^n} p_{n-1}(o) \sum_{\delta \in \Delta} Pr(\delta|o)(\log p_{n-1}(o) + \log Pr(\delta|o))
\]
= - \sum_{o \in \Delta^{n-1}} p_n-1(o) \sum_{\delta \in \Delta} Pr(\delta|o) \log Pr(\delta|o)
- \sum_{o \in \Delta^{n-1}} p_n-1(o) \sum_{\delta \in \Delta} Pr(\delta|o) \log p_n-1(o)
= - \sum_{\delta \in \Delta} Pr(\delta|o) \sum_{o \in \Delta^{n-1}} p_n-1(o) \log p_n-1(o)
- \sum_{\delta \in \Delta} Pr(\delta|o) \sum_{o \in \Delta^{n-1}} p_n-1(o)
= - \sum_{\delta \in \Delta} p_n-1(o) \sum_{o \in \Delta^{n-1}} p_n-1(o) \log Pr(\delta|o) + H(p_{n-1}).

Thus Lemma 1 is established.

**Proof for Lemma 2:** By expanding (5), we have

\[ D(p^{x_0}, p^{c_0}) = H(\lambda^{x_0} p^{x_0} + \lambda^{c_0} p^{c_0}) \]

\[ + \sum_{\delta \in \Delta} \lambda^{x_0} p^{x_0}(\delta) \log \lambda^{x_0} p^{x_0}(\delta) \]

\[ + \sum_{\delta \in \Delta} \lambda^{c_0} p^{c_0}(\delta) \log \lambda^{c_0} p^{c_0}(\delta) \]

\[ = H(\lambda^{x_0} p^{x_0} + \lambda^{c_0} p^{c_0}) \]

\[ + \sum_{\delta \in \Delta} \lambda^{x_0} p^{x_0}(\delta) \log \lambda^{x_0} p^{x_0}(\delta) \]

\[ + \sum_{\delta \in \Delta} \lambda^{c_0} p^{c_0}(\delta) \log \lambda^{c_0} p^{c_0}(\delta) \]

\[ - \lambda^{x_0} \log \lambda^{x_0} \left( \sum_{\delta \in \Delta} p^{x_0}(\delta) \right) \]

\[ - \lambda^{c_0} \log \lambda^{c_0} \left( \sum_{\delta \in \Delta} p^{c_0}(\delta) \right). \]

Since \((\sum_{\delta \in \Delta} p^{x_0}(\delta)) = (\sum_{\delta \in \Delta} p^{c_0}(\delta)) = 1\), we have

\[ D(p^{x_0}, p^{c_0}) = H(\lambda^{x_0} p^{x_0} + \lambda^{c_0} p^{c_0}) \]

\[ + \sum_{\delta \in \Delta} \lambda^{x_0} p^{x_0}(\delta) \log \lambda^{x_0} p^{x_0}(\delta) \]

\[ + \sum_{\delta \in \Delta} \lambda^{c_0} p^{c_0}(\delta) \log \lambda^{c_0} p^{c_0}(\delta) \]

\[ - \lambda^{x_0} \log \lambda^{x_0} - \lambda^{c_0} \log \lambda^{c_0} \]

\[ = H(\lambda^{x_0} p^{x_0} + \lambda^{c_0} p^{c_0}) + H(\{\lambda^{x_0}, \lambda^{c_0}\}) \]

\[ - H(\lambda^{x_0} p^{x_0} - H(\lambda^{c_0} p^{c_0}). \]

Thus Lemma 2 is established.

**Proof for Lemma 3:** We first define some notations, for simplicity of presentation, as follows:

\[ \tilde{p}_n(o) := Pr(s \in K \cap M^{-1}(o)) = \lambda^s p_n(o) \]

\[ \tilde{p}_n(o) := Pr(s \in (L - K) \cap M^{-1}(o)) = \lambda^s p_n(o) \]

\[ \tilde{p}(\delta|o) := \frac{Pr(\delta \in (L - K) \cap M^{-1}(o|d))}{Pr(o)} = \lambda^{x_o} p^{x_0}(\delta) \]

\[ \tilde{p}(\delta|o) := \frac{Pr(s \in (L - K) \cap M^{-1}(o|d))}{Pr(o)} = \lambda^{x_o} p^{x_0}(\delta) \]

\[ p(\delta|o) := \lambda^{x_o} p^{x_0}(\delta) + \lambda^{c_o} p^{c_0}(\delta) = Pr(\delta|o). \]

We start by deriving a recursive computation for \(H(p^s_n)\) as defined in (2), as follows:

\[ H(p^s_n) = - \sum_{o \in \Delta} p_n^s(o) \log p_n^s(o) \]

\[ = - \frac{1}{\lambda_n^s} \sum_{o \in \Delta} \tilde{p}_n(o) \log \tilde{p}_n(o) \]

\[ = - \frac{1}{\lambda_n^s} \sum_{o \in \Delta} \sum_{\delta \in \Delta} \tilde{p}_n(o) \log \tilde{p}_n(o) \lambda^{x_o}(\delta) \]

\[ = - \frac{1}{\lambda_n^s} \sum_{o \in \Delta} \sum_{\delta \in \Delta} \tilde{p}_n(o) \log \tilde{p}_n(\delta|o) \log p_{n-1}(o) \tilde{p}(\delta|o) \lambda^{x_0}(\delta) \]

\[ = - \frac{1}{\lambda_n^s} \sum_{o \in \Delta} \sum_{\delta \in \Delta} \tilde{p}_n(o) \log \tilde{p}_n(\delta|o) \lambda^{x_0}(\delta) \]

\[ + \frac{1}{\lambda_n^s} \sum_{o \in \Delta} \sum_{\delta \in \Delta} \tilde{p}_n(o) \log \lambda_n^s. \]

Similarly, \(H(p^c_n)\) as defined in (3) can be recursively characterized as:

\[ H(p^c_n) = - \frac{1}{\lambda_n^c} \sum_{o \in \Delta} p_{n-1}(o) \log p_{n-1}(o) \lambda^{c_0}(o) + \log \lambda_n^c \]

\[ + \frac{1}{\lambda_n^c} \sum_{o \in \Delta} p_{n-1}(o) \lambda^{x_0}(o) H(\lambda^{x_0} p^{x_0}). \]

Expand (4) using the above recursion and Lemma 1, yielding the follows:

\[ D(p^s_n, p^c_n) = H(p_n) - \lambda_n^s H(p^s_n) - \lambda_n^c H(p^c_n) \]

\[ = H(p_n) - \lambda_n^s \log \lambda_n^s - \lambda_n^c \log \lambda_n^c \]

\[ + \sum_{o \in \Delta} p_{n-1}(o) \log p_{n-1}(o) \lambda^{x_0} \]

\[ + \sum_{o \in \Delta} p_{n-1}(o) \log p_{n-1}(o) \lambda^{c_0} \]

\[ + \sum_{o \in \Delta} p_{n-1}(o) \lambda^{c_0} H(\lambda^{x_0} p^{x_0}). \]
Finally, by substituting (6) in Lemma 2, we have
\[
+ \sum_{o \in \Delta^{n-1}} p_{n-1}(o) \left[ -H(\lambda^a_{\alpha}^{o}, p^{o}_{\alpha}) - H(\lambda^c_{\alpha}^{o}, p^{c}_{\alpha}) \right]
= H(\{\lambda^a_{\alpha}, \lambda^c_{\alpha}\})
+ \sum_{o \in \Delta^{n-1}} p_{n-1}(o) \left[ -\sum_{\delta \in \Delta} p(\delta | o) \log p(\delta | o) \right]
- H(\lambda^a_{\alpha}^{o}, p^{o}_{\alpha}) - H(\lambda^c_{\alpha}^{o}, p^{c}_{\alpha})
= H(\{\lambda^a_{\alpha}, \lambda^c_{\alpha}\})
+ \sum_{o \in \Delta^{n-1}} p_{n-1}(o) \left[ H(\lambda^a_{\alpha}^{o} p^{o}_{\alpha} + \lambda^c_{\alpha}^{o} p^{c}_{\alpha}) \right]
- H(\lambda^a_{\alpha}^{o} p^{o}_{\alpha}) - H(\lambda^c_{\alpha}^{o} p^{c}_{\alpha})
\]

Finally, by substituting (6) in Lemma 2, we have
\[
D(p^a_{\alpha}, p^c_{\alpha}) = H(\{\lambda^a_{\alpha}, \lambda^c_{\alpha}\})
+ \sum_{o \in \Delta^{n-1}} p_{n-1}(o) \left[ -H(\{\lambda^a_{\alpha}, \lambda^c_{\alpha}\}) + D(p^a_{\alpha}, p^c_{\alpha}) \right].
\]

Thus Lemma 3 is established.

REFERENCES


