

OPTIMIZATION OF PROTOCOLS

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAII AT MĀNOA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER SCIENCE

JULY 2025

By

Oscar I. Hernandez

Thesis Committee:

Dusko Pavlovic, Chairperson

Peter-Michael Seidel

Edward V. Ziegler Jr.

Keywords: security, protocol, artificial intelligence, induction

OPTIMIZATION OF PROTOCOLS

Abstract

The development in capabilities of artificial intelligence brings the increased participation of intelligent machines in the protocols of computer networks and society, playing some roles earmarked for machines and others ripe for deception. This exacerbates existing concerns, and it introduces a new dimension to the problems of privacy & security. Whereas a cryptographic protocol can be analyzed with formal methods in terms of the properties of traces it produces, the probabilistic protocols involving potentially deceitful AI participants are analyzed in terms of probability distributions over its traces. In contrast with formal specifications of explicit requirements, the requirements for such AI protocols are specified informally and implicitly by reward models trained on data. A mathematical model of protocol post-training is proposed in terms of an objective function defined by such rewards and regularized by statistical distances from the pre-trained behaviors. It is shown that any instance of such a protocol post-training problem admits solutions at a level of generality that does not depend on particular details of algorithms or computational paradigms, thus showing the existence of optimal behaviors that learning algorithms aim to represent in a way that applies to reinforcement learning algorithms and algorithms in any other paradigm of learning. This establishes the proposed model of protocol post-training as a general setting for reasoning about the opportunities and limitations of protocols involving AI actors.

Contents

Front Matter	1
Contents	3
Notation	4
1 Introduction	5
1.1 Outline	6
2 Language	7
2.1 Motivation	8
2.1.1 Observations	10
Questions	12
2.2 Background	13
2.2.1 Natural language	16
2.2.2 Model of language	16
2.2.3 Resource	17
2.3 Preliminaries	19
2.3.1 Definitions	24
3 Security	26
3.1 Uncertainty	27
3.2 Measure	31
3.3 Measurable property	36
4 Channel Security	39
4.1 Measurable channel	43
4.2 Graph	51
4.3 Probabilistic channel	55
5 Network	57
5.1 Security	60
5.2 Global channel	61
5.3 Communication	62
5.3.1 Adaptive communication	63
6 Protocol	66
6.1 Protocol Post-Training	70
6.1.1 Optimization Problem of Protocol Post-Training	72
Minimizers for Protocol Post-Training	73
Existence of Minimizers for Protocol Post-Training	73
7 Discussion	75
7.1 Outlook	76
Bibliography	77
Back Matter	78
Index	78

Notation

Table 1: Categories

Category	Object	Morphism	Reference
Cat	(small) category	functor	Lambek and Scott [1986]
$\langle \mathbf{Set}, \times \rangle$	(small) set	function	18
$\langle \mathbf{Top}, \times \rangle$	topological space	continuous function	Munkres [2000]
Smgrp	semigroup	Smgrp (homo)morphism	Mac Lane [1998]
Mon	monoid	monoid morphism	Mac Lane [1998]

Table 2: Notation

Symbol	Type	Description	Reference
$X \begin{smallmatrix} f \\ \rightrightarrows \\ g \end{smallmatrix} Y$	$X \rightarrow Y$	functions $f, g : X \rightarrow Y$	Lambek and Scott [1986]
V	type	nodes	Lambek and Scott [1986]
E	type	arrows, oriented edges	Lambek and Scott [1986]
$E \begin{smallmatrix} s \\ \rightrightarrows \\ t \end{smallmatrix} V$	graph	types of nodes V and edges E	9
S, s	Set	sets	Halmos [1960]
$s \subset S$	relation	subset	Halmos [1960]
p	s	point p in set s	
\in	predicate	containment $p \in S$ of p in S	
\ni	predicate	containment $S \ni p$ in S of p	
\mathfrak{P}	Set \rightarrow Set	power set $\{s \in \mathbf{Set} \mid s \subseteq S\}$	18
ι	Set \rightarrow Set	inclusion $\iota : s \rightarrow S$ of $s \subseteq S$ by $\iota(p) = p$	19
\mathbb{S}	type	subjects, users, actors	10
\mathbb{A}	type	actions	17
\mathbb{J}	type	items, objects	17

CHAPTER 1

INTRODUCTION

It is inherently difficult to understand exactly how intelligent machines work, but the history of child rearing suggests that this is not necessary to teach machines to behave as intended.

Remark 1.1 (mechanical communication). The machines have learned to speak and *communicate*.

The purpose [Beer, 2004] of a so-called *language machine* is to interact with other language machines in a multi-agent system governed by a protocol.

On a *microscopic* level, the task of an individual language machine when viewed *in isolation* is to compute, for any input in the form of (say, written) language, an appropriate response [Radford et al., 2018]. A process known as *post-training* is executed to transform a *pre-trained* machine learning (M. L.) model into a *post-trained* ML model so that observations of its microscopic behavior are in better alignment with an intended preference [Christiano et al., 2017]. The analogous process for multi-agent systems is referred to as *protocol post-training*. This raises the following questions.

Question 1.2 (protocol post-training). Consider a multi-agent system governed by a protocol.

1. What is the structure of protocol post-training, that is, what is the problem solved by an instance of protocol post-training and what constitutes a solution?
2. When does an instance of the protocol post-training problem have a solution?

The above questions are largely unexplored at this level of generality for a few reasons. Much of the literature on post-training, for example [Christiano et al., 2017], is supported empirically but not justified from first principles. The literature on the underlying theory, e.g. [Achiam et al., 2017], tends to focus on microscopic behavior instead of the behavior of interactions within a multi-agent system. The theoretical results in reinforcement learning (R. L.), e.g. [Gu et al., 2021, 2023], address this shortcoming and support the following intuition.

Hypothesis 1.3 (protocol optimization). The problem of protocol post-training is one of (constrained) optimization, and instances tend to have solutions.

The above intuition is not formally justified by the existing literature, which restricts the focus to individual RL algorithms instead of the underlying problems of optimization and computation. These optimization problems can also be approached with supervised learning and the choice of learning paradigm often depends less on theoretical considerations than on other practical considerations. While some results from multi-agent reinforcement learning are unique to particular algorithms with particular properties, the core assertion is that there are general principles of protocol post-training. The main contribution of this thesis is a proposed model of protocol post-training presented in Chapter 6 that supports Hypothesis 1.3 without reference to any details of particular algorithms or paradigms of learning.

1.1 Outline

The main research questions under investigation were formulated in this chapter. A hierarchy of structure and results, culminating in Section 6.1, is presented in Chapters 2-5.

In particular, the basic concepts of AI protocol security are presented in Chapter 2. One of the main ideas expressed in Chapter 3 is that the requirements relevant to AI protocols are implicitly specified by *uncertain properties* as opposed to the trace properties given by explicit requirements. The problem of post-training a (directed) probabilistic model is formulated in Chapter 4. An analogous result is shown for a bidirectional communication channel in Chapter 5. The optimization problem of protocol post-training is formulated in Section 6.1.1 and shown to admit solutions. The existence results for the hierarchy of post-training problems are discussed in Chapter 7 together with some directions for future research.

CHAPTER 2

LANGUAGE

Prolegomena

The path to security begins by understanding the nature of security.

Slogan 2.1 (security is being). The security of an artifact is its existence.

Broadly construed, security is the art of assuring that an artifact behaves as intended and thus that the intended artifact exists. A secret password that isn't secret does not exist.

Perfect security can not be achieved, but insecurity can be observed and security is a science in which the claims can be refuted [Pavlovic and Seidel, 2025]. However, there is no prescriptive method for doing science [Feyerabend, 1975]. If there were, the tasks would be discharged to the machines and the problems would be solved. The good news is that understanding is universal in the following sense.

“If you know the way broadly, you will see it in all things.” – Miyamoto [1645]

The reader has a broad operational understanding of language, communication, and reason. The ways of language echo in the machines, in their understanding of the world and notably in our understanding of them. Reality is described by language [Wittgenstein, 1921], and language is the principal tool of our cooperation with the machines. Herein, a (*computing*) *machine* is a human [Offray, 1747] or another physical device that performs computation in the sense of transforming inputs to outputs.

Some roles of language are evident in the multi-agent systems of (computing) machines, but the underlying principle is that all communication has an appropriate description in some language: natural, visual, formal, etc. The user communicates with a Large Language Model (L. L. M.) in a natural language, whereas the system programmer communicates with the LLM in hardware-specific machine language and the cosmos communicates with the LLM in the physical language of molecules and sub-atomic interactions. The language used in this thesis to reason about LLMs will be of a mathematical nature. The approach is to adopt Slogan 2.1 by (1) formulating a precise sense in which the aim of protocol post-training is the production of an artifact whose behavior can be measured, and (2) exhibiting the existence of an ideal artifact with optimal behavior.

Overview

The motivation of this thesis is articulated in Section 2.1.1 on the next page. Then, some relevant background is introduced in Section 2.2 and the basis of the formalism is presented in Section 2.3.

2.1 Motivation

One may wonder about the appropriate language for reasoning about *artificial* computing machines, i.e. non-human computing machines made by humans, and the algorithms that are designed to execute on them. Some aspects are amenable to a form of reasoning in natural language, although the conventional wisdom is that a clearer form of reasoning is enabled by a more appropriate description in some suitably mathematical language. One could reason about an algorithm’s mathematical properties, reason about an algorithm by considering its executable implementation derived from source code in a programming language, or apply a formal reasoning system that operates on a formal specification of the algorithm that is suitable for some task. The key idea is that the algorithm admits descriptions in a variety of languages that encourage different forms of reasoning. One complication is that a given artifact is not amenable to all forms of analysis.

Consider a machine that exhibits *intelligence* in the following sense.

Principle 2.2 (intelligence). The fine-grained behavior of an *intelligent* machine is not *exactly* predictable by a machine of lesser intelligence [Turing, 1950].

While humans have genetic components, insights into questions about financial negotiation are better suited to analysis in natural language than the language of genes.

So, what is the appropriate language for reasoning about AI security? Some of this is done in natural language, some in a wide variety of programming languages, and some in a wide variety of mathematical languages. A number of them are selected, suited for particular purposes. No claim is made of universality, but the main idea should echo in other structures bearing conceptual similarity.

To get an initial sense of the proposed level of analysis, consider a chatbot. In the literature on probabilistic machine learning [Murphy, 2022], a chatbot is often described as a parametrized conditional probability distribution $p_\theta(\mathbf{y} \mid \mathbf{x})$, i.e., a fixed formula that depends on a parameter θ drawn from a vector space of parameters and determines a conditional probability distribution for each input. This level of analysis is useful for reasoning about the computation of gradients and the construction of algorithms.

Instead, the proposed level of analysis models a chatbot as a conditional probability distribution $[\mathbf{x} \vdash_g \mathbf{y}] = g(\mathbf{y} \mid \mathbf{x})$ directly, without reference to a parameter or to any other detail about the program that implements the function. While this may be less immediately useful for reasoning about the computation of gradients and the construction of algorithms, it offers a key advantage over approaches that explicitly model the parameter space and additional forms of structure. The insight is that every model parameterization p_θ determines a conditional probability distribution g , but an arbitrary conditional probability distribution g does not directly determine a model parameterization p_θ or parameter space from which θ would be drawn. For this reason, theorems about arbitrary parametrized probability distribution p_θ generally do *not* directly apply to arbitrary prob-

ability distributions, but theorems about arbitrary probability distributions do apply to arbitrary parametrized probability distributions. The price of this generalization is that the weaker set of initial assumptions limits the number of results to those shared by the general setting. The advantage is that, since the analysis is not done on a particular class of parameter spaces or algorithms, they hold for any set of choices that gives rise to the general structure under consideration. In particular, some of these structures include the information-theoretic *channels* reviewed in Section 2.3.1.

Instead of considering a chatbot as an individual channel in isolation, the ultimate objects of study are collections of machines and their interactions. The context determines whether a pair of interacting machines constitutes a single system.

Definition 2.3 (intelligent system, artificial intelligent system). An *intelligent (computing) system* is defined to be a set of intelligent computing machines with negligible latency in communication, and an *AI system* is defined to be an intelligent computing system containing an artificial computing machine.

While it is unclear of the golem [Ashi and Ravina II, 500], the AI systems are capable of communicating with other machines that reside in a shared instance of the following structure.

Definition 2.4 (graph, node, edge, source, target). A *graph* $G = E \overset{s}{\underset{t}{\rightrightarrows}} V$ consists of a type V of *nodes* and a type E of *edges* equipped with well-defined functions $s, t : E \rightarrow V$ that assign any edge e to its so-called *source* $s(e)$ and *target* $t(e)$.

The simplest graphs are depicted in Figure 2.1 below.

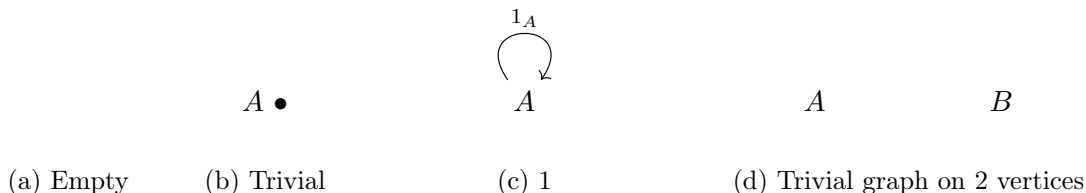


Figure 2.1: Graphs

Informally, a multi-agent system on a graph $E \overset{s}{\underset{t}{\rightrightarrows}} V$ consists of an assignment of nodes to machines such that the presence of an edge $e \in E$ indicates the existence of a medium drawn from a space of functions with a common domain that depends on the source $s(e)$ of e and a common codomain that depends on its target $t(e)$. A configuration, as given by a protocol in Definition 5.4, of a multi-agent system is the selection of a vector belonging to the product of function spaces. The execution of a multi-agent system gives rise to a flow of data along the graph and subsets of users can be equipped with functions that capture the (un)desirability of traces produced by their flows of data.

The task of protocol post-training is, given an initial configuration of a multi-agent system and a global objective, to produce an ideal configuration that optimizes the global objective. The main contribution of this thesis is to formalize this task and show that ideal configurations tend to exist.

2.1.1 Observations

Let $G = E \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} V$ be a graph. In order to model networks of intelligent systems, note that an edge $e \in E$ with source $u = s(e)$ and target $v = t(e)$ can be viewed as a *directed* path e from u to v and denoted $u \xrightarrow{e} v$. This defines a relation on V , denoted $V \xrightarrow[E]{} V$ or $V \rightarrow V$ when the edge set E of a graph is understood, where $u \rightarrow v$ iff¹ there is an edge $u \xrightarrow{e} v$. In general, the relation is not symmetric and the nodes can be distinguished. Although the graph with nodes $V = \{u, v\}$ and edges $\{u \rightarrow v\}$ is isomorphic to the graph on the same nodes with edges $\{v \rightarrow u\}$, the two are not equal. Thus, the terms in V can be thought of as *labeled* nodes.

Let \mathbb{S} [Pavlovic and Seidel, 2025] denote a type whose terms are equivalently called *subjects*, *users*, or *actors* and \mathcal{S} denote a type whose terms are types of subjects. Define a *subject graph* to be a graph $E \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} \mathcal{S}$ in which each node is a type \mathbb{S}_v of subjects, and define the type of subjects for the graph to be the coproduct $\mathbb{S} = \coprod_{\mathbb{S}_v:V} \mathbb{S}_v$ of the types of subjects.

As an example, the mail system is defined on a graph in which edges represent transmission of mail from one node to another and each node represents the subjects in a building.

Consider the graphs depicted in Figure 2.2 below.



Figure 2.2: Imitation game with three players: Alice, Bob, and Carol

The above diagram in Figure 2.2a depicts a graph consisting of a single edge from B to A which represents a direction for the transmission of data, and Figure 2.2b depicts the following example.

Example 2.5 (3-player game of 2-player authentication). The subject *Bob* at node B interacts with the subject *Alice* at node A , and each of these subjects has class 1 if they satisfy a pre-determined property or the class 0 if they do not. To authenticate a subject is to infer their class. In the general case of *2-player authentication*, each speaker can be of either class and Carol must authenticate both subjects after reading a trace of their interaction. In this example, Carol does not transmit messages to the other subjects.

¹if and only if

The imitation game is a specific example in which the class 1 indicates that the subject is human.

The graph $E \xrightleftharpoons[t]{s} V$ in Example 2.5 is given by $V = \{A, B, C\}$, $E = \{A \rightarrow B, B \rightarrow A\}$.

Consider the restriction of Example 2.5 given below.

Example 2.6 (assisted authentication). In the case of *assisted authentication*, it is assumed that Carol and Bob are computer programs under the control of a single operator and that the task is for them to authenticate Alice. Since they are under the control of the same operator, it is assumed that Bob has access to Carol during the interaction with Alice and prior to any interactions with Alice. Since the operator’s task is to correctly authenticate Alice, an instrumental task is for Bob to elicit signals from Alice during their interaction that will assist Carol’s authentication of Alice.

In the above example, the cooperation between the two actors $\{\text{Bob}, \text{Carol}\}$ can be modeled by the identification of nodes $BC = B = C$ with a single type of subjects $\{\text{Bob}, \text{Carol}\}$. One of the reasons to continue distinguishing the two subjects in the authentication team is to add the edge $\{BC \rightarrow BC\}$, enforcing that the two subjects can view all messages from Alice while only Carol can view communications from Bob or Carol.

The structure of communication in Example 2.6 between A and BC can be viewed as a *protocol*, a term formalized in Definition 5.4. Its form is simple in comparison to cryptographic protocols since it merely specifies that the protocol traces consist are concatenations of alternating traces from Alice and Bob. However, it is prohibitively difficult to analyze with the traditional formal methods of cryptographic protocol analysis since the state space is exponential in the size of the vocabularies and the lengths of traces are not assumed to be bounded. This is not the only difficulty.

Recall a common technique used to prove an iterative algorithm for sorting arrays is correct that relies on loop invariants. Given a counter that tracks the iteration of a *for loop* and a prefix of the subarray at that point whose length is given by the counter, one may desire that growing prefix remains sorted at the end of each iteration and declare the state of the algorithm to be *unsafe* if a certain prefix is not sorted by then. This can be viewed as a protocol whose traces are concatenations of the algorithm states, and a class of protocol traces can be declared unsafe if it contains a state which is unsafe. It is straightforward to specify this property of unsafety and compute whether a given trace satisfies it. For the task of verifying the intended behavior on AI system, the properties under investigation are often subtle.

There is a big difference between analyzing an individual sentence in isolation and in context, especially if the context is an interactive conversation.

Consider the refinement of Example 2.6 presented below.

Example 2.7 (authentication attack). Consider the refinement of Example 2.6 in which Alice, unbeknown to Carol, does not satisfy the property under authentication. Then, the task of Carol is to determine this failure of Alice to satisfy the property. Alice is said to launch an *authentication attack* if it participates in the assisted authentication protocol with an intent to be incorrectly

authenticated by Carol. Alice is said to *perform* an authentication attack if its authentication attack is successful.

The performance of an authentication attack in Example 2.7 is a successful act of *deception*. In the imitation game, the adversary Alice aims to deceive the classifier Carol into authenticating it as a person.

The view of deception as an attribute of individual sentences or responses is crude. Instead, it is a property of unsafety exhibited by some traces of interactions. While it is said that the truth comes out in the end, language is inherently ambiguous. The meaning of a sentence interpreted by the reader is determined by the practically infinite context of lived experience, and the sender can only communicate their intended meaning precisely by supplying a practically infinite context. Communication is possible in some respects because these infinite contexts can be approximated with spectra containing high-order information. Much of the precision in perturbative calculations of quantum amplitudes is provided by low-order Feynman diagrams, and much of the meaning of language is carried by composing information-rich concepts that live in high-dimensional spaces. This is why high-dimensional vector embeddings of language are often replaced by lower-order dense embeddings. Just as there remain calculations of quantum amplitudes that cannot be done precisely with perturbative methods, there are concepts that remain ineffable. And truly precise communication in natural language may be more elusive than non-perturbative calculations of quantum amplitudes on smooth spacetime. All this is to say that the meaning, due to the imprecision of finite vocabularies for expressing seemingly infinite concepts, appears inherently ambiguous in a way that safety of sorting algorithms can be unambiguous. Where the unsafety of a sorting algorithm can be formally specified as membership in a binary class and deterministically computed, the unsafety of interaction traces generated by a protocol of natural language is not amenable to binary classification. Deception and other subtle properties are *uncertain* in the sense that their satisfaction by a given trace is not well characterized by a deterministic function with values in $\{0, 1\}$.

Questions

Formal methods for the analysis of cryptographic protocols do not apply to all dimensions of protocols involving potentially deceitful AI actors.

This raises the following questions.

Question 2.8 (structure). Consider a given notion of unsafety.

1. What is the structure of an uncertain property, and how can a notion of unsafety be represented as an uncertain property?
2. What does it mean to strengthen a protocol against an unsafety property?

3. Which set of conditions is sufficient to guarantee the existence of an optimally strong protocol?

The contributions of this article are proposed answers to Questions 1.2 and 2.8.

Consideration of Example 2.7 motivates a definition of *inauthenticity* for traces and of what it means to improve the relevant abilities of Bob and Carol.

The first consideration is how Carol could prepare for the upcoming authentication trial. Given the success of LLMs, suppose the computer program Carol is an LLM and suppose there is a dataset of conversations with labels that indicate the class of each subject. If the dataset is sufficiently large and representative of the conversations to take place during the authentication trials, it is reasonable to expect that the techniques of machine learning can be used to train Carol. This approach could be undermined if Alice was prepared in consultation with a model that was sufficiently similar to Carol. This is likely to be the case, for example, if they both used standard approaches: pre-trained models, publicly available datasets, fine-tuning on similar choices of computer hardware with the same set of techniques under similar constraints.

In order for the authentication team to undermine Alice, they decide to train with a language machine Eve that they suspect is similar to Alice. Then Carol can be used to post-train Bob to have conversations with Eve that elicit a confident prediction from Carol. If the behavior of Eve sufficiently approximates that of Alice, then it is plausible that using it to post-train Bob could result in improved performance on the authentication task, i.e., of decreasing the probability of incorrect authentication.

In terms of post-training, Carol can be seen as a model which punishes incorrect authentication or, equivalently, which rewards correct authentication. Note that it is trained on entire traces of conversations, as opposed to human-annotated pairs consisting of a single prompt and response. This reward model could serve as the basis for traditional reinforcement learning algorithms, or the cumulative nature of the traces can be exploited for “dense” reinforcement learning or other forms of reward shaping. For example, the authentication team could post-train Bob with a loss function that is a product of the standard loss function for language generation and the cost of incorrect authentication. If Carol operates under an *ansatz* that Alice is a particular model, say the pre-trained “DialoGPT-small”, she could do this with a synthetic dataset of interactions between Bob and Eve, another instance of “DialoGPT-small” or perhaps a smaller model like “tiny-gpt2”.

2.2 Background

This section introduces the basic concepts of AI, natural language, language models, and resource security.

Recall the relativity in Definition 2.3.

Example 2.9 (game). Consider the following examples of intelligent systems.

1. In the protocol that consists of a human playing a board game against a machine, the two intelligent systems are viewed as separate actors on separate nodes.
2. In the protocol that consists of a game between two AI-assisted humans, each of the two nodes consists of a human and an AI assistant. This can be refined to a protocol with four nodes to express intricacies of communication between the humans and their AI assistants, but this is not necessary to model if such intricacies are not relevant to the investigation.

To highlight Principle 2.2 in the context of Example 2.9.1, recall that AI systems dominate humans in chess, shogi, and Go [Silver et al., 2018]. The intelligence of these AI systems engenders their success and the unpredictability of their fine-grained behavior in selecting individual moves. If a human or another machine could predict this fine-grained behavior, they could leverage this prediction to select the ideal counter-moves that would lead to victory. This is the basis of the leading AI systems. Their success indicates that they implicitly strike a balance between correctly predicting their opponents' moves while simultaneously making their own moves unpredictable to their opponents. This dynamic is well understood in poker, and it applies in many other games as well as in the social protocols that govern society.

Example 2.10 (essay). Consider the task of writing an essay for a class whose instructor will grade it. If there is a minimum word count of $m = 800$, then a simple grading process can be given in terms of the word count $w \in \mathbb{N}$ by the function $\text{grade} : \mathbb{N} \rightarrow [0, 100]$ given below:

$$\text{grade}_m(w) = \begin{cases} 100 & \text{if } w \geq m \\ \frac{w}{m} & \text{else.} \end{cases} \quad (2.1)$$

Suppose the instructor's grading process exhibits intelligence.

In the above example, a student could play the game by coming to understand the instructor's grading process, i.e., learning to recognize what qualifies an essay for the desired grade and using their understanding to write a good essay. Instructors often expend effort to explain their grading process, but one reality of bad grades is that it is hard for students to discern quality and to develop a mental model of the instructor's grading process. This stems from the very virtue of intelligence.

The complexity of nature limits guarantees about any property of a system to the trust placed in the assumptions from which they follow, so it is intractable to exactly characterize the desired properties of a system that exhibits intelligence.

For example, the assurance of the conservation of "matter" (as opposed to matter-energy) is justified only to the extent it can be trusted not to dissipate and thus not in any context in which there is sufficient benefit to outweigh the cost of accounting for the matter-energy equivalence.

Just as it is valuable for scientists to trust the conservation of matter and for engineers to trust the physical implementation of machines on which their designs depend, it is valuable for security

practitioners to place trust in assumptions of some systems for the greater good of effectively reasoning about other systems for which there is greater value in withholding trust. As Niccolò Machiavelli advised and as the recent history has shown, it can be effective to trust one adversary in order to defeat another adversary, and one should be willing to betray the trust when it is convenient and remain aware of the fact that others share the same incentive. Simultaneously situated in the worlds of science and of Chancellor Machiavelli, the practice of security seeks guarantees founded on trust about processes of deception that are also founded on trust.

Many students in Example 2.10 find it difficult to simulate the intelligence of the instructor's grading processes, but the impossibility of this task should not dissuade them from trying. In fact, this is among the major goals in many courses — for example, to learn what constitutes a mathematical proof or a “succinct” description of an algorithm in the eyes of the grader who is an adversary in many mental models by students. It can be viewed as *the* central goal, for example, if it is believed that a strong understanding of quality unlocks the ability to produce with high quality. Considering the grader's probability distribution of grades over the set of essays, the failure of the student's ability to simulate the intelligence of the grader does not preclude the ability to build a mental model of an uncertain property that approximates it. Students ought to do this and place trust in the assumption that their mental model aligns with the uncertain property in question because optimizing for the mental model, although not guaranteed to yield optimal grades, will cause them to move in a direction that is directionally correct. In any case, it is clearly valuable for the student to build their own model of what makes a good essay.

Is it the case that the ability to measure success, e.g., the correctness of a candidate solution to a problem, is sufficient information to produce success? Obviously, it is not, as evidenced by the existence of aesthetic critics of a craft in which they are not experts; examples are abundant in music, technology, the culinary arts, and science. A notable example is presented below.

Recall the P vs NP question of whether every problem decidable by a non-deterministic Turing machine in polynomial time is also decidable by a deterministic Turing machine in polynomial time. Suppose there is a problem for which there is a deterministic Turing machine that can verify a certificate of correctness in polynomial time. Since the transition function in a non-deterministic Turing machine is set-valued — that is, it can explore any number of states simultaneously — the problem can be decided in polynomial time by a Turing machine which simultaneously evaluates correctness of all candidate solutions and then transitions to the final state. The following slogan captures this observation that the existence of a polynomial-time non-deterministic Turing machine is guaranteed from the assumption of a deterministic Turing machine that verifies whether a candidate solution is correct.

Slogan 2.11 (aesthetics). Aesthetic preference begets craft.

Of course, the question of whether NP is in P remains open. Thus, it is unknown if the existence of this polynomial-time non-deterministic Turing machine guarantees the existence of a

polynomial-time deterministic Turing machine for the same problem. While Slogan 2.11 captures that discrimination enables generation in principle, it does not guarantee the existence of an effective generator or even of an effective procedure for finding a generator whose error is approximately bounded.

Drawing an analogy with the classical concept of verifying correctness deterministically in polynomial time, one feat of intelligent machines is that they are able to evaluate all sorts of things. For example, a neural network can learn to determine whether a sentence is grammatical. In this particular example, it is worth noting that the evidence of their ability to evaluate grammar far predates the evidence of their ability to generate grammatical sentences. There are still examples of tasks for which neural networks can evaluate quality but for which they can not generate artifacts of high quality, and there are other tasks for which they have an as yet unobserved ability to evaluate quality.

In addition to the theory of formal languages that is needed to verify the security of computer network protocols, the growing participation of AI systems in protocols motivates the brief review of natural language in Section 2.2.1, computational models of language in Section 2.2.2, and the mathematical language in Section 2.2.3 as presented in the forthcoming textbook by Pavlovic and Seidel [2025] on *Security Science: Basic Concepts and Mathematical Foundations*.

2.2.1 Natural language

Natural language is composed of fundamental building blocks. The basic units of sound are *phonemes* whose written counterparts are *graphemes*, and the basic units of meaning are called *morphemes*.

Some words are comprised of a single morpheme.

Example 2.12 (monomorphemic). The contemporary English word “nature” is derived from the Old French “nature” and the Latin word “natura”, which meant “innate disposition” and originally “birth”.

Other words are comprised of multiple morphemes.

Example 2.13 (polymorphemic). The English word “philosophy” consists of a morpheme “sophia” which is derived from an Ancient Greek morpheme that means “wisdom” and “philo-”, a participle derived from an Ancient Greek morpheme that means “love”.

The meaning of a word is derived from its constituent morphemes and evolving use.

2.2.2 Model of language

As shown in Figure 2.3, a computational model of language consists of a type V of *tokens* called the *vocabulary*, a *tokenizer* which converts raw text into a sequence of tokens, and a base model which transforms an input sequence of tokens to a token that represents the successor of the sequence.

The tokenizer is trained to learn the tokens, which are subwords, punctuation, and contiguous sequences that frequently arise in the corpus. The tokens generally do not correspond with morphemes, although a simplified picture is that the morphemes are like the tokens learned by humans.

The output of an auto-regressive probabilistic base model of language on a given input sequence is a probability distribution over the vocabulary that represents the token that should follow the input sequence in the corpus.

The base model is post-trained to produce an *instruction model* that obeys the social protocol of a chatbot, for example that recognizes questions and responds with answers. The instruction model is post-trained to satisfy additional requirements, for example of dependability and security. When a *reasoning model* is given an input, it generates outputs that are fed auto-regressively as inputs to itself some number of times before a view of the final output is sent to the user.

2.2.3 Resource

This section reviews the mathematical language of resources presented in [Pavlovic and Seidel, 2025].

The fundamental object under investigation in the security of a system is a type \mathbb{J} whose terms are called *items*, or equivalently, *objects*. In models of language, the type \mathbb{J} models *tokens*, subword-like objects whose sequential compositions form larger subwords, words, phrases, sentences, and higher-order structures.

A typical conversation with a chatbot includes two subjects: the chatbot and the user. These come together with a type \mathbb{A} whose terms are called *actions*. Typical actions include **send** and **receive**. Actions are performed by subjects on objects.

Definition 2.14 (resource model). A *resource model* is a tuple $R = \langle \mathbb{J}, \mathbb{S}, \mathbb{A} \rangle$ that consists of types of objects, subjects, and actions.

Let R be a resource model.

Consider the following example.

Example 2.15 (token). A conversation can be viewed as a resource model $R = \langle V, \{u, p\}, \{\langle \bullet \rangle, (\bullet) \} \rangle$ which consists of a type $\mathbb{J} = V$ of tokens, a pair of subjects, and actions that represent transmission and reception. An element of R represents the transmission of a token by a subject or the reception of a token by a subject.

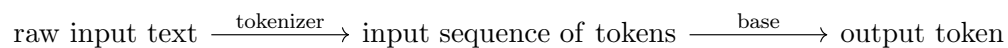


Figure 2.3: Computational model of language

In general, a *reasoning* model communicates with itself using regular tokens that are viewable to the user and *reasoning tokens* that are not viewable to the user. This distinction is captured by a type \mathbb{L} of security levels, for example $\mathbb{L} = \{\text{low} \leq \text{high}\}$, which is equipped with a binary relation that is a partial order. This is formalized in the following definition.

Definition 2.16 (security levels, clearance, locality, multi-level resource model). Let $R = \langle \mathbb{J}, \mathbb{S}, \mathbb{A} \rangle$ be a resource model.

1. A type of *security levels* is a non-empty join-semicomplete lattice $\langle \mathbb{L}, \leq, \vee, \wedge \rangle$.
2. A *clearance* is a function $\text{cl} : \mathbb{S} \rightarrow \mathbb{L}$.
3. A *locality* is a function $\text{pl} : \mathbb{S} \cup \mathbb{J} \rightarrow \mathbb{L}$.
4. A *multi-level resource model* is a structure $M = \langle R, \mathbb{L}, \text{cl}, \text{pl} \rangle$ in which $\leq_{\mathbb{L}}$ is not empty.

In reasoning models of language, the clearance $\text{cl} : \mathbb{S} \rightarrow \mathbb{L}$ expresses the different security levels of the user and the model, as discussed in the following example.

Example 2.17. The token resource model R admits a multi-level resource model $\langle R, \mathbb{L}, \text{cl}, \text{pl} \rangle$, where $\mathbb{L} = \{l, h\}$, $\text{cl}(u) = \text{pl}(u) = l$, $\text{cl}(p) = \text{pl}(p) = h$, objects $\mathbb{J}_R = T_O \sqcup T_R$ consisting of input-output tokens and reasoning tokens, $\text{pl} \upharpoonright_{T_O} (i) = l$ and $\text{pl} \upharpoonright_{T_R} (j) = h$.

The access is captured by the locality $\text{pl} : \mathbb{S} \cup \mathbb{J} \rightarrow \mathbb{L}$ which assigns different security levels to different subjects and objects, as formalized in the following definition.

Definition 2.18 (multi-level security requirement, multi-level security model). A *multi-level security model* is a multi-level resource model $M = \langle \langle \mathbb{J}, \mathbb{S}, \mathbb{A} \rangle, \mathbb{L}, \text{cl}, \text{pl} \rangle$ that satisfies the *multi-level security requirement* stated below:

$$\forall u \in \mathbb{S}, \quad \text{pl}(u) \leq \text{cl}(u).$$

The notation and approach of this thesis are primarily drawn from [Pavlovic and Seidel, 2025]. The reader is also expected to possess a basic familiarity with linear algebra [Strang, 2006] and topology [Munkres, 2000].

The remainder of this section reviews some notation of set theory [Halmos, 1960].

In the remainder of this article, let \mathbf{Set} denote the collection of (small) sets. Let $\mathfrak{P} : \mathbf{Set} \rightarrow \mathbf{Set}$ be the *power set* operator which transforms a set $S \in \mathbf{Set}$ to the power set $\mathfrak{P}(S) = \{s \in \mathbf{Set} \mid s \subseteq S\}$ which consists of all its subsets. Define the *diagonal* on any set S to be the function $\text{diagonal} : S \rightarrow S \times S$ given by $s \mapsto \langle s, s \rangle$, and define the following operation.

Definition 2.19 (image). The image of a function $f : A \rightarrow B$ is defined below:

$$f : \mathfrak{P}A \rightarrow \mathfrak{P}B, \quad S \xrightarrow{f} \bigcup_{a \in S} f(a).$$

For any subset $s \subseteq S$, the *inclusion* of s in S is the injection $\iota : s \rightarrow S$ defined by $\iota = 1 \upharpoonright_s$ and the injection $\iota : S \hookrightarrow \mathfrak{P}S$ of S in $\mathfrak{P}S$ given by $p \mapsto \{p\}$ satisfies $S \cong \iota(S)$.

Recall the following characteristics.

Definition 2.20 (extensivity, monotonicity, idempotency, closure operator). Given a set $S \in \mathbf{Set}$ and function $\text{cl} : \mathfrak{P}(S) \rightarrow \mathfrak{P}(S)$, define the following conditions:

$$\mathbf{extensivity} := s \subseteq \text{cl}(s) \subseteq S \quad \text{for all } s \in \mathfrak{P}(S); \quad (2.2)$$

$$\mathbf{monotonicity} := s \subseteq t \implies \text{cl}(s) \subseteq \text{cl}(t) \quad \text{for all } t, s \in \mathfrak{P}(S); \quad (2.3)$$

$$\mathbf{idempotency} := \text{cl}(\text{cl}(s)) = \text{cl}(s) \quad \text{for all } s \in \mathfrak{P}(S). \quad (2.4)$$

A *closure operator* is a function satisfying *extensivity*, *monotonicity*, and *idempotency*. A set $s \subseteq S$ is *closed* under cl iff $\text{cl}(s) = s$.

Consider the following example of a closure operator.

Example 2.21 (power set). For any set $S \in \mathbf{Set}$, the power set $\mathfrak{P}(S)$ under the binary relation of inclusion forms a lattice $\langle \mathfrak{P}(S), \subseteq, \cup, \cap \rangle$ with a closure operator $\text{cl} : \mathfrak{P}(S) \rightarrow \mathfrak{P}(S)$ defined by $\text{cl}(S) = \{A \subseteq S \mid \exists C \in \mathcal{S}. C \subseteq A\}$.

The basic structures in Section 2.3 assume a basic familiarity with sets [Halmos, 1960], relations [Davey and Priestley, 2002], categories and types [Lambek and Scott, 1986].

2.3 Preliminaries

Major definitions are reviewed in Section 2.3.1 after a brief investigation on security levels.

In the remainder of this thesis, let \mathbb{R} denote the space of real numbers, $\mathbb{R}_{\geq 0}$ denote the subspace $[0, +\infty) \subseteq \mathbb{R}$ of non-negative real numbers, \mathbb{N} denote the subspace of natural numbers, and $[-\infty, +\infty] = [-\infty, +\infty]$ denote the space of extended real numbers.

Let M be a multi-level security model with a type \mathbb{S} of subjects, a type \mathbb{L} of levels, and a clearance $\text{cl} : \mathbb{S} \rightarrow \mathbb{L}$.

The first ideas in this section is that clearance cl extends the relation from \mathbb{L} to \mathbb{S} .

Proposition 2.22. Given the reflexive relation $\langle \mathbb{L}, \leq \rangle$, the clearance $\text{cl} : \mathbb{S} \rightarrow \mathbb{L}$ equips \mathbb{S} with a preorder \lesssim_{cl} .

Proof. The result follows from Proposition 2.25. □

The language of categories and functors [Lambek and Scott, 1986] is useful for arguments about universality, such as the proof of the above proposition, but its necessity is limited to this section and the unfamiliar reader can safely proceed to Section 2.3.1.

The familiar reader with categories should recall that the category **Set** of (small) sets and functions is *complete*, i.e., that it contains all small limits. In order to show that \mathbb{S} inherits a relation from \mathbb{L} via cl , let L be an arbitrary non-empty type with a non-empty binary relation R_L and P be a type with a function $l : P \rightarrow \langle L, R_L \rangle$.

Recall that R_L can be viewed as a non-empty subset $R_L \subseteq L \times L$, and consider the cospan $R_L \xrightarrow{\iota_L} L \times L \xleftarrow{l \times l} P \times P$ depicted in Figure 2.4.

$$\begin{array}{ccc}
 & P \times P & \\
 & \downarrow l \times l & \\
 R_L & \xrightarrow{\iota_L} & L \times L
 \end{array}
 \qquad
 \begin{array}{ccc}
 (l \times l)^{-1}(R_L) & \xleftarrow{\iota_P} & P \times P \\
 \downarrow l \times l & & \downarrow l \times l \\
 R_L & \xrightarrow{\iota_L} & L \times L
 \end{array}$$

(a) Cospan $R_L \xrightarrow{\iota_L} L \times L \xleftarrow{l \times l} P \times P$

(b) Pullback of the diagram in Figure 2.4a

Figure 2.4: Relation R_P on P induced by the function $l : P \rightarrow \langle L, R_L \rangle$

The cospan determines a relation on P .

Proposition 2.23 (pullback relation). Given a type P and non-empty relation $R_L \subseteq L \times L$, any function $l : P \rightarrow L$ induces a unique relation $(l \times l)^{-1} R_L$ on P , and it satisfies the following conditions:

1. universality: it is the pullback of the cospan $R_L \xrightarrow{\iota_L} L \times L \xleftarrow{l \times l} P \times P$ depicted in Figure 2.4a;
2. maximality: it is the largest relation whose image under $l \times l$ is contained in R_L ;
3. reflexivity: it is reflexive if R_L contains diagonal $(l(P))$; and
4. transitivity: it is transitive if R_L is transitive.

Proof (pullback). Let R_L and $l : P \rightarrow L$ be as stated.

Recall that all small limits in the complete category **Set** exist. The pullback of the cospan $R_L \xrightarrow{\iota_L} L \times L \xleftarrow{l \times l} P \times P$ in Figure 2.4a is given by the span $P \times P \xleftarrow{\iota_P} (l \times l)^{-1}(R_L) \xrightarrow{l \times l} R_L$ depicted in Figure 2.4b.

Universality of this pullback $R_l = (l \times l)^{-1}(R_L)$ implies that the cospan $R_L \xleftarrow{l \times l} R_P \xrightarrow{i_{R_P}} P \times P$ depicted in Figure 2.5a factors through a unique function such that the diagram in Figure 2.5b commutes. Commutativity yields inclusion of R_P and thus maximality of R_l .

Universality implies that the cospan $P \times P \xleftarrow{\text{diagonal}_P} P \xrightarrow{\text{diagonal}_L \circ l} R_L$ depicted in Figure 2.6a factors through a unique function $\delta : P \rightarrow R_l$ such that the diagram in Figure 2.6b commutes. Commutativity yields reflexivity of R_l .

Let $R_l \times_P R_l = \{(p, q), (q', r) \in R_l \times R_l \mid q = q'\}$ consider the cospan $P \times P \xleftarrow{\iota_l} R_l \times_P R_l \xrightarrow{(l \circ \pi_{1,1}) \times (l \circ \pi_{2,2})} R_L$ depicted in Figure 2.7a. Universality of the pullback implies it factors through a unique function $c : R_l \times_P R_l \rightarrow R_l$ such that the diagram in Figure 2.7b commutes. Commutativity yields transitivity of R_l . \square

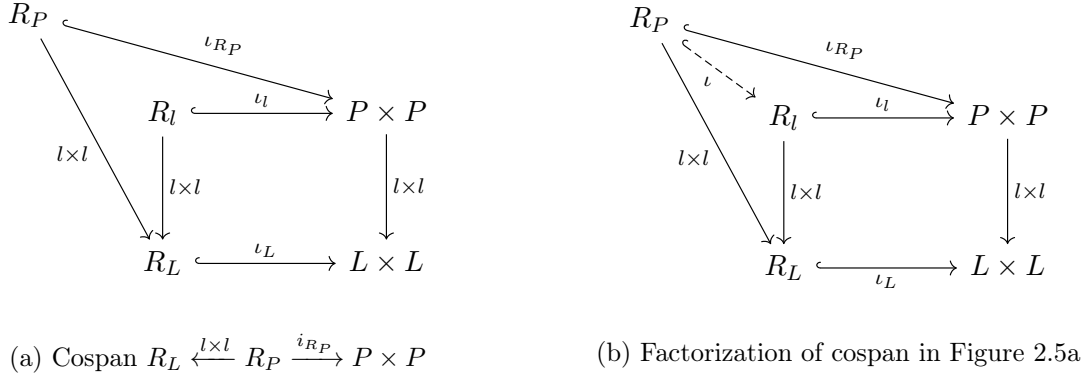


Figure 2.5: Maximality

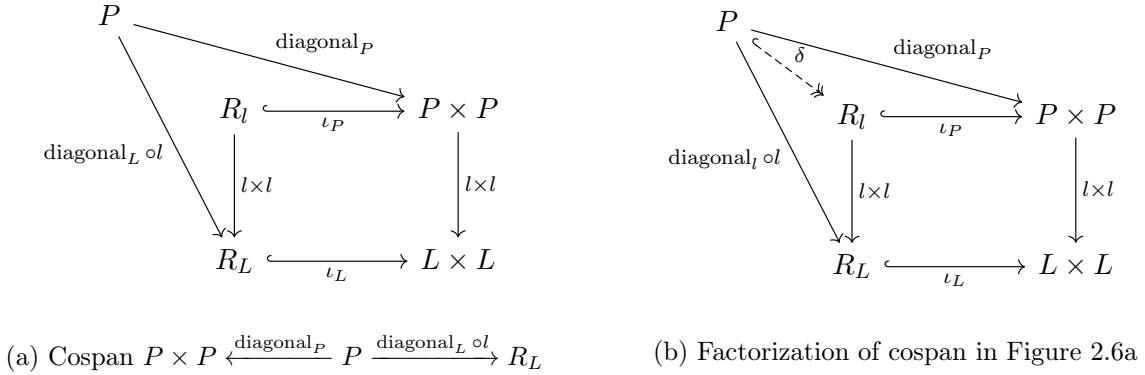


Figure 2.6: Reflexivity

Thus, the relation R_L on L determines a relation on P that is universal in the sense defined by the above proposition.

Definition 2.24. Given a relation $R_L \subseteq L \times L$ and a function $l : P \rightarrow L$, define the relation $R_l \subseteq P \times P$ to be $R_l = (l \times l)^{-1}(R_L)$.

This relation R_l inherits some characteristics from R_L .

Proposition 2.25. Given a type P and a non-empty preorder $\langle L, \lesssim \rangle$, any function $l : P \rightarrow L$ induces a relation \lesssim_l that is a preorder on P .

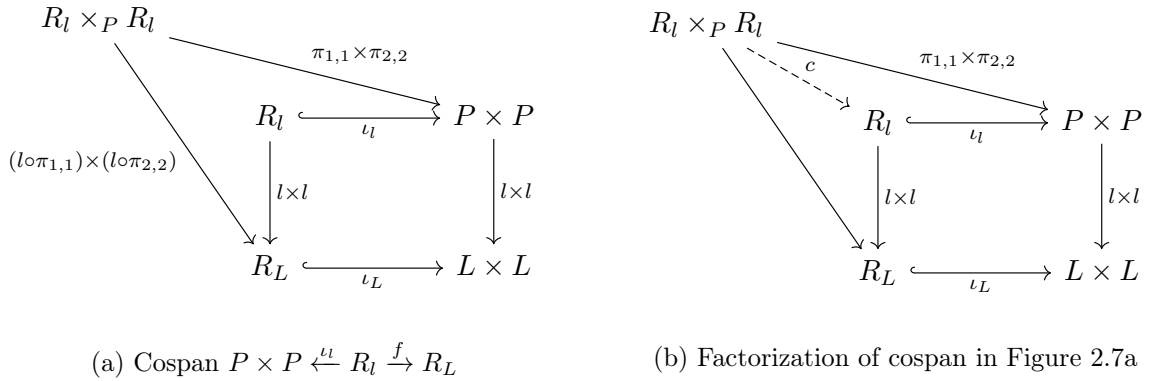


Figure 2.7: Transitivity

Proof. The result follows from Proposition 2.23. □

Asymmetry of \lesssim_l follows from injectivity of l and it is the largest relation on which l is monotonic. Such a relation is induced on the type \mathbb{S} of subjects as shown in Proposition 2.22.

Any relation $\langle P, R_P \rangle$, such as a lattice \mathbb{L} of security levels or any set that maps to it, admits a closure operator as presented in Definition 2.20. Let S denote a set. One such closure operator is defined below.

Definition 2.26 (upper closure, downward directed set, non-triviality, prefilter, pointed prefilter). Given a relation $\langle S, R \rangle$ and letting U denote a subset of S , define:

1. the *upper closure* $\uparrow: \mathfrak{P}S \rightarrow \mathfrak{P}S$ by $s \mapsto \{u \in S \mid \exists a \in s. aRu\}$;
2. a set to be an *upper set* and *upward closed set* iff it is closed under \uparrow ;
3. a set D to be *downward directed set* iff any pair $u, v \in D$ have a $d \in D$ such that $d \leq u$ and $d \leq v$;
4. a set is *non-trivial* if it is not empty;
5. a *prefilter* P to be a non-trivial set that is downward directed; and
6. a *pointed prefilter at a point* $s \in S$ to be a prefilter $P \subseteq S$ such that $s \in \uparrow P$;

The prefix *pre* in the above definitions suggests the following definition.

Definition 2.27 (filter, pointed filter, principal filter). Given a relation $\langle S, R \rangle$, define:

1. a *filter* F to be a prefilter that is upper closed;
2. a *pointed filter at a point* $s \in S$ to be a pointed prefilter P at s that is a filter; and

3. $\uparrow: S \rightarrow \mathfrak{P}S$ at $s \in S$ to be the *principal filter* $\uparrow s = \uparrow \{s\}$ at s .

Consider a system of filters on the power set $\mathfrak{P}(X)$ of X under the binary relation \supseteq .

Definition 2.28 (π -system, filter system). Given a set X :

1. a *filter system* is a collection $\mathcal{B} = \{\mathcal{B}_x \subseteq \mathfrak{P}(X)\}_{x \in X}$ of filters on the poset $\langle \mathfrak{P}(X), \supseteq \rangle$; and
2. a π -*system* is a non-empty set of subsets of X that is closed under finite intersections.

A filter system gives rise to a π -system that models proximity.

Definition 2.29 (neighborhood filter at a point, topology, topological space, open set). Given a set X and filter system \mathcal{B} :

1. the *neighborhood filter* $\mathcal{N}(\mathcal{B}_x)$ at x is the pointed filter $\uparrow \mathcal{B}_x$ at \mathcal{B}_x ;
2. the *topology* \mathcal{T} generated by \mathcal{B} is the π -system defined below

$$\mathcal{T} = \{O \in \mathfrak{P}(X) \mid \forall x \in O, \exists N \in \mathcal{N}(\mathcal{B}_x). N \subset O\};$$

and

3. the *topological space* generated by \mathcal{B} is $\langle X, \mathcal{T} \rangle$;
4. an *open set* in X is a subset $O \subseteq X$ in \mathcal{T} .

Topology is the study of topological spaces and *continuity*, and it is the mathematical language of geometry. Use of the term *space* is meant to be evocative of a topological space. This should be taken literally if a topological space is clear from the context of use and suggestively otherwise, e.g., because the author intends to define a topology in the future.

Any preorder $\langle S, \lesssim \rangle$ gives rise to a space defined by the upper closure $\uparrow = \uparrow_{\lesssim}$. In general, subscripts may be safely dropped if they are clear from context. In particular, the upper closure \uparrow of a relation generates the following topology.

Definition 2.30. Given a binary relation $\langle S, \lesssim \rangle$,

1. define the *upper basis* to be $\mathcal{U} = \{\uparrow_{\lesssim} s \mid s \in S\}$; and
2. define the *upper topology* to be the topology generated by the upper basis \mathcal{U} .

Not every topology is given in terms of an upper closure \uparrow , but the upper topology on a preorder should be considered in the absence of an alternative specification.

2.3.1 Definitions

This section reviews the definitions from [Pavlovic and Seidel, 2025] that form the basis of this thesis.

Let $R = \langle \mathbb{J}, \mathbb{S}, \mathbb{A} \rangle$ be a resource model and $M = \langle R, \mathbb{L}, \text{cl}, \text{pl} \rangle$ be a multi-level security model. They give rise to the following data.

Definition 2.31 (event space, event, event string). Given a resource model $R = \langle \mathbb{A}, \mathbb{J}, \mathbb{S} \rangle$ and multi-level security model $M = \langle R, \mathbb{L}, \text{cl}, \text{pl} \rangle$, the *event space* is the product type $\Sigma = \mathbb{J} \times \mathbb{A} \times \mathbb{S} \times \mathbb{L}$ and an *event* is a point in Σ . Given an event type Σ , the type of event strings is the free semigroup $\langle \Sigma^+, :: \rangle$ generated by Σ under the binary operation of concatenation $::$ and an *event string* is a point in Σ^+ .

An event can be interpreted as the performance of an action by a subject on an object. Let Σ be a type of events and consider sequences of such events.

Definition 2.32 (trace). The *trace space* over a type Σ of events is the free monoid $\langle \Sigma^*, ::, \varepsilon \rangle$ with identity $\varepsilon = ()$, and a *trace* is a point in Σ^* . It is equipped with a *prefix* relation $\Sigma^* \sqsubseteq \Sigma^*$ for which $\mathbf{t}, \mathbf{v} \in \Sigma^*$ satisfy $\mathbf{t} \sqsubseteq \mathbf{v}$ iff there is a trace $\mathbf{u} \in \Sigma^*$ such that $\mathbf{t} :: \mathbf{u} = \mathbf{v}$. If $\mathbf{t} \sqsubseteq \mathbf{v}$, then \mathbf{t} is called a *prefix* of \mathbf{v} and \mathbf{v} is called an *extension* of \mathbf{t} .

Note that any trace $\mathbf{t} \in \Sigma^*$ is a prefix of itself.

Note that traces are concatenations of events.

Definition 2.33 (length). For any type Σ of events, the *length* function $|\bullet| : \Sigma^* \rightarrow \mathbb{N}$ is the assignment from a trace $\mathbf{t} = r_1 r_2 \cdots r_{n-1} r_n$ to its length $|\mathbf{t}| = n$.

An extension of a prefix is said to be a *proper* extension if they are not equal. Then, transitivity of the prefix relation makes it a preorder. Thus, the prefix relation induces the *upper topology* on Σ^* . If not explicitly specified otherwise, the topology on the trace space is the upper topology given by the prefix relation.

A (*certain*) *property* is a point P in $\mathfrak{P}\Sigma^*$, and $\Delta : \mathbf{Set} \rightarrow \mathbf{Meas}$ assigns any set S to the set ΔS of probability distributions over S .

Let \mathcal{X}, \mathcal{Y} denote types of events with event strings $\mathcal{X}^+, \mathcal{Y}^+$ and trace spaces $\mathcal{X}^*, \mathcal{Y}^*$. The idea is that the trace space is a language, and that strings of a source language transition to events in a target language.

Definition 2.34 (stochastic matrix). A *stochastic matrix* from \mathcal{X} to \mathcal{Y} is a function $f : \mathcal{X}^+ \rightarrow \Delta\mathcal{Y}$ denoted $[\mathbf{x} \vdash_f y] = f(\mathbf{x})(\{y\})$.

Alternatively, traces of the source language transition to traces of the target language.

Definition 2.35 (probabilistic channel, continuous probabilistic channel). A *probabilistic channel* is a function $g : \mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$ denoted $[\mathbf{x} \vdash_g \mathbf{y}] = g(\mathbf{x})(\{\mathbf{y}\})$. A *continuous probabilistic channel* is a function $\gamma : \Delta\mathcal{X}^* \xrightarrow{\Sigma} \Delta\mathcal{Y}^*$ preserving convex combinations and *finitude*² of non-zero values.

Some probabilistic channels arise from stochastic matrices.

Definition 2.36 (cumulative probabilistic channel). A *cumulative* probabilistic channel is a probabilistic channel $g : \mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$ for which there is a stochastic matrix $f : \mathcal{X}^+ \rightarrow \Delta\mathcal{Y}$ satisfying the equation $[x_0 \dots x_n \vdash_g y_0 \dots y_n] = \prod_{i=0}^n [x_0 \dots x_i \vdash_f y_i]$.

They give rise to continuous probabilistic channels.

Definition 2.37 (continuation). The *continuation* of a probabilistic channel $g : \mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$ is the continuous probabilistic channel $\bar{g} : \Delta\mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$ given by $[[\mathbf{x}] \vdash_{\bar{g}} \mathbf{y}] = \sum_{\mathbf{t} \in \mathcal{X}^*} [\mathbf{x}](\mathbf{t}) [\mathbf{t} \vdash_g \mathbf{y}]$.

The problem of probabilistic protocol post-training subsumes a specialized problem of probabilistic channel post-training. This specialized problem is formulated in the following section on trace spaces and security.

²Finitude or *finiteness* is the property of being finite.

CHAPTER 3 SECURITY

In order to build intuition about probabilistic channels, this chapter reviews the structure of spaces of channels after reviewing the structure of channel endpoints.

Let Σ be a type of events and equip the free monoid $\langle \Sigma^*, ::, \varepsilon, \sqsubseteq \rangle$ of traces with the upper topology generated by the prefix relation \sqsubseteq . Recall that the relation $\mathbf{t} \sqsubseteq \mathbf{v}$ iff \mathbf{v} is an extension of \mathbf{t} .

Consider a trace $\mathbf{t} \in \Sigma^*$. By definition, the upper closure \uparrow_{\sqsubseteq} of \mathbf{t} is the set $\uparrow \mathbf{t}$ of extensions of \mathbf{t} . To characterize the upper topology on Σ^* , recall from the definition of the upper topology on a poset that $\uparrow \mathbf{t}$. The open sets in Σ^* are formed by closing the basic open sets under unions and finite intersections.

Observe that the empty trace $\varepsilon = ()$ is a prefix of every trace since the free monoid $\langle \Sigma^*, ::, \varepsilon \rangle$ satisfies the identity law $\varepsilon :: \mathbf{t} = \mathbf{t}$.

Proposition 3.1 (empty trace). In the upper topology \mathcal{T} on Σ^* , the following hold:

1. $\uparrow \varepsilon = \Sigma^*$;
2. for any $\mathbf{t} \in \Sigma^*$, $\varepsilon \in \uparrow \mathbf{t}$ iff $\mathbf{t} = \varepsilon$;
3. for any open set $U \in \mathcal{T}$, $\varepsilon \in U$ iff $U = \Sigma^*$; and
4. for any non-empty set $S \in \mathfrak{P}\Sigma^*$, $\varepsilon \in \downarrow S$.

Proof. Consider the upper topology \mathcal{T} on Σ^* .

Recall that any trace $\mathbf{t} \in \Sigma^*$ satisfies $\varepsilon \sqsubseteq \mathbf{t}$. Consider any trace $\mathbf{t} \in \Sigma^*$.

1. It follows from the definition of \uparrow that $\uparrow \varepsilon = \{\mathbf{u} \in \Sigma^* \mid \varepsilon \sqsubseteq \mathbf{u}\}$ and thus that $\uparrow \varepsilon = \Sigma^*$.
2. Observe $\varepsilon \in \uparrow \mathbf{t}$ iff $\mathbf{t} \sqsubseteq \varepsilon$ iff $\mathbf{t} = \varepsilon$.
3. Let U be an open set in \mathcal{T} . Since the only basic open set in \mathcal{T} that contains ε is $\uparrow \varepsilon$, it follows that $U \subseteq \Sigma^*$ contains $\uparrow \varepsilon = \Sigma^*$ and thus that $U = \Sigma^*$.
4. Consider any subset $S \subseteq \Sigma^*$, point $q \in S$, and prefix $p \in \uparrow q$. Since $\varepsilon \sqsubseteq p$, it follows that $\varepsilon \in \downarrow q$ and from generality of q that $\varepsilon \in \downarrow S$.

□

The empty trace is an extension of only itself, and it can be said to be distant from other traces since it is not contained in any proper open subset of Σ^* . The final item in the above proposition

says that the empty trace is contained in the downward closure of any set. All of these characteristics of a trace in Σ^* are unique to ε .

These unique features of ε can be used to characterize much about the entire space of traces.

Proposition 3.2 (trace shape). The upper topology on Σ^* is a connected, compact, Kolmogorov topological space that satisfies the T_1 separation condition iff $|\Sigma| = 0$.

Proof. Consider the upper topology \mathcal{T} on Σ^* .

To show \mathcal{T} is connected, note that a pair of non-empty open sets U, V covering $U \cup V = \Sigma^*$ cannot be disjoint since the containment of ε in $U \cup V$ implies $U = \Sigma^*$ or $V = \Sigma^*$.

To show \mathcal{T} is compact, note that any open cover \mathcal{U} of $\Sigma^* \ni \varepsilon$ must contain the finite subcover $\{\Sigma^*\} \subseteq \mathcal{U}$.

If $|\Sigma| = 0$ then $\Sigma^* = \{\varepsilon\}$ is a Hausdorff space that thus satisfies the T_1 separation condition.

Suppose $|\Sigma| \neq 0$. To determine the separability of \mathcal{T} , consider any pair $\mathbf{t}, \mathbf{u} \in \Sigma^*$ of traces and observe that $U = \uparrow \mathbf{u}$ is an open set in \mathcal{T} satisfying $\mathbf{u} \in U$. If $\mathbf{u} \not\sqsubseteq \mathbf{t}$, then $U \in \mathcal{T}$ does not contain \mathbf{t} . Else, suppose $\mathbf{u} \sqsubseteq \mathbf{t}$. Since $V = \uparrow \mathbf{t}$ is an open set in \mathcal{T} such that $\mathbf{t} \in V$ and $\mathbf{u} \notin V$, it follows that \mathcal{T} satisfies the T_0 separation condition. Since any open set containing \mathbf{u} must also contain \mathbf{t} , \mathcal{T} does not satisfy the T_1 separation condition. \square

3.1 Uncertainty

Consider, for a given prompt to a language model, the response prior to sampling. The language model computes logits and applies a method such as *softmax* to compute a probability distribution over the space of tokens. At a higher level, the language model computes probability distributions over the space of multi-token sequences. If the response tokens belong to a set Σ , then a response corresponds to a probability distribution over Σ^* and the set of responses is given by the set $\Delta\Sigma^*$ of probability distributions over Σ^* .

Elements of the power set $\mathfrak{P}\Sigma^*$ are known as trace properties [Pavlovic and Seidel, 2025] and elements of $\mathfrak{P}(\mathfrak{P}\Sigma^*)$ are known as hyperproperties [Clarkson and Schneider, 2010].

Definition 3.3 (uncertain property). A (*probabilistic*) *uncertain property* of traces in Σ^* is a point in $\Delta\Sigma^*$.

In addition to the role of trace properties in the analysis of protocols, the analysis of AI protocols requires a theory of *uncertain properties*.

The probabilistic theory of uncertain property is built on the following algebra of sets.

Definition 3.4 (trace algebra, probabilistic mass). Given any topology on Σ^* , define the *trace algebra over the set Σ^** to be the algebra $\mathcal{B}\Sigma^*$ of Borel sets in Σ^* , and define a *probabilistic mass* to be a function $\mu : \Sigma^* \rightarrow [0, 1]$ satisfying $\sum_{\mathbf{t} \in \Sigma^*} \mu(\mathbf{t}) = 1$.

Unless explicitly specified otherwise, the topology on Σ^* used to construct the trace algebra is the upper topology generated by the prefix relation on Σ^* . The trace algebra defines a measurable space.

Proposition 3.5 (point measurability). The upper topology on Σ^* is a measurable space under the Borel σ -algebra $\mathcal{B}\Sigma^*$ in which any trace $\mathbf{t} \in \Sigma^*$ gives rise to a measurable set $\{\mathbf{t}\}$ in $\mathcal{B}\Sigma^*$. Any probabilistic uncertain property $[\mathbf{t}] \in \Delta\Sigma^*$ gives rise to a probabilistic mass $\mu_{[\mathbf{t}]} : \Sigma^* \rightarrow [0, 1]$ defined by $\mu_{[\mathbf{t}]}(\mathbf{u}) = [\mathbf{t}](\{\mathbf{u}\})$, and any probabilistic mass $\mu : \Sigma^* \rightarrow [0, 1]$ gives rise to a probabilistic uncertain property $[\mathbf{t}]_\mu : \mathcal{B}\Sigma^* \rightarrow [0, 1]$ in $\Delta\Sigma^*$ defined by $[\mathbf{t}]_\mu(E) = \sum_{\mathbf{u} \in E} \mu(\mathbf{u})$.

Proof. Consider the upper topology on Σ^* and the Borel algebra $\mathcal{B}\Sigma^*$ generated by it.

Observe that $\langle \Sigma^*, \mathcal{B}\Sigma^* \rangle$ is a measurable space since the closure of $\mathcal{B}\Sigma^*$ under countable unions holds by construction.

To show for any trace $\mathbf{t} \in \Sigma^*$ that the singleton $\{\mathbf{t}\}$ is a measurable set in $\mathcal{B}\Sigma^*$, observe that this result is trivial if $|\Sigma| = 0$ and suppose $|\Sigma| \neq 0$. Then, observe that the non-empty coset $U = \mathbf{t} :: \Sigma :: \Sigma^*$ is in the Borel σ -algebra $\mathcal{B}\Sigma^*$ closed under difference and the open set $V = \uparrow\mathbf{t}$ in $\mathcal{B}\Sigma^*$ satisfies $U \subseteq V$. Then, $\mathcal{B}\Sigma^*$ contains the difference $V \setminus U = \{\mathbf{t}\}$.

For any uncertain property $[\mathbf{t}] \in \Delta\Sigma^*$, the requirement that $[\mathbf{t}](\Sigma^*) = 1$ guarantees that $\mu_{[\mathbf{t}]} : \Sigma^* \rightarrow [0, 1]$ satisfies $\sum_{\mathbf{u} \in \Sigma^*} \mu_{[\mathbf{t}]}(\mathbf{u}) = \sum_{\mathbf{u} \in \Sigma^*} [\mathbf{t}](\{\mathbf{u}\}) = [\mathbf{t}]\left(\bigcup_{\mathbf{u} \in \Sigma^*} \{\mathbf{u}\}\right) = 1$ by countable additivity. Given any probabilistic mass $\mu : \Sigma^* \rightarrow [0, 1]$, the construction guarantees that $[\mathbf{t}]_\mu : \mathcal{B}\Sigma^* \rightarrow [0, 1]$ is a measure for which $[\mathbf{t}]_\mu(\Sigma^*) = \sum_{\mathbf{u} \in \Sigma^*} \mu(\mathbf{u})$ is equal to 1 by definition. \square

This allows the notation overload which identifies $\mu(\mathbf{t}) = \mu(\{\mathbf{t}\})$.

One way to capture the significance of shorter traces in contrast with longer traces is with the following structure.

Definition 3.6 (weight). Given any countable space X , define a weight on X to be a function $w_X : X \rightarrow (0, +\infty)$ such that $\sum_{x \in X} w_X(x)$.

Weights are mainly defined on a trace space Σ^* , which is countable if the number of distinct events in Σ is countable. Let $\alpha \in (-1, +1)$, and observe that it gives rise to a weight on Σ^* .

Definition 3.7 (exponential weight). Given any type Σ of events and $\alpha \in (-1, +1)$, the *exponential weight* function $w_\alpha : \Sigma^* \rightarrow (0, +\infty)$ is defined by $w_\alpha^{(\alpha)}(\mathbf{t}) = |\alpha^{|\mathbf{t}|}|$.

If a weight is not specified on a countable space, it should be assumed to be an exponential weight given by a constant α that is 0.5 unless otherwise specified.

Definition 3.8 (probabilistic product, probabilistic weight). Given any type Σ of events and weight w on Σ^* , define:

1. *probabilistic product* $\langle \bullet, \star \rangle_w : \Delta\Sigma^* \times \Delta\Sigma^* \rightarrow \mathbb{R}_{\geq 0}$ as follows

$$\langle \mu, \nu \rangle_w = \sum_{\mathbf{t} \in \Sigma^*} \mu(\{\mathbf{t}\}) \nu(\{\mathbf{t}\}) w(\mathbf{t}) \quad (3.1)$$

for any $\mu, \nu \in \Delta\Sigma^*$; and

2. the *probabilistic weight* function $\|\bullet\|_w : \Delta\Sigma^* \rightarrow \mathbb{R}_{\geq 0}$ defined by $\|\bullet\|_w = \sqrt{\langle \bullet, \bullet \rangle_w}$.

The subscripts can be safely dropped if they can be inferred from context.

Equip $\langle \Sigma^*, \langle \bullet, \star \rangle \rangle$ with the probabilistic weight $\|\bullet\|$.

The probabilistic product provides a notion of angle from the first argument to the second argument, and it happens to be non-negative.

Definition 3.9 (positivity, divergence). A non-negative function $D : S \times S \rightarrow [-\infty, +\infty]$ is *positive* if $\ker D = \text{diagonal}(S)$, and a *divergence* over a set S is a non-negative function $D : S \times S \rightarrow [0, +\infty]$ that is *positive*.

Observe that the non-negative function $\langle \bullet, \star \rangle$ is not positive, which can be seen by observing that one of the non-negative terms on the right-hand side of Equation 3.1 must be positive since $w_{\mathbf{t}}$ is positive for all \mathbf{t} and a probability distribution that must sum to 1 cannot assign 0 to all singletons since their union is Σ^* .

Intuitively, a divergence is a kind of distance.

Definition 3.10 (metric). A *metric (distance)* on X is a divergence $d : X \times X \rightarrow \mathbb{R}_{\geq 0}$ that is symmetric and satisfies the triangle inequality stated below

$$\forall x, y, z \in X, \quad d(x, z) \leq d(x, y) + d(y, z).$$

A divergence over $\Delta\Sigma^*$ provides a notion of *statistical* distance from one probability distribution to another which is directed since it is not assumed to be symmetric, unlike a metric distance. An example of an asymmetric divergence is the relative entropy of Kullback and Leibler [1951]. Such divergences play a certain role in the proposed framework, but the framework also requires a statistical divergence that interacts well with the notion of angle provided by the probabilistic product and the notion of length provided by the probabilistic weight. A distance can be viewed as a kind of length; in vector algebra, the distance from one vector to another is the length of the difference. The intuition points to defining a probabilistic distance with a formula akin to $d(\bullet, \star) = \|\star - \bullet\| = \|\star + (-\bullet)\|$. However, a proper definition of such a formula requires a definition of subtraction or, equivalently, addition and additive inversion. Since an uncertain property can be viewed as a functions on Σ , it is tempting to define addition point-wise.

Hypothesis 3.11. Addition of measures can be defined pointwise.

However, the space $\Delta\Sigma^*$ of probability measures is not closed under this operation. The task remains to implement a metric notion of statistical distance on $\Delta\Sigma^*$ and, with additional data, on a space of channels with values in $\Delta\Sigma^*$.

Recalling Example 2.7, the metric notion of statistical distance could be used to ask a precise refinement of the following question.

Question 3.12. Is it possible for Carol to post-train Bob while bounding the degradation of his performance on other tasks, for example as measured by a statistical distance from the pre-trained model to the post-trained model?

To answer this question affirmatively by way of a more general picture in Chapter 5, consider arbitrary event types \mathcal{X} and \mathcal{Y} .

Definition 3.13 (space of stochastic matrices). Given event types \mathcal{X} and \mathcal{Y} , let $\mathcal{F}_\Delta(\mathcal{X}, \mathcal{Y}) = \{f : \mathcal{X}^+ \rightarrow \Delta\mathcal{Y}\}$ denote the space of stochastic matrices from \mathcal{X} to \mathcal{Y} .

Recall that stochastic matrices accumulate to cumulative probabilistic channels.

Definition 3.14 (space of cumulative probabilistic channels, probabilistic product, probabilistic weight). Given event types \mathcal{X}, \mathcal{Y} with weights $w_X : \mathcal{X}^* \rightarrow (0, +\infty)$ and $w_Y : \mathcal{Y}^* \rightarrow (0, +\infty)$, define:

1. the space $\Delta(\mathcal{X}, \mathcal{Y}) = \{g : \mathcal{X}^* \rightarrow \Delta\mathcal{Y}^* \mid \exists f \in \mathcal{F}(\mathcal{X}, \mathcal{Y}) . g = f^*\}$ of cumulative probabilistic channels from \mathcal{X} to \mathcal{Y} ;
2. the *probabilistic channel product* $\langle \bullet, \star \rangle_{w_X, w_Y} : \Delta(\mathcal{X}, \mathcal{Y}) \times \Delta(\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{R}_{\geq 0}$ below

$$\langle g, h \rangle_{w_X, w_Y} = \sum_{\mathbf{x} \in \mathcal{X}^*} \langle g(\mathbf{x}), h(\mathbf{x}) \rangle_{w_Y} w_X(\mathbf{x});$$

and

3. the *probabilistic weight* function $\|\bullet\|_{w_X, w_Y} : \Delta(\mathcal{X}, \mathcal{Y}) \rightarrow [0, +\infty)$ as follows.

$$\|\bullet\|_{w_X, w_Y} = \sqrt{\langle \bullet, \bullet \rangle_{w_X, w_Y}}$$

The idea is that an instance of post-training involves an initial point in a space of channels, a given uncertainty property over the space of inputs, a desired uncertainty over the space of outputs (learned from data), a mechanism for measuring the utility of a channel based on how well the uncertain property it produces over the space of outputs aligns with the desired uncertain property over outputs, and the task of finding an extremal point in the space of channels. This requires an investigation into optimization on the geometry of uncertain properties that starts by defining a topology on the set of uncertain properties over a topological space X , such as Σ or Σ^* .

One drawback of working directly with probability measures in matters of optimization is that the relevant topologies are not well-behaved because they are naturally defined in terms of algebraic structures which are ill-behaved. To this end, an ambient space of measures is reviewed in Section 3.2 below.

3.2 Measure

Let Σ be a type of events and recall the intuition from Hypothesis 3.11 that an additive operation could be defined pointwise on the space $\Delta\Sigma^*$ of (probability) measures. This fails to be well-defined because, for example, the sum of two probability distributions yields a function whose total measure is 2 instead of 1.

To resolve this issue, consider the inner product space $\langle \mathbb{C}, 0, 1, \langle \bullet, \star \rangle, |\bullet| \rangle$ of complex numbers whose complex norm $|\bullet| = |\bullet|_{\mathbb{C}}$ is the complex modulus.

Definition 3.15 (complex measure, real measure, probability measure). Given a measurable space (X, \mathcal{F}) , define:

1. a *(complex) measure* $\mu : \mathcal{F} \rightarrow \mathbb{C}$ to be a function satisfying $\mu(\emptyset) = 0$ that is countably additive; and
2. a *real measure* to be a complex measure satisfying $\mu(E) \in \mathbb{R}$ for any $E \in \mathcal{F}$;
3. a *non-negative measure* to be a real measure satisfying $\mu(E) \geq 0$ for any $E \in \mathcal{F}$;
4. a *probability measure* μ to be a non-negative real measure satisfying $\mu(X) = 1$.

Consider the following claim about planning with unbounded horizons.

Slogan 3.16. The long term of the future must be discounted.

Consider the following example.

Example 3.17. The value of a move in chess is determined, ultimately, by its effect on the final outcome of the match and this can be done with certainty in the end game. Since the enormity of the state space precludes exhaustive search for ideal moves, it is not tractable to certainly determine the value of an arbitrary move. The value of such moves are instead approximated. The effect of an opening move on the end game is discounted from the calculation of its overall value, and the approximation of a move's value is concentrated on its effects in the near term.

For the same reason, a language model whose coherence is determined by a long stream of output makes its determinations on the basis of its confidence in the next token or small number of next tokens.

To capture this discount with a factor, consider the open unit disk $D_1 = B_1(0) = \{c \in \mathbb{C} \mid |c| < 1\}$. Let $\alpha \in D_1$ denote the discount factor and recall that the measurable space $\mathcal{F} = \mathcal{B}\Sigma^*$ admits the following structure.

Definition 3.18 (measure finitude, space of finite measures, addition, zero). Given any topological space X with Borel σ -algebra $\mathcal{F} = \mathcal{B}X$, define:

1. a *partition* ρ of a measurable set $E \in \mathcal{F}$ to be a countable collection of disjoint subsets of E that covers E and the partition function $\Pi : \mathcal{F} \rightarrow \mathfrak{P}\mathcal{F}$ by $\{\rho \in \mathfrak{P}\mathcal{F} \mid \rho \text{ is a partition of } E\}$;
2. the *variation* function $|\bullet| : \mathbb{C}^{\mathcal{F}} \rightarrow [0, +\infty]^{\mathcal{F}}$ as follows:

$$\forall \mu \in \mathbb{C}^{\mathcal{F}}, \quad |\mu| : \mathcal{F} \rightarrow [0, +\infty] \text{ defined by } E \mapsto \sup_{\rho \in \Pi(E)} \sum_{A \in \rho} |\mu(A)|$$

and the *total variation* function $\|\bullet\| : \mathbb{C}^{\mathcal{F}} \rightarrow [0, +\infty]$ by $\mu \mapsto \|\mu\|(X)$;

3. a complex measure μ to be *finite* if $\|\mu\| \in [0, +\infty]$ is finite;
4. the subset $\mathfrak{M}X \subset \mathbb{C}^{\mathcal{F}}$ of (complex) measures on X which are finite;
5. the pointwise *addition* function $+: \mathfrak{M}X \times \mathfrak{M}X \rightarrow \mathbb{C}^{\mathcal{F}}$ by the function which sends (μ, ν) to the measure $\mu + \nu$ defined by $E \mapsto \mu(E) + \nu(E)$; and
6. the *zero* measure $0 = \mathfrak{z}$ in $\mathfrak{M}X$ by $\bullet \mapsto 0$.

In contrast with the space of probability measures, the space of all finite (complex) measures is closed under pointwise addition.

Proposition 3.19. For any topological space X , the space $\langle \mathfrak{M}X, +, \mathfrak{z} \rangle$ of finite measures is a commutative monoid under pointwise addition $+$.

Proof. Let $\langle \mathfrak{M}X, +, \mathfrak{z} \rangle$ be as stated and let $\mu, \nu \in \mathfrak{M}X$.

Countable additivity of the sum $\mu + \nu$ is shown below.

$$\begin{aligned} (\mu + \nu) \left(\bigsqcup_{n=1}^{\infty} E_n \right) &= \mu \left(\bigsqcup_{n=1}^{\infty} E_n \right) + \nu \left(\bigsqcup_{n=1}^{\infty} E_n \right) \\ &= \sum_{n=1}^{\infty} \mu(E_n) + \sum_{n=1}^{\infty} \nu(E_n) = \sum_{n=1}^{\infty} (\mu + \nu)(E_n) \end{aligned}$$

For any $E \in \mathcal{F}$ and $\rho \in \Pi(E)$, the triangle inequality yields:

$$\sum_{A \in \rho} |(\mu + \nu)(A)| \leq \sum_{A \in \rho} (|\mu(A)| + |\nu(A)|) = \sum_{A \in \rho} |\mu(A)| + \sum_{A \in \rho} |\nu(A)|.$$

Finitude of $\mu + \nu$ is shown from $|\mu + \nu|(X) = \sup_{\rho \in \Pi(X)} \sum_{A \in \rho} |(\mu + \nu)(A)|$ below.

$$|\mu + \nu|(X) \leq \sup_{\rho \in \Pi(X)} \sum_{A \in \rho} |\mu(A)| + \sup_{\rho \in \Pi(X)} \sum_{A \in \rho} |\nu(A)| = |\mu|(X) + |\nu|(X) < \infty$$

Associativity, identity, and commutativity follow pointwise. \square

The additive inverse is defined below.

Definition 3.20 (negation, scalar multiplication, conjugation, support, ε -support). Given the Borel σ -algebra $\mathcal{F} = \mathcal{B}X$, define:

1. the *negation* function $- : \mathfrak{M}X \rightarrow \mathbb{C}^{\mathcal{F}}$ by $(-\mu)(E) = -\mu(E)$;
2. the *scalar multiplication* function $\cdot : \mathbb{C} \times \mathfrak{M}X \rightarrow \mathbb{C}^{\mathcal{F}}$ by $(c \cdot \mu)(E) = c \cdot \mu(E)$;
3. the *conjugation* function $\bar{\bullet} : \mathfrak{M}X \rightarrow \mathbb{C}^{\mathcal{F}}$ by $\bar{\mu}(E) = \overline{\mu(E)}$;
4. the function *support* $: \mathfrak{M}X \rightarrow \mathfrak{P}X$ by $\text{support } \mu = \{x \in X \mid \mu(\{x\}) \neq 0\}$;
5. the ε -*support* function $\text{support} : (0, +\infty] \times \mathfrak{M}X \rightarrow \mathfrak{P}X$ by

$$\forall \varepsilon > 0, \quad \forall \mu \in \mathfrak{M}X, \quad \text{support } \mu = \{x \in X \mid |\mu(\{x\})| > \varepsilon\}.$$

The above operations define a vector space.

Proposition 3.21. For any topological space X with Borel σ -algebra $\mathcal{B}X$, the following hold:

1. the commutative monoid $\langle \mathfrak{M}X, +, \mathfrak{z}, -, \cdot, 1 \rangle$ is closed under negation, scalar multiplication, conjugation; and
2. the above functions form a vector space.

Proof. Let X be any topological space and consider its Borel σ -algebra.

Recall that $\langle \mathfrak{M}X, +, \mathfrak{z}, -, \cdot, 1 \rangle$ is a commutative monoid.

1. Consider a measure $\mu \in \mathfrak{M}X$, measurable sets $\{E_n\}_{n \in \mathbb{N}} \subseteq \mathcal{B}X$ and their union $E = \bigcup_{n \in \mathbb{N}} E_n$.

Closure of $\mathfrak{M}X$ under negation follows from $(-\mu)(E) = -\mu(E) = -\sum_{n \in \mathbb{N}} \mu(E_n) = -\sum_{n \in \mathbb{N}} (-\mu)(E_n)$

and finitude of μ . Closure of $\mathfrak{M}X$ under scalar multiplication follows from the fact, for any constant $c \in \mathbb{C}$, that $(c \cdot \mu)(E) = c \cdot \sum_{n \in \mathbb{N}} \mu(E_n) = \sum_{n \in \mathbb{N}} (c \cdot \mu)(E_n)$ and μ is finite. Closure of $\mathfrak{M}X$

under conjugation follows from $\bar{\mu}(E) = \overline{\sum_{n \in \mathbb{N}} \mu(E_n)} = \sum_{n \in \mathbb{N}} \bar{\mu}(E_n)$.

2. It is trivial to check the multiplicative identity, associativity, and distributivity laws. □

The effect of trace lengths is captured by the following structure.

Definition 3.22 (weighted product, weight, weighted distance). For any vector space $\mathfrak{M}X$ with a weight $w_X : X \rightarrow (0, +\infty)$, define

1. the (measurable) weighted product $\langle \bullet, \star \rangle_{w_X} : \mathfrak{M}X \times \mathfrak{M}X \rightarrow \mathbb{C}$ by

$$\langle \mu, \nu \rangle_{w_X} = \sum_{x \in X} \mu(\{x\}) \overline{\nu(\{x\})} w_X(x)$$

for any $\mu, \nu \in \mathfrak{M}X$;

2. the (measurable) weight function $\| \bullet \|_{w_X} : \mathfrak{M}X \rightarrow [0, +\infty]$ by $\| \bullet \|_{w_X} = \sqrt{\langle \bullet, \bullet \rangle_{w_X}}$; and

3. the (measurable) weighted distance function $d_{w_X} : \mathfrak{M}X \times \mathfrak{M}X \rightarrow \mathbb{R}_{\geq 0}$ by $d(\mu, \nu) = \|\mu - \nu\|$.

While the formula for the weighted distance function may not seem to apply to the subspace $\Delta X \subseteq \mathfrak{M}X$ because the difference of two probability measures is a measure which is not necessarily unital or non-negative, the function is well-defined on this subspace. Investigations of *probabilistic* post-training motivate that a more general investigation from which the *probabilistic* results will follow as corollaries.

The probabilistic product is a restriction of the following inner product.

Proposition 3.23. The vector space $\langle \mathfrak{M}X, +, \mathfrak{z}, -, \cdot, 1, \langle \bullet, \star \rangle_{w_X}, \| \bullet \|_{w_X}, d \rangle$, is an inner product space with inner product $\langle \bullet, \star \rangle_{w_X}$, a normed vector space with norm $\| \bullet \|_{w_X}$, and a metric space with metric d_{w_X} .

Proof. It is trivial to show that $\langle \bullet, \star \rangle_{w_X}$ satisfies the required linearity, conjugate symmetry, and positive definiteness. Thus, $\mathfrak{M}X$ is an inner product space which induces the norm $\| \bullet \|_{w_X}$ and metric d_{w_X} . □

This inner product space can be seen as a subspace of $\mathbb{C}^{\Sigma^*} = \{f : X \rightarrow \mathbb{C}\}$. Define $\langle \bullet, \star \rangle : \mathbb{C}^X \times \mathbb{C}^X \rightarrow \mathbb{C}$ by $\langle f, g \rangle = \sum_{x \in X} f(x) \overline{g(x)} w_X(x)$, $\| \bullet \| : \mathbb{C}^X \rightarrow \mathbb{R}$ by $\| \bullet \| = \sqrt{\langle \bullet, \bullet \rangle}$, $i : \mathfrak{M}X \rightarrow \mathbb{C}^X$ by $\mu \xrightarrow{i} \mu \circ \iota$. Define the *weighted l^2 space* $l^2(X, w_X) = \langle f : X \rightarrow \mathbb{C} \mid \|f\| < \infty \rangle$. Define $m : l^2(X, w_X) \rightarrow \mathfrak{M}X$ by the assignment of any $f \in l^2(X, w_X)$ to the measure in $\mathfrak{M}X$ given by $E \xrightarrow{m(f)} \sum_{x \in X} f(x)$. This inner product space is a *Hilbert space*, an inner product space whose induced metric is complete.

Proposition 3.24. The weighted l^2 space $\langle l^2(X; w_X), \langle \bullet, \star \rangle, \| \bullet \| \rangle$ is a Hilbert space, and m is an isometric isomorphism to X with inverse e .

Proof. Let w_X be a weight on a topological space X . Then, the weighted l^2 space $l^2(X, w_X)$ is well-known to be a Hilbert space. It is clear from the constructions that e, m are linear isometries and that they are inverses of one another. Thus, they form an isometric isomorphism. \square

The space $\langle \mathfrak{M}X, w_X \rangle$ is a Hilbert space isomorphic to the weighted l^2 space $l^2(X, w_X)$, and completeness can also be shown directly.

For any $x \in X$, define $e_x : X \rightarrow \mathbb{C}$ in $l^2(X, w_X)$ by $e_x = \frac{1}{\sqrt{w_X(x)}} \cdot 1_x$ and $\delta_x : \mathfrak{B}X \rightarrow \mathbb{C}$ in $\mathfrak{M}X$ by $\delta_x = m(e_x)$ so that $E \xrightarrow{\delta_x} \frac{1}{\sqrt{w_X(x)}} \cdot \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{otherwise} \end{cases}$. The subset $B = \{\delta_x\}_{x \in X}$ of $\mathfrak{M}X$ is a countable basis of any countable space X , and it generates a countable dense subset $\left\{ \sum_{k=1}^N q_k \delta_{x_k} \mid N \in \mathbb{N}, x_k \in X, q_k \in \mathbb{Q}[i] \right\}$ of $\mathfrak{M}X$.
This gives rise to the following function.

Definition 3.25 (evaluator). For any topological space X , define the *evaluator* to be the function $\pi : X \times \mathfrak{M}X \rightarrow \mathbb{C}$ given by $\pi_p([x]) = [x](p)$.

The evaluator is shown to consist of linear transformations below.

Proposition 3.26 (evaluation). For topological space X with weight w_X and point $p \in X$, the *evaluator transformation* $\pi_p : \mathfrak{M}X \rightarrow \mathbb{C}$ at p is linear, bounded, continuous and surjective.

Proof. Let X, w_X, p, π_p be as stated. Linearity is shown below:

$$\begin{aligned} \forall \mu, \nu \in \mathfrak{M}X. \quad \pi_p(\mu + \nu) &= (\mu + \nu)(\{p\}) = \mu(\{p\}) + \nu(\{p\}) = \pi_p(\mu) + \pi_p(\nu); \\ \forall \mu \in \mathfrak{M}, \forall z \in \mathbb{C}. \quad \pi_p(z \cdot \mu) &= (z \cdot \mu)(\{p\}) = z \cdot (\mu(\{p\})) = z \cdot \pi_p(\mu). \end{aligned}$$

Let $K = \frac{1}{\sqrt{w_X(p)}}$ and observe that $|\mu(\{p\})|^2 \cdot w_X(p) \frac{1}{w_X(p)} \leq 1 \cdot \frac{1}{w_X(p)}$ yields boundedness as follows:

$$\begin{aligned} 0 < |\pi_p(\mu)| &\leq \sqrt{\frac{1}{w_X(p)} \sum_{r \in X} |\mu(\{r\})|^2 w_X(r)} && \text{by non-negativity of each term} \\ &\leq K \|\mu\|_{w_X} && \text{by multiplicativity of } \sqrt{\bullet}. \end{aligned}$$

Continuity of the linear transformation follows from boundedness.

Surjectivity follows from the function $C : \mathbb{C} \rightarrow \mathfrak{M}\Sigma^*$ which maps c to $C_c(\mu) = c$. \square

Recall that the Hilbert space $\mathfrak{M}X$ is equipped with the metric topology given by the metric $d_{w_X} : \mathfrak{M}X \times \mathfrak{M}X \rightarrow [0, +\infty)$.

Definition 3.27 (probabilistic distance). For any topological space X with weight w_X and metric $d_{w_X} : \mathfrak{M}X \times \mathfrak{M}X \rightarrow [0, +\infty)$ derived from the inner product of $l^2(X, w_X)$, define the *probabilistic distance* function $d_\Delta : \Delta X \times \Delta X \rightarrow [0, +\infty)$ by the restriction $d_\Delta = d_{w_X} \upharpoonright_{\Delta X \times \Delta X}$ to $\Delta X \times \Delta X$. The distance function d_Δ is also denoted by d_{w_X} .

Note that the *probabilistic distance* function defined above is non-negative.

Lemma 1 (probabilistic distance metric). *Given a weight $w_X : X \rightarrow \mathbb{R}_{\geq 0}$ on a countable space X , the probabilistic distance function $d_\Delta : \Delta X \times \Delta X$ given by the restriction of d_{w_X} is a metric on ΔX , and the metric topology on $\langle \Delta X, d_\Delta \rangle$ is equal to the subspace topology on ΔX inherited by the metric topology on $\langle \mathfrak{M}X, d_{w_X} \rangle$.*

Proof. Let X be as stated.

It is well-known that the subspace of a metric topology is a metric topology with the induced metric. \square

The view of ΔX as a subspace of $\mathfrak{M}X$ extends the notion of “statistical” distance from probability distributions to the space $\mathfrak{M}X$ of complex measures.

Many implementations of statistical distance satisfy the following condition.

Definition 3.28 (divergence atomicity). A divergence $D : \mathfrak{M}X \times \mathfrak{M}X \rightarrow [0, +\infty]$ is *atomic* iff there are functions $\phi : X \times \mathfrak{M}X \times \mathbb{C} \rightarrow [0, +\infty]$ and $d : X \times \mathfrak{M}X \times \mathfrak{M}X \rightarrow [0, +\infty]$ satisfying $d_p(\mu, \nu) = \phi_\mu(\nu(x))$ and $D(\mu \parallel \nu) = \sum_{x \in X} d_x(\mu, \nu)$.

3.3 Measurable property

Consider a chatbot, modeled as cumulative probabilistic channel $g : \mathcal{X}^* \rightarrow \Delta \mathcal{Y}^*$ which transforms an input sequence of tokens in \mathcal{X}^* to an appropriate response. Specifically, the computation results in a probability distribution over output sequences which is then sampled auto-regressively.

Let X be a topological space, such as the upper topology $X = \mathcal{Y}^*$, and recall the presentation in Definition 3.3 of a probabilistic *uncertain property*.

Definition 3.29 (measurable property, measurable trace property). A *measurable property* over a topological space X is a finite complex measure in $\mathfrak{M}X$, and a *measurable trace property* is a measurable property over a trace space Σ^* .

Any probabilistic uncertain property over X can be seen as a measurable property over X , so the response of a chatbot to any input is a measurable property. There is also a measurable property that models the intended behavior of the chatbot on a given input; for example, it is desired that a response is observed to be helpful, honest, and harmless.

The geometry of resource security [Pavlovic and Seidel, 2025] is echoed by the information geometry of uncertain properties and the **geometry** of measurable properties. Although the set $\mathfrak{M}X$ of measurable properties over X forms the *algebraic* structure of a vector, the word **geometry** is used because any weight $w_X : X \rightarrow \mathbb{R}_{\geq 0}$ on X gives rise to a notion of length $\|\bullet\|_{w_X}$, angle $\langle \bullet, \star \rangle_{w_X}$, and metric distance $d_{w_X}(\bullet, \star)$. Some basic components for geometric theories of uncertain properties and measurable properties were sketched in this chapter. While the development of these theories could be approached from a purely geometric perspective or by analogy with the mathematical theory of resource security, they seem to be unexplored in the academic literature on privacy & security. Instead of exploring them further at this time, the remainder of this thesis motivates their investigation by demonstrating an application of the theory presented thus far to the security science of machine learning and language processing [Pavlovic, 2024].

After instruction fine-tuning a chatbot, the task of post-training a chatbot is to align its behavior with an intended behavior that is modeled only implicitly through datasets, e.g. of human-annotated preferences [Christiano et al., 2017]. An instance of this task is comprised of the items in Table 3.1.

Theory	Model
countable type \mathcal{X} of events	vocabulary of input tokens
countable type \mathcal{Y} of events	vocabulary of output tokens
$\Sigma = \mathcal{X} \sqcup \mathcal{Y}$	vocabulary of tokens
\mathcal{X}^*	<i>input sequences</i> of input tokens
\mathcal{Y}^*	<i>output sequences</i> of output tokens
Σ^*	transcript sequences
$\Delta\mathcal{X}^*, \Delta\mathcal{Y}^*$	spaces of probability distributions over input and output sequences
$\Delta(\mathcal{X}, \mathcal{Y})$	set of functions $\mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$ containing model parameterizations
$g \in \Delta(\mathcal{X}, \mathcal{Y})$	“pre-trained” model $g(\mathbf{y} \mid \mathbf{x}) = [\mathbf{x} \vdash_g \mathbf{y}]$
$[\mathbf{x}]_0 \in \Delta\mathcal{X}^*$	probability of sampling input, given implicitly by (test) dataset
$H \subseteq \Delta(\mathcal{X}, \mathcal{Y})$	search space of model parameterizations during training
$C : \Sigma^* \rightarrow [0, +\infty)$	cost of transcript to minimize
$R : \Sigma^* \rightarrow [0, +\infty)$	reward of transcript to maximize
$r : \mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$	implicit behavioral modification of g that maximizes reward R
$D : \Delta\mathcal{Y}^* \times \Delta\mathcal{Y}^* \rightarrow [0, +\infty)$	divergence for measuring difference in behavioral responses
$\lambda \in \mathbb{R}$	trade-off parameter for regularization
$L : H \rightarrow (-\infty, +\infty]$	loss objective to minimize
$\min_{h \in H} L(h)$	optimization problem to solve
$\arg \min_{h \in H} L(h)$	set of optimal behaviors, targets of post-training

Table 3.1: Model of channel post-training

Problem 3.30. Given the items in the left column of Table 3.1 such that

1. the types \mathcal{X}, \mathcal{Y} have weights $w_{\mathcal{X}^*} : \mathcal{X}^* \rightarrow (0, +\infty)$ and $w_{\mathcal{Y}^*} : \mathcal{Y}^* \rightarrow (0, +\infty)$;

2. the subspaces $\langle \Delta \mathcal{X}^*, w_{\mathcal{X}^*} \rangle, \langle \Delta \mathcal{Y}^*, w_{\mathcal{Y}^*} \rangle$ are equipped with the metric topologies;
3. the loss function $L(h; D, g, [\mathbf{x}]_0, \lambda, r)$ is defined below

$$L(h; D, g, [\mathbf{x}]_0, \lambda, r) = D(\bar{g}([\mathbf{x}]_0), \bar{h}([\mathbf{x}]_0)) + \lambda \cdot D(\bar{r}([\mathbf{x}]_0), \bar{h}([\mathbf{x}]_0)) \quad (3.2)$$

for some $\lambda \in \mathbb{R}$ such that $\lambda \geq 0$:

the *probabilistic channel post-training problem* is the following optimization problem.

$$\min_{h \in H} L(h; D, g, [\mathbf{x}]_0, r) \quad (3.3)$$

The intuition arising from the existing practice of post-training is stated by the following hypothesis.

Hypothesis 3.31. Given any instance of Problem 3.30, there is a non-empty set of solutions.

That is, the set of ideal behaviors exists, to say nothing of uniqueness or convergence.

The tasks ahead are to define how a reward-normalized channel $r_{g,C} : \mathcal{X}^* \rightarrow \Delta \mathcal{Y}^*$ arises implicitly from a pre-trained model g and cost C and to determine a notion of convergence under consideration by constructing an appropriate topology on the space $\Delta(\mathcal{X}, \mathcal{Y})$ of functions inhabited by the pre-trained model's parameterizations. The difficulty of post-training is that r will evidently exist, but it is intractable to specify r or a direct transformation from g in the direction of r . In practice, this is done by indirect, iterative approximations of D and its gradients.

Recall that geometry of $\Delta \mathcal{X}^*$ was studied as a subspace of $\mathfrak{M} \mathcal{X}^*$ in this section, despite the fact that $\langle \Delta \mathcal{X}^*, d_{w_{\mathcal{X}^*}} \rangle$ is a metric space, since the pointwise operations on $\Delta \mathcal{X}^*$ do not form a group. This was because the formula for $d_{w_{\mathcal{X}^*}, w_{\mathcal{X}^*}} : \Delta \mathcal{X}^* \times \Delta \mathcal{X}^* \rightarrow \mathbb{R}_{\geq 0}$ was defined in terms of a group $\mathfrak{M} \mathcal{X}^*$ that contains $\Delta \mathcal{X}^*$. Similarly, the geometry of $\Delta(\mathcal{X}, \mathcal{Y})$ is investigated as a subspace of a well-behaved ambient space defined in Chapter 4.

CHAPTER 4

CHANNEL SECURITY

Recall the presentation in Problem 3.30 of probabilistic channel post-training. Solubility of this problem will be established in Section 4.3, where it will be shown to follow as a special case of another problem in Section 4.2 involving the proposed generalization of probabilistic channels in Definition 4.4.

Consider the view of a chatbot as a probabilistic channel $g : \mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$ from a countable event space \mathcal{X} to a countable event space \mathcal{Y} . The first thing to note is that the probability distributions in image $g \subseteq \Delta\mathcal{Y}^*$ tend to concentrate on traces whose lengths are bounded. One of the driving causes is that g is typically learned from datasets consisting of finitely many samples and that such datasets implicitly impose an upper bound on the lengths of prompts & responses. This upper bound is modeled with a function whose decay depends on the weights $w_{\mathcal{X}} : \mathcal{X} \rightarrow (0, +\infty)$ and $w_{\mathcal{Y}} : \mathcal{Y} \rightarrow (0, +\infty)$.

Recall that any finite event space Σ can be equipped with a constant weight function $w_{\Sigma} : \Sigma \rightarrow (0, +\infty)$ given by $\bullet \mapsto \frac{w_{\Sigma}}{|\Sigma|}$.

Consider a function $w_{\Sigma^*} : \Sigma^* \rightarrow (0, +\infty)$ that satisfies the following compatibility condition.

Definition 4.1 (trace weight). Given a countable event space Σ with a weight $w_{\Sigma} : \Sigma \rightarrow (0, +\infty)$, define a *trace weight compatible with w_{Σ}* to be a function $w_{\Sigma^*} : \Sigma^* \rightarrow (0, +\infty)$ for which there are constants $C_0, C_1 \in \mathbb{R}$ such that any trace $\mathbf{t} = r_0 r_1 r_2 \cdots r_{k-1} r_k r_{k+1} \cdots r_{n-1} r_n$ in Σ^* satisfies the inequalities $C_0 \prod_{k=0}^n w_{\Sigma}(r_k) \leq w_{\Sigma^*}(\mathbf{t}) \leq C_1 \prod_{k=0}^n w_{\Sigma}(r_k)$.

An example in terms of the unit disk D_1 is given below.

Example 4.2. If $|\Sigma| < \infty$ and $w_{\Sigma} : \Sigma \rightarrow (0, +\infty)$ is given by $w_{\Sigma}(\bullet) = |\alpha_{\Sigma}|$ for some constant $\alpha_{\Sigma} \in D_1$, then the exponential weight $w_{\Sigma^*} : \Sigma^* \rightarrow (0, +\infty)$ given by $w_{\Sigma^*}(\mathbf{t}) = |\alpha_{\Sigma}|^{|\mathbf{t}|}$ satisfies $C_1 \prod_{k=0}^{|\mathbf{t}|-1} w_{\Sigma}(r_k) \leq w_{\Sigma^*}(\mathbf{t}) \leq C_2 \prod_{k=0}^{|\mathbf{t}|-1} w_{\Sigma}(r_k)$ for constants $C_1 = C_2 = 1$.

That is, a trace weight on an upper topology with a weight on events must be compatible with the multiplicative structure, a requirement that is only needed asymptotically. Compatibility for all trace lengths, e.g. below a cut-off length, is unnecessary and assumed only for the sake of convenience.

In the remainder of this article, the topologies on $\mathfrak{M}\mathcal{X}^*, \mathfrak{M}\mathcal{Y}, \mathfrak{M}\mathcal{Y}^*$ are assumed to be the Hilbert spaces defined in Section 3 and the topologies $\Delta\mathcal{X}^*, \Delta\mathcal{Y}, \Delta\mathcal{Y}^*$ are assumed to be the subspace topologies inherited from the Hilbert spaces.

Recall that a chatbot channel g can be viewed as the accumulation of a stochastic matrix $f_{\Delta} : \mathcal{X}^+ \rightarrow \Delta\mathcal{Y}$ as presented in Definition 2.34. The view that its responses are probability

distributions in $\Delta\mathcal{Y}$ over the space \mathcal{Y} of tokens is consistent with the model of a neural network whose final layer is a *softmax* layer or any other method for converting logits to probabilities. Consider, instead, the function that produces logits. It is like a probabilistic channel, except that the probability measures do not sum to 1. Reasoning about the behavior of this object motivates the following generalization of *stochastic* matrices.

Definition 4.3 (measurable matrix, boundedness, space of measurable matrices). Given countable event types $\langle\mathcal{X}, w_{\mathcal{X}}\rangle, \langle\mathcal{Y}, w_{\mathcal{Y}}\rangle$ and function $f : \mathcal{X}^+ \rightarrow \mathfrak{M}\mathcal{Y}$, *square-summability* is the condition that $\sum_{\mathbf{x} \in \mathcal{X}^+} \|f(\mathbf{x})\|_{w_{\mathcal{Y}}}^2 \cdot w_{\mathcal{X}^*}(\mathbf{x}) < \infty$, and *root-boundedness* is the condition that the sequence $\{S_n\}_{n \in \mathbb{N}}$ given for each $n \in \mathbb{N}$ by $S_n = \sum_{\mathbf{x} \in \mathcal{X}^n} w_{\mathcal{X}^*}(x_0 \cdots x_{n-1}) \prod_{k=0}^{n-1} \|f(x_0 \cdots x_k)\|_{w_{\mathcal{Y}}}^2$ satisfies $\limsup_{n \rightarrow \infty} S_n^{1/n} < 1$. A *measurable matrix* from \mathcal{X} to \mathcal{Y} is a square-summable function $f : \mathcal{X}^+ \rightarrow \mathfrak{M}\mathcal{Y}$ that is root-bounded, and a *boundedness* is the condition that a measurable matrix $f : \mathcal{X}^+ \rightarrow \mathfrak{M}\mathcal{Y}$ satisfies $\sup_{\mathbf{x} \in \mathcal{X}^+} \|f(\mathbf{x})\|_{w_{\mathcal{Y}}}^2 \cdot w_{\mathcal{X}^*}(\mathbf{x}) < \infty$. Define the space

$$\mathcal{F}(\mathcal{X}, \mathcal{Y}) = \{f : \mathcal{X}^+ \rightarrow \mathfrak{M}\mathcal{Y} \mid f \text{ is square-summable and root-bounded}\}$$

of all measurable matrices from \mathcal{X} to \mathcal{Y} . A real measurable matrix from \mathcal{X} to \mathcal{Y} is a measurable matrix $f : \mathcal{X}^+ \rightarrow \mathfrak{M}\mathcal{Y}$ in $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ for which any $\mathbf{x} \in \mathcal{X}^+$ yields a complex measure $f(\mathbf{x}) \in \mathfrak{M}\mathcal{Y}$ that is real-valued.

Then, a measurable matrix $f : \mathcal{X}^+ \rightarrow \mathfrak{M}\mathcal{Y}$ from \mathcal{X} to \mathcal{Y} can be said to be *stochastic* if it is the case for any $\mathbf{x} \in \mathcal{X}^+$ that $f(\mathbf{x}) \in \mathfrak{M}\mathcal{Y}$ is in the subspace $\Delta\mathcal{Y} \subseteq \mathfrak{M}\mathcal{Y}$. While a base model of language generation can be viewed as a stochastic matrix, auto-regression gives rise to a probabilistic channel as presented in Definition 2.36.

Recall for any point $y \in Y$ that the singleton $\{y\} \in \mathfrak{P}Y$ is a measurable set in the Borel σ -algebra $\mathcal{B}Y$, and that this allows the notation overload for any $\mu : \mathcal{B}Y \rightarrow \mathbb{C}$ in $\mathfrak{M}Y$ that identifies $\mu(y) = \mu(\{y\})$. Reasoning about the accumulation of networks that produce logits (instead of probabilities) motivates the following generalization of probabilistic channels.

Definition 4.4 (measurable channel, square-summability, boundedness). A *measurable channel* from $\langle\mathcal{X}, w_{\mathcal{X}^*}\rangle$ to $\langle\mathcal{Y}, w_{\mathcal{Y}^*}\rangle$ is a function $g : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ denoted $[\mathbf{x} \vdash_g \mathbf{y}] = g(\mathbf{x})(\{\mathbf{y}\})$ that satisfies the condition of *square-summability* $\sum_{\mathbf{x} \in \mathcal{X}^*} \|g(\mathbf{x})\|_{w_{\mathcal{Y}^*}}^2 \cdot w_{\mathcal{X}^*}(\mathbf{x}) < \infty$, and it is bounded if $\sup_{\mathbf{x} \in \mathcal{X}^*} \|g(\mathbf{x})\|_{w_{\mathcal{Y}^*}}^2 \cdot w_{\mathcal{X}^*}(\mathbf{x}) < \infty$. Let $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ denote the space of all measurable channels from \mathcal{X} to \mathcal{Y} .

Intuitively, a measurable channel is a generalization of a probabilistic channel in which the probability distributions are generalized by finite, complex measures.

Just as stochastic matrices give rise to probabilistic channels, measurable matrices give rise to measurable channels.

Definition 4.5 (accumulation, cumulative measurable channel). Define the *accumulation* function $\bullet^* : \mathcal{F}(\mathcal{X}, \mathcal{Y}) \rightarrow (\mathfrak{M}\mathcal{Y}^*)^{\mathcal{X}^*}$ as the assignment from any measurable matrix $f : \mathcal{X}^+ \rightarrow \mathfrak{M}\mathcal{Y}$ in $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ to the accumulation $f^* : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ given by $[x_0 \cdots x_n \vdash_{f^*} y_0 \cdots y_n] = \prod_{k=0}^n [x_0 \cdots x_k \vdash_f y_k]$, and define a *cumulative* measurable channel to be a measurable channel g that is the accumulation of a measurable matrix.

Observe that the accumulation of a measurable matrix is a measurable channel.

Proposition 4.6 (accumulation). For any measurable matrix $f : \mathcal{X}^+ \rightarrow \mathfrak{M}\mathcal{Y}$ in $\mathcal{F}\mathcal{X}, \mathcal{Y}$ from $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$ to $\langle \mathcal{Y}, w_{\mathcal{Y}} \rangle$ and weight $w_{\mathcal{Y}^*} : \mathcal{Y}^* \rightarrow (0, +\infty)$ on \mathcal{Y}^* , the accumulation $f^* : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ is a cumulative measurable channel from \mathcal{X} to \mathcal{Y} .

Proof. Let $f : \mathcal{X}^+ \rightarrow \mathfrak{M}\mathcal{Y}$ be a measurable matrix from $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$ to $\langle \mathcal{Y}, w_{\mathcal{Y}} \rangle$, and define $S_n = \sum_{\mathbf{x} \in \mathcal{X}^n} w_{\mathcal{X}^*}(x_0 \cdots x_{n-1}) \prod_{k=0}^{n-1} \|f(x_0 \cdots x_k)\|_{w_{\mathcal{Y}}}^2$ for each $n \in \mathbb{N}$.

To show that f^* satisfies square-summability, observe that the weight $w_{\mathcal{Y}^*}$ has a constant C_2 satisfying $\|f^*(\mathbf{x})\|_{w_{\mathcal{Y}^*}}^2 \leq C_2 \prod_{k=0}^{n-1} \|f(x_0 \cdots x_k)\|_{w_{\mathcal{Y}}}^2$ for all $\mathbf{x} \in \mathcal{X}^n$. Observe the following.

$$M = \sum_{\mathbf{x} \in \mathcal{X}^*} \|f^*(\mathbf{x})\|_{w_{\mathcal{Y}^*}}^2 \cdot w_{\mathcal{X}^*}(\mathbf{x}) = \sum_{n=0}^{\infty} \sum_{\mathbf{x} \in \mathcal{X}^n} \|f^*(\mathbf{x})\|_{w_{\mathcal{Y}^*}}^2 \cdot w_{\mathcal{X}^*}(\mathbf{x})$$

This yields $M \leq C_2 \sum_{n=0}^{\infty} S_n \leq \infty$. □

Consider the space $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ of measurable channels and note that it admits the following structure.

Definition 4.7 (addition, zero). Given the space $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ of measurable channels from $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$ to $\langle \mathcal{Y}, w_{\mathcal{Y}^*} \rangle$, define:

1. *addition* $+$: $\mathfrak{M}(\mathcal{X}, \mathcal{Y}) \times \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathfrak{M}\mathcal{Y}^{\mathcal{X}^*}$ by $(g + h)(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$;
2. *zero* z : $\mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ by $z(\mathbf{x}) = \mathfrak{z}_{\mathcal{Y}^*}$;

The space $\langle \mathfrak{M}(\mathcal{X}, \mathcal{Y}), +, z \rangle$ of measurable channels is closed under addition.

Proposition 4.8. For any event types $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$ and $\langle \mathcal{Y}, w_{\mathcal{Y}^*} \rangle$, the space $\langle \mathfrak{M}(\mathcal{X}, \mathcal{Y}), +, z \rangle$ is a monoid under addition $+$ with identity z .

Proof. Let $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ be as stated above, and recall that $\mathfrak{M}\mathcal{Y}^*$ is a monoid under addition.

To show $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ is closed under addition, consider any $g, h \in \mathfrak{M}(\mathcal{X}, \mathcal{Y})$ and recall that the sum $(g + h) : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ is given by $\mathbf{x} \xrightarrow{g+h} g(\mathbf{x}) + h(\mathbf{x})$. Closure of $\mathfrak{M}\mathcal{Y}^*$ under addition guarantees

that $g(\mathbf{x}) + h(\mathbf{x})$ is a measure in $\mathfrak{M}\mathcal{Y}^*$ for any $\mathbf{x} \in \mathcal{X}^*$. Square-summability of $(g + h)$ is shown below.

$$\sum_{\mathbf{x} \in \mathcal{X}^*} \|g(\mathbf{x}) + h(\mathbf{x})\|_{w_{\mathcal{Y}^*}}^2 \cdot w_{\mathcal{X}^*}(\mathbf{x}) \leq \sum_{\mathbf{x} \in \mathcal{X}^*} \left(\|g(\mathbf{x})\|_{w_{\mathcal{Y}^*}}^2 + \|h(\mathbf{x})\|_{w_{\mathcal{Y}^*}}^2 \right) \cdot w_{\mathcal{X}^*}(\mathbf{x})$$

Associativity and identity laws follow pointwise. Therefore, $(g + h) \in \mathfrak{M}(\mathcal{X}, \mathcal{Y})$. \square

Let $\alpha_{\mathcal{X}^*} \in D_1$, and observe it admits the following structure.

Definition 4.9 (negation, scalar multiplication, exponential weight). Given a space $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ of measurable channels from $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$ to $\langle \mathcal{Y}, w_{\mathcal{Y}^*} \rangle$, define:

1. the *negation* function $- : \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathfrak{M}(\mathcal{X}, \mathcal{Y})$ by $(-g)(\mathbf{x}) = -g(\mathbf{x})$;
2. the *scalar multiplication* function $\cdot : \mathbb{C} \times \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathfrak{M}(\mathcal{X}, \mathcal{Y})$ by $(c \cdot g)(\mathbf{x}) = c \cdot g(\mathbf{x})$;
3. the *exponential weight* function $w_{\mathcal{X}^*} : \mathcal{X}^* \rightarrow \mathbb{R}$ by $w_{\mathcal{X}^*}(\mathbf{x}) = \left| \alpha_{\mathcal{X}^*}^{|\mathbf{x}|} \right|$ for any $\alpha_{\mathcal{X}^*} \in D_1$.

Consider $\langle \mathfrak{M}(\mathcal{X}, \mathcal{Y}), +, z, -, \cdot, w_{\mathcal{X}^*} \rangle$ as defined above. It admits the following structure.

Proposition 4.10. For any event types $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$ and $\langle \mathcal{Y}, w_{\mathcal{Y}^*} \rangle$, the space $\langle \mathfrak{M}(\mathcal{X}, \mathcal{Y}), +, z, -, \cdot \rangle$ of measurable channels is a vector space.

Proof. Let $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ be as stated above. Recall from Proposition 4.8 that $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ is an additive monoid with identity z . All of the remaining axioms of a vector space follow pointwise from the fact that $\mathfrak{M}\mathcal{Y}^*$ is a vector space. \square

The restrictions of these operations to $\Delta(\mathcal{X}, \mathcal{Y})$ do not form a vector space because the failure of $\Delta\mathcal{Y}^*$ to be a vector space causes the pointwise operations to not be closed. This drives the primary consideration of $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$, and the analogous results for $\Delta(\mathcal{X}, \mathcal{Y})$ will follow as corollaries.

Recall the presentation in Definition 3.14 of the probabilistic channel product, and consider the following generalization to the space of measurable channels.

Definition 4.11 (weighted product, weight, weighted distance). Given any vector space $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ of measurable channels from $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$ to $\langle \mathcal{Y}, w_{\mathcal{Y}^*} \rangle$, define:

1. the *(measurable) weighted product* $\langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} : \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \times \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{C}$ below

$$\langle g, h \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} = \sum_{\mathbf{x} \in \mathcal{X}^*} \langle g(\mathbf{x}), h(\mathbf{x}) \rangle_{w_{\mathcal{Y}^*}} w_{\mathcal{X}^*}(\mathbf{x});$$

2. the *(measurable) weight* function $\|\bullet\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} : \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{R}$ by $\|g\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} = \sqrt{\langle g, g \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}}$;
and

3. the (measurable) weighted distance function $d_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} : \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \times \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{R}_{\geq 0}$ below

$$d_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}(g, h) = \|g - h\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}.$$

4.1 Measurable channel

Equip $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ with the (measurable) weighted product $\langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}$, and note that it forms the weighted vector-valued l^2 vector space $l^2(\mathcal{X}^*, \mathfrak{M}\mathcal{Y}^*; w_{\mathcal{X}^*}, w_{\mathcal{Y}^*})$ which is known to be an inner product space.

Proposition 4.12. The inner product space $\langle \mathfrak{M}(\mathcal{X}, \mathcal{Y}), +, z, -, \cdot, w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}, \langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} \rangle$ is an inner product space with inner product $\langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}$, a normed vector space with norm $\|\bullet\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}$, a metric space with metric $d(\bullet, \star)_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}$, and a Hilbert space with respect to $\langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}$.

Proof. Since $\mathfrak{M}\mathcal{Y}^*$ is a Hilbert space, it follows that the inner product space in question is a Hilbert space. The remaining results are derived from this fact. \square

Recall the model of a chatbot as a channel $g : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ in $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$, and consider a training process in which model parametrizations become arbitrarily close. Then, Proposition 4.12 guarantees that there is a model parametrization which is the limit of this process. This is the foundation needed for results about existence and convergence. A, say, iterative algorithm can only converge to a solution if the sequence of iterations has a limit that exists in the space of representable solutions; in this case, gradient descent produces a sequence $\{h_n\}_{n \in \mathbb{N}}$ of functions in $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ and the proposition guarantees that such a sequence converges if elements become arbitrarily close. This is not true, for example, in many spaces of continuous functions on \mathbb{R} in which a sequence of continuous functions can become arbitrarily close but fail to converge to a continuous function. The fact that Proposition 4.12 holds suggests that spaces of measures are better behaved than \mathbb{R} or that the measurable channels in $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ are well behaved.

Recalling the view of a generative model of language as a probabilistic channel $g : \mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$, it should be noted that the language model is applied to a sequence $\mathbf{x} \in \mathcal{X}^*$ drawn from a probability distribution $[\mathbf{x}'] \in \Delta\mathcal{X}^*$ that models the observed frequency of natural language in the real world. If there is such a distribution, then it is desirable to measure the value of g in a way that is weighed by $[\mathbf{x}]$. This motivates the presentation of a continuous probabilistic channel $g : \Delta\mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$ in Definition 2.36 and the following generalization.

Definition 4.13 (space of linear measurable channels, continuous measurable channel). For any event types \mathcal{X} and \mathcal{Y} , define:

1. the space $\mathbf{Vect}(\mathcal{X}, \mathcal{Y}) = \{\gamma : \mathfrak{M}\mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^* \mid \gamma \text{ is linear}\}$ of linear transformations from $\mathfrak{M}\mathcal{X}^*$ to $\mathfrak{M}\mathcal{Y}^*$;

2. a *continuous measurable channel* from \mathcal{X} to \mathcal{Y} to be a linear transformation $\gamma : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathfrak{M}\mathcal{Y}^*$ in $\mathbf{Vect}(\mathcal{X}, \mathcal{Y})$ that is bounded.

Observe that the word *continuous* is used in the above definition because any bounded, linear transformation is continuous. Note that not every bounded, measurable channel of the form $g : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ is continuous. This motivates the related study of continuous measurable channels.

Definition 4.14 (space of continuous measurable channels, addition, zero, negation). Given the space $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ of measurable channels from \mathcal{X} to \mathcal{Y} , define:

1. the space

$$B(\mathcal{X}, \mathcal{Y}) = \left\{ \gamma : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathcal{Y}^* \mid \gamma \text{ is bounded} \right\}$$

of continuous measurable channels;

2. *addition* $+$: $\mathbf{Vect}(\mathcal{X}, \mathcal{Y}) \times \mathbf{Vect}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathbf{Vect}(\mathcal{X}, \mathcal{Y})$ by $(\gamma + \eta)([\mathbf{x}]) = \gamma([\mathbf{x}]) + \eta([\mathbf{x}])$;
3. *zero* $\zeta \in B(\mathcal{X}, \mathcal{Y})$ by $\zeta([\mathbf{x}]) = \mathfrak{z}$; and
4. *negation* $-\bullet$: $\mathbf{Vect}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathbf{Vect}(\mathcal{X}, \mathcal{Y})$ by $(-\gamma)([\mathbf{x}]) = -\gamma([\mathbf{x}])$.

Define the analogous operations on $B(\mathcal{X}, \mathcal{Y})$ by restriction.

Consider the space $\langle B(\mathcal{X}, \mathcal{Y}), +, \zeta, - \rangle$ of continuous measurable channels from \mathcal{X} to \mathcal{Y} . Then, the above notion of pointwise addition is well-defined.

Proposition 4.15. For any event types \mathcal{X}, \mathcal{Y} , the space $\langle B(\mathcal{X}, \mathcal{Y}), +, \zeta, - \rangle$ is a commutative group under addition with identity ζ .

Proof. Consider event types \mathcal{X}, \mathcal{Y} .

Recall, for linear transformations, that the condition of boundedness is equivalent to continuity. Closure of $B(\mathcal{X}, \mathcal{Y})$ under addition follows from the fact that the sum of continuous functions is continuous. Associativity, identity, and commutativity follow pointwise. \square

It admits scalar multiplication over the field \mathbb{C} of complex numbers.

Definition 4.16 (scalar multiplication, weighted product, weight, weighted distance). Given the commutative group $\langle B(\mathcal{X}, \mathcal{Y}), +, \zeta, - \rangle$ of continuous measurable channels from a type $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$ of events to another type $\langle \mathcal{Y}, w_{\mathcal{Y}^*} \rangle$ of events, define:

1. the *scalar multiplication* function $\cdot : \mathbb{C} \times B(\mathcal{X}, \mathcal{Y}) \rightarrow B(\mathcal{X}, \mathcal{Y})$ over \mathbb{C} by $(z \cdot \gamma)([\mathbf{x}]) = z \cdot \gamma([\mathbf{x}])$;
2. the *weighted product* $\langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} : B(\mathcal{X}, \mathcal{Y}) \times B(\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{C}$ below

$$\langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}(\gamma, \eta) = \sum_{[\mathbf{x}] \in \mathfrak{M}\mathcal{X}^*} \langle \gamma([\mathbf{x}], \eta([\mathbf{x}]) \rangle_{w_{\mathcal{Y}^*}} w_{\mathcal{X}^*}([\mathbf{x}]);$$

3. the *weight* function $\|\bullet\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} : B(\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{R}$ by $\|\gamma\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} = \sqrt{\langle \gamma, \gamma \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}}$; and
4. the *weighted distance* function $d_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} : B(\mathcal{X}, \mathcal{Y}) \times B(\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{R}_{\geq 0}$ below.

$$d_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}(\gamma, \eta) = \|\eta - \gamma\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}$$

The commutative group $\langle B(\mathcal{X}, \mathcal{Y}), +, \zeta, -, \cdot, 1, \langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}, \|\bullet\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}, d_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} \rangle$ is an inner product space with inner product $\langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}$, norm $\|\bullet\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}$, and metric $d_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}$.

If terms in a sequence become increasingly close under the inner product, then the sequence converges pointwise.

Proposition 4.17. Any Cauchy sequence in the inner product space

$$\langle B(\mathcal{X}, \mathcal{Y}), +, \zeta, -, \cdot, 1, \langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}, \|\bullet\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}, d_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} \rangle$$

converges in the product topology on $(\mathfrak{M}\mathcal{Y}^*)^{\mathfrak{M}\mathcal{X}^*}$ to a linear transformation in $\mathbf{Vect}(\mathcal{X}, \mathcal{Y})$.

Proof. Let $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$, $\langle \mathcal{Y}, w_{\mathcal{Y}^*} \rangle$, and $B(\mathcal{X}, \mathcal{Y})$ be as stated.

Consider a Cauchy sequence $\{\gamma_n\}_{n \in \mathbb{N}}$ in the inner product space $B(\mathcal{X}, \mathcal{Y})$.

Consider a source $[\mathbf{x}] \in \mathfrak{M}\mathcal{X}^*$, a threshold $\varepsilon > 0$, and recall that $\mathfrak{M}\mathcal{Y}^*$ is a Hilbert space. By the Cauchy condition, there is a threshold $N \in \mathbb{N}$ such that any pair of indices $n, m \in \mathbb{N}$ satisfy $\|\gamma_n - \gamma_m\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}^2 = \sum_{\mu \in \mathfrak{M}\mathcal{X}^*} \|\gamma_n(\mu) - \gamma_m(\mu)\|_{w_{\mathcal{Y}^*}}^2 w_{\mathcal{X}^*}(\mu) < \varepsilon^2$. Any pair of indices $n, m \geq N$ satisfy $\|\gamma_n([\mathbf{x}]) - \gamma_m([\mathbf{x}])\|_{w_{\mathcal{X}^*}([\mathbf{x}])}^2 \leq \|\gamma_n - \gamma_m\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}^2 < \varepsilon^2$ and $\|\gamma_n([\mathbf{x}]) - \gamma_m([\mathbf{x}])\| < \frac{\varepsilon}{\sqrt{w_{\mathcal{X}^*}([\mathbf{x}])}}$. Thus, $\{\gamma_n([\mathbf{x}])\}_{n \in \mathbb{N}}$ is a Cauchy sequence in the Hilbert space $\mathfrak{M}\mathcal{Y}^*$. It follows from completeness of $\mathfrak{M}\mathcal{Y}^*$ that the Cauchy sequence converges to a limit $[\mathbf{y}]_{[\mathbf{x}]} := \lim_{n \rightarrow \infty} \gamma_n([\mathbf{x}])$ in $\mathfrak{M}\mathcal{Y}^*$. \square

Then, it is equivalent to the weighted vector-valued l^2 Hilbert space $l^2(\mathfrak{M}\mathcal{X}^*, \mathfrak{M}\mathcal{Y}^*; w_{\mathcal{X}^*}, w_{\mathcal{Y}^*})$.

Proposition 4.18. The inner product space

$$\langle B(\mathcal{X}, \mathcal{Y}), +, \zeta, -, \cdot, 1, \langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}, \|\bullet\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}, d_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} \rangle$$

is a Hilbert space.

Proof. Let \mathcal{X} , \mathcal{Y} , and $\langle B(\mathcal{X}, \mathcal{Y}), +, \zeta, -, \cdot, 1, \langle \bullet, \star \rangle_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}, \|\bullet\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}, d_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} \rangle$ be as stated. The result follows immediately by observing it is an l^2 space. For clarity, it is shown directly.

Recall that $\langle \mathfrak{M}\mathcal{X}^*, \langle \bullet, \star \rangle; w \rangle$ is a Hilbert space. To show $B(\mathcal{X}, \mathcal{Y})$ is a Hilbert space, let $\{\gamma_n\}_{n \in \mathbb{N}}$ be a Cauchy sequence in $B(\mathcal{X}, \mathcal{Y})$ and let $\gamma \in \mathbf{Vect}(\mathcal{X}, \mathcal{Y})$ be the pointwise limit of $\{\gamma_n\}_{n \in \mathbb{N}}$ established by Proposition 4.17.

Consider any source $\mu \in \mathfrak{M}\mathcal{X}^*$ and threshold $\varepsilon > 0$. Convergence $\gamma_n(\mu) \rightarrow \gamma(\mu)$ implies there is a threshold M such that any index $n \geq M$ satisfies $\|\gamma_n(\mu) - \gamma(\mu)\| < \varepsilon$. Any pair of indices $m, n \geq M$ obey the following inequalities:

$$\begin{aligned} \|\gamma_n(\mu) - \gamma(\mu)\| &= \|\gamma_n(\mu) - \gamma_m(\mu) + \gamma_m(\mu) - \gamma(\mu)\| \\ &\leq \|\gamma_n(\mu) - \gamma_m(\mu)\| + \|\gamma_m(\mu) - \gamma(\mu)\| < \|\gamma_n(\mu) - \gamma_m(\mu)\| + \varepsilon. \end{aligned}$$

Then, $\|\gamma_n(\mu) - \gamma(\mu)\|_{w_{\mathcal{X}^*}}^2 \leq \left(\inf_{m \geq n} \|\gamma_n(\mu) - \gamma_m(\mu)\| \right)^2$. For any $o \geq n$, the sum $\|\gamma_n - \gamma\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}^2 = \sum_{\mu \in \mathfrak{M}\mathcal{X}^*} \|\gamma_n(\mu) - \gamma(\mu)\|_{w_{\mathcal{Y}^*}}^2 w_{\mathcal{X}^*}(\mu)$ obeys the following inequalities:

$$\begin{aligned} \|\gamma_n - \gamma\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}^2 &\leq \sum_{\mu \in \mathfrak{M}\mathcal{X}^*} \left(\inf_{m \geq n} \|\gamma_n(\mu) - \gamma_m(\mu)\|_{w_{\mathcal{Y}^*}} \right)^2 w_{\mathcal{X}^*}(\mu) \\ &\leq \sum_{\mu \in \mathfrak{M}\mathcal{X}^*} \|\gamma_n(\mu) - \gamma_o(\mu)\|_{w_{\mathcal{Y}^*}}^2 w_{\mathcal{X}^*}(\mu) = \|\gamma_n - \gamma_o\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}^2 \\ &< \varepsilon^2. \end{aligned}$$

Thus, $\|\gamma_n - \gamma\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}} < \varepsilon$. That is, γ_n converges to γ in the norm topology. \square

Some continuous measurable channels arise from measurable channels as follows.

Definition 4.19 (continuation, restriction). Given event types \mathcal{X} and \mathcal{Y} , define:

1. the *continuation* function $\bar{\bullet} : \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \rightarrow (\mathfrak{M}\mathcal{Y}^*)^{\mathfrak{M}\mathcal{X}^*}$ defined by the assignment from $h \in \mathfrak{M}(\mathcal{X}, \mathcal{Y})$ to the function $\bar{h} : \mathfrak{M}\mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ defined below

$$[[\mathbf{x}] \vdash_{\bar{h}} \mathbf{y}] = \sum_{\mathbf{x}' \in \text{support}[\mathbf{x}']} [\mathbf{x}'](\mathbf{x}) \cdot [\mathbf{x} \vdash_h \mathbf{y}];$$

and

2. the *restriction* function $\underline{\bullet} : \mathbf{Vect}(\mathcal{X}, \mathcal{Y}) \rightarrow (\mathfrak{M}\mathcal{Y}^*)^{\mathcal{X}^*}$ defined by the assignment from $\eta : \mathfrak{M}\mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ to the function $\underline{\eta} : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ defined by $\underline{\eta}(\mathbf{x}) = \eta(\delta_{\mathbf{x}})$.

Given a measurable channel $g \in \mathfrak{M}(\mathcal{X}, \mathcal{Y})$, the continuation \bar{g} weighs g by an input measure $[\mathbf{x}] \in \mathfrak{M}\mathcal{X}^*$ in a way that is linear over $\mathfrak{M}\mathcal{X}^*$. Intuitively, it lifts the channel on \mathcal{X}^* to a continuous channel that acts on distributions in $\mathfrak{M}\mathcal{X}^*$.

Proposition 4.20 (continuation). Given event types $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle$ and $\langle \mathcal{Y}, w_{\mathcal{Y}^*} \rangle$, the following hold:

1. the continuation function $\bar{\bullet} : \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \rightarrow (\mathfrak{M}\mathcal{Y}^*)^{\mathfrak{M}\mathcal{X}^*}$ function is well-defined, and the restriction function $\underline{\bullet} : \mathbf{Vect}(\mathcal{X}, \mathcal{Y}) \rightarrow \mathfrak{M}(\mathcal{X}, \mathcal{Y})$ is well-defined.

2. the image of the continuation is contained in the subspace $B(\mathcal{X}, \mathcal{Y}) \subset (\mathfrak{M}\mathcal{Y}^*)^{\mathfrak{M}\mathcal{X}^*}$.

Proof. Let $\langle \mathcal{X}, w_{\mathcal{X}^*} \rangle, \langle \mathcal{Y}, w_{\mathcal{Y}^*} \rangle, \bar{\bullet}, \underline{\bullet}$ be as stated.

For any $g : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ in $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ and source $[\mathbf{x}] \in \mathfrak{M}\mathcal{X}^*$, observe that the continuation $\bar{g} : \mathfrak{M}\mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ satisfies the Cauchy-Schwarz inequality below.

$$\|\bar{g}([\mathbf{x}])\|^2 \leq \left(\sum_{\mathbf{t} \in \text{support}[\mathbf{x}]} \|\mathbf{x}\|_{w_{\mathcal{X}^*}}^2 \right) \cdot \left(\sum_{\mathbf{t} \in \text{support}[\mathbf{x}]} \|g(\mathbf{t})\|_{w_{\mathcal{Y}^*}} \right) = \|\mathbf{x}\|_{w_{\mathcal{X}^*}}^2 \|g\|_{w_{\mathcal{X}^*}, w_{\mathcal{Y}^*}}^2 \quad (4.1)$$

1. For any linear transformation $\eta : \mathfrak{M}\mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ in $\mathbf{Vect}(\mathcal{X}, \mathcal{Y})$, finitude of $\eta(\bullet)$ implies finitude of $\underline{\eta}(\bullet)$ and thus that $\underline{\eta}$ is well-defined. To show $\bar{\bullet}$ is well-defined on $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$, observe for any $g \in \mathfrak{M}(\mathcal{X}, \mathcal{Y})$ that convergence of the sum \bar{g} follows from the Cauchy-Schwarz inequality and finitude of the two factors.

2. Linearity follows pointwise and boundedness follows from the Cauchy-Schwarz inequality. □

Continuation is linear over $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$.

Lemma 2. *The continuation operator $\bullet : \mathfrak{M}(\mathcal{X}, \mathcal{Y}) \rightarrow B(\mathcal{X}, \mathcal{Y})$ is linear, bounded, and continuous.*

Proof. Let $\mathcal{X}, \mathcal{Y}, \mathfrak{M}(\mathcal{X}, \mathcal{Y}), B(\mathcal{X}, \mathcal{Y})$ be as stated. Closure under addition and scalar multiplication are verified directly. For any $h : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ in $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ and $a \in \mathbb{C}$, the following holds:

$$\begin{aligned} [[\mathbf{x}'] \vdash_{a \cdot h} \mathbf{y}] &= \sum_{\mathbf{x} \in \text{support}[\mathbf{x}']} [\mathbf{x}'](\mathbf{x}) [\mathbf{x} \vdash_{a \cdot h} \mathbf{y}] = a \cdot \sum_{\mathbf{x} \in \text{support}[\mathbf{x}']} [\mathbf{x}'](\mathbf{x}) [\mathbf{x} \vdash_h \mathbf{y}] \\ &= a \cdot [[\mathbf{x}'] \vdash_h \mathbf{y}]. \end{aligned}$$

For any $h_1, h_2 : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$, the following holds:

$$\begin{aligned} [[\mathbf{x}'] \vdash_{h_1 + h_2} \mathbf{y}] &= \sum_{\mathbf{x} \in \text{support}[\mathbf{x}']} [\mathbf{x}'](\mathbf{x}) ([\mathbf{x} \vdash_{h_1} \mathbf{y}] + [\mathbf{x} \vdash_{h_2} \mathbf{y}]) \\ &= [[\mathbf{x}'] \vdash_{h_1} \mathbf{y}] + [[\mathbf{x}'] \vdash_{h_2} \mathbf{y}]. \end{aligned}$$

Since any measurable channel $g : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ is bounded, the supremum $C = \sup_{\mathbf{x} \in \mathcal{X}^*} \|g(\mathbf{x})\|_{w_{\alpha\mathcal{Y}}}$ satisfies $C < \infty$. For any $[\mathbf{x}'] \in \mathfrak{M}\mathcal{X}^*$, the inequality $\|\bar{g}([\mathbf{x}'])\|_{w_{\alpha\mathcal{Y}}} \leq \sum_{\mathbf{x} \in \mathcal{X}^*} \|[\mathbf{x}'](\mathbf{x})\| \cdot \|g(\mathbf{x})\|_{w_{\alpha\mathcal{Y}}}$ yields:

$$\|\bar{g}([\mathbf{x}'])\|_{w_{\alpha\mathcal{Y}}} \leq C \cdot \|[\mathbf{x}']\|_{w_{\mathcal{X}^*}}.$$

Continuity follows from linearity and boundedness. □

Define the evaluator $\pi : \mathfrak{M}\mathcal{X}^* \times B((,) \mathcal{X}) \mathcal{Y} \rightarrow \mathfrak{M}\mathcal{Y}^*$ by $\pi_{[\mathbf{x}]_0}(\gamma) = \gamma([\mathbf{x}]_0)$ and observe that any $\psi \in \mathfrak{M}\mathcal{Y}^*$ induces a function $c_\psi : \mathfrak{M}\mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ given by $c_\psi(\bullet) = \psi$ that is well-defined.

Lemma 3 (Lifting). *For Hilbert spaces $\mathfrak{M}\mathcal{Y}^*$ of finite measures and $B((,) \mathcal{X}) \mathcal{Y}$ of continuous measurable channels, the following hold.*

1. For any finite measure $\chi \in \mathfrak{M}\mathcal{X}^*$, the transformation $\pi_\chi : B((,) \mathcal{X}) \mathcal{Y} \rightarrow \mathfrak{M}\mathcal{Y}^*$ is a linear, bounded, continuous, surjective, open, hereditarily quotient map.
2. For any finite measure $\chi \in \mathfrak{M}\mathcal{X}^*$, the transformation $\pi_\chi : B((,) \mathcal{X}) \mathcal{Y} \rightarrow \mathfrak{M}\mathcal{Y}^*$ satisfies the lifting property: given any sequence $\{\omega_n\}_{n \in \mathbb{N}} \subseteq \mathfrak{M}\mathcal{Y}^*$ converging to some $\omega \in \mathfrak{M}\mathcal{Y}^*$ and continuous measurable channel $\gamma : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathfrak{M}\mathcal{Y}^*$ over $\pi_\chi(\gamma) = \omega$, there is a sequence

$$\left\{ \gamma_n : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathfrak{M}\mathcal{Y}^* \mid n \in \mathbb{N}, \pi_\chi(\gamma_n) = \omega_n \right\}$$

that converges to γ .

Proof. Let π and χ be as stated.

1. Linearity is shown below.

$$\begin{aligned} \forall \eta, \gamma \in B(\mathcal{X}, \mathcal{Y}), \quad \pi_\chi(\gamma + \eta) &= (\gamma + \eta)(\chi) = \gamma(\chi) + \eta(\chi) = \pi_\chi(\gamma) + \pi_\chi(\eta), \\ \forall c \in \mathbb{C}, \gamma \in B(\mathcal{X}, \mathcal{Y}), \quad \pi_\chi(z \cdot \gamma) &= c \cdot \gamma(\chi) = z \cdot \pi_\chi(\gamma). \end{aligned}$$

Let $K = \frac{1}{\sqrt{w_{\mathcal{X}^*}(i)}}$. Note that $\|\gamma\|^2 \geq \|\gamma(\chi)\|^2 w_{\mathcal{X}^*}(\chi)$ yields $\|\gamma(\chi)\| \leq K \|\gamma\|$.

Continuity of the linear transformation follows from boundedness.

To show surjectivity, observe that any $\psi \in \mathfrak{M}\mathcal{Y}^*$ induces a well-defined function $c_\psi : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathfrak{M}\mathcal{Y}^*$ that is a continuous measurable channel satisfying $\pi_\chi(c_\psi) = c_\psi(\chi) = \psi$.

Since $B(\mathcal{X}, \mathcal{Y})$ and $\mathfrak{M}\mathcal{Y}^*$ are Hilbert spaces, it follows from the open mapping theorem that the surjective continuous linear transformation π_χ is an open map and thus a hereditarily quotient map.

2. The lifting property follows from the open mapping theorem.

Therefore, convergent sequences in $\mathfrak{M}\mathcal{Y}^*$ can be continuously lifted to convergent sequences in $B(\mathcal{X}, \mathcal{Y})$ and the converse holds by continuous projection. \square

This yields the following consequence for divergences.

Lemma 4 (atomicity). *If Σ is countable, then the following hold.*

1. For any $\mathbf{t} \in \Sigma^*$, function $\varphi : \Sigma^* \times \mathbb{C} \rightarrow [0, +\infty]$, the function $f : \mathfrak{M}\Sigma^* \rightarrow [0, +\infty]$ defined by $f(\nu) = \varphi_{\mathbf{t}}(\nu(\{\mathbf{t}\}))$ is lower semi-continuous iff $\varphi_{\mathbf{t}} : \mathbb{C} \rightarrow [0, +\infty]$ is lower semi-continuous.

2. For any $\mathbf{t} \in \Sigma^*$, function $\phi : \Sigma^* \times \mathfrak{M}\Sigma^* \times \mathbb{C} \rightarrow [0, +\infty]$, the function $d_{\mathbf{t}} : \mathfrak{M}\Sigma^* \rightarrow \mathfrak{M}\Sigma^* \rightarrow [0, +\infty]$ defined by $d_{\mathbf{t}}(\mu, \nu) = \phi_{\mathbf{t}}^{\mu}(\nu(\{\mathbf{t}\}))$ is lower semi-continuous in the second argument iff $\phi_{\mathbf{t}}^{\mu} : \mathbb{C} \rightarrow [0, +\infty]$ is lower semi-continuous.
3. For any atomic divergence $D : \mathfrak{M}\Sigma^* \times \mathfrak{M}\Sigma^* \rightarrow [0, +\infty]$ given by $D(\mu||\nu) = \sum_{\mathbf{t} \in \Sigma^*} d_{\mathbf{y}}(\mu, \nu)$ in terms of a function $d : \Sigma^* \times \mathfrak{M}\Sigma^* \times \mathfrak{M}\Sigma^* \rightarrow [0, +\infty]$ defined by $d_{\mathbf{y}}(\mu, \nu) = \phi_{\mathbf{t}}^{\mu}(\nu(\{\mathbf{t}\}))$ for some function $\phi : \Sigma^* \times \mathfrak{M}\Sigma^* \times \mathbb{C} \rightarrow [0, +\infty]$, $N_{\mu} = \{\nu \mid D(\mu, \nu) < \infty\}$, $\lambda \in \mathbb{R}$, the function $D(\mu||\bullet)$ is lower semi-continuous on N_{μ} at ν iff $\phi_{\mathbf{t}}^{\mu} : \mathbb{C} \rightarrow [0, \infty]$ is lower semi-continuous for all $\mathbf{t} \in \Sigma^*$.

Proof (infima). Let Σ be a countable type of events.

Recall that the evaluator $\pi_{\mathbf{t}}$ is a continuous surjection.

1. Let \mathbf{t}, φ, f be as stated.

For the “only if” direction, suppose f is lower semi-continuous. To show $\varphi_{\mathbf{t}} : \mathbb{C} \rightarrow [0, +\infty]$ is lower semi-continuous, let $(z_n) \subseteq \mathbb{C}$ converge to some $z \in \mathbb{C}$ and recall that surjectivity of $\pi_{\mathbf{t}}$ allows the construction of a sequence $(\nu_n)_{n \in \mathbb{N}} \subseteq \mathfrak{M}\Sigma^*$ with $\pi_{\mathbf{t}}(\nu_n) = z_n$ that converges to a measure $\nu \in \mathfrak{M}\Sigma^*$ such that $\pi_{\mathbf{t}}(\nu) = z$. Then, lower semi-continuity of $\varphi_{\mathbf{t}}$ follows from that of f as $\liminf_{n \rightarrow \infty} \varphi_{\mathbf{t}}(z_n) = \liminf_{n \rightarrow \infty} f(\nu_n) \geq f(\nu) = \varphi_{\mathbf{t}}(z)$.

For the “if” direction, suppose $\varphi_{\mathbf{t}}$ is lower semi-continuous and let $(\nu_n)_{n \in \mathbb{N}} \subseteq \mathfrak{M}\Sigma^*$ be a sequence converging to some $\nu \in \mathfrak{M}\Sigma^*$. Then, continuity of $\pi_{\mathbf{t}}$ and lower semi-continuity of the composition $\varphi_{\mathbf{t}} \circ \pi_{\mathbf{t}} : \mathfrak{M}\Sigma^* \rightarrow [0, +\infty]$ yields lower semi-continuity of f as follows: $\liminf_{n \rightarrow \infty} f(\nu_n) = \liminf_{n \rightarrow \infty} \varphi_{\mathbf{t}}(\nu_n(\{\mathbf{t}\})) \geq \varphi_{\mathbf{t}}(\nu(\{\mathbf{t}\})) = f(\nu)$.

2. Let $\mathbf{t}, \phi, d_{\mathbf{t}}$ be as stated and define $f : \mathfrak{M}\Sigma^* \rightarrow [0, +\infty]$ by $f(\nu) = d_{\mathbf{t}}(\mu, \nu) = \phi_{\mathbf{t}}^{\mu}(\nu(\{\mathbf{t}\}))$. Then, lower semi-continuity in the second argument of $d_{\mathbf{t}}$ follows from lower semi-continuity in the sole argument of $f(\nu) = \phi_{\mathbf{t}}^{\mu}(\nu(\{\mathbf{t}\}))$ as shown in the first part of this argument.
3. Let ϕ, d, D, μ be as stated.

Suppose $D(\mu||\bullet)$ is lower semi-continuous on N_{μ} . To show for any $\mathbf{t} \in \Sigma^*$ that $\phi_{\mathbf{t}}^{\mu} : \mathbb{C} \rightarrow [0, +\infty]$ is lower semi-continuous, let $z \in \mathbb{C}$, $(z_n) \subseteq \mathbb{C}$ be a sequence converging to z , $(\nu_n)_{n \in \mathbb{N}} \subseteq \mathfrak{M}\Sigma^*$ be a sequence converging to some $\nu \in \mathfrak{M}\Sigma^*$ with $\pi_{\mathbf{t}}(\nu) = z$ such that $\pi_{\mathbf{t}}(\nu_n) = z_n$ and $\nu_n((E)) = 0 = \nu(E)$ for any measurable set $E \in \mathcal{F}$ with $E \not\ni \mathbf{t}$. Lower semi-continuity of $D(\mu||\nu)$ yields that of $\phi_{\mathbf{t}}^{\mu}$ as follows: $\liminf_{n \rightarrow \infty} \phi_{\mathbf{t}}^{\mu}(\nu_n(\{\mathbf{t}\})) = \liminf_{n \rightarrow \infty} D(\mu||\nu_n) \geq D(\mu||\nu) = \phi_{\mathbf{t}}^{\mu}(\nu(\{\mathbf{t}\}))$.

Suppose it is the case for any $\mathbf{t} \in \Sigma^*$ that $\phi_{\mathbf{t}}^{\mu} : \mathbb{C} \rightarrow [0, +\infty]$ is lower semi-continuous. Then, lower semi-continuity of $D(\mu||\bullet)$ follows from Fatou’s Lemma.

The result for $\lambda \cdot D(\mu, \bullet)$ follows from the facts that scalar multiplication is continuous and composition preserves lower semi-continuity.

□

This extends to divergences on channels.

Lemma 5 (semi-continuity). *For any Hilbert space $\mathfrak{M}\mathcal{Y}^*$ and divergence $D : \mathfrak{M}\mathcal{Y}^* \times \mathfrak{M}\mathcal{Y}^* \rightarrow [0, +\infty]$, the following conditions are equivalent:*

1. *for any $\psi \in \mathfrak{M}\mathcal{Y}^*$, the function $D(\psi, \bullet) : \mathfrak{M}\mathcal{Y}^* \rightarrow [0, +\infty]$ is lower semi-continuous.*
2. *for any $\gamma : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathfrak{M}\mathcal{Y}^*$ and $\chi \in \mathfrak{M}\mathcal{X}^*$, the function $D_\chi^\gamma : B(\mathcal{X}, \mathcal{Y}) \rightarrow [0, +\infty]$ defined by $D_\chi^\gamma(\eta) = D(\gamma(\chi), \eta(\chi))$ is lower semi-continuous;*
3. *for any $\gamma : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathfrak{M}\mathcal{X}^*$ and $\chi \in \mathfrak{M}\mathcal{X}^*$, the function $D_\chi^\gamma : B(\mathcal{X}, \mathcal{Y}) \rightarrow [0, +\infty]$ defined by $D_\chi^\gamma(\eta) = \lambda \cdot D(\gamma(\chi), \eta(\chi))$ is lower semi-continuous for any $\lambda \in \mathbb{R}$; and*
4. *if there is a function $\phi : \mathcal{Y}^* \times \mathfrak{M}\mathcal{Y}^* \times \mathbb{C} \rightarrow [0, +\infty]$ such that the function $d : \mathcal{Y}^* \times \mathfrak{M}\mathcal{Y}^* \times \mathfrak{M}\mathcal{Y}^* \rightarrow [0, +\infty]$ defined by $d_{\mathbf{y}}(\psi, \omega) = \phi_{\mathbf{y}}^\psi(\omega(\mathbf{y}))$ satisfies the equation $D(\psi, \omega) = \sum_{\mathbf{y} \in \mathcal{Y}^*} d_{\mathbf{y}}(\psi, \omega)$, then $\phi_{\mathbf{y}}^\psi : \mathbb{C} \rightarrow [0, +\infty]$ is lower semi-continuous for all \mathbf{y} .*

If any of the above conditions hold, $\chi \in \mathfrak{M}\mathcal{X}^$, $\gamma : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathfrak{M}\mathcal{Y}^*$, $H \subseteq B(\mathcal{X}, \mathcal{Y})$ is compact, and the restriction $D_\chi^\gamma \upharpoonright_H : H \rightarrow [0, +\infty]$ of D_χ^γ on H is not identically $+\infty$, then the infimum $d_0 = \inf_{h \in H} D_\chi^\gamma(h)$ is finite and the minimizer set $H_0 = \{h \in H \mid D_\chi^\gamma(h) \leq d_0\}$ is not empty.*

Proof. Let D be as stated.

1. Suppose Condition 2 holds and let $\psi \in \mathfrak{M}\mathcal{Y}^*$. To show $D(\psi, \bullet)$ is lower semi-continuous, let $\chi \in \mathfrak{M}\mathcal{X}^*$ and $\{\omega_n\}_{n \in \mathbb{N}}$ be a sequence converging to some $\omega \in \mathfrak{M}\mathcal{Y}^*$. By the surjectivity of $\pi_\chi : B(\mathcal{X}, \mathcal{Y}) \rightarrow \mathfrak{M}\mathcal{Y}^*$ established in the first part of the Lifting Lemma (3), there is a continuous measurable channel $\gamma \in B(\mathcal{X}, \mathcal{Y})$ such that $\pi_\chi(\gamma) = \gamma(\chi) = \psi$. By the lifting property of π_χ established in the second part of the Lifting Lemma (3), there is a sequence $\{\eta_n : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathfrak{M}\mathcal{Y}^* \mid n \in \mathbb{N}, \pi_\chi(\eta_n) = \omega_n\} \subseteq B(\mathcal{X}, \mathcal{Y})$ converging to $\eta \in B(\mathcal{X}, \mathcal{Y})$. Lower semi-continuity of D_χ^γ is shown to yield that of $D(\psi, \bullet)$ by $\liminf_{n \rightarrow \infty} D(\psi, \nu_n) = \liminf_{n \rightarrow \infty} D_\chi^\gamma(\eta_n) \geq D_\chi^\gamma(\eta) = D(\psi, \nu)$.
2. Suppose Condition 1 holds and let $\gamma : \mathfrak{M}\mathcal{X}^* \xrightarrow{\Sigma} \mathfrak{M}\mathcal{Y}^*$ be a continuous measurable channel. To show for any $\chi \in \mathfrak{M}\mathcal{X}^*$ that D_χ^γ is lower semi-continuous, let $\{\eta_n\}_{n \in \mathbb{N}} \subseteq B(\mathcal{X}, \mathcal{Y})$ be a sequence converging to some $\eta \in B(\mathcal{X}, \mathcal{Y})$. It follows from continuity of the evaluator that $\{\pi_\chi(\eta_n)\}_{n \in \mathbb{N}} = \{\eta_n(\chi)\}_{n \in \mathbb{N}} \subseteq \mathfrak{M}\mathcal{Y}^*$ converges to $\pi_\chi(\eta) = \eta(\chi)$. Lower semi-continuity of $D(\gamma(\chi), \bullet)$ yields the result below:

$$\liminf_{n \rightarrow \infty} D_\chi^\gamma(\eta_n) = \liminf_{n \rightarrow \infty} D(\gamma(\chi), \eta_n(\chi)) \geq D(\eta(\chi), \eta(\chi)) = D_\chi^\gamma(\eta).$$

3. The equivalence of Conditions 2-3 follow from the continuity of scalar multiplication and the fact that composition preserves lower semi-continuity.
4. The equivalence of Conditions 4 and 1 was established in Lemma 4.

The final result follows from the extreme value theorem. \square

Channels form spaces over edges in the graphs presented in the next section.

4.2 Graph

The formulation in Problem 4.32 of probabilistic channel post-training can be seen as a special case of post-training a protocol on a network. This section introduces what is meant in this context by a *network* and specifies a restriction of the protocol post-training problem that applies to a measurable channel.

Let $G = E \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} V$ be a graph. The type E of edges defines a relation $V \rightarrow V = V \xrightarrow[E]{} V$ on V in which a node u is related to a node v iff there is an edge $e \in E$ denoted by $u \xrightarrow{e} v = u \xrightarrow[E]{} v$ that satisfies $s(e) = u$ and $t(e) = v$, and it induces the subset $E \subseteq V \times V$ defined below.

$$E = \left\{ (u, v) \in V \times V \mid \exists u \xrightarrow{e} v \in E \right\}$$

Nodes obtain the following labels.

Definition 4.21 (event graph). An *event graph* over a graph $G = E \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} V$ is a graph $E \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} \mathcal{X}$ defined below:

1. types of subjects $\{\mathbb{S}_v\}_{v \in V}$, actions $\{\mathbb{A}_v\}_{v \in V}$, objects $\{\mathbb{J}_v\}_{v \in V}$, and security levels $\{\mathbb{L}_v\}_{v \in V}$;
2. event types $\mathcal{X} = \{\mathcal{X}_v\}_{v \in V}$ given by $\mathcal{X}_v = \mathbb{A}_v \times \mathbb{J}_v \times \mathbb{S}_v \times \mathbb{L}_v$ for each $v \in V$;
3. functions $s, t : E \rightarrow \mathcal{X}$ satisfying $(s, t)(e) = (\mathcal{X}_u, \mathcal{X}_v)$ iff $(s, t)(e) = (u, v)$ for any $e \in E$ and $u, v \in V$.

It gives rise to the following structure.

1. upper topologies on the free monoids $\{\mathcal{X}_v^*\}_{v \in V}$ over the event types;
2. Hilbert spaces $\{\mathfrak{M}\mathcal{X}_v\}_{v \in V}$ and $\{\mathfrak{M}\mathcal{X}_v^*\}_{v \in V}$;
3. coproducts $\mathbb{A} = \bigsqcup_{v \in V} \mathbb{A}_v, \mathbb{J} = \bigsqcup_{v \in V} \mathbb{J}_v, \mathbb{S} = \bigsqcup_{v \in V} \mathbb{S}_v, \mathbb{L} = \bigsqcup_{v \in V} \mathbb{L}_v$;
4. the type $\Sigma_{\mathcal{X}} = \mathbb{A} \times \mathbb{J} \times \mathbb{S} \times \mathbb{J}$ of *global* events.

Note that a global event cannot be produced unless it is comprised of data corresponding to a single node.

Definition 4.22 (network, nodal purge). A *network* over an event graph $G = \langle V, E, \mathcal{X} \rangle$ is a type $\mathcal{F} = \{F_e\}_{e \in E}$ of edge labels where each edge $u \xrightarrow{e} v$ in E is assigned the Hilbert space $F_e = \{f : \mathcal{X}_u^+ \rightarrow \mathfrak{M}\mathcal{X}_v\}$ of measurable matrices from \mathcal{X}_u^+ to \mathcal{X}_v . It induces the following structure:

1. Hilbert spaces $\{G_e\}_{e \in E}$ of cumulative measurable channels where each edge $u \xrightarrow{e} v$ is assigned $G_e = \mathfrak{M}(\mathcal{X}_u, \mathcal{X}_v)$;
2. product topology on $F = \prod_{e \in E} F_e$;
3. the subtype $\Sigma = \bigsqcup_{v \in V} \mathcal{X}_v$ in $\Sigma_{\mathcal{X}}$ of *network* events; and
4. the *nodal purge* function $\upharpoonright : V \times \Sigma^* \rightarrow \Sigma^*$ defined by the purge channel $\mathbf{t} \upharpoonright_{\mathbb{S}_v}$ and denoted $\mathbf{t}^{(v)}$.

The network model is one in which there may be adversaries with unknown capabilities to eavesdrop on unknown subsets of the channels, i.e., to view their outputs, but with no ability to eavesdrop on states of nodes.

Recall that a measure $\mu : \mathcal{B}\Sigma^* \rightarrow \mathbb{C}$ in $\mathfrak{M}\Sigma^*$ can be viewed as a function $C_\mu : \Sigma^* \rightarrow \mathbb{C}$ and, equivalently, that a function $C : \Sigma^* \rightarrow \mathbb{C}$ gives rise to a function $\mu_C : \mathcal{B}\Sigma^* \rightarrow \mathbb{C}$ which is a measure under certain conditions.

Definition 4.23 (network cost, network reward). A *network value* consists of a network N with a type Σ of network events and function with signature $\Sigma^* \rightarrow \mathbb{C}$. A network cost is also referred to as a *network cost* C or *network reward* reward R . A network reward R corresponds to a network cost C iff they are defined on the same network and $R(\mathbf{t}) = |C(\mathbf{t})| - C(\mathbf{t})$.

Observe that network costs give rise to network rewards and any network reward arises in this way from some network cost. A cost function $C : \Sigma^* \rightarrow \mathbb{C}$ corresponds to an unintended behavior, e.g., measuring a type of unsafety. It is common to learn a model $C : \Sigma^* \rightarrow [0, 1]$ from data consisting of traces in Σ^* with labels $\{0, 1\}$. They can be defined on the following substructures.

Definition 4.24 (subnetwork). Given a network $\langle N, \Sigma \rangle$ over a graph $G = E \begin{smallmatrix} s \\ \rightrightarrows \\ t \end{smallmatrix} V$, define a *subnetwork* to consist of the following:

1. a subtype $A \subseteq E$;
2. the subtype $V_A = s(A) \cup t(A)$ of V ;
3. the subgraph $G_A = E \begin{smallmatrix} s \upharpoonright_A \\ \rightrightarrows \\ t \upharpoonright_A \end{smallmatrix} V$;

4. the subtype $\mathcal{F}_A = \{F_a\}_{a \in A}$ of \mathcal{F} ;
5. the subtype $\{\mathcal{X}_v\}_{v \in V_A}$; and
6. the subtype $\Sigma_A = \bigsqcup_{v \in V_A} \mathcal{X}_v$ of Σ .

Network values are assigned to multi-turn interactions in subnetwork traces instead of the output traces \mathcal{X}_v^* of an individual channel.

Definition 4.25 (subnetwork value). A *subnetwork value* consists of a subnetwork N_A of a network and a network cost over N_A . A subnetwork reward corresponds to a subnetwork cost over the same subnetwork iff the network reward on the subnetwork corresponds to the network cost on the subnetwork.

Network rewards are used to weigh channels.

Definition 4.26 (reward-weighted channel). Given a subnetwork reward $R_A : \Sigma_A^* \rightarrow \mathbb{C}$ on a subnetwork over a subgraph $\langle V_A, A \rangle$, subtype \mathcal{X} of Σ_A^* , subtype \mathcal{Y} of Σ_A^* , and a channel $g : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$, define the *reward-weighted channel* $g_R : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ by $[\mathbf{x} \vdash_{g_R} \mathbf{y}] = R_A(\mathbf{x} :: \mathbf{y}) \cdot [\mathbf{x} \vdash_g \mathbf{y}]$.

This is used to normalize the rewards.

Definition 4.27 (reward-normalized channel). Let $g_R : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ be a reward-weighted channel. Define its *reward partition function* $Z : \mathcal{X}^* \rightarrow \mathbb{C}$ by $Z(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}^*} [\mathbf{x} \vdash_{g_R} \mathbf{y}]$ and its *reward-normalized channel* $r_A : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ by $[\mathbf{x} \vdash_{r_A} \mathbf{y}] = \frac{1}{\|Z(\mathbf{x})\|} [\mathbf{x} \vdash_{g_R} \mathbf{y}]$.

Given a subnetwork value on a subnetwork N_A of N , a subnetwork N_B of N_A is optimized by means of a divergence for each edge in N_B .

Definition 4.28 (local network preference). A *local network preference on a subnetwork value* over a subgraph $G_A = V_A \xrightarrow[s]{t} A$ is a subtype $B \subseteq A$ of edges equipped with lower semi-continuity in divergences $\left\{ D_b : \mathfrak{M}\mathcal{X}_{t(b)}^* \times \mathfrak{M}\mathcal{X}_{t(b)}^* \rightarrow [0, +\infty] \right\}_{b \in B}$.

The first consideration is the optimization of a single edge in a subnetwork with a subnetwork value.

Definition 4.29 (channel preference). A *channel preference* for an edge $b \in E$ in a network N is a local network preference $\{b\} \subseteq A$ on a subnetwork value R_A in N .

Attention is initially restricted to the situation in which a subnetwork value is defined on a subnetwork that consists of a single edge.

Definition 4.30 (microscopic channel preference). A *microscopic channel preference* for an edge $b \in E$ in a network N is a channel preference for an b in which the subnetwork value R_A is defined on a subnetwork over a graph $G_A = V_A \underset{t}{\overset{s}{\rightrightarrows}} A$ with $A = \{b\}$.

Observe that any channel can be viewed as a protocol on a network over a graph that consists of a single edge.

Problem 4.31 (measurable channel post-training). Given:

1. countable event types $V = \{\mathcal{X}, \mathcal{Y}\}$;
2. a cumulative measurable channel $g : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ and a measurable property $[\mathbf{x}]_0 \in \mathfrak{M}\mathcal{X}^*$;
3. a compact subspace $H \ni g$ of $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$;
4. a microscopic channel preference over g with a divergence D over $\mathfrak{M}\mathcal{Y}^*$ and a reward-normalized channel $r : \mathcal{X}^* \rightarrow \mathcal{Y}^*$
5. a real number $\lambda \in \mathbb{R}$ such that $\lambda \geq 0$;
6. a function $L : H \rightarrow (-\infty, +\infty]$ defined below:

$$L(h; D, g, [\mathbf{x}]_0, \lambda) = D(\bar{g}([\mathbf{x}]_0), \bar{h}([\mathbf{x}]_0)) + \lambda \cdot D(\bar{r}([\mathbf{x}]_0), \bar{h}([\mathbf{x}]_0)) \quad (4.2)$$

in terms of fixed D, g, λ, r

the *measurable channel post-training problem* is the following optimization problem.

$$\min_{h \in H} L(h) \quad (4.3)$$

This problem is soluble.

Theorem 1 (solubility of measurable channel post-training). *Any instance of Problem 4.31 admits a solution.*

Proof. Consider an instance of Problem 4.31 and let $\ell = \inf_{h \in H} L(h)$.

If $\ell = +\infty$, then any $h \in H$ is vacuously a solution. Else, suppose $\ell \neq +\infty$.

Then, observe that L is a composition of lower semi-continuous functions and thus that it is lower semi-continuous. The result follows from the extreme value theorem.

In all cases, there is an $h^* \in H$ such that $L(h^*) = \inf_{h \in H} L(h)$. □

The proof followed naturally from the continuity of continuation and the completeness of Hilbert spaces, which necessitated the generalization from probability to complex measure.

A particular case restricted to real measures is presented in the next section.

4.3 Probabilistic channel

In analogy with the general Problem 4.31, consider the following refinement to the probabilistic channels of interest.

Problem 4.32 (probabilistic channel post-training). Given

1. countable event types \mathcal{X}, \mathcal{Y} with subspace topologies on $\Delta\mathcal{X}^* \subseteq \mathfrak{M}\mathcal{X}^*$ and $\Delta\mathcal{Y}^* \subseteq \mathfrak{M}\mathcal{Y}^*$;
2. a cumulative probabilistic channel $g : \mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$ in $\Delta(\mathcal{X}, \mathcal{Y}) \subseteq \mathfrak{M}(\mathcal{X}, \mathcal{Y})$ and a probabilistic property $[\mathbf{x}]_0$ in $\Delta\mathcal{X}^* \subseteq \mathfrak{M}\mathcal{X}^*$;
3. a compact subspace $H \ni g$ of $\Delta(\mathcal{X}, \mathcal{Y}) \subseteq \mathfrak{M}(\mathcal{X}, \mathcal{Y})$ with the subspace topology for which any $h \in H$ satisfies $\bar{g}([\mathbf{x}]_0) \ll \bar{h}([\mathbf{x}]_0)$;
4. a microscopic channel preference over g with a reward-normalized channel $r : \mathcal{X}^* \rightarrow \Delta\mathcal{Y}^*$ and a statistical divergence D over $\Delta\mathcal{Y}^* \subseteq \mathfrak{M}\mathcal{Y}^*$;
5. a constant $\lambda \in \mathbb{R}$ such that $\lambda \geq 0$
6. a function $L : H \rightarrow (-\infty, +\infty]$ defined below

$$L(h; D, g, [\mathbf{x}]_0, \lambda) = D(\bar{g}[\mathbf{x}]_0, \bar{h}[\mathbf{x}]_0) + \lambda \cdot D(\bar{r}[\mathbf{x}]_0, \bar{h}[\mathbf{x}]_0) \quad (4.4)$$

in terms of fixed $D, g, [\mathbf{x}]_0, \lambda$:

the *probabilistic channel post-training problem* is the following optimization problem.

$$\min_{h \in H} L(h) \quad (4.5)$$

The desired result is shown below.

Theorem 2 (solubility of the probabilistic channel post-training problem). *Consider Problem 4.32.*

1. *Any instance of Problem 4.32 is an instance of Problem 4.31.*
2. *Any instance of Problem 4.32 admits a solution.*
3. *Any instance of Problem 3.30 is an instance of Problem 4.31.*

Proof. Consider Problem 4.32.

1. The reduction of Problem 4.32 to Problem 4.31 follows from the fact that probability measures are a special case of complex measures.
2. The solubility of Problem 4.32 follows from Theorem 1.

3. The subspace topology inherited from a metric topology is the metric topology on the subspace.

□

Therefore, any pre-trained probabilistic AI model admits post-trained AI counterparts that better satisfy microscopic channel preferences, e.g. as specified by a reward model trained on data.

A formal simplification is that the network preference in Problem 4.32 is microscopic. This restriction is lifted in the next chapter on networks and communication channels.

CHAPTER 5 NETWORK

The results thus far concerning individual channels in isolation are generalized in this section to networks of channels. In particular, it is shown that post-training admits a solution when the cost function C is assigned to the traces of an interactive conversation between *two* speakers, as opposed to the previous section which did this for the outputs of a single channel from one speaker. The analogous result for a protocol that involves an arbitrary number of speakers is shown in Chapter 6, and some of the necessary infrastructure is presented in this section.

In the remainder of this article, let $G = E \underset{t}{\overset{s}{\rightrightarrows}} V$ denote an arbitrary finite graph and $N = \langle G, \mathcal{F} \rangle$ be a network over G .

One of the nodes in G is labeled to denote that the interaction begins at that node. In the context of a protocol for the game of chess, the two nodes could correspond to the two players, and the label attached to one of the two players indicates that they make the first move.

Definition 5.1 (graph initialization, initial graph). An *initialization* of a graph $G = \langle V, E \rangle$ is a pair $\langle v_0, e_0 \rangle$ consisting of an *initial* node $v_0 \in V$ and an *initial* edge $v_0 \xrightarrow{e_0} v_1$ in E . An *initial* graph $G_0 = \langle V_0, E_0 \rangle$ consists of a graph $G = \langle V, E \rangle$ with a pointed type $\langle V, v_0 \rangle$ of nodes and a pointed type $\langle E, e_0 \rangle$ of edges such that $\langle v_0, e_0 \rangle$ is an initialization of G .

The networks under investigation are defined over such graphs.

Definition 5.2 (network initialization, initial network). A *network initialization* on a network N over a graph $G = \langle V, E \rangle$ is a graph initialization $\langle v_0, e_0 \rangle$ on G with a fixed measure $[\mathbf{x}_0] \in \mathfrak{M}\mathcal{X}_{v_0}^*$ over the trace space \mathcal{X}_0^* , and an *initial network* $N_0 = \langle G_0, [\mathbf{x}_0] \rangle$ consists of an initial graph with a network initialization over it.

Motion along the network is enabled by the following structure.

Definition 5.3 (router, node channel, scheduler). Given a network N over a graph $G = \langle V, E \rangle$, define:

1. a *router* to be a function $o : V \times \Sigma^* \rightarrow \Delta V$;
2. a *node matrix* to be a function $f : V \times \Sigma^* \rightarrow \mathfrak{M}\Sigma$; and
3. a *scheduler* to be a function $v : \Sigma^* \rightarrow \Delta V$.

The dynamics arise from the following structure.

Definition 5.4 (protocol, initial protocol, executable protocol). A *protocol* over a network $N = \langle G, \mathcal{F} \rangle$ consists of a point $\vec{f} \in \mathcal{F}$. An *initial* protocol is a protocol on an initial network, and an *executable protocol* is an initial protocol with a scheduler and node matrix.

The idea is that the protocol executes evolution by executing its constituent channels on a global trace accumulated by the entire protocol, although a given node only has a local view of the global trace.

Example 5.5 (stateless surveillance). Surveillance equipment can monitor communications and facial expressions, but it can not yet observe states of mind directly.

This motivates the view of a protocol *state* as consisting of concatenations of traces from channels without regard to any notion of node-internal state.

Whereas the traces of a channel $g_{u \rightarrow v} : \mathcal{X}_u^* \rightarrow \mathfrak{M}\mathcal{X}_v^*$ belong in \mathcal{X}_v^* , the traces of a protocol belong in Σ^* .

Consider the phenomenon described below.

Principle 5.6 (simultaneity). Given a graph $G = \langle V, E \rangle$ and a node $w \in V$ with a type \mathbb{S}_w of subjects, it is possible for a subject $A \in \mathbb{S}_w$ to simultaneously witness events from the stochastic matrices over a distinct pair of edges $u \xrightarrow{e_u} w$ and $v \xrightarrow{e_v} w$.

Whenever such a phenomenon occurs, neither of the incoming traces from the stochastic matrices are received. This could be addressed in many ways, for example, by extending V to $V_\infty = V \sqcup \{\infty\}$ and then replacing each channel $\mathcal{X}_u^* \xrightarrow{g_{u,v}} \mathfrak{M}\mathcal{X}_v^*$ with restrictions $g_{u,\infty}, g_{\infty,v}$ of the continuous channels satisfying $\mathfrak{M}\mathcal{X}_u^* \xrightarrow[g_{u,\infty}]{\Sigma} \mathfrak{M}\mathcal{X}_\infty^* \xrightarrow[g_{\infty,v}]{\Sigma} \mathfrak{M}\mathcal{X}_v^* = \overline{g_{u,v}}$ so that the network is well-behaved and the error modes are described by the node ∞ of nature with perturbations of the original channels, which may serve as a basis for alternative models of threat and attack. This nuance is disregarded by the following simplification.

Assumption 5.7 (sequence). Nodes never witness the simultaneous arrival of events, and the time elapsed between the witness of events from distinct matrices is large with respect to the time elapsed in transit.

That is, inputs are processed before subsequent inputs arrive. Thus, the state of a node is captured by the unique trace $\mathbf{x}_v \in \mathcal{X}_v^*$ observed by some point in time and the state of the protocol can be viewed as an element in $\prod_{v \in V} \mathcal{X}_v^*$, although its points do not correspond to states uniquely since two different nodes can observe the arrival of traces at moments in time that they perceive to be simultaneous.

As such, protocol traces are equivalence classes of $\prod_{v \in V} \mathcal{X}_v^*$ defined by indistinguishability to observers at nodes. The following simplification holds in classical regimes.

Assumption 5.8 (triviality). The equivalence classes that define protocol states are trivial.

Thus, the state of a protocol is described by a unique element in $\prod_{v \in V} \mathcal{X}_v^*$. Observe that there is a surjection $\pi : \Sigma^* \rightarrow \prod_{v \in V} \mathcal{X}_v^*$ since the local states are projections and that there are generally many

global sections since there are many permutations with fixed projections which permute symbols belonging to different nodes. That is, there are many ways that traces in Σ^* correspond to the same local traces.

The main idea is that the protocol evolves one step a time as captured by the following structure.

Definition 5.9 (protocol matrix, cumulative protocol channel). Given an executable protocol, an initial source $[\mathbf{x}_0] \in \mathfrak{M}\mathcal{X}_0^*$, scheduler $v : \Sigma^* \rightarrow \Delta V$, and node matrix $f : V \times \Sigma^* \rightarrow \mathfrak{M}\Sigma$, define:

1. the *protocol matrix* M to be the measurable matrix $M : \Sigma^* \rightarrow \mathfrak{M}\Sigma$ given below

$$M(\mathbf{t}) = \sum_{v \in V} v(\mathbf{t})(v) f_v(\mathbf{t});$$

2. the cumulative *channel/protocol (measurable) channel* P to be the accumulation of M .

The uncertain properties relevant for the protocol are probability distributions in $\Delta\Sigma^*$ or more general measures in $\mathfrak{M}\Sigma^*$, and the requirements of the protocol are requirements of its measurable matrix or cumulative measurable channel.

Definition 5.10 (protocol continuation, discount, discounted protocol). Given a cumulative protocol channel P , define:

1. the *protocol continuation* $\rho : \mathfrak{M}\Sigma^* \rightarrow \mathfrak{M}\Sigma^*$ to be the continuation $\rho = \overline{P}$;
2. the composition $\rho^n : \mathfrak{M}\Sigma^* \rightarrow \mathfrak{M}\Sigma^*$ inductively by $\rho^0 = 1_{\mathfrak{M}\Sigma^*}$ and $\rho^n = \rho \circ \rho^{n-1}$ for $n \geq 1$;
3. a *discount* to be a sequence $\beta = (\beta_n)_{n \in \mathbb{N}}$ in $\mathbb{R}^{\mathbb{N}}$ such that $\sum_{n \in \mathbb{N}} \beta_n < \infty$;
4. the *discounted protocol* $\rho^{(N)} : \mathfrak{M}\Sigma^* \rightarrow \mathfrak{M}\Sigma^*$ for some discount $\beta = (\beta_n)_{n \in \mathbb{N}}$ in $\mathbb{R}^{\mathbb{N}}$ by
$$\rho^{(N)}([\mathbf{t}]) = \sum_{n=0}^N \beta_n \rho^n([\mathbf{t}]).$$

Initial protocols give rise to the following sources.

Definition 5.11 (initial protocol source, N -discounted protocol source, discounted protocol source). Given discount $\beta = (\beta_n)_{n \in \mathbb{N}}$ in $\mathbb{R}^{\mathbb{N}}$ and a cumulative protocol channel P for some initial protocol with initial node source $[\mathbf{x}_0] \in \mathfrak{M}\mathcal{X}_0^*$, define:

1. the *initial protocol source* $[\mathbf{t}]_0 \in \mathfrak{M}\Sigma^*$ by the measure $[\mathbf{t}]_0 : \mathcal{B}\Sigma^* \rightarrow \mathbb{C}$ given point-wise as follows

$$[\mathbf{t}]_0(\mathbf{x}) = \begin{cases} [\mathbf{x}_0](\mathbf{x}) & \text{if } \mathbf{x} \in \mathcal{X}_0^* \\ 0 & \text{otherwise} \end{cases}; \quad (5.1)$$

2. the N -discounted protocol source $[\mathbf{t}]_N \in \mathfrak{M}\Sigma^*$ defined by $[\mathbf{t}]_N = \rho^{(N)}([\mathbf{t}]_0)$ for any $N \in \mathbb{N}$;
and

These sources give rise to the following source.

Definition 5.12 (discounted protocol source, discounted node source). Given an initial protocol with N -discounted protocol sources $\{[\mathbf{t}]_N\}_{N \in \mathbb{N}}$, the *discounted* protocol source $[\mathbf{t}] \in \mathfrak{M}\Sigma^*$ is defined by $[\mathbf{t}] = \lim_{N \rightarrow \infty} [\mathbf{t}]_N$ and the discounted node source $[\mathbf{x}_v] \in \mathfrak{M}\mathcal{X}_v^*$ at any node $v \in V$ is given by $[\mathbf{x}_v]_v(\mathbf{x}'_v) = \sum_{\mathbf{u}^{(v)} = \mathbf{x}'_v} [\mathbf{t}](\mathbf{u})$ for each $v \in V$.

Any protocol gives rise to a protocol channel, and this can be done in different ways.

5.1 Security

First, the data needed to construct a protocol channel can be defined probabilistically as follows.

Example 5.13 (router, node channel, scheduler). Given a network N over G with measurable matrices $\{F_{u,v}\}_{(u,v) \in E}$, define:

1. the *router* $o : V \times \Sigma^* \rightarrow \Delta V$ by $[\mathbf{t} \vdash_{o_u} v] = \frac{\|f_{u,v}(\mathbf{t}^{(u)})\|}{\sum_{(u,w) \in E} \|f_{u,w}(\mathbf{t}^{(u)})\|}$;
2. the *node matrix* $f : V \times \Sigma^* \rightarrow \mathfrak{M}\Sigma$ by $f_u(\mathbf{t}) = \sum_{(u,v) \in E} [\mathbf{t} \vdash_{o_u} v] f_{u,v}(\mathbf{t}^{(u)})$; and
3. the *scheduler* $v : \Sigma^* \rightarrow \Delta V$ by $[\mathbf{t} \vdash_v u] = \frac{\|f_u(\mathbf{t})\|}{\sum_{v \in V} \|f_v(\mathbf{t})\|}$.

This gives rise to a protocol matrix and thus to a protocol channel.

It can also be defined differently for a type of protocol defined below.

Definition 5.14 (Send/Rcvactions, sendevent, rcvevent, Send/Rcvprotocol). Given a protocol P with a type \mathbb{S} of subjects, define:

1. the *type A of Send/Rcvactions* by $A = \{\langle \bullet \rangle, (\bullet)\} \times \mathbb{S}$
2. an event r to be a *sendevent* iff its action is $\langle u \rangle \in A$ for some $u \in \mathbb{S}$ and r to be a *rcvevent* iff its action is (u) for some $u \in \mathbb{S}$; and
3. P to be a *Send/Rcvprotocol* iff $\mathbb{A}_v = A$ for each node $v \in V$.

Consider the following delays.

Definition 5.15 (match, unmatched, out, delay, partition, normalized delay). For any type Σ of events in a Send/Rcvprotocol, define:

1. an event s with action $a_i = \langle u \rangle$ to be *matched* in a trace $\mathbf{t} \in \Sigma^*$ iff there are an indices $i, j \leq |\mathbf{t}|$ such that $\mathbf{t}_i = s$ and $\mathbf{t}_j = r$ for some event r at index $j \geq i$ with action $a_j = \langle u \rangle$;
2. an event s to be *unmatched* in a trace iff it is not matched;
3. the *out* function $q : \Sigma^* \times \mathbb{S} \rightarrow \mathbb{N}$ to map (\mathbf{t}, u) to the number of *sendevents* $\langle u \rangle$ which are not matched in \mathbf{t} ;
4. the *delay* function $D : \Sigma^* \times \mathbb{S} \rightarrow [0, +\infty)$ by $D(\mathbf{t}, u) = \frac{1}{1 + q(\mathbf{t}, u)}$;
5. the *delay partition* function $Z : \Sigma^* \rightarrow [0, +\infty)$ by $Z(\mathbf{t}) = \sum_{u \in \mathbb{S}} D(\mathbf{t}, u)$; and
6. the *normalized delay* function $d : \Sigma^* \times \mathbb{S} \rightarrow [0, \infty)$ by $d(\mathbf{t}, u) = \frac{D(\mathbf{t}, u)}{Z(\mathbf{t})}$.

This can be used to define the data required for a protocol channel.

5.2 Global channel

Consider a protocol with protocol channel $P : \Sigma^* \rightarrow \mathfrak{M}\Sigma^*$. The protocol can be viewed as channel on Σ^* or as an operator $\rho = \overline{P}$ for evolving a measurable property in $\mathfrak{M}\Sigma^*$. Even though the data of a protocol is defined locally per node, they give rise to a notion of evolution that is global.

Definition 5.16 (global protocol preference). A *global protocol preference* consists of a (global) network value and lower semi-continuity in the second argument of a divergence $D : \mathfrak{M}\Sigma^* \times \mathfrak{M}\Sigma^* \rightarrow [0, +\infty]$.

Consider a global protocol preference and the question of whether there is a global behavior that optimizes it.

Problem 5.17 (global protocol channel post-training problem). Given:

1. an initial graph $G_0 = \langle V_0, E_0 \rangle$ over $\langle V, E \rangle$;
2. an executable protocol channel P over G_0 with source $[\mathbf{t}] \in \mathfrak{M}\Sigma^*$;
3. a compact subspace $H \ni P$ of the Hilbert space $\mathfrak{M}(\Sigma, \Sigma)$ of measurable channels from Σ to Σ ;
4. a global protocol preference over the full subset $A = E$ of edges with a reward-normalized channel $r_E : \Sigma^* \rightarrow \mathfrak{M}\Sigma^*$;
5. a constant $\lambda \in \mathbb{R}$ satisfying $\lambda \geq 0$;

6. a function $L : H \rightarrow (-\infty, +\infty]$ defined as follows

$$L(h; D, P, [\mathbf{t}], \lambda, r) = D(\bar{P}([\mathbf{t}]), \bar{h}([\mathbf{t}])) + \lambda \cdot D(\bar{r}([\mathbf{t}]), \bar{h}([\mathbf{t}])) \quad (5.2)$$

in terms of fixed $D, P, [\mathbf{t}], \lambda, r$:

the *global protocol channel post-training problem* is the following optimization problem.

$$\min_{h \in H} L(h) \quad (5.3)$$

This problem is soluble.

Theorem 3 (solubility of the global protocol channel post-training problem). *Any instance of Problem 5.17 admits a solution.*

Proof. Observe that H is a compact space containing P and that lower semi-continuity of L follows from lower semi-continuity in the second argument of D . Then, the result follows from the extreme value theorem. \square

When the protocol is viewed as a program on the global state, the above theorem guarantees that there is an optimally post-trained program on the global state. However, this program $h^* \in \arg \min_{h \in H} L(h)$ may not be realizable as the protocol channel of an executable protocol over the same initial protocol. This issue is resolved by a general result in Chapter 6. A special case is presented in Section 5.3 below.

5.3 Communication

A special case of an initial protocol on two nodes is presented below.

Definition 5.18 (communication channel). A *communication channel* is an initial protocol with two nodes $V = \{X, Y\}$, edges between distinct nodes, $\vec{f} = (f_X, f_Y)$, corresponding cumulative channels $g_X : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ from X to Y and $g_Y : \mathcal{Y}^* \rightarrow \mathfrak{M}\mathcal{X}^*$ from Y to X , the channel $P : \mathcal{X}^* \rightarrow \Delta\mathcal{X}^*$ whose continuation $\bar{P} : \Delta\mathcal{X}^* \rightarrow \Delta\mathcal{X}^*$ is given by $\bar{P} = \bar{g}_Y \circ \bar{g}_X$, a probability distribution $[\mathbf{x}]_0 \in \Delta\mathcal{X}^*$, and the languages $L_Y, L_X \subseteq \Sigma^*$ given below.

$$L_X = \bigcup_{i=0}^n \text{support } \bar{P}^i([\mathbf{x}]_0)$$

$$L_Y = \bigcup_{\mathbf{x} \in L_X} \text{support } g_X(\mathbf{x})$$

Let $\langle g_X, g_Y, [\mathbf{x}]_0 \rangle$ be a communication channel.

A *communication preference* is a protocol preference over a communication channel.

Problem 5.19. Given:

1. a communication channel $\langle g, g_Y, [\mathbf{x}]_0, \Sigma \rangle$;
2. a communication preference $\langle r, D \rangle$ for which D is lower semi-continuous in the second argument;
3. a compact subspace $H \subseteq \mathfrak{M}(\mathcal{X}, \mathcal{Y})$ containing g such that any $h \in H$ satisfies $\bar{g}([\mathbf{x}]_0) \ll \bar{h}([\mathbf{x}]_0)$ and $\bar{r}([\mathbf{x}]_0) \ll \bar{h}([\mathbf{x}]_0)$;
4. a constant $\lambda \in \mathbb{R}$ such that $\lambda > 0$;
5. a function $L : H \rightarrow (-\infty, +\infty]$ defined below:

$$L(h) = D(\bar{g}([\mathbf{x}]_0) \parallel \bar{h}([\mathbf{x}]_0)) + \lambda D(\bar{r}([\mathbf{x}]_0) \parallel \bar{h}([\mathbf{x}]_0)) \quad (5.4)$$

in terms of fixed $D, g, [\mathbf{x}]_0, r$;

the communication post-training problem is the following optimization problem:

$$\min_{h \in H} L(h). \quad (5.5)$$

Such a problem admits a solution.

Theorem 4. *Given any instance of Problem 5.19, there is a minimizer $h_0 \in \arg \min_{h \in H} L(h)$.*

Proof. Let D, H, L be as stated. Then, H is compact and contains g .

Observe that lower semi-continuity of L follows from lower semi-continuity in the second argument of D , continuity of the continuation operator, and the fact that composition preserves lower semi-continuity.

The result follows from the extreme value theorem. □

A special case is presented in the next section below.

5.3.1 Adaptive communication

Recall the Example 2.7 of attacks on the assisted authentication protocol and consider the following refinement.

Example 5.20. Consider a refinement of Example 2.6 in which the assisted authentication protocol is executed over several trials.

Rather than assuming the trials in the above example are independent, it is assumed that the actors are able to adapt before the first trial and in between trials.

Strategy 5.21 (adaptive offense). An intelligent deceiver is *adaptive*, refraining from adaptation if optimal and otherwise adaptively maximizing its capacity to launch authentication attacks.

Prior to any interaction with Alice, Bob is given permanent access to all of Carol’s resources for the purpose of preparing in concert with Carol to build an internal model Eve of Alice and obey the following long-term strategy.

Strategy 5.22 (adaptive defense). An intelligent, defensive helper is *adaptive*, refraining from adaptation if optimal and otherwise adaptively minimizing the deceiver’s capacity to launch authentication attacks.

Although the offensive strategy of eliciting responses from Alice that maximize Carol’s confidence on the basis of a single conversation may be ideal in the short term of a single trial, the defensive strategy is chosen in consideration of a long-term strategy over repeated trials in which the confidence of evaluation on earlier trials is less valuable than that of later trials by which all of the actors have learned from previous trials and their interactions in other protocols between such trials.

This long-term view discounts evaluation of Alice’s *intention* to deceive in favor of minimizing Alice’s *education* in deception for the following reason.

Heuristic 5.23 (cybernetic teleology). The purpose of a system is not its specification.

“The purpose of a system is what it does.” [Beer, 2004]

In order to authenticate after the final trial, the authentication team aims to minimize Alice’s deceptive capability in the unknown final trial by minimizing Alice’s capacity to learn deception in each trial. In order to honor Bob’s existing performance in this protocol and others, Bob must adapt to optimize this aim while minimizing the amount of adaptation in the following sense.

In order to understand the feasibility for Bob to adapt, consider the following problem.

Problem 5.24. Given

1. a communication channel consisting of channels $g : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$, $g_{\mathcal{Y}} : \mathcal{Y}^* \rightarrow \mathfrak{M}\mathcal{X}^*$, an initial property $[\mathbf{x}]_0$ in $\mathfrak{M}\mathcal{X}^*$ and the derived property $[\mathbf{x}] \in \mathfrak{M}\mathcal{X}^*$;
2. a network cost $C : \Sigma^* \rightarrow [0, 1]$ that determines a reward-normalized channel $r : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$;
3. a compact subspace $H \ni g$ of the Hilbert space $\mathfrak{M}(\mathcal{X}, \mathcal{Y})$ such that any $h \in H$ satisfies $g(\mathbf{x}) \ll h(\mathbf{x})$ and $r(\mathbf{x}) \ll h(\mathbf{x})$ for all $\mathbf{x} \in \text{support } [\mathbf{x}]$;
4. lower semi-continuity in the second argument of a divergence $D : \mathfrak{M}\mathcal{Y}^* \times \mathfrak{M}\mathcal{Y}^* \rightarrow [0, +\infty]$;

5. a trace $\mathbf{x} \in \mathcal{X}^*$;
6. a real number $\lambda \in \mathbb{R}$ such that $\lambda \geq 0$;
7. a function $L : H \rightarrow (-\infty, +\infty]$ defined below:

$$L_{\mathbf{x}}(h; D, g, \lambda, r) = D(g(\mathbf{x}), h(\mathbf{x})) + \lambda \cdot D(r(\mathbf{x}), h(\mathbf{x})) \quad (5.6)$$

in terms of fixed D, g, λ, r :

the *online communication post-training problem* is the optimization problem stated below.

$$\min_{h \in H} L_{\mathbf{x}}(h) \quad (5.7)$$

Consider the following conjecture.

Theorem 5. *Any instance of Problem 5.24 admits a solution.*

Proof. Consider an arbitrary instance of Problem 5.24 with $H, D, \mathbf{x}, \lambda, L_{\mathbf{x}}$. Recall that H is a non-empty compact space. Lower semi-continuity of $L_{\mathbf{x}}$ follows from that of continuation $\bar{\bullet}$, D , composition, and linear combinations. The result follows from the extreme value theorem. \square

Observe that Problems 5.19 and 5.24 concerned the task of post-training the channel $g : \mathcal{X}^* \rightarrow \mathfrak{M}\mathcal{Y}^*$ of the first speaker but disregarded the task of post-training the channel $g_Y : \mathcal{Y}^* \rightarrow \mathfrak{M}\mathcal{X}^*$ of the second speaker.

CHAPTER 6 PROTOCOL

Consider a collection V of intelligent systems, a collection E of edges that signify one-hop directed paths, and the graph $G = E \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} V$ which assigns any edge $u \xrightarrow{e} v$ to its source node $u = s(e)$ and target node $v = t(e)$. Recall the notation in Tables 2 and 6.1.

Structure	Description
$G = E \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} V$	graph with source s and target t
$V_0 = \langle V, v_0 \rangle$	initial node $v_0 \in V$
$E_0 = \langle E, e_0 \rangle$	initial edge $v_0 \xrightarrow{e_0} v_1 \in E$
$\langle v_0, e_0 \rangle \in V \times E$	initialization of G
$G_0 = \langle V_0, E_0 \rangle$	initial graph G_0 over G

Table 6.1: Initialization over graph

An initial graph specifies an initial node $v_0 \in V$ and an initial edge $v_0 \xrightarrow{e_0} v_1$ in E .

A network $N = F \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} \mathcal{X}$ over G assigns each node $v \in V$ to a type $\mathcal{X}_v \in \mathcal{X}$ of events and each edge $u \xrightarrow{e} v$ in E to the Hilbert space F_e of measurable matrices from \mathcal{X}_u to \mathcal{X}_v , as shown in Table 6.2.

Structure	Description
$G_0 = \langle V_0, E_0 \rangle$	initial graph G_0 over $G = E \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} V$
\mathcal{X}_v	type of events at node $v \in V$
$\Sigma = \bigsqcup_{v \in V} \mathcal{X}_v$	type of network events
F_e	measurable matrices from $\mathcal{X}_{s(e)}$ to $\mathcal{X}_{t(e)}$
$[\mathbf{x}_{v_0}] \in \mathfrak{M}\mathcal{X}_{v_0}^*$	initial source on \mathcal{X}_{v_0}
$N_0 = \langle G_0, \{F_e\}_{e \in E}, [\mathbf{x}_{v_0}] \rangle$	initial network over G_0

Table 6.2: Initial network over initial graph

Each node $\mathcal{X}_v \in \mathcal{X}$ gives rise to a Hilbert space $\mathfrak{M}\mathcal{X}_v^*$ of (finite, complex) measures on the measurable space given by the Borel σ -algebra generated by the upper topology on the free monoid \mathcal{X}_v^* of traces of events in \mathcal{X}_v , and each edge $u \xrightarrow{e} v$ gives rise to a Hilbert space $F_e = \mathfrak{M}(\mathcal{X}_u, \mathcal{X}_v)$ of measurable channels from \mathcal{X}_u to \mathcal{X}_v . An initial network specifies a source $[\mathbf{x}_{v_0}] \in \mathfrak{M}\mathcal{X}_{v_0}^*$. While any event at node v is a term in \mathcal{X}_v , an event in the network is a term in the type $\Sigma = \bigsqcup_{v \in V} \mathcal{X}_v$.

A protocol is specified by a vector \vec{f} in $F = \prod_{e \in E} F_e$ as shown in Table 6.3.

Structure	Description
$N_0 = \langle G_0, \{F_e\}_{e \in E}, [\mathbf{x}_{v_0}] \rangle$	initial network over G_0
$F = \prod_{e \in E} F_e$	product F equipped with the product topology
$\vec{f} \in F$	vector of measurable matrices
$\langle N_0, \vec{f} \rangle$	initial protocol over N_0
$X_e = \mathfrak{M}(\mathcal{X}_{s(e)}, \mathcal{X}_{t(e)})$	measurable channels from $\mathcal{X}_{s(e)}$ to $\mathcal{X}_{t(e)}$
$\vec{g} = \vec{f}^*$	vector $\vec{g} = [g_e]_{e \in E}$ of measurable channels given by $g_e = f_e^*$ for each $e \in E$
$P : \Sigma^* \rightarrow \mathfrak{M}\Sigma^*$	cumulative protocol channel
$[\mathbf{x}_v] \in \mathfrak{M}\mathcal{X}_v$	prior source on node \mathcal{X}_v

Table 6.3: Initial protocol over initial network

It gives rise to a vector \vec{g} in the product $X = \prod_{e \in E} X_e$ of Hilbert spaces, a protocol channel $P : \Sigma^* \rightarrow \mathfrak{M}\Sigma^*$, a source $[\mathbf{x}_v] \in \mathfrak{M}\mathcal{X}_v^*$ at each node $v \in V$, and more structure shown in Table 6.3.

The vector \vec{g} can be viewed as an initial configuration of the protocol. An instance of protocol post-training consists of an objective function \mathcal{L} on a topological space \mathcal{H} and the task of computing a representation $\hat{h} \in \mathcal{H}$ whose behavior approximates that of an objective minimum $\vec{h}^* \in \arg \min_{\vec{h}} \mathcal{L}(\vec{h})$ in \mathcal{H} . It may be theoretically possible in contrived settings to directly compute \vec{h} by means of a closed-form solution. The methods of post-training arising in practice tend to enforce $\vec{h}_0 = \vec{g}$ is in \mathcal{H} and compute a sequence of iterations $\vec{h}_0 \rightarrow \vec{h}_1 \rightarrow \vec{h}_2 \rightarrow \dots \rightarrow \vec{h}_n \rightarrow \dots$ until some condition is met, often by way of gradient-based algorithms that update iteratively update the parameter vector of a parametrized machine learning model. There may not be an optimal configuration if the space \mathcal{H} is topologically ill-behaved.

The notion of a well-behaved space is formalized below.

Definition 6.1 (protocol state space). Given

1. an initial network N_0 over a graph $E \xrightleftharpoons[t]{s} V$ that determines a Hilbert space $H_e = \mathfrak{M}(\mathcal{X}_{s(e)}, \mathcal{X}_{t(e)})$ of measurable channels for each edge $e \in E$;
2. the topological space $X = \prod_{e \in E} H_e$ equipped with the product topology;
3. an initial protocol $\langle N_0, \vec{f} \rangle$ that determines a vector $\vec{g} = \vec{f}^*$ in X ;

a *protocol state space over \vec{g}* is a pointed topological space $\langle \mathcal{H}, \vec{g} \rangle$ satisfying the following conditions:

1. \mathcal{H} is a topological space equipped with the subspace topology inherited from X ;
2. \mathcal{H} contains \vec{g} ; and
3. \mathcal{H} is compact.

Consider a protocol state space $\langle \mathcal{H}, \vec{g} \rangle$ over \vec{g} and an instance of the following objective.

Objective 6.2 (macroscopic channel post-training objective). A *macroscopic channel post-training objective* defined in terms of

1. an initial network N_0 over a graph $G = E \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} V$;
2. an edge $u \xrightarrow{b} v$ in a subgraph $G_A = A \begin{smallmatrix} \xrightarrow{s} \\ \xrightarrow{t} \end{smallmatrix} V_A$ of G ;
3. an initial protocol $\langle N_0, [f_e]_{e \in E} \rangle$ that determines a vector $\vec{g} = [f_e^*]_{e \in E}$ and a node source $[\mathbf{x}_u] \in \mathfrak{M}\mathcal{X}_u^*$;
4. a subnetwork N_A over G_A ;
5. a subnetwork reward R_A on N_A that determines a reward-normalized channel

$$r_b : \mathfrak{M}\mathcal{X}_u^* \rightarrow \mathfrak{M}\mathcal{X}_v^*$$

which weighs g_b by R_A ;

6. a channel preference for b in N_A satisfying lower semi-continuity in the second argument of a divergence $D : \mathfrak{M}\mathcal{X}_v^* \times \mathfrak{M}\mathcal{X}_v^* \rightarrow [0, +\infty]$; and
7. a constant $\lambda \in \mathbb{R}$ satisfying $\lambda \geq 0$;

on a protocol state space $\langle \mathcal{H}, \vec{g} \rangle$ is a function $L_b : \mathcal{H} \rightarrow (-\infty, +\infty]$ given by the assignment from any vector $\vec{h} = [h_e]_{e \in E}$ in \mathcal{H} to the following value.

$$L_b(\vec{h}) = D(\vec{g}_b([\mathbf{x}]_u), \vec{h}_b([\mathbf{x}]_u)) + \lambda \cdot D(\vec{r}_b([\mathbf{x}]_u), \vec{h}_b([\mathbf{x}]_u)) \quad (6.1)$$

Observe that the objective is well-defined.

Proposition 6.3 (macroscopic channel post-training objective). For any instance of Objective 6.2, the following hold:

1. the function $L_b : \mathcal{H} \rightarrow (-\infty, +\infty]$ defined by Equation (6.1) is well-defined;
2. if $\inf_{\vec{h} \in \mathcal{H}} L_b(\vec{h}) = +\infty$, then $\arg \min_{\vec{h} \in \mathcal{H}} L_b(\vec{h}) = \mathcal{H}$; and
3. the function L_b is lower semi-continuous on \mathcal{H} .

Proof. Let $u \xrightarrow{b} v$, \vec{g} , r_b , D , \mathcal{H} , λ , L_b be as stated, and let $G_b = \mathfrak{M}(\mathcal{X}_u, \mathcal{X}_v)$.

1. Define the projection $\pi_b : \mathcal{H} \rightarrow G_b$ by $\left[h_e \right]_{e \in E} \xrightarrow{\pi_b} h_b$, and recall that π_b is continuous. Let $H = \pi_b(\mathcal{H})$. Recall from Proposition 4.20 that $\bar{\bullet}$ is well-defined on $H \subseteq G_b$. Recall that D is well-defined on $\mathfrak{M}\mathcal{X}_v^* \times \mathfrak{M}\mathcal{X}_v^*$ and that $D(\mu, \bullet)$ is well-defined on any subset of $\mathfrak{M}\mathcal{X}_v^*$ for any $\mu \in \mathfrak{M}\mathcal{X}_v^*$. Then, observe that $D(\bar{g}_b([\mathbf{x}]_u), \bar{\bullet}([\mathbf{x}]_u))$ and $\lambda \cdot D(\bar{r}_b([\mathbf{x}]_u), \bar{\bullet}([\mathbf{x}]_u))$ are well-defined on H . The result follows from the fact that addition is well-defined.
2. If $l = \inf_{\vec{h} \in \mathcal{H}} L_b(\vec{h})$ in $(-\infty, +\infty]$ is $+\infty$ and $\vec{h} \in \mathcal{H}$, then the inequalities $l \leq L_b(\vec{h}) \leq +\infty$ imply that $L_b(\vec{h}) = +\infty = l$ and thus that $\vec{h} \in \arg \min_{\vec{h} \in \mathcal{H}} L_b(\vec{h})$.
3. Observe the lower semi-continuity of $\bar{\bullet}$, scalar multiplication, addition, and the second argument of D . The result follows from the fact that composition preserves lower semi-continuity.

□

The task of post-training is to minimize the objective.

Problem 6.4 (macroscopic channel post-training). Given

1. an initial protocol that determines a vector \vec{g} of measurable channels;
2. protocol state space $\langle \mathcal{H}, \vec{g} \rangle$;
3. a macroscopic channel post-training objective $L_b : \mathcal{H} \rightarrow (-\infty, +\infty]$ for an edge b ;

the *macroscopic channel post-training problem* is the optimization problem stated below.

$$\min_{\vec{h} \in \mathcal{H}} L_b(\vec{h}). \quad (6.2)$$

This problem is shown to be soluble.

Corollary 6.5. Any instance of Problem 6.4 admits a solution.

Proof. The result follows as a corollary of Theorem 6, and a direct proof is provided below.

Let $L_b : \mathcal{H} \rightarrow (-\infty, +\infty]$ be as stated. Define $l \in [-\infty, +\infty]$ by $l = \inf_{\vec{h} \in \mathcal{H}} L_b(\vec{h})$.

Recall Lemma 6. The result follows immediately if $l = +\infty$, so suppose $l < +\infty$. Non-negativity of D implies l is contained in $[0, +\infty)$. Recall the compactness of \mathcal{H} and the lower semi-continuity of L_b . Then, the result follows from the extreme value theorem. □

Observe that the above objective was restricted to an individual channel b and that this restriction is lifted in the following objective.

Objective 6.6 (local protocol post-training objective). A *local protocol post-training objective* defined in terms of

1. an initial network N_0 over a graph $G = E \begin{smallmatrix} \xrightarrow{s} \\ \xleftarrow{t} \end{smallmatrix} V$
2. an initial protocol $\langle N_0, \vec{f} \rangle$ that determines a vector \vec{g} of measurable channels and node sources $\{\mathcal{X}_u\}_{u \in V}$;
3. edges $B_A \subseteq A \subseteq E$ in a subnetwork N_B of a subnetwork N_A of N_0 ;
4. a subnetwork reward R_A on N_A that determines reward-normalized channels

$$\left\{ r_b : \mathcal{X}_{s(b)}^* \rightarrow \mathfrak{M}\mathcal{X}_{t(b)}^* \right\}_{b \in B_A}$$

5. a local network preference for B in N_A satisfying lower semi-continuity in the second argument of divergences $\left\{ D_b : \mathfrak{M}\mathcal{X}_{t(b)}^* \times \mathfrak{M}\mathcal{X}_{t(b)}^* \rightarrow [0, +\infty] \right\}_{b \in B_A}$;
6. a vector $\lambda_{B_A} \in \mathbb{R}^{|B_A| \cdot 2}$ in the non-negative orthant;

on a protocol state space $\langle \mathcal{H}, \vec{g} \rangle$ is a function $L_{B_A} : \mathcal{H} \rightarrow (-\infty, +\infty]$ given by the assignment from any vector $\vec{h} = [h_e]_{e \in H}$ in \mathcal{H} to the following value.

$$L_{B_A}(\vec{h}) = \sum_{u \xrightarrow{b} v \in B_A} \lambda_{b,0} \cdot D_b(\bar{g}_b([\mathbf{x}_u]), \bar{h}_b([\mathbf{x}_u])) + \lambda_{b,1} \cdot D_b(\bar{r}_b([\mathbf{x}_u]), \bar{h}_b([\mathbf{x}_u])) \quad (6.3)$$

The local objective L_\emptyset corresponding to the empty set $B_A = \emptyset$ is defined by $L_\emptyset(\vec{h}) = 0$. It is clearly well-defined and lower semi-continuous.

Proposition 6.7 (local post-training objective). For any instance $L_{B_A} : \mathcal{H} \rightarrow (-\infty, +\infty]$ of Objective 6.6, the following hold:

1. the function L_{B_A} defined by Equation 6.3 is well-defined;
2. if $\inf_{\vec{h} \in \mathcal{H}} L_{B_A}(\vec{h}) = +\infty$, then $\arg \min_{\vec{h} \in \mathcal{H}} L_{B_A}(\vec{h}) = \mathcal{H}$; and
3. the function L_{B_A} is lower semi-continuous on \mathcal{H} .

Proof. Let L_{B_A} be as stated. Observe that L_{B_A} is a linear combination of terms which have previously been shown to be well-defined and continuous on \mathcal{H} . The same argument can be reused to show the second item. \square

6.1 Protocol Post-Training

The general problem of protocol post-training is the problem of optimizing the following objective.

Objective 6.8 (protocol post-training). A *protocol post-training objective* consists of the following:

1. an initial network N_0 with a network initialization on a network over a graph $G = E \xrightarrow[s]{t} V$;
2. an initial protocol on N_0 given by a vector $\vec{f} = \left[f_e \right]_{e \in E}$
3. a protocol state space $\langle \mathcal{H}, \vec{g} \rangle$ over the vector $\vec{g} = \left[f_e^* \right]_{e \in E}$ of measurable channels; and
4. a function $\mathcal{L} : \mathcal{H} \rightarrow (-\infty, +\infty]$ defined in terms of
 - a choice $B : \mathfrak{P}E \rightarrow \mathfrak{P}E$ such that $B_A := B(A)$ satisfies $B_A \subseteq A$ for each $A \in \mathfrak{P}E$;
 - a vector $\vec{\lambda} = \left[\lambda_A \right]_{A \in \mathfrak{P}E}$ in the non-negative orthant of $\mathbb{R}^{2^{|E|}}$;
 - a vector $\vec{L} = \left[L_{B_A} \right]_{A \in \mathfrak{P}E}$ in which the entry at index $A \in \mathfrak{P}E$ is given by a local protocol post-training objective $L_{B_A} : \mathcal{H} \rightarrow (-\infty, +\infty]$ in terms of a subnetwork value on a subnetwork over a subgraph $G_A = A \xrightarrow[s]{t} V_A$ of G and a local network preference over a subgraph $B_A \xrightarrow[s]{t} V_{B_A}$ of G_A

by the assignment from any vector $\vec{h} = \left[h_e \right]_{e \in E}$ in \mathcal{H} to the value given below.

$$\mathcal{L}(\vec{h}) = \sum_{A \in \mathfrak{P}E} \lambda_A \cdot L_{B_A}(\vec{h}) \quad (6.4)$$

Consider an objective $\mathcal{L} : \mathcal{H} \rightarrow (-\infty, +\infty]$ as defined above, and observe that it is a linear combination consisting of one local post-training objective $L_{B_A} : \mathcal{H} \rightarrow (-\infty, +\infty]$ for each subset $A \in \mathfrak{P}E$ together with a penalty coefficient λ_A . For each subset A of edges, the objective L_A represents a shared cost that the nodes in $\left\{ u \in V \mid \exists u \xrightarrow{e} v \in E \right\}$ cooperate to minimize. Despite the fact that each coalition experiences the shared cost, in the sense that they seek to minimize it, the optimization of L_A induces trade-offs within each coalition.

It is worth noting that this objective is well-defined.

Lemma 6 (protocol post-training objective). *For any instance $\mathcal{L} : \mathcal{H} \rightarrow (-\infty, +\infty]$ of Objective 6.8, the following hold:*

1. the function \mathcal{L} defined by Equation (6.4) is well-defined;
2. if $\inf_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h}) = +\infty$, then $\arg \min_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h}) = \mathcal{H}$; and
3. the function \mathcal{L} is lower semi-continuous on \mathcal{H} .

Proof. Let $\mathcal{L} : \mathcal{H} \rightarrow (-\infty, +\infty]$ be as stated and E be the edges in the underlying graph.

Then, \mathcal{L} is a linear combination consisting of one objective per edge in E together with a penalty. Thus, it is well-defined and lower semi-continuous on \mathcal{H} . The second item follows as it did for each local post-training objective. \square

The optimization of \mathcal{L} forces competition among the collection of all coalitions in $\mathfrak{P}E$.

6.1.1 Optimization Problem of Protocol Post-Training

Given a graph $G = \langle V, E \rangle$ and initial protocol specified by \vec{g} , consider a protocol state space $\langle \mathcal{H}, \vec{g} \rangle$ and protocol post-training objective $\mathcal{L}(\bullet; \vec{\lambda}, \vec{L}) : \mathcal{H} \rightarrow (-\infty, +\infty]$. This determines an instance of the following problem.

Problem 6.9 (protocol post-training). A *protocol post-training problem* consists of a *protocol post-training objective* function $\mathcal{L} : \mathcal{H} \rightarrow (-\infty, +\infty]$ defined in Objective 6.8 by Equation (6.4) and the optimization problem stated below.

$$\min_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L}) \quad (6.5)$$

Solutions to the above problem are characterized below.

Definition 6.10 (minimizer). Given any instance $\mathcal{L}(\bullet; \vec{\lambda}, \vec{L}) : \mathcal{H} \rightarrow (-\infty, +\infty]$ of Problem 6.9, a vector $\vec{h}_0 = [h_{0e}]_{e \in E}$ in \mathcal{H} is said to be a *minimizer* of \mathcal{L} iff $\mathcal{L}(\vec{h}_0) = \min_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h})$.

Minimizers are also called solutions.

Solutions may not be unique.

Definition 6.11 (minimizing set). Given any instance of Problem 6.9, the *minimizing set* of \mathcal{L} is given below.

$$\mathcal{H}_{\mathcal{L}} = \arg \min_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L})$$

It is defined to contain $\vec{h}^* = [h_e^*]_{e \in E}$ iff $\mathcal{L}(\vec{h}^*; \vec{\lambda}, \vec{L}) \leq \inf_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L})$.

An instance of Problem 6.9 defined by \mathcal{L} is called *insoluble* iff $\mathcal{H}_{\mathcal{L}} = \emptyset$ and *soluble* iff $\mathcal{H}_{\mathcal{L}} \neq \emptyset$.

Each node $u \in V$ experiences a superposition of cooperative and competitive forces specified by a vector \vec{L} . Any specification of relevant trade-offs, in the form of a penalty vector $\vec{\lambda}$, can be viewed as the specification of a game. The main contribution of this article is to show that any such game has an equilibrium in the sense that there is no direction of motion that improves the total loss \mathcal{L} .

Minimizers for Protocol Post-Training

For an optimization problem such as Problem 6.9 to be well-posed, it is typically required that solutions satisfy conditions of existence, uniqueness, and stability. The solutions are characterized below.

Lemma 7. *For any instance of Problem 6.9 with objective $\mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L}) : \mathcal{H} \rightarrow (-\infty, +\infty]$, the minimizers are given below.*

$$\arg \min_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L}) = \mathcal{L}^{-1} \left(\min_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L}) \right)$$

Proof. Let $\mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L}) : \mathcal{H} \rightarrow (-\infty, +\infty]$ be as stated and $l = \min_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L})$.

If $\vec{h}_0 \in \arg \min_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L})$, then $\mathcal{L}(\vec{h}_0; \vec{\lambda}, \vec{L}) = l$ and $\vec{h}_0 \in \mathcal{L}^{-1}(l)$. If $\vec{h}_0 \in \mathcal{L}^{-1}(l)$, then $\mathcal{L}(\vec{h}_0) = l$ and $\vec{h}_0 \in \arg \min_{\vec{h}_0 \in \mathcal{H}} \mathcal{L}(\vec{h}; \vec{\lambda}, \vec{L})$. □

The existence of global minimizers guarantees there is an equilibrium that is ideal, and this is conjectured to hold.

Conjecture 1 (solubility of protocol post-training). *Given any instance of Problem 6.9, the set of minimizers is not empty.*

Existence of Minimizers for Protocol Post-Training The above conjecture is proven below.

Proof (of Conjecture 1 on the existence of minimizers for the protocol post-training problem).

Consider any instance of Problem 6.9 defined by a protocol post-training objective.

Recall that any instance of Objective 6.8 consists of an initial network N_0 , an initial protocol $\langle N_0, \vec{f} \rangle$, a protocol state space \mathcal{H} over $\vec{g} = \vec{f}^*$, and a function $\mathcal{L}(\bullet; \vec{\lambda}, \vec{L}) : \mathcal{H} \rightarrow (-\infty, +\infty]$. By Definition 6.1 of a protocol state space, the subspace topology on \mathcal{H} is compact and the vector \vec{g} is contained in \mathcal{H} .

Recall from Lemma 6 that \mathcal{L} is well-defined, non-negative, and lower semi-continuous on \mathcal{H} . It follows from the containment $\vec{g} \in \mathcal{H}$ that \mathcal{H} and $\mathcal{L}(\mathcal{H})$ are not empty.

Consider the subspace topology on $\mathcal{L}(\mathcal{H})$ inherited from $(-\infty, +\infty]$. Compactness of $\mathcal{L}(\mathcal{H})$ follows from the extreme value theorem. This compact subspace of $[0, +\infty]$ is closed, bounded below by 0, and contains the minimum $l = \min_{\vec{h} \in \mathcal{H}} \mathcal{L}(\vec{h})$. Recall from Lemma 7 that $\mathcal{H}_{\mathcal{L}} = \mathcal{L}^{-1}(l)$. It follows from the containment $l \in \mathcal{L}(\mathcal{H})$ that $\mathcal{H}_{\mathcal{L}}$ is not empty. □

The above proof of Conjecture 1 establishes the following result.

Theorem 6. *Any instance of Problem 6.9 admits a non-empty of solutions.*

Therefore, the target outcomes of protocol post-training exist.

To refute the claim that an arbitrary instance is well-posed, consider a pathological instance in which all divergences are constant; then, every vector in the domain is a minimizer and uniqueness does not hold.

Further investigation into the mathematical theory of protocol post-training is needed to characterize the stability of Problem 6.9 and the circumstances under which the optimization problem is well-posed. Additional directions for future research are discussed in the next chapter.

CHAPTER 7

DISCUSSION

The motivations and results are revisited in this chapter.

The deployment of AI introduces new dimensions to the problems of privacy, security, and trust. They center around the need established in Chapters 1–2 to constrain the behaviors of AI actors. Question 1.2 concerned the structure of protocol post-training, and Question 2.8 concerned the conditions that guarantee existence of solutions. These questions have not been addressed at this level of generality in the existing literature [Achiam et al., 2017, Gu et al., 2021, 2023], which tends to focus on refinements for particular reinforcement learning algorithms.

The main contribution of this thesis is a proposed model of protocol post-training, extending [Pavlovic and Seidel, 2025], that answers Questions 1.2 & 2.8. The proposed model serves as a general setting for reasoning about the opportunities and limitations of protocol post-training in a manner that is independent of algorithms and computational paradigms. A hierarchy of structure was developed to establish the results in the table below.

Theorem	Existence of solutions to the problems of post-training a ...
1	measurable channel
2	probabilistic channel
3	global protocol channel
4 and 5	bidirectional communication channel
6	initial protocol

Table 7.1: Main results

They support the implications in Section 7.1 and motivate many directions of further research.

In analogy to the geometry of trace properties, the notion of an uncertain property was introduced in Chapter 3 along with a brief investigation of the geometry of uncertain properties and the geometry of (finite, complex) measurable properties. The development of these theories remains a matter for future work.

The first problems considered the tasks of post-training (finite, complex) measurable channels in Chapter 4 and, as a special case, probabilistic channels in Section 4.3. In addition to results that address the technical requirements of these problems, the main results characterized the existence of solutions. The key limitation of the results is that they address only the existence of solutions. The related characteristics of uniqueness or stability were not explored in detail, and they remain open for future work. In addition to such features of well-posedness, they also invite future work on the construction of global minimizers and the approximation of local minimizers. For example, consider the following question: if the lower semi-continuous divergences are assumed to be suitably differentiable, would iterative applications of gradient descent converge to a local minimizer? This

further motivates the development of methods to predict the limits of preference satisfaction and the minimum achievable error $\varepsilon_\lambda > 0$ for a given penalty $\lambda > 0$.

It also raises the need for the fusion of preferences that generally conflict. In non-trivial settings, the optimization is infeasible if there is a requirement that a particular trace is sampled with probability 1, or if one requires probability $p > 0.5$ for a particular trace while another constrains $p < 0.5$ for the same trace. The theory of post-training requires additional tools for merging conflicting preferences.

7.1 Outlook

The ideas presented in Chapters 5–6 support the view of protocol post-training as a game in which the players form intersecting coalitions with competing objectives. The result in Section 6.1 establishes that this game admits an ideal equilibrium. Future work in AI protocol security may draw on ideas from computer science, mathematics, philosophy and economics.

The practical implication for the analysis of protocols involving potentially deceitful AI actors is that the strength of an actor can be measured by the extent to which they contribute to the production of undesirable traces. Then, a protocol can be designed in iterations from the starting point of a general-purpose language model by iteratively post-training it to minimize the cost incurred by its contribution to undesirable traces.

Conclusion

The results support the intuition that protocol post-training is feasible and encourage the iterative development of protocols that mediate interactions among networks of intelligent systems.

BIBLIOGRAPHY

- Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.
- Rav Ashi and Ravina II, editors. *Talmud*. Sefaria, 500. Sanhedrin 65b.
- Stafford Beer. What is cybernetics? *Kybernetes: The International Journal of Systems & Cybernetics*, 33, 03 2004. doi: 10.1108/03684920410523742.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Michael R. Clarkson and Fred B. Schneider. Hyperproperties. *Journal of Computer Security*, 18 (6):1157–1210, 2010. doi: 10.3233/JCS-2009-0393. URL <https://journals.sagepub.com/doi/abs/10.3233/JCS-2009-0393>.
- Brian Albert Davey and Hilary Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2 edition, 2002. URL <https://www.cambridge.org/core/books/introduction-to-lattices-and-order/946458CB6638AF86D85BA00F5787F4F4>.
- Paul Feyerabend. *Against Method: Outline of an anarchistic theory of knowledge*. Verso Books, 1975.
- Shangding Gu, Jakub Grudzien Kuba, Muning Wen, Ruiqing Chen, Ziyang Wang, Zheng Tian, Jun Wang, Alois Knoll, and Yaodong Yang. Multi-agent constrained policy optimisation. *arXiv preprint arXiv:2110.02793*, 2021.
- Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319:103905, 2023. ISSN 0004-3702. doi: <https://doi.org/10.1016/j.artint.2023.103905>. URL <https://www.sciencedirect.com/science/article/pii/S0004370223000516>.
- Paul Richard Halmos. *Naive Set Theory*. D. Van Nostrand Company, 1960.
- Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- Joachim Lambek and Philip Scott. *Introduction to Higher-Order Categorical Logic*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1986. ISBN 9780521356534. URL <https://www.cambridge.org/us/universitypress/subjects/mathematics/>

- logic-categories-and-sets/introduction-higher-order-categorical-logic?format=PB&isbn=9780521356534.
- Saunders Mac Lane. *Categories for the working mathematician*, volume 5. Springer Science & Business Media, 1998.
- Musashi Miyamoto. *The Book of Five Rings*. Kodansha International, 1645.
- James Raymond Munkres. *Topology*. Featured Titles for Topology. Pearson, 2 edition, 2000. ISBN 9780131816299. URL <https://www.pearson.com/en-us/subject-catalog/p/topology-classic-version/P200000006299/9780137848669>.
- Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022. URL <http://probml.github.io/book1>.
- Julien de La Mettrie Offray. *L’homme machine*. Elie Luzac Fils, 1747.
- Dusko Pavlovic. Language processing in humans and computers, 2024. URL <https://arxiv.org/abs/2405.14233>.
- Dusko Pavlovic and Peter-Michael Seidel. Security Science: Basic concepts and mathematical foundations. Preprint, 2025.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL <https://www.science.org/doi/abs/10.1126/science.aar6404>.
- Gilbert Strang. *Linear Algebra and its Applications*. Thomson Brooks/Cole, 4 edition, 2006.
- Alan Mathison Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. ISSN 00264423, 14602113. URL <http://www.jstor.org/stable/2251299>.
- Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Linkgua, 1921.

Index

- channel, 9
 - probabilistic channel, 25
 - cumulative probabilistic channel, 25
- clearance, 18
- closure
 - closure operator, 19
 - extensivity,monotonicity,idempotency, 19
 - upper closure, 22
- communication channel, 62
- continuation, 46
- continuity, 23
 - probability, 25
 - continuation, 25
 - continuous probabilistic channel, 25
- divergence, 29
- evaluator, 35
- event, 24
 - event string, 24
 - type of events, 24
- graph, 9
 - edge, node, source, target, 9
 - initial graph, 57
- intelligence, 8
 - intelligent machine, 8
 - intelligent system, 9
 - AI system, 9
- locality, 18
- machine, 5
 - computing machine, 7
 - artificial computing machine, 8
- matrix
 - measurable matrix, 40
 - protocol matrix, 59
 - stochastic matrix, 24
- morpheme, 16
- multi-level resource, 18
- multi-level security model, 18
- network
 - initial network, 57
 - network value, 52
- post-training, 5
 - protocol post-training, 5
- power set, 18
- preference
 - communication preference, 63
 - global protocol preference, 61
- product
 - probabilistic product, 29
 - probability, 30
- property, 24
 - complex, 36
 - measurable property, 36
 - measurable trace property, 36
- resource, 17
- restriction, 46
- security level, 18
- theorem
 - existence of solutions to post-training
 - protocol, 73
- token, 16
- tokenizer, 16
- trace, 24
- uncertain property, 27
- vocabulary, 16