EPFL

Thèse n° 9905

Opinion Formation over Adaptive Networks

Présentée le 14 novembre 2022

Faculté des sciences et techniques de l'ingénieur
Laboratoire de Systèmes Adaptatifs
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Virginia BORDIGNON

Acceptée sur proposition du jury

Prof. P. Frossard, président du jury
Prof. A. H. Sayed, directeur de thèse
Prof. H. Bölcskei, rapporteur
Prof. S. Marano, rapporteur
Prof. N. Kiyavash, rapporteuse

École
polytechnique
fédérale
de Lausanne

2022

Our opinions do not really blossom into fruition
until we have expressed them to someone else.
— Mark Twain

To my family…

Acknowledgements

First of all, I would like to thank Professor Ali H. Sayed, for giving me this life-changing opportunity. I thank him for sharing with me interesting and challenging research problems, and for his guidance while I developed tools and competencies to solve them. His work ethics, and attention to detail are some of the lessons I intend to always carry with me.

I must also express my gratitude to other professors with whom I have had the pleasure of collaborating and who played an important role in my growth as a researcher. I thank Professor Vincenzo Matta for his teachings and for being patient, especially at the beginning of my studies. I thank my former post-doctoral colleague in the lab now Professor Roula Nassif for her valuable professional advice as I was still learning to navigate my PhD years. I thank my former lab colleague now Professor Stefan Vlaski for his constant willingness to discuss research and to brainstorm ideas.

I would also like to thank Professor Pascal Frossard, Professor Negar Kiyavash, Professor Helmut Bölcskei, and Professor Stefano Marano for their valuable feedback and participation on the thesis jury.

I have had the pleasure of collaborating and interacting with excellent colleagues and friends from the Adaptive Systems Laboratory at EPFL, to whom I am grateful: Augusto Santos, Professor Ricardo Merched, Lucas Cassano, Elsa Rizk, Mert Kayaalp, Konstantinos Ntemos, Ainur Zhaikhan, Valentina Shumovskaia, Ping Hu, Ying Cao, and Flávio Pavan. The daily lab discussions, lunch times, and coffee breaks constitute important building elements of this thesis work. I am also grateful to Patricia Vonlanthen for her help when dealing with administrative tasks.

Being an expat is not always easy, hence the importance of having friends that support you in difficult times. I would like to thank Roula, Elsa, and their respective families for welcoming me into their homes and showing me their beautiful country. I thank Karen, Ahmad, Jacques, and Marie-Line for all the fun moments we shared in Lausanne. I am also grateful to all friends spread across the world, who have been present during the past years. I thank my partner Roman for his continuous love, patience, and support. I also thank Claudia, Jean-Pierre, and Linda for treating me as if we were family.

I do not have enough words to thank my parents, Assis and Inete, who did their best to give me every opportunity they did not have themselves. I thank my sister, Cláudia, for being a reliable role model ever since my childhood. Distance may separate us physically, but has never kept us

from supporting and loving each other unconditionally.

This dissertation is based upon work partially supported by the Swiss National Science Foundation (SNSF) under Grant 205121-184999.

Lausanne, September 2022 V. B.

Abstract

An *adaptive network* consists of multiple communicating agents, equipped with sensing and learning abilities that allow them to extract meaningful information from measurements. The objective of the network is to solve a *global* inference problem in a decentralized manner, i.e., by exchanging only *local* information with neighboring agents.

Such adaptive networks find inspiration in real-world networks, e.g., power networks, biological networks, and social networks. Decentralized solutions allow the network to outperform stand-alone strategies by yielding improved performance and robustness. They also enable agents to overcome their individual limitations by leveraging collaboration during the learning process.

Several of these solutions draw on *social learning* paradigms, through which individuals form opinions (or *beliefs*) by observing the world and communicating within their social group. The world is explained by a discrete-valued state, and agents discover the unknown *state of the world* while updating their beliefs regarding a set of plausible hypotheses. Many such solutions result in consistent truth learning at fast convergence rates. Existing works however fail to account for more realistic assumptions such as the exchange of incomplete information, adaptation under nonstationary conditions, and the use of imperfect private statistical models.

This thesis aims to address the aforementioned problems and answer questions regarding the behavior and performance of social learning strategies under more realistic conditions. This is carried out by exploiting four key elements to learning over adaptive networks, namely, **i)** *network* topology, **ii)** *exchanged information*, **iii)** surrounding *world*, and **iv)** private *models*, which we divide in two parts.

In the first part, we focus on the *stationary* setting, i.e., where world conditions are static. **i)** The social network is represented by a weakly connected graph, which results in a power asymmetry among network clusters. To estimate the level of influence from influential clusters toward specific agents, we formulate the *reverse learning* problem. We characterize the feasibility of this problem and show that a certain statistical diversity among components is sufficient for it to be feasible. **ii)** We consider a strongly connected social network, with constrained communication, i.e., where only *partial beliefs* are shared with neighbors. We show how different learning regimes arise and under which conditions the agents can learn the truth or, on the other hand, be misled.

In the second part, we address the *nonstationary* setting. **iii)** Existing social learning strategies are limited in their ability to *adapt* under changing world conditions. We propose an adaptive

social learning formulation and characterize its performance both in the steady-state and the transient phases. We show that the approach enables a trade-off between learning accuracy and adaptation capability. **iv)** Social learning agents use statistical models that are assumed to be perfectly known a priori. We propose a social *machine learning* framework, where the models are first trained from a finite set of labeled samples and then deployed in a collaborative implementation to classify streaming unlabeled (possibly nonstationary) observations. We show that the proposed fully data-based strategy results in consistent learning, despite the imprecise models, and in improved accuracy as the number of unlabeled observations grows.

Keywords: Social learning, Bayesian update, diffusion strategies, influence recovery, partial information sharing, adaptive network, distributed classification.

Résumé

Un *réseau adaptatif* se compose de plusieurs agents communicants, dotés de capacités de détection et d'apprentissage qui leur permettent d'extraire des informations utiles à partir des mesures. L'objectif du réseau est de résoudre un problème d'inférence *global* de manière décentralisée, i.e., en n'échangeant que des informations *locales* avec des agents voisins.

Ces réseaux adaptatifs s'inspirent de réseaux réels tels que les réseaux électriques, les réseaux biologiques et les réseaux sociaux. Les solutions décentralisées permettent au réseau de surpasser les stratégies non coopératives en performance et en robustesse. Elles permettent également aux agents de surmonter leurs limites individuelles en tirant parti de la collaboration au cours du processus d'apprentissage.

Plusieurs de ces solutions s'appuient sur des paradigmes d'apprentissage social (ou *social learning*) à travers lesquels les individus forment des opinions (ou *beliefs*) en observant le monde et en communiquant au sein de leur groupe social. Le monde est expliqué par un état à valeurs discrètes, et les agents découvrent l'*état du monde* en mettant à jour leurs opinions concernant un ensemble d'hypothèses plausibles. Plusieurs de ces solutions aboutissent à un apprentissage cohérent de la vérité à des taux de convergence rapides. Les travaux existants ne tiennent cependant pas compte d'hypothèses plus réalistes tels que l'échange d'informations incomplètes, l'adaptation dans des conditions non stationnaires et l'utilisation de modèles statistiques privés imparfaits.

Cette thèse vise à aborder les problèmes susmentionnés et à répondre à des questions concernant le comportement et la performance des stratégies d'apprentissage social dans des conditions plus réalistes. Ceci est réalisé en exploitant quatre éléments clés de l'apprentissage sur les réseaux adaptatifs, à savoir, **i)** topologie du *réseau*, **ii)** *informations échangées*, **iii)** *monde* environnant, et **iv)** *modèles* privés, que nous divisons en deux parties.

Dans la première partie, nous nous concentrons sur le cas *stationnaire*, i.e., où les conditions du monde sont statiques. **i)** Le réseau social est représenté par un graphe faiblement connexe, ce qui entraîne une asymétrie de pouvoir entre les clusters du réseau. Pour estimer le niveau d'influence des clusters influents envers des agents spécifiques, nous formulons le problème d'*apprentissage inverse*. Nous caractérisons la faisabilité de ce problème et montrons qu'une certaine diversité statistique entre les composants du graphe est suffisante pour qu'il soit réalisable. **ii)** Nous considérons un réseau social fortement connexe, avec une communication limitée, i.e., où seules des *opinions partielles* sont partagées avec les voisins. Nous montrons comment différents régimes d'apprentissage apparaissent et sous quelles conditions les agents

peuvent apprendre la vérité ou, au contraire, être induits en erreur.

Dans la deuxième partie, nous abordons le cas *non stationnaire*. **iii)** Les stratégies d'apprentissage social existantes sont limitées dans leur capacité à *s'adapter* aux conditions variables du monde. Nous proposons une formulation d'apprentissage social adaptatif et caractérisons sa performance à la fois dans la phase permanente et dans la phase transiente. Nous montrons que l'approche donne lieu à un compromis entre précision d'apprentissage et capacité d'adaptation. **iv)** Les agents d'apprentissage social utilisent des modèles statistiques supposés parfaitement connus a priori. Nous proposons un système d'*apprentissage automatique* social, dans lequel les modèles sont d'abord entraînés à partir d'un ensemble fini d'exemples étiquetés, puis déployés dans une implémentation collaborative pour classer les observations en continu non étiquetées—possiblement non stationnaires. Nous montrons que la stratégie proposée entièrement basée sur les données se traduit par un apprentissage cohérent, malgré les modèles imprécis, et par une précision améliorée à mesure que le nombre d'observations non étiquetées augmente.

Mots-clés : Apprentissage social, mise à jour bayésienne, stratégies de diffusion, apprentissage d'influences, échange d'informations partielles, réseau adaptatif, classification distribuée.

Notation

a	Normal font denotes a deterministic variable
$\boldsymbol{a}$	Boldface font denotes a random variable
$\mathbb{E}$	Expected value operator
$\mathbb{E}_f$	Expected value operator under distribution f
$\mathbb{P}$	Probability measure operator
$D(f\|\|g)$	Kullback-Leibler divergence of distribution f from distribution g
$\xrightarrow{\text{a.s.}}$	Almost sure convergence as $i \to \infty$
$\xrightarrow{\text{p}}$	Convergence in probability as $i \to \infty$
$\xrightarrow{\text{d}}$	Convergence in distribution as $i \to \infty$
$\overset{\text{d}}{=}$	Equality in distribution
$\doteq$	Equality to the leading exponential order
$\succ, \succcurlyeq, \prec, \preccurlyeq$	Element-wise inequalities
$\|a\|$	Absolute value of scalar a
A^{T}	Transpose of matrix A
$\|A\|_2$	Spectral norm of matrix A
A^{-1}	Inverse of matrix A
$A^{\dagger}$	Moore-Penrose inverse of matrix A
$[A]_{\ell k}$	$(\ell, k)-$th element of matrix A
col$\{a, b\}$	Column vector with elements a and b
blkdiag$\{A, B\}$	Block diagonal matrix with blocks A and B
$\mathbb{1}$, $\mathbb{1}_D$	Vector of ones, vector of ones with dimensions $D \times 1$
I, I_D	Identity matrix, identity matrix with dimensions $D \times D$
$\mathbb{I}[\mathcal{A}]$	Indicator function, i.e., it equals 1 if event $\mathcal{A}$ holds, 0 otherwise

$\mathcal{A} \cap \mathcal{B}$	Intersection of sets (or events) $\mathcal{A}$ and $\mathcal{B}$
$\mathcal{A} \cup \mathcal{B}$	Union of sets (or events) $\mathcal{A}$ and $\mathcal{B}$
$\lvert\mathcal{A}\rvert$	Cardinality of set $\mathcal{A}$
$\overline{\mathcal{A}}$	Complement of set $\mathcal{A}$
$\text{rect}(x)$	Rectangle function, i.e., it is equal to 1 if $x \in (-\frac{1}{2}, \frac{1}{2})$, 0 otherwise
$\text{sign}(x)$	Sign function, i.e., it is equal to 1 if $x \geq 0$, -1 otherwise

Contents

1 Introduction

An essential part of human learning relies on exchanging opinions with a social group. A single individual cannot have access to all existing evidence about any particular phenomenon. They can however incorporate the opinions and beliefs of others in their social clique as a way to make up for—or complement—their limited observation ability.

For example, say a traveler is planning a vacation to Rio. They have been there once before during summer and learned that, due to the rainy season, the period is not ideal for traveling in the region. To choose a more suitable season, they consult three friends who went to Rio on different occasions, namely, in winter, spring and autumn. Collecting all three accounts allows the traveler to make a more educated choice, than relying on their unique prior experience.

Human learning is tied to the concept of *social learning*, or learning in a group. The manner in which a group of individuals is able to aggregate dispersed information is a historical subject of study. As early as [1], the work studied how a large group of individuals can combine their information in an honest manner to learn some underlying truth. Similarly, in [2], a social experiment illustrated the wisdom-of-the-crowd argument, where the estimate of a parameter of interest by a group turned out to be more reliable than the estimate by a single individual.

In recent years, several works on social learning have investigated how to model opinion dynamics in groups [3]–[8]. From a *behavioral* perspective, these works examined how beliefs evolve in response to various learning strategies. The methods are not only expected to bring individuals closer to the *underlying truth*, but they should also help reveal other interesting social phenomena that may arise such as manipulation among individuals, stubbornness, and herding behavior. From a *design* perspective, the works examined the quality of the decision process and whether one can infer the truth from the evolving beliefs with sufficient accuracy and speed of convergence.

An example of an engineering system whose design is inspired by social learning is a collection of sensors recording data from a common observed scene. These could be, for example, meteorological sensors measuring different attributes such as air pressure, humidity, and so on. The goal of the network of sensors is to discover or predict the state of the weather in the geographic region under observation. It is usually the case that information at each individual sensor is insufficient to allow it to make a correct inference. However, through interactions with

their neighboring sensors, all sensors would be able to arrive at more informed predictions.

This thesis is dedicated to studying social learning strategies and their performance, while proposing solutions that enable their application in more realistic settings.

Across the different social learning strategies, a key ingredient used to incorporate new information into beliefs (or opinions) is Bayesian processing. The Bayes rule has multiple interpretations and uses. For example, it can be formulated as an optimal information processing rule [9]. It can also be exploited as a model for how the brain learns from sensory input [10]. It is also considered to be the *rational* approach to solving inference problems from data [11], since it takes into account the uncertainty of the observed information to update the probability of a certain event of interest. In the next two sections we review basic concepts about Bayesian and social learning in order to explain the state of the art and to highlight the problems that exist. We start by motivating the Bayesian way of thinking.

1.1 Bayesian Thinking

A Bayesian decision system starts with an *agent* and the *world* surrounding it. From the point of view of the agent, the world exists through observations. Specifically, the agent perceives its environment by means of an *observation* denoted by $\boldsymbol{\xi}$, belonging to a set $\mathcal{X}$. The observation is a random variable—hence the bold font—and embodies a piece of evidence on the current *state of the world*.

The agent considers a set of possible hypotheses that could explain the underlying state of the world. For example, the probability distribution of $\boldsymbol{\xi}$ could be dependent on some discrete parameter $\theta \in \Theta$, say, $L(\xi|\theta)$. The set of hypotheses is denoted by the discrete set Θ with cardinality H, say:

$$\Theta = \{1, 2, \ldots, H\}. \tag{1.1}$$

For illustration, consider a weather forecast system equipped with an air pressure sensor. In this case, $\boldsymbol{\xi}$ is the air pressure measurement, and the set of hypotheses could be binary, i.e., $H = 2$, corresponding to the weather states “clear” and “cloudy”.

Real-world measurements include uncertainty. For example, if the air pressure is low, then it is *likely* that the sky is covered by clouds. In this uncertain world, the agent must reason in a *Bayesian* way, where the perception of “likely” can be represented probabilistically. The agent would be endowed with a *belief* $\boldsymbol{\mu}(\theta)$, which is a probability mass function (pmf) over the set of hypotheses $\theta \in \Theta$. The belief is written in bold font, since it is considered to inherit the random nature of the observation.

To form its belief, the Bayesian agent assumes (or possesses) a model for the observation $\boldsymbol{\xi}$ given the different hypotheses, denoted by

$$L(\xi|\theta), \quad \theta \in \Theta, \quad \xi \in \mathcal{X}, \tag{1.2}$$

and referred to as the *likelihood model*. If $\boldsymbol{\xi}$ is a continuous (or discrete) random variable, the likelihood model will be a probability density (or mass) function. It is a probability distribution

when seen as a function of ξ, and a likelihood function when seen as a function of θ.

The agent also has a *prior belief*, which we denote by $\mu_0(\theta)$ for $\theta \in \Theta$, reflecting its opinion prior to the observation of $\boldsymbol{\xi}$. The prior belief summarizes the agent's biases and past experiences. We write it in normal font and consider it to be a deterministic variable[1]. Based on the Bayes rule, the belief is updated according to

$$\boldsymbol{\mu}(\theta) \propto L(\boldsymbol{\xi}|\theta)\mu_0(\theta), \quad \theta \in \Theta \tag{1.3}$$

where the entries of $\boldsymbol{\mu}(\theta)$ should be normalized to add up to 1.

Assume an air pressure measurement $\boldsymbol{\xi}$ corresponds to the following likelihood values:

$$L(\boldsymbol{\xi}|\text{clear}) = 0.5, \quad L(\boldsymbol{\xi}|\text{cloudy}) = 0.4. \tag{1.4}$$

Disregarding the agent's biases, i.e., considering an uninformative prior belief $\mu_0(\theta) = 1/2$ for $\theta \in \Theta$, the Bayesian update would result in the following beliefs:

$$\boldsymbol{\mu}(\text{clear}) = 0.56, \quad \boldsymbol{\mu}(\text{cloudy}) = 0.44, \tag{1.5}$$

showing a slight preference for the hypothesis "clear". Some external knowledge, however, may suggest that it is a rainy season, in which case, the agent's prior belief would reflect some bias towards a cloudy weather, say, as

$$\mu_0(\text{clear}) = 0.2, \quad \mu_0(\text{cloudy}) = 0.8. \tag{1.6}$$

leading to the posterior beliefs:

$$\boldsymbol{\mu}(\text{clear}) = 0.24, \quad \boldsymbol{\mu}(\text{cloudy}) = 0.76. \tag{1.7}$$

Now that we have motivated Bayesian processing, we will turn our attention to the problem of processing streaming observations over time.

1.1.1 Bayesian Inference with Streaming Data

In Bayesian inference, agents rely on a growing amount of evidence over time. This corresponds to the situation in which the agent receives streaming observations such as

$$\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i, \tag{1.8}$$

where i represents the time index and $\boldsymbol{\xi}_i \in \mathcal{X}$. Recall that $\mathcal{X}$ is the set of all possible observations. We can use the Bayesian update in (1.3) to update the prior belief to the posterior belief as follows:

$$\boldsymbol{\mu}_i(\theta) \propto \mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta)\mu_0(\theta), \tag{1.9}$$

where $\mathcal{L}(\cdot|\theta)$ denotes the joint likelihood model for $\theta \in \Theta$. If we assume that, conditioned on knowledge of θ, the samples $\boldsymbol{\xi}_i$ are independent and identically distributed (iid), then the joint

[1] We could also consider $\boldsymbol{\mu}_0$ to be a random variable that is independent of $\boldsymbol{\xi}$.

model can be decomposed into a product form, i.e.,

$$\mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta) = \prod_{m=1}^{i} L(\boldsymbol{\xi}_m|\theta), \tag{1.10}$$

with the *same* conditional pdf $L(\xi|\theta)$. Substituting into (1.9) we arrive at the recursive form:

$$\boldsymbol{\mu}_i(\theta) \propto L(\boldsymbol{\xi}_i|\theta)\boldsymbol{\mu}_{i-1}(\theta). \tag{1.11}$$

This expression allows us to update the belief vector recursively over time. A natural follow-up question concerns the asymptotic behavior of (1.11), when i grows. To answer this question, we assume that the samples $\boldsymbol{\xi}_i$ are iid and generated from some unknown *true model* $f(\xi)$:

$$\boldsymbol{\xi}_i \sim f(\xi), \tag{1.12}$$

where $f(\xi)$ is either a pmf or a pdf, depending on whether the observations are discrete or continuous random variables, respectively.

The asymptotic convergence of the Bayesian recursion in (1.11) for a stream of observations is a classical problem of interest. In [12], the problem is more generally formulated for a continuous set of hypotheses. It can also be formulated for the case when the true model $f(\xi)$ belongs to the set of likelihood models $\{L(\xi|\theta)\}$ [13]–[15].

We are interested in the scenario where the set of hypotheses is discrete and the true model does not necessarily belong to the set of likelihoods. In the next paragraphs, we introduce the classical convergence result for this specific case, which we consider relevant for two reasons: **i)** It helps to clarify the operation of the learning process; **ii)** It serves as a counterpoint to the belief evolution in the multi-agent social learning problem, which will be introduced in Chapter 2.

First, we define the Kullback-Leibler (KL) divergence [16] between $f(\xi)$ and $L(\xi|\theta)$:

$$D(f||L(\theta)) = \mathbb{E}_f \left(\log \frac{f(\boldsymbol{\xi})}{L(\boldsymbol{\xi}|\theta)} \right), \tag{1.13}$$

where $\mathbb{E}_f$ denotes the expectation computed with respect to distribution $f(\xi)$. We omit the argument $\boldsymbol{\xi}$ on the LHS of (1.13) from both $f(\cdot)$ and $L(\cdot|\theta)$ for ease of notation. Then, we introduce the following technical assumptions.

Assumption 1.1 (Finite KL divergences). Assume that, for all $\theta \in \Theta$, $D(f||L(\theta)) < \infty$.

Assumption 1.2 (Positive initial beliefs). Assume that, for all $\theta \in \Theta$, $\mu_0(\theta) > 0$.

Assumption 1.3 (Unique minimizer). Assume that the KL divergence between $f(\xi)$ and

$L(\xi|\theta)$ is minimized at a unique *target* hypothesis $\theta^\star$, i.e.,

$$\theta^\star \triangleq \arg\min_{\theta\in\Theta} D(f||L(\theta)). \tag{1.14}$$

The target hypothesis corresponds to the likelihood model $L(\xi|\theta^\star)$ that best approximates the true model, $f(\xi)$, in the sense of minimizing the KL divergence between the true model and the likelihood models.

We can now state the convergence result for the recursive Bayesian update described in (1.11), whose proof is patterned after the one in [17].

Theorem 1.1 (Belief convergence of a single Bayesian agent). *Under Assumptions 1.1, 1.2, and 1.3, the recursive Bayesian update* (1.11) *enables learning the target hypothesis in the limit:*

$$\boldsymbol{\mu}_i(\theta^\star) \xrightarrow{\text{a.s.}} 1 \tag{1.15}$$

Proof. We first rewrite the recursive Bayesian update (1.11) in a more explicit form:

$$\boldsymbol{\mu}_i(\theta) = \frac{L(\boldsymbol{\xi}_i|\theta)\boldsymbol{\mu}_{i-1}(\theta)}{\sum_{\theta'\in\Theta} L(\boldsymbol{\xi}_i|\theta')\boldsymbol{\mu}_{i-1}(\theta')}, \tag{1.16}$$

where the belief components $\boldsymbol{\mu}_i(\theta)$ are normalized so that they sum up to 1. Using (1.16), we write the ratio of $\boldsymbol{\mu}_i(\theta^\star)$ to $\boldsymbol{\mu}_i(\theta)$ for any $\theta \neq \theta^\star$ as

$$\frac{\boldsymbol{\mu}_i(\theta^\star)}{\boldsymbol{\mu}_i(\theta)} = \frac{\boldsymbol{\mu}_{i-1}(\theta^\star)L(\boldsymbol{\xi}_i|\theta^\star)}{\boldsymbol{\mu}_{i-1}(\theta)L(\boldsymbol{\xi}_i|\theta)}. \tag{1.17}$$

Taking the $\log$ of the above expression leads to a linear expression of log-ratio terms, i.e.,

$$\log\frac{\boldsymbol{\mu}_i(\theta^\star)}{\boldsymbol{\mu}_i(\theta)} = \log\frac{\boldsymbol{\mu}_{i-1}(\theta^\star)}{\boldsymbol{\mu}_{i-1}(\theta)} + \log\frac{L(\boldsymbol{\xi}_i|\theta^\star)}{L(\boldsymbol{\xi}_i|\theta)}. \tag{1.18}$$

Developing the recursion over time yields:

$$\log\frac{\boldsymbol{\mu}_i(\theta^\star)}{\boldsymbol{\mu}_i(\theta)} = \log\frac{\mu_0(\theta^\star)}{\mu_0(\theta)} + \sum_{m=1}^{i}\log\frac{L(\boldsymbol{\xi}_m|\theta^\star)}{L(\boldsymbol{\xi}_m|\theta)}. \tag{1.19}$$

The following arguments are adapted from [17]. Dividing the above expression by i and studying its limit as i goes to infinity leads to:

$$\lim_{i\to\infty}\frac{1}{i}\log\frac{\boldsymbol{\mu}_i(\theta^\star)}{\boldsymbol{\mu}_i(\theta)} = \lim_{i\to\infty}\frac{1}{i}\log\frac{\mu_0(\theta^\star)}{\mu_0(\theta)} + \lim_{i\to\infty}\frac{1}{i}\sum_{m=1}^{i}\log\frac{L(\boldsymbol{\xi}_m|\theta^\star)}{L(\boldsymbol{\xi}_m|\theta)}. \tag{1.20}$$

The first term on the RHS of (1.20) vanishes. Moreover, the iid property of $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots$ and the finiteness condition in Assumption 1.1 allow us to use the strong law of large numbers (SLLN) [18] to establish the convergence of the second term on the RHS of (1.20) in the following

manner:

$$\frac{1}{i}\sum_{m=1}^{i}\log\frac{L(\boldsymbol{\xi}_m|\theta^\star)}{L(\boldsymbol{\xi}_m|\theta)} \xrightarrow{\text{a.s.}} \mathbb{E}_f\left(\log\frac{f(\boldsymbol{\xi}_m)}{L(\boldsymbol{\xi}_m|\theta)}\right) - \mathbb{E}_f\left(\log\frac{f(\boldsymbol{\xi}_m)}{L(\boldsymbol{\xi}_m|\theta^\star)}\right)$$
$$= D(f||L(\theta)) - D(f||L(\theta^\star)). \tag{1.21}$$

Since $\theta^\star$ satisfies (1.14), it follows that

$$D(f||L(\theta)) - D(f||L(\theta^\star)) > 0 \tag{1.22}$$

for all $\theta \neq \theta^\star$, and therefore we have that

$$\frac{1}{i}\log\frac{\boldsymbol{\mu}_i(\theta^\star)}{\boldsymbol{\mu}_i(\theta)} \xrightarrow{\text{a.s.}} D(f||L(\theta)) - D(f||L(\theta^\star)) > 0$$
$$\Rightarrow \log\frac{\boldsymbol{\mu}_i(\theta^\star)}{\boldsymbol{\mu}_i(\theta)} \xrightarrow{\text{a.s.}} +\infty \Rightarrow \frac{\boldsymbol{\mu}_i(\theta^\star)}{\boldsymbol{\mu}_i(\theta)} \xrightarrow{\text{a.s.}} +\infty, \tag{1.23}$$

for all $\theta \neq \theta^\star$. Since the belief vector is a pmf over the set of hypotheses, each individual component is upper bounded by 1. Thus, (1.23) implies that

$$\boldsymbol{\mu}_i(\theta) \xrightarrow{\text{a.s.}} 0, \tag{1.24}$$

for all $\theta \neq \theta^\star$. We arrive at the desired result in (1.15) by noting that the entries of the belief vector must add up to 1. □

From Theorem 1.1, the recursive Bayesian update results in a belief distribution whose mass is concentrated at the single target hypothesis $\theta^\star$. From Assumption 1.3, this hypothesis is associated with the likelihood model $L(\xi|\theta^\star)$ that best approximates the true model, $f(\xi)$, using the KL divergence metric. In a nutshell, the asymptotic belief indicates which *hypothesis best explains the received observations*.

An important element of this convergence result is Assumption 1.3, which requires that the minimizer of the KL divergence between $f(\xi)$ and $L(\xi|\theta)$ be *unique*, namely, hypothesis $\theta^\star$. This assumption avoids the following singular behavior. Suppose multiple hypotheses minimize the KL divergence, i.e., there exists a target subset $\Theta^\star \subset \Theta$ such that

$$\Theta^\star = \arg\min_{\theta\in\Theta} D(f||L(\theta)). \tag{1.25}$$

In this case, similar arguments to the ones used in the proof of Theorem 1.1 could be repeated to conclude that

$$\boldsymbol{\mu}_i(\theta) \xrightarrow{\text{a.s.}} 0, \quad \theta \notin \Theta^\star. \tag{1.26}$$

In other words, the hypotheses that do not belong to $\Theta^\star$ will be asymptotically discarded. Since the entries of the belief should add up to 1, this implies that,

$$\sum_{\theta\in\Theta^\star} \boldsymbol{\mu}_i(\theta) \xrightarrow{\text{a.s.}} 1. \tag{1.27}$$

In other words, the beliefs corresponding to the hypotheses in $\Theta^\star$ *do not* necessarily converge individually. They instead evolve randomly over time and the agent finds itself in a state of everlasting *confusion* among the hypotheses in $\Theta^\star$.

1.1.2 MAP and ML Inference

We can gain further insight into the Bayesian construction by examining the maximum a-posteriori (MAP) and maximum likelihood (ML) formulations. To begin with, the MAP estimator for the discrete hypothesis θ given the sequence of i iid observations $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i$ is constructed as follows:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{MAP},i} \triangleq \arg\max_{\theta\in\Theta} \mathcal{P}(\theta|\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i), \tag{1.28}$$

where $\mathcal{P}(\theta|\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i)$ models the posterior probability of θ given the observations.

If we again let $\mu_0(\theta)$ denote the prior probability over θ and $\mathcal{L}(\xi_1, \xi_2, \ldots, \xi_i|\theta)$ the joint likelihood of the observations given θ, then, from the Bayes rule, we have that

$$\mathcal{P}(\theta|\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i) \propto \mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta)\mu_0(\theta). \tag{1.29}$$

This allows us to rewrite the MAP estimator as

$$\widehat{\boldsymbol{\theta}}_{\mathsf{MAP},i} = \arg\max_{\theta\in\Theta} \mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta)\mu_0(\theta). \tag{1.30}$$

From (1.9) and (1.30), we conclude that

$$\widehat{\boldsymbol{\theta}}_{\mathsf{MAP},i} = \arg\max_{\theta\in\Theta} \boldsymbol{\mu}_i(\theta). \tag{1.31}$$

That is, the hypothesis that maximizes the belief $\boldsymbol{\mu}_i(\theta)$ at each instant i can be seen as the MAP estimator of θ given the sequence of i iid observations $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i$. As such, the $\theta^\star$ in (1.14) and (1.15) is the asymptotic MAP estimator of θ given infinitely many iid observations.

Now, consider the ML estimator defined by:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{ML},i} \triangleq \arg\max_{\theta\in\Theta} \mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta). \tag{1.32}$$

We show next that the asymptotic ML estimator converges according to:

$$\lim_{i\to\infty} \widehat{\boldsymbol{\theta}}_{\mathsf{ML},i} = \arg\min_{\theta\in\Theta} D(f||L(\theta)) \tag{1.33}$$

almost surely, which, in view of (1.15), corresponds again to $\theta^\star$. Eq. (1.33) is in accordance with [19]–[21], where a similar result is established for a more general continuous set of hypotheses. In the case of a finite set of hypotheses and under Assumptions 1.1 and 1.3, we can provide a simpler proof for (1.33) motivated by the proof of Theorem 1.1.

First, note that, in view of (1.32), we can rewrite the ML estimator as:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{ML},i} = \arg\min_{\theta\in\Theta} \frac{1}{i} \log \frac{\mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta^\star)}{\mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta)}, \tag{1.34}$$

for $i > 0$, where $\theta^\star$ is the target hypothesis.

Now, using the decomposition of $\mathcal{L}$ in (1.10), we can write the following ratio of likelihoods:

$$\log \frac{\mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta^\star)}{\mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta)} = \sum_{m=1}^{i} \log \frac{L(\boldsymbol{\xi}_m|\theta^\star)}{L(\boldsymbol{\xi}_m|\theta)}. \tag{1.35}$$

Dividing both sides of (1.35) by i, under Assumption 1.1 and using similar arguments as in (1.21), the ratio converges according to:

$$\frac{1}{i} \log \frac{\mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta^\star)}{\mathcal{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_i|\theta)} \xrightarrow{\text{a.s.}} D(f||L(\theta)) - D(f||L(\theta^\star)), \tag{1.36}$$

which, from (1.34), implies that

$$\widehat{\boldsymbol{\theta}}_{\mathsf{ML},i} \xrightarrow{\text{a.s.}} \arg\min_{\theta\in\Theta} \Big\{ D(f||L(\theta)) - D(f||L(\theta^\star)) \Big\}. \tag{1.37}$$

From Assumption 1.3, we know that the RHS of (1.37) is uniquely minimized at $\theta^\star$. Therefore, we conclude that asymptotically

$$\widehat{\boldsymbol{\theta}}_{\mathsf{ML},i} \xrightarrow{\text{a.s.}} \theta^\star \overset{(1.14)}{=} \arg\min_{\theta\in\Theta} D(f||L(\theta)). \tag{1.38}$$

We can thus conclude two properties of the recursive Bayesian update in (1.9): **i)** It tracks the *instantaneous* MAP estimator; **ii)** It corresponds *asymptotically* to the ML estimator.

1.2 Social Learning

In our treatment of Bayesian processing in Section 1.1, we assumed the existence of a *single* Bayesian agent in the world. A more meaningful and realistic scenario is when *multiple agents* coexist. Besides interacting with the surrounding environment, these agents are allowed, and also encouraged, to interact with each other, forming a *social network*. The network connectivity dictates the communication links, which are assumed to connect neighboring agents in a sparse *network*.

The interaction with the environment occurs in the following manner. At each instant i, the group of K agents senses the world through the observation profile:

$$\boldsymbol{\xi}_{1,i}, \boldsymbol{\xi}_{2,i}, \ldots, \boldsymbol{\xi}_{K,i}. \tag{1.39}$$

Each observation $\boldsymbol{\xi}_{k,i}$ belongs to a set $\mathcal{X}_k$ and is *private* to agent k. This setup allows for a significant heterogeneity in the observation profiles across agents. For example, one agent can observe RGB images captured from a scene of interest, while another might record audio waves

from the scene. This data constitutes evidence about the state of the world θ, belonging to a common set of hypotheses Θ, and it can be dependent across different agents.

Since the observations are private, they cannot be exchanged during interactions among agents. Yet these interactions must contain essential information to solve the following inference problem:

> **Social Learning Problem**: *Find the hypothesis $\theta \in \Theta$ that best explains the observations received by the network.*

A multitude of strategies have been proposed in the literature under the umbrella of *social learning*. These strategies can be split in two main categories: Bayesian and non-Bayesian social learning.

1.2.1 Bayesian Social Learning

Following the discussion in Section 1.1 concerning Bayesian learning, we can extend the single-agent approach and consider a fully Bayesian strategy to solving the inference problem. Such approach would take the form of a *centralized* Bayesian update:

$$\boldsymbol{\mu}_i(\theta) \propto L(\boldsymbol{\xi}_{1,i}, \boldsymbol{\xi}_{2,i}, \ldots, \boldsymbol{\xi}_{K,i}|\theta)\boldsymbol{\mu}_{i-1}(\theta), \tag{1.40}$$

where $L(\xi_1, \xi_2, \ldots, \xi_K|\theta)$ is the joint likelihood model of the observation profile in (1.39) from across all agents in the network given hypothesis θ. A fully Bayesian strategy would thus require knowledge about this *joint* likelihood model, in addition to centralized processing of the joint information. Both requirements cannot be fulfilled in our scenario. First, agents do not possess knowledge of the dependencies between different sources of data. They only have access to models for the marginal distribution of their local observations, i.e., each agent k possesses its own marginal likelihood model $L_k(\xi|\theta)$. Second, their observations are private, and agents would not want to share their raw observations with neighbors. Inspired by real-life social dynamics, we will instead limit the agents to sharing instantaneous opinions or *beliefs* with neighbors. We denote the belief of agent k at instant i by $\boldsymbol{\mu}_{k,i}$. The belief of an agent acts as a summary of its observations until that point in time. It embodies not only the likelihood of that observation given different hypotheses, but also the agent's prior belief.

Example 1.1 (Multi-agent Bayesian processing). To illustrate the complexity of a fully Bayesian solution, we consider the following simple example. Consider a set of 4 agents, namely $\mathcal{N} \triangleq \{k, \ell, m, n\}$, connected in a directed graph according to Figure 1.1. For simplicity, we assume that only three steps of Bayesian processing take place, which we describe next.

Step 1: Agent k observes $\boldsymbol{\xi}_k$. It updates its belief $\boldsymbol{\mu}_k$ and exchanges it with agents ℓ and m.

Step 2: Agents ℓ and m observe respectively $\boldsymbol{\xi}_\ell$ and $\boldsymbol{\xi}_m$. They update independently their beliefs $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\mu}_m$ and exchange them with agent n.

Step 3: Agent n observes $\boldsymbol{\xi}_n$. It updates its belief $\boldsymbol{\mu}_n$.

Note that *not all* agents interact at every step and that the updates take place in a sequential

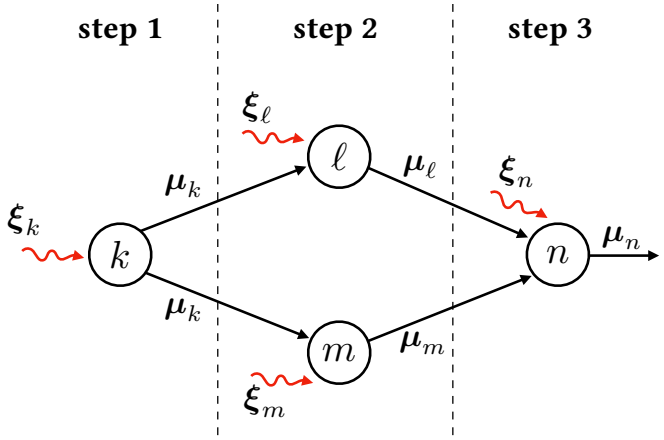


Figure 1.1: Diagram showing the communication graph between agents k, ℓ, m and n.

manner. This is done to keep the example as simple as possible for illustration purposes. Since observations are not present at every step i for all agents, we choose to omit the subindex i from the following description for ease of notation.

Assume a binary hypothesis space $\Theta = \{0, 1\}$ and that each agent $j \in \mathcal{N}$ receives binary observations as well, i.e., $\boldsymbol{\xi}_j \in \{a, b\}$. Agents have furthermore likelihoods of the form:

$$L_j(\xi|\theta) = p_\theta^{(j)}\, \mathbb{I}\,[\xi = a] + (\,1 - p_\theta^{(j)}\,)\mathbb{I}\,[\xi = b], \tag{1.41}$$

with $p_1^{(j)} \neq p_0^{(j)}$, which are further assumed to be common knowledge to *all* agents. We assume that observations are independent across agents, such that

$$L(\boldsymbol{\xi}_k, \boldsymbol{\xi}_\ell, \boldsymbol{\xi}_m, \boldsymbol{\xi}_n|\theta) = \prod_{j\in\mathcal{N}} L_j(\boldsymbol{\xi}_j|\theta), \tag{1.42}$$

which implies that from knowledge of the marginal likelihoods, each agent can compute the joint likelihood as well. Moreover, we assume that all agents start from the same initial belief μ_0.

We describe next the computations performed at each of the three steps, which allow agents to update their beliefs in a fully Bayesian manner.

Step 1: Agent k updates its belief

When agent k receives a new observation $\boldsymbol{\xi}_k$, it updates its prior belief using the Bayesian update (1.3), i.e.,

$$\boldsymbol{\mu}_k(\theta) \propto L_k(\boldsymbol{\xi}_k|\theta)\mu_0(\theta). \tag{1.43}$$

Agent k then sends its updated belief $\boldsymbol{\mu}_k$ to its two neighbors ℓ and m.

Step 2: Agents ℓ and m update their beliefs

We first focus on agent ℓ, which received the belief $\boldsymbol{\mu}_k$. If agent ℓ wishes to update its belief

using its own observation $\boldsymbol{\xi}_\ell$ *and* the information received from k in a fully Bayesian way, then, in view of the independence property in (1.42), it would need to compute:

$$\begin{aligned}\boldsymbol{\mu}_\ell(\theta) &\propto L_\ell(\boldsymbol{\xi}_\ell|\theta)\boldsymbol{\mu}_k(\theta)\\ &\overset{(1.43)}{\propto} L_\ell(\boldsymbol{\xi}_\ell|\theta)L_k(\boldsymbol{\xi}_k|\theta)\mu_0(\theta),\end{aligned} \tag{1.44}$$

which corresponds to the Bayesian way of updating the prior belief μ_0 using the independent observations $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_\ell$.

Agent m could perform a similar procedure as agent ℓ to update its belief according to:

$$\begin{aligned}\boldsymbol{\mu}_m(\theta) &\propto L_m(\boldsymbol{\xi}_m|\theta)\boldsymbol{\mu}_k(\theta)\\ &\overset{(1.43)}{\propto} L_m(\boldsymbol{\xi}_m|\theta)L_k(\boldsymbol{\xi}_k|\theta)\mu_0(\theta).\end{aligned} \tag{1.45}$$

In the next step, agents ℓ and m send their beliefs $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\mu}_m$ to their common neighbor n.

Step 3: Agent n updates its belief

Agent n receives beliefs $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\mu}_m$, both containing redundant information about the observation $\boldsymbol{\xi}_k$. In order to incorporate the information in $\boldsymbol{\mu}_\ell$, $\boldsymbol{\mu}_m$, *and* its private observation $\boldsymbol{\xi}_n$ in a Bayesian way, agent n needs to disentangle the observations $\boldsymbol{\xi}_k$, $\boldsymbol{\xi}_\ell$, and $\boldsymbol{\xi}_m$ from the received beliefs. This can be performed by agent n in the following manner.

First, consider the task of extracting information on $\boldsymbol{\xi}_\ell$ and $\boldsymbol{\xi}_k$ from the belief $\boldsymbol{\mu}_\ell$, which are related through Eq. (1.44). From (1.44), we can write

$$\frac{\boldsymbol{\mu}_\ell(1)}{\boldsymbol{\mu}_\ell(0)} = \frac{L_\ell(\boldsymbol{\xi}_\ell|1)L_k(\boldsymbol{\xi}_k|1)}{L_\ell(\boldsymbol{\xi}_\ell|0)L_k(\boldsymbol{\xi}_k|0)}\frac{\mu_0(1)}{\mu_0(0)} \Leftrightarrow \frac{L_\ell(\boldsymbol{\xi}_\ell|1)L_k(\boldsymbol{\xi}_k|1)}{L_\ell(\boldsymbol{\xi}_\ell|0)L_k(\boldsymbol{\xi}_k|0)} = \frac{\boldsymbol{\mu}_\ell(1)}{\boldsymbol{\mu}_\ell(0)}\frac{\mu_0(0)}{\mu_0(1)} \triangleq \boldsymbol{\alpha}. \tag{1.46}$$

Agent n can compute $\boldsymbol{\alpha}$, since the belief terms in (1.46) are known. Since (1.41) is also known, then agent n can recover $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_\ell$ by comparing $\boldsymbol{\alpha}$ against the following 4 possibilities:

$$\{\boldsymbol{\xi}_\ell,\, \boldsymbol{\xi}_k\} = \begin{cases} \{a,a\}, & \text{if } \boldsymbol{\alpha} = \dfrac{p_1^{(\ell)}}{(\,1-p_0^{(\ell)}\,)}\dfrac{p_1^{(k)}}{(\,1-p_0^{(k)}\,)},\\[2ex] \{a,b\}, & \text{if } \boldsymbol{\alpha} = \dfrac{p_1^{(\ell)}}{(\,1-p_0^{(\ell)}\,)}\dfrac{(1-p_1^{(k)})}{p_0^{(k)}},\\[2ex] \{b,a\}, & \text{if } \boldsymbol{\alpha} = \dfrac{(1-p_1^{(\ell)})}{p_0^{(\ell)}}\dfrac{p_1^{(k)}}{(\,1-p_0^{(k)}\,)},\\[2ex] \{b,b\}, & \text{if } \boldsymbol{\alpha} = \dfrac{(1-p_1^{(\ell)})}{p_0^{(\ell)}}\dfrac{(1-p_1^{(k)})}{p_0^{(k)}}. \end{cases} \tag{1.47}$$

A similar procedure, with 4 different comparisons, can be applied to belief $\boldsymbol{\mu}_m$ to recover information about $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_m$.

Upon recovery of $\boldsymbol{\xi}_k$, $\boldsymbol{\xi}_\ell$, and $\boldsymbol{\xi}_m$, agent n can update its belief in a Bayesian way according to:

$$\boldsymbol{\mu}_n(\theta) \propto L_n(\boldsymbol{\xi}_n|\theta)L_m(\boldsymbol{\xi}_m|\theta)L_\ell(\boldsymbol{\xi}_\ell|\theta)L_k(\boldsymbol{\xi}_k|\theta)\mu_0(\theta). \tag{1.48}$$

□

Even in this simple example, with only two hypotheses and two possible observations, we see that the dynamics between four Bayesian agents is complex and requires that each agent have significant knowledge about the operation of the other. The example also does not consider *synchronous* interactions among agents, i.e., a scenario where *all* agents update their beliefs and interact at every instant.

In a *distributed* and synchronous setup, in which multiple agents exchange their beliefs with neighbors at each instant, the complexity of a fully Bayesian solution makes its implementation computationally prohibitive. Agent ℓ has to disentangle the different sources of information present in the belief propagated by agent k. The information can come from the neighbors of k and the observation of k. To do that, further knowledge of the social dynamics within the network is necessary at agent k, and even under a complete knowledge assumption, the computations have been shown to be NP-hard [22].

The implementation of a fully Bayesian solution can be manageable in some simplified scenarios. For example, in [23]–[25] the authors propose a model in which agents learn *sequentially*, as opposed to synchronously. At each instant, a different agent updates its belief using *actions*[2] by all previous agents and the current observation. Observations are assumed to be independent across agents. Through this sequential dynamics, the authors in [23]–[25] avoid the entangling of different sources of information.

Under the sequential structure, these works are able to reproduce a *herding* behavior, i.e., from a certain period onward the agents disregard local observations in favor of just repeating previous agents' actions. A similar sequential framework is investigated in [26], where agents observe instead the actions of a subset of previous agents, denoted as *neighbors*. The concept of neighborhood introduces the possibility that some neighbors, if visited often enough, can affect and *influence* the asymptotic convergence behavior.

Other tractable scenarios for Bayesian strategies are found in [27], [28] for communication structures such as trees and fully connected networks. Although tractable solutions exist, all multi-agent Bayesian strategies require extensive knowledge of the statistical models and connectivity paths in the network.

In face of such communication and knowledge limitations, researchers have pursued non-Bayesian social learning strategies to model opinion formation over networked systems. A precursor for these models is the concept of *bounded rationality* in Economics. Bounded rationality acknowledges that human cognition is limited, and that therefore deliberating and making decisions are costly activities [29], [30]. A classical example is chess. Given that the search space is finite, there exists a well defined optimal strategy at every move. It is however

[2] In some of these models, instead of beliefs agents share actions. These actions are the result of maximizing the agents' expected utility. A particular choice of utility function can result optimally in the sharing of beliefs [22].

impossible for a player to compute it in full extent, thus players content themselves to calculate up to a few moves ahead before choosing the next move. Motivated by these arguments, we motivate the concept of non-Bayesian social learning in the next sections.

1.2.2 Non-Bayesian Social Learning

The study of non-Bayesian social learning is motivated by the understanding that social interactions are an important aspect of decision-making, but they do not necessarily happen in a fully rational manner. Some of the earlier attempts to model social interactions were based on a non-Bayesian paradigm, i.e., they employed heuristic combination protocols to solve the opinion pooling problem. One such example can be found in the social experiment described in [2], where people at a fair were asked to guess the weight of an ox. The surprising result was that, while individual estimates varied, the median value from 787 guesses approached the true weight of the animal. The success of aggregating estimates in this experiment reinforced the idea of the *wisdom of the crowd*, where a collective of agents could combine opinions to reach a more robust conclusion.

One of the pioneering opinion pooling strategies was proposed in [31], in which agents start from initial opinions $\mu_{k,0}(\theta)$ regarding an unknown parameter θ and proceed to update their opinions over time (indexed by i) using a linear pooling operation:

$$\mu_{k,i}(\theta) = \sum_{\ell=1}^{K} a_{\ell k}\mu_{\ell,i-1}(\theta), \tag{1.49}$$

where the weights $a_{\ell k}$ are nonnegative and add up to 1:

$$a_{\ell k} \geq 0, \quad \sum_{\ell=1}^{K} a_{\ell k} = 1, \quad a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k. \tag{1.50}$$

Here, the notation $\mathcal{N}_k$ denotes the set of neighbors of k. The weights can be assembled into a matrix $A = [a_{\ell k}]$. Assume the following conditions hold for matrix A:

$$A^{\mathsf{T}}\mathbb{1} = \mathbb{1} \tag{1.51}$$

$$\rho\Big(A^{\mathsf{T}} - \mathbb{1}\pi^{\mathsf{T}}\Big) < 1, \tag{1.52}$$

where π is the right eigenvector of A associated with eigenvalue 1, whose elements are positive and add up to 1, and where $\rho(X)$ denotes the spectral radius of matrix X. Note that the first condition means that the entries on each column of A add up to one. It is obvious that every such matrix has an eigenvalue at 1. Under the above conditions, it is known that as time grows agents reach *consensus* around the *weighted* average of the initial opinion vector [31]–[33]:

$$\lim_{i\to\infty} \mu_{k,i}(\theta) = \sum_{\ell=1}^{K} \pi_\ell \mu_{\ell,0}(\theta) \tag{1.53}$$

for all $k = 1, 2, \ldots, K$, whose weights π_k—elements of vector π—determine the amount of

weight given to the initial opinion at each agent k.

Many non-Bayesian social learning strategies have been subsequently inspired by this consensus result. We will illustrate in the following some non-Bayesian strategies that allow the network to learn from *streaming data* in a *synchronous* manner. In a synchronous distributed operation, all agents process information and exchange beliefs with neighboring agents simultaneously.

Arithmetic-Average Combination

Inspired by the single-agent non-Bayesian update proposed in [34] and by consensus-type aggregation procedures [35], [36], the authors of [4] proposed an iterative multi-agent strategy, wherein at every instant i, two steps are performed by each agent k: **i)** In view of the new private observation $\boldsymbol{\xi}_{k,i}$, the agent updates its belief using a local Bayesian update and a local set of likelihood models, $L_k(\xi|\theta)$ for $\theta \in \Theta$, resulting in an intermediate belief $\boldsymbol{\psi}_{k,i}(\theta)$; **ii)** The agent combines the result of the first step, namely, its own intermediate belief, with the *past* beliefs of neighboring agents using a weighted arithmetic average operation. The resulting algorithm is given by:

$$\boldsymbol{\psi}_{k,i}(\theta) \propto L_k(\boldsymbol{\xi}_{k,i}|\theta)\boldsymbol{\mu}_{k,i-1}(\theta) \qquad \text{(Bayesian update)} \tag{1.54}$$

$$\boldsymbol{\mu}_{k,i}(\theta) = a_{kk}\boldsymbol{\psi}_{k,i}(\theta) + \sum_{\substack{\ell=1 \\ \ell\neq k}}^{K} a_{\ell k}\boldsymbol{\mu}_{\ell,i-1}(\theta) \qquad \text{(combination rule)} \tag{1.55}$$

The symbol $\propto$ in the first line means that the entries of $\boldsymbol{\psi}_{k,i}(\theta)$ are normalized to add up to one, as is characteristic of belief distributions. The scalar $a_{\ell k}$ is a nonnegative weight used by agent k to scale the belief received from ℓ, such that $\sum_{\ell=1}^{K} a_{\ell k} = 1$, with K being the number of agents in the network. If ℓ is not a neighbor of k, the weight $a_{\ell k}$ is set to zero. These weights represent a trust score given to the information received from neighboring agents and reveal the underlying network topology. In this case, the network is assumed to be strongly connected, i.e., there exists a communication path between every two agents and at least one agents possesses a self-loop—more details are discussed in Chapter 2.

In [4], the observations $\boldsymbol{\xi}_{k,i}$ arise from *stationary* world conditions and are distributed according to one of the likelihoods $L_k(\xi|\theta_0)$ for some hypothesis $\theta_0 \in \Theta$, which denotes the true state of the world. The fact that the true distribution of observations is fixed over time qualifies the world as stationary. The main result of [4] states that, using the protocol (1.54)–(1.55), agents in a strongly connected network are able to recover the truth almost surely as i grows. More precisely, they show that

$$\boldsymbol{\mu}_{k,i}(\theta_0) \xrightarrow{\text{a.s.}} 1, \tag{1.56}$$

for all agents. This result shows that agents are able to recover the truth with strong convergence guarantees under *limited* knowledge about data dependencies among agents, i.e., agents do not know the underlying joint true model of observations. They also do not know the global network structure, i.e., they only interact with their local neighborhood, or the statistical models used by other agents in the network.

Algorithm (1.54)–(1.55) treats beliefs asymmetrically as is evident from (1.55). Observe that the right-hand side of (1.55) involves a combination of past beliefs, represented by $\boldsymbol{\mu}_{\ell,i-1}(\theta)$, and

one updated belief represented by $\boldsymbol{\psi}_{k,i}(\theta)$. The asymmetry in consensus-type implementation of this form has been shown to lead to degraded performance in the context of learning over networks [36], [37]. Motivated by the superior performance and stability of *diffusion* strategies for learning over networks [36], [38], the authors in [5] proposed the following alternative *diffusion* social learning strategy:

$$\boldsymbol{\psi}_{k,i}(\theta) \propto L_k(\boldsymbol{\xi}_{k,i}|\theta)\boldsymbol{\mu}_{\ell,i-1}(\theta) \tag{1.57}$$

$$\boldsymbol{\mu}_{k,i}(\theta) = \sum_{\ell=1}^{K} a_{\ell k}\boldsymbol{\psi}_{\ell,i}(\theta) \tag{1.58}$$

Observe from (1.58) that the diffusion implementation relies solely on the updated beliefs. The results in [5] establish asymptotic truth learning similar to (1.56), however with an improved rate of convergence with respect to the consensus-type implementation in [4].

The implementation (1.57)–(1.58) corresponds to what is known as the Adapt-Then-Combine (ATC) form of diffusion [36], [37]. One can also consider a Combine-Then-Adapt (CTA) form where the Bayesian and combination steps are reversed:

$$\boldsymbol{\psi}_{k,i-1}(\theta) = \sum_{\ell=1}^{K} a_{\ell k}\boldsymbol{\mu}_{\ell,i-1}(\theta) \tag{1.59}$$

$$\boldsymbol{\mu}_{k,i}(\theta) \propto L_k(\boldsymbol{\xi}_{k,i}|\theta)\boldsymbol{\psi}_{k,i-1}(\theta) \tag{1.60}$$

Geometric-Average Combination

The consensus and diffusion social learning strategies (1.54)–(1.55) and (1.57)–(1.58) combine the belief vectors in their second equations in order to propagate the $\boldsymbol{\mu}_{k,i}$. An alternative approach is to combine the logarithmic values of the belief entries, rather than the belief entries themselves. Since the belief entries are nonnegative and constrained to add up to one, it follows that the logarithms of beliefs are unconstrained real values. Motivated by these considerations, subsequent works have focused on arithmetic averaging in the *logarithmic* domain, which corresponds to geometric averaging in the original domain [39]–[42]. For instance, the social learning protocol proposed by [40] takes the following geometric CTA form:

$$\boldsymbol{\psi}_{k,i-1}(\theta) \propto \prod_{\ell=1}^{K} \Big(\boldsymbol{\mu}_{\ell,i-1}(\theta)\Big)^{a_{\ell k}} \tag{1.61}$$

$$\boldsymbol{\mu}_{k,i}(\theta) \propto L_k(\boldsymbol{\xi}_{k,i}|\theta)\boldsymbol{\psi}_{\ell,i-1}(\theta) \tag{1.62}$$

where $a_{\ell k}$ is again a nonnegative weight given by each agent k to neighbor ℓ, such that $\sum_{\ell=1}^{K} a_{\ell k} = 1$. The network topology is again assumed to be strongly connected.

The corresponding ATC form appears in [41], namely,

$$\boldsymbol{\psi}_{k,i}(\theta) \propto L_k(\boldsymbol{\xi}_{k,i}|\theta)\boldsymbol{\mu}_{\ell,i-1}(\theta) \tag{1.63}$$

$$\boldsymbol{\mu}_{k,i}(\theta) \propto \prod_{\ell=1}^{K} \Big(\boldsymbol{\psi}_{\ell,i}(\theta)\Big)^{a_{\ell k}} \tag{1.64}$$

In the logarithm domain, these updates can be rewritten as linear operations. To see that, define the following auxiliary variables for any distinct pair $\theta, \theta' \in \Theta$:

$$\boldsymbol{\lambda}_{k,i} \triangleq \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')}, \quad \boldsymbol{\eta}_{k,i} \triangleq \log \frac{\boldsymbol{\psi}_{k,i}(\theta)}{\boldsymbol{\psi}_{k,i}(\theta')}, \quad \boldsymbol{x}_{k,i} \triangleq \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta)}{L_k(\boldsymbol{\xi}_{k,i}|\theta')}, \tag{1.65}$$

whose dependence on $\{\theta, \theta'\}$ is omitted. Next, let us focus on (1.61)–(1.62). Computing the ratio of beliefs between any two hypotheses θ and θ', applying the log operation, and using the definitions in (1.65), we can rewrite (1.61)–(1.62) as a diffusion strategy with respect to the iterate $\boldsymbol{\lambda}_{k,i}$ [37], [38]:

$$\boldsymbol{\eta}_{k,i-1} = \sum_{\ell=1}^{K} a_{\ell k} \boldsymbol{\lambda}_{\ell,i-1} \tag{1.66}$$

$$\boldsymbol{\lambda}_{k,i} = \boldsymbol{\eta}_{k,i-1} + \boldsymbol{x}_{k,i} \tag{1.67}$$

where agents first *combine* their log-ratio of beliefs $\boldsymbol{\lambda}_{k,i}$ with neighbors and then *adapt* them using the new information in the form of the log-ratio of likelihoods $\boldsymbol{x}_{k,i}$. Similarly, the strategy in (1.63)–(1.64) can be rewritten according to:

$$\boldsymbol{\eta}_{k,i} = \boldsymbol{\lambda}_{k,i-1} + \boldsymbol{x}_{k,i} \tag{1.68}$$

$$\boldsymbol{\lambda}_{k,i} = \sum_{\ell=1}^{K} a_{\ell k} \boldsymbol{\eta}_{\ell,i} \tag{1.69}$$

where agents first *adapt* their log-ratio of beliefs using new information and then *combine* them with neighbors.

Both strategies employ similar steps, but in different orders. The algorithm in (1.66) and (1.67) takes the form of a CTA rule, while the one in (1.68) and (1.69) takes the form of an ATC rule. In the context of learning over networks, such as in the least-mean-squares problem described in [38], ATC strategies are known to yield superior steady-state performance.

As in the study of the earlier social learning strategies based on arithmetic averaging, the world conditions are generally assumed to be *stationary* [40], [41]. Moreover, in [41] observations are distributed according to one of the likelihoods $L_k(\xi|\theta_0)$ for some true hypothesis $\theta_0 \in \Theta$. The convergence result from [41] shows asymptotic truth learning at an exponential rate. In [40], the true model generating the observations does not necessarily belong to the set of likelihood models, therefore resulting in more diversified convergence behavior.

The choice of using geometric- or arithmetic-average pooling is motivated in [43] by following an axiomatic approach. Based on some behavioral assumptions, one form or the other may follow as the preferred implementation for non-Bayesian social learning. The geometric-average implementation can also be motivated as a distributed stochastic mirror descent solution using a variational interpretation [44]. In terms of rates of convergence, the works [40], [41] show that empirically the geometric-average rule converges faster than the arithmetic-average rule. This result is confirmed theoretically by the recent work [45].

All results described in Section 1.2.2 concern *strongly connected* networks. A variation of this

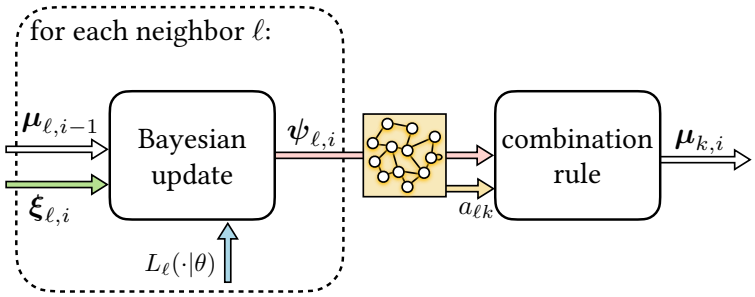


Figure 1.2: Block diagram of social learning under streaming observations. Bayesian update takes as input the previous belief, observations of the world, likelihood models, and outputs the intermediate beliefs. The combination rule takes as input the intermediate beliefs shared by neighbors and network weights, and outputs the updated belief.

communication strategy is considered in [46], where agents choose at random one neighbor to exchange beliefs with at each time. Similar performance guarantees can be achieved in steady-state with this sparser communication scheme. Other interesting extensions for *weakly connected* networks [47] can be found in [48] and [49]. In these works, we observe that parts of the network can exert full control over the belief convergence of the remaining agents—this phenomenon will be detailed in Chapter 2 when we discuss weakly connected networks. In [44], [50], the authors extend the space of hypotheses to a continuum space Θ.

Throughout this thesis we focus on the social learning formulation in the ATC form described by (1.63) and (1.64) and variations thereof.

1.3 From Models to the World

We represent the ATC form of social learning in block diagram in Figure 1.2. It is clear from the figure that the two main building blocks of social learning are the "Bayesian update" and the "combination rule". The first block is motivated by the desire to characterize agents as locally rational, meaning that every observation is incorporated into the belief in a Bayesian way. The second block allows flexibility in choosing the type of opinion pooling operation. With these two blocks fixed, the social learning framework consists of four building elements—highlighted in different colors in Figure 1.2. These elements are described below from the lowest to the highest level of abstraction:

1. *Models:* At the lowest level of the social learning framework, the models of an agent quantify how it perceives the observations of the world. These models exist in the form of *likelihood models* for the observations given different hypothetical states of the world and are used by agents to perform their local Bayesian update. The exactness of these models in representing the observations given different states has an important impact on the belief evolution of agents.

2. *Exchanged information:* The exchange of information between neighbors enables agents to solve the inference problem collaboratively and thus to overcome the limitations of their individual models. This is achieved through the exchange of *intermediate beliefs*

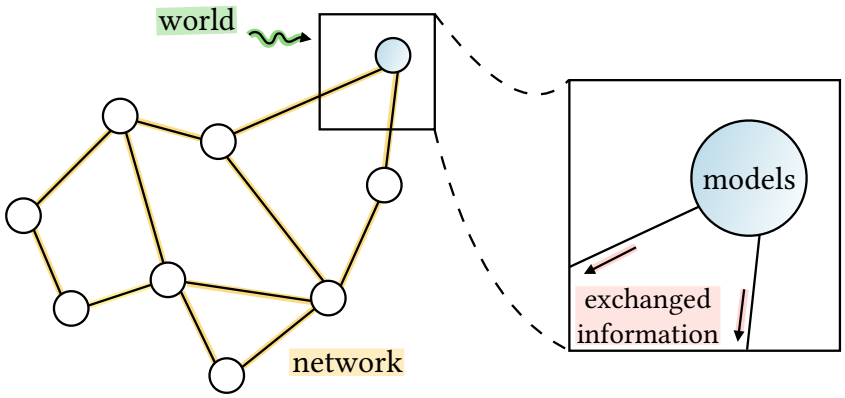


Figure 1.3: Alternative view of the social learning framework, with the four building elements highlighted in color: Models, exchanged information, network and world.

between neighboring agents, but could also be designed to account for other types of information, as long as the observations remain private.

3. *Network:* The network structure determines the communication paths and the flow of the exchanged information across agents. It consists of the underlying *inter-agent connectivity*, e.g., whether the network is strongly connected, connected, or weakly connected, and the weights given by each agent to the information received from its neighbors. Different connectivity and weight patterns give rise to influence dynamics and richer belief formation scenarios.

4. *World:* At the highest level of social learning, we have the effect of the world on the network. This effect is represented by the *observations* or measurements of the world, which are received by agents in a private manner. The quality of these measurements, the way they are distributed across agents and over time are of utmost importance for the learning outcome.

A complementary view of the social learning framework can be seen in Figure 1.3, where the aforementioned four elements are highlighted in different colors. The individual models, the exchanged information among agents, the network topology, and the surrounding world constitute the key elements of learning over adaptive networks. The remainder of this research work is dedicated to investigating and exploiting different aspects of these elements.

1.4 Outline and Main Contributions

The objective of this thesis is to exploit each of the four social learning elements—highlighted in Figure 1.3—to better understand the features and limits of social learning strategies and to address existing issues in their formulation and bring them closer to real-world settings. At the end of this research work, we will be able to answer questions regarding the influence between different components of the network and questions concerning the performance of social learning in view of the exchange of partial information, nonstationary world conditions, and imperfectly trained models.

As detailed in Section 1.2.2, an important aspect of social learning is the existence of a source of *streaming* observations [4], [5]. This feature allows for real-time measurement and processing of information, and thus has attracted the interest of engineers and economists toward a strategy that can continually *learn over time*. The drawback is that social learning is designed to operate in a *stationary* environment, that is, where the conditions of the world do not change over time—a phenomenon that will be thoroughly explained in Chapter 5. This situation is neither realistic from a social network point of view nor desirable in real-time processing systems. For example, the conditions of the financial market are always shifting around decision-making agents. Another example is a network of meteorological sensors trying to infer the current weather state, which must *adapt* and provide reliable results over time. We consider the study of social learning in a *nonstationary world* essential to bridge the gap between theoretical models and applications, and we believe that such strategies should not only learn but also continually *adapt over time*.

For that reason, we choose to organize this thesis in two parts. Part I, entitled "Stationary World", which addresses two contributions under the assumption of stationary conditions. Part II, entitled "Non-Stationary World", which addresses two contributions suitable for nonstationary conditions. Before delving into the research contributions of this work, we formally introduce the social learning problem and its mathematical notation in Chapter 2 and review some of the available results in the literature for both strongly and weakly connected networks and provide essential results on their asymptotic belief convergence. In Chapter 2, the reader will also find simulation examples to illustrate the behavior of these strategies as well as a detailed proof of the belief convergence for strongly connected networks.

1.4.1 Part I: Stationary World

Part I contains the following contributions under a stationary environment.

Chapter 3: Recovering Influences in Weak Graphs
We consider the role of the *network influence* in social learning. More specifically, we model the social network as a weakly connected network consisting of a receiving subnetwork and multiple sending subnetworks, which are described in Chapter 3. In view of existing results, which show that sending agents control the beliefs of the receiving agents [49], we address the reverse learning problem, i.e., to estimate the amount of influence that each sending network exerts on the beliefs of receiving agents [42], [51]. We establish sufficient and necessary conditions for the reverse learning problem to be feasible. Our analysis reveals that the reverse learning problem can be solved if there exists a sufficient degree of diversity in the statistical models of the sending sub-networks. The discussion and results in this chapter are useful in describing influence patterns over weakly connected social networks.

Chapter 4: Exchange of Partial Information
In this chapter, we consider the role of the *exchange of partial information* in social learning. In other words, instead of exchanging the entirety of their beliefs, this chapter assumes that communication among agents is constrained. Agents only exchange their confidence regarding

one hypothesis of interest, reflecting a desire to retain part of its private knowledge for reasons such as social dynamics, limited bandwidth, or regulation. The goal of the network is to ascertain the validity of said hypothesis. We propose two approaches for sharing partial information, depending on whether agents behave in a self-aware manner or not. The results show how different learning regimes arise, depending on the approach employed and on the inherent characteristics of the inference problem.

1.4.2 Part II: Non-Stationary World

Part II contains the following contributions under a nonstationary environment.

Chapter 5: Adaptive Social Networks
In this chapter, we consider the role of a *nonstationary world* in social learning. Although tailored for working with streaming observations, social learning strategies do not perform well under nonstationary conditions. To address this issue, we propose in this chapter an Adaptive Social Learning strategy, which relies on a small step-size parameter to tune the adaptation degree. We provide a detailed characterization of the learning performance, namely, the probability of making a wrong inference, during both steady-state and transient phases. We show that the step-size parameter plays a key role in determining the trade-off between adaptation and learning accuracy, and study its influence on the adaptation time during the transient phase. Our conclusions are key to enable social learning to be used in actual online learning settings.

Chapter 6: Learning with Imperfect Models
In this chapter, we consider the role of *imperfect models* in social learning. Traditional social learning strategies rely on the assumption that each agent has significant knowledge of the underlying models of the observations. In this chapter we overcome this issue by introducing a machine learning framework, referred to as Social Machine Learning (SML), which involves a training phase, where the models are trained from finite data, and a prediction phase, where these imperfectly trained models are deployed in a collaborative manner, inspired by social learning strategies, to classify unlabeled observations. We show that the SML strategy enables the agents to learn consistently under a highly-heterogeneous setting and allows the network to continue improve performance during the prediction phase. These results allow the social learning mechanism to be used as a fully data-based solution in realistic learning tasks.

Finally, Chapter 7 summarizes the main contributions of this work and suggests future research directions.

2 The Social Learning Model

In this chapter, in order to place our contributions in context, we first review some of the available results on the convergence behavior of the diffusion social learning algorithm described in Section 1.2.2. The network is assumed to exist in a *stationary world*, that is, observations $\boldsymbol{\xi}_{k,i}$ are distributed according to a fixed agent-dependent distribution $f_k(\xi)$:

$$\boldsymbol{\xi}_{k,i} \sim f_k \tag{2.1}$$

We refer to $f_k(\xi)$ as the *true model* of agent k with support $\mathcal{X}_k$. Observations $\boldsymbol{\xi}_{k,i}$ are iid over time, i.e., over i, but can be dependent across agents, i.e., across k. They can be either continuous or discrete random variables, in which cases $f_k(\xi)$ is either a pdf or a pmf.

The goal of the network of agents is to discover the state of the world from a finite set of H discrete hypotheses denoted by $\Theta \triangleq \{1, 2, \ldots, H\}$. With each hypothesis, agents associate *likelihood models*, which act as candidate models for the unknown true model. More precisely, for each θ, agent k possess a likelihood model $L_k(\xi|\theta)$, which is a pdf (or pmf if the observations are of discrete nature) when seen as a function of ξ. We stress that f_k does not need to belong to the set of likelihood models available at agent k, namely, the set of functions $L_k(\xi|\theta)$ for $\theta \in \Theta$.

The social learning algorithm based on geometric ATC diffusion is reproduced here for ease of reference:

$$\boldsymbol{\psi}_{k,i}(\theta) = \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta)\boldsymbol{\mu}_{k,i-1}(\theta)}{\sum_{\theta' \in \Theta} L_k(\boldsymbol{\xi}_{k,i}|\theta')\boldsymbol{\mu}_{k,i-1}(\theta')} \tag{2.2}$$

$$\boldsymbol{\mu}_{k,i}(\theta) = \frac{\prod_{\ell \in \mathcal{N}_k} [\boldsymbol{\psi}_{\ell,i}(\theta)]^{a_{\ell k}}}{\sum_{\theta' \in \Theta} \prod_{\ell \in \mathcal{N}_k} [\boldsymbol{\psi}_{\ell,i}(\theta)]^{a_{\ell k}}} \tag{2.3}$$

where weights $a_{\ell k}$ are the elements of a left-stochastic combination matrix A, associated with the underlying communication graph. In this chapter, we report the convergence behavior of the algorithm in the context of *strongly* and *weakly connected networks*. We will see that while the former promotes *consensus* across agents, the latter stimulates *disagreement* and *influence dynamics* within the network.

2.1 Reaching Consensus

From classical arguments such as the *law of large numbers* and the *wisdom of the crowds*, we have confidence that the average of many samples can serve as a good estimator for some parameter or variable of interest. For example, pre-election polls provide average results over limited samples that indicate the ongoing state of elections, and statistical sampling theory tells us how close this average is from the yet-unknown true result.

The concept of averaging is ubiquitous in distributed processing. It allows systems to achieve *agreement* around a more reliable estimate by combining local opinions and estimates. The use of *consensus*-based strategies can be traced back to the works [31], [32], where authors dealt with the problem of averaging estimates over graphs. Since then, several works [37], [52]–[57] in the areas of distributed optimization and estimation consider more general formulations and broader contexts. In social learning, averaging is present in the combination step, i.e., Eq. (2.3), where it takes the form of a weighted geometric average. In the next sections, we describe how consensus is relevant in the context of social learning.

2.1.1 Strongly Connected Networks

We consider a strongly connected network with K agents. A network is said to be strongly connected if there exists a path linking any pair of agents in *both* directions and if at least one agent possesses a self-loop [37]. A diagram of a strongly connected network can be seen in Figure 2.1.

The network can be mathematically represented by a graph, where its nodes are agents and edges represent communication links, which can be directed or not. A directed edge from agent ℓ to agent k indicates that agent ℓ can send information to agent k (equivalently, agent k receives information from agent ℓ). In this case, agent ℓ is said to be a *neighbor* of agent k. Each agent k attributes a nonnegative weight (confidence score) to the information received from its neighbor ℓ, namely, $a_{\ell k} \in (0, 1]$, such that

$$\sum_{\ell=1}^{K} a_{\ell k} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} = 1, \tag{2.4}$$

where we denote the set of neighbors of k by $\mathcal{N}_k$. We can therefore define $A \triangleq [a_{\ell k}]$ as the *combination matrix* associated with our graph. From (2.4), the entries on each column of A should add up to one, such that A is left stochastic, i.e.,

$$A^{\mathsf{T}}\mathbb{1} = \mathbb{1}, \quad a_{\ell k} > 0 \Rightarrow \ell \in \mathcal{N}_k. \tag{2.5}$$

Since the network is strongly connected, the combination matrix is a left-stochastic primitive matrix, i.e., there exists some finite integer n_o, such that the n_o-power of A has strictly positive entries, i.e., $A^{n_o} \succ 0$, where $\succ$ denotes element-wise inequality. From the Perron-Frobenius theorem [37], [58], these properties ensure that the spectral radius of A is equal to 1, i.e., $\rho(A) = 1$ and that the eigenvalue 1 is simple. Moreover, we can associate with the eigenvalue

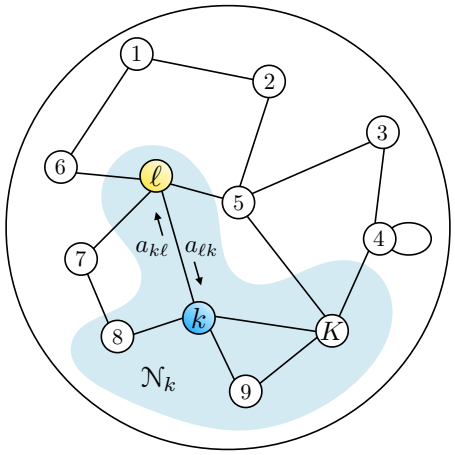


Figure 2.1: Network with K agents with the set of neighbors $\mathcal{N}_k$ highlighted in blue.

at 1 an eigenvector denoted by π whose entries are positive and add up to 1, namely,

$$A\pi = \pi, \quad \pi \succ 0, \quad \pi^{\mathsf{T}}\mathbb{1} = 1. \tag{2.6}$$

We refer to π as the Perron vector. If the combination matrix is furthermore right-stochastic, that is, the entries on each row of A add up to one as well, then we say that A is *doubly stochastic*. As a result, the associated Perron eigenvector will be a uniform vector, that is,

$$A^{\mathsf{T}}\mathbb{1} = \mathbb{1} \quad \text{and} \quad A\mathbb{1} = \mathbb{1} \quad \Rightarrow \quad \pi = \frac{1}{K}\mathbb{1}. \tag{2.7}$$

The aforementioned properties imply that the columns of the matrix powers A^m converge, as $m \to \infty$, to the Perron eigenvector at an exponential rate governed by the second largest-magnitude eigenvalue of A, as stated in the following property [59].

Property 2.1 (Convergence of matrix powers). Let A be a left-stochastic matrix, where its second largest-magnitude eigenvalue is denoted by β_2. Then, for any positive β such that $|\beta_2| < \beta < 1$, there exists a positive constant κ (depending only on A and β), such that, for all $\ell, k = 1, 2, \ldots, K$, and for all $m = 1, 2, \ldots$, we have that:

$$\left| [A^m]_{\ell k} - \pi_\ell \right| \leq \kappa \beta^m. \tag{2.8}$$

2.1.2 Convergence Behavior

Social learning is first and foremost a collaborative effort to discover which hypothesis, belonging to the set Θ, provides the best explanation for the observations $\boldsymbol{\xi}_{k,i}$. This is carried out by seeking which likelihood $L_k(\xi|\theta)$ provides the best approximation for the unknown true distribution $f_k(\xi)$ on *average* across the network.

For this purpose, we will use the concept of KL divergence [16] as a measure of dissimilarity

between the true model $f_k(\xi)$ and the likelihood $L_k(\xi|\theta)$:

$$D_k(f_k||L_k(\theta)) = \mathbb{E}_{f_k}\left(\log \frac{f_k(\boldsymbol{\xi})}{L_k(\boldsymbol{\xi}|\theta)}\right), \tag{2.9}$$

where $\mathbb{E}_{f_k}$ denotes the expectation computed with respect to distribution $f_k(\xi)$. Note that we are dropping the argument ξ from the left-hand side for simplicity of notation. The KL divergence will be a recurring instrument throughout our analysis and results. To ensure that these quantities are well posed, we introduce the following condition.

Assumption 2.1 (Finite KL divergences). We assume that, for all $k = 1, 2, \ldots, K$, and all $\theta \in \Theta$:

$$D(f_k||L_k(\theta)) < \infty. \tag{2.10}$$

Assumption 2.1 implies that the support of the true model $f_k(\xi)$, i.e., the range of values over which $f_k(\xi)$ is strictly positive, is contained in the support of each likelihood $L_k(\xi|\theta)$.

We further assume that at the initial time $i = 0$ agents have no reason to discard any hypothesis, and therefore they should have positive initial beliefs across all hypotheses.

Assumption 2.2 (Positive initial beliefs). We assume that, for all $k = 1, 2, \ldots, K$, and all $\theta \in \Theta$, $\mu_{k,0}(\theta) > 0$[1].

Note that, in view of (2.2), if $\boldsymbol{\mu}_{k,i-1}(\theta) > 0$, then $\boldsymbol{\psi}_{k,i}(\theta) > 0$ since $L_k(\boldsymbol{\xi}_{k,i}|\theta) > 0$ (since the support of each likelihood contains the support of the true model). Thus, $\boldsymbol{\mu}_{k,i}(\theta)$ is strictly positive, in view of (2.3), since the combination weights are nonnegative. And, more generally, from Assumption 2.2, $\boldsymbol{\mu}_{k,i}(\theta) > 0$ for all $i = 1, 2, \ldots$.

We also introduce the following classical identifiability condition meant to avoid the confusion state described in Section 1.1.1, where the beliefs of an agent do not converge but can wander around randomly over time. Let the *network KL divergence* be defined as the weighted quantity:

$$D(\theta) \triangleq \sum_{\ell=1}^{K} \pi_\ell D(f_\ell||L_\ell(\theta)). \tag{2.11}$$

Assumption 2.3 (Unique minimizer in strongly connected graphs). The function $D(\theta)$ has a unique minimizer:

$$\theta^\star \triangleq \operatorname*{argmin}_{\theta\in\Theta} D(\theta). \tag{2.12}$$

We state next a theorem characterizing the convergence of beliefs within strongly connected networks. Similar results are found in [40] and [43] for the Combine-Then-Adapt version of

[1]The initial beliefs can also be taken as random quantities assumed to be independent of the observations at subsequent instants, in which case it is denoted by $\boldsymbol{\mu}_{k,0}(\theta)$. This setting will appear further ahead in Chapter 5.

the social learning algorithm. We choose to report here an alternative result and its proof, for the Adapt-Then-Combine version (2.2)–(2.3), as we find the arguments to be material to understanding the inner workings of social learning.

Theorem 2.1 (Belief convergence in strongly connected networks). *Consider a strongly connected network with a left-stochastic combination matrix and Perron eigenvector π. Under Assumptions 2.1, 2.2, and 2.3, the beliefs of strategy* (2.2)–(2.3) *converge almost surely to the unique minimizer $\theta^\star$ of the network divergence* (2.11)*, namely,*

$$\boldsymbol{\mu}_{k,i}(\theta^\star) \xrightarrow{\text{a.s.}} 1. \tag{2.13}$$

Proof. In view of (2.2), we can write, for any distinct pair $\theta, \theta' \in \Theta$,

$$\log \frac{\boldsymbol{\psi}_{k,i}(\theta)}{\boldsymbol{\psi}_{k,i}(\theta')} = \log \frac{\boldsymbol{\mu}_{k,i-1}(\theta)}{\boldsymbol{\mu}_{k,i-1}(\theta')} + \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta)}{L_k(\boldsymbol{\xi}_{k,i}|\theta')}, \tag{2.14}$$

and similarly, from (2.3), we can write

$$\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')} = \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\psi}_{\ell,i}(\theta)}{\boldsymbol{\psi}_{\ell,i}(\theta')}. \tag{2.15}$$

Replacing (2.14) into (2.15) yields a recursion in terms of *log-ratio* quantities:

$$\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')} = \sum_{\ell=1}^{K} a_{\ell k} \left(\log \frac{\boldsymbol{\mu}_{\ell,i-1}(\theta)}{\boldsymbol{\mu}_{\ell,i-1}(\theta')} + \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta')} \right). \tag{2.16}$$

Iterating the recursion over i, and dividing the resulting expression by i, allows us to write:

$$\frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')} = \frac{1}{i} \sum_{\ell=1}^{K} [A^i]_{\ell k} \log \frac{\mu_{\ell,0}(\theta)}{\mu_{\ell,0}(\theta')} + \frac{1}{i} \sum_{m=1}^{i} \sum_{\ell=1}^{K} [A^m]_{\ell k} \underbrace{\log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i-m+1}|\theta)}{L_\ell(\boldsymbol{\xi}_{\ell,i-m+1}|\theta')}}_{\triangleq \boldsymbol{x}_{\ell,i-m+1}}. \tag{2.17}$$

From Assumption 2.2, the first term on the RHS of (2.17) converges to 0, as i goes to infinity. The second term can be split in two terms:

$$\begin{aligned} \frac{1}{i} \sum_{m=1}^{i} \sum_{\ell=1}^{K} [A^m]_{\ell k} \boldsymbol{x}_{\ell,i-m+1} &= \frac{1}{i} \sum_{m=1}^{i} \sum_{\ell=1}^{K} \Big([A^m]_{\ell k} - \pi_\ell \Big) \boldsymbol{x}_{\ell,i-m+1} \\ &\quad + \frac{1}{i} \sum_{m=1}^{i} \sum_{\ell=1}^{K} \pi_\ell \boldsymbol{x}_{\ell,i-m+1}. \end{aligned} \tag{2.18}$$

We establish in the remainder of the proof the following claims: The first term on the RHS of (2.18) vanishes while the second term converges to a finite value.

1. First term on the RHS of (2.18) **vanishes**

We verify that the first term on the RHS of (2.18) goes almost surely to 0, using the property of

convergence of powers of A. Since A is a left-stochastic strongly connected matrix, from the Perron-Frobenius theorem, we have the following limit:

$$\lim_{m\to\infty} A^m = \pi \mathbb{1}^\mathsf{T}, \tag{2.19}$$

which implies that for some $\varepsilon > 0$, there exists an index i_0 such that for $m > i_0$,

$$|[A^m]_{\ell k} - \pi_\ell| < \epsilon. \tag{2.20}$$

We split the first term on the RHS of (2.18) into two summations:

$$\begin{aligned}\sum_{m=1}^{i}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k} - \pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1} = &\sum_{m=1}^{i_0-1}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k} - \pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1} \\ &+ \sum_{m=i_0}^{i}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k} - \pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1}.\end{aligned} \tag{2.21}$$

We take the absolute value of (2.21), divide it by i, and use the triangle inequality to write:

$$\begin{aligned}\frac{1}{i}\left|\sum_{m=1}^{i}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k} - \pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1}\right| \le &\frac{1}{i}\left|\sum_{m=1}^{i_0-1}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k} - \pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1}\right| \\ &+ \frac{1}{i}\left|\sum_{m=i_0}^{i}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k} - \pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1}\right|.\end{aligned} \tag{2.22}$$

The first term on the RHS of (2.22) can be upper bounded in the following manner:

$$\frac{1}{i}\left|\sum_{m=1}^{i_0-1}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k} - \pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1}\right| \overset{\text{(a)}}{\le} \frac{1}{i}\sum_{m=1}^{i_0-1}\sum_{\ell=1}^{K}|[A^m]_{\ell k} - \pi_\ell|\,|\boldsymbol{x}_{\ell,i-m+1}| \tag{2.23}$$

where (a) follows from the triangle inequality. Since A is left stochastic, we have that

$$|[A^m]_{\ell k} - \pi_\ell| \le 1 \tag{2.24}$$

which implies that:

$$\begin{aligned}\frac{1}{i}\left|\sum_{m=1}^{i_0-1}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k} - \pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1}\right| &\le \frac{1}{i}\sum_{m=1}^{i_0-1}\sum_{\ell=1}^{K}|\boldsymbol{x}_{\ell,i-m+1}| \\ &\overset{\text{(a)}}{=} \sum_{\ell=1}^{K}\sum_{j=i-i_0}^{i}\frac{|\boldsymbol{x}_{\ell,j}|}{i}\end{aligned} \tag{2.25}$$

where in (a) the terms in the summation are reordered. From[2]

$$\frac{\boldsymbol{x}_{\ell,j}}{i} \xrightarrow{\text{a.s.}} 0, \tag{2.27}$$

we have that the inner summands on the RHS of (2.25) vanish a.s., which implies that

$$\frac{1}{i}\left|\sum_{m=1}^{i_0-1}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k}-\pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1}\right| \xrightarrow{\text{a.s.}} 0. \tag{2.28}$$

The second term on the RHS of (2.22) can be upper bounded by:

$$\frac{1}{i}\left|\sum_{m=i_0}^{i}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k}-\pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1}\right| \leq \frac{1}{i}\sum_{m=i_0}^{i}\sum_{\ell=1}^{K}\varepsilon\left|\boldsymbol{x}_{\ell,i-m+1}\right|, \tag{2.29}$$

where (a) follows from the triangle inequality and the property in (2.20). Therefore, we can bound (2.29) as:

$$\begin{aligned}\frac{1}{i}\left|\sum_{m=1}^{i}\sum_{\ell=1}^{K}\Big([A^m]_{\ell k}-\pi_\ell\Big)\boldsymbol{x}_{\ell,i-m+1}\right| &\overset{\text{(a)}}{\leq} \frac{1}{i}\sum_{m=i_0}^{i}\sum_{\ell=1}^{K}\varepsilon\left|\boldsymbol{x}_{\ell,i-m+1}\right| \\ &\xrightarrow{\text{a.s.}} 0,\end{aligned} \tag{2.30}$$

where the convergence to 0 follows from the arbitrariness of ε and the fact that[3]

$$\frac{1}{i}\sum_{m=i_0}^{i}\left|\boldsymbol{x}_{\ell,i-m+1}\right| \xrightarrow{\text{a.s.}} \mathbb{E}_{f_\ell}\left|\boldsymbol{x}_{\ell,i}\right| < \infty. \tag{2.32}$$

2. Second term on the RHS of (2.18) **converges**

In view of Assumption 2.1 and the iid property of observations $\boldsymbol{\xi}_{k,i}$ over time, we can use the strong law of large numbers [18] to establish that the second term on the RHS of (2.18) converges almost surely to its mean, i.e.,

$$\frac{1}{i}\sum_{m=1}^{i}\sum_{\ell=1}^{K}\pi_\ell\boldsymbol{x}_{\ell,i-m+1} \xrightarrow{\text{a.s.}} \sum_{\ell=1}^{K}\pi_\ell\mathbb{E}_{f_\ell}(\boldsymbol{x}_{\ell,i})$$

[2]This convergence can be seen by writing the following sum of i samples:

$$\frac{1}{i}\sum_{m=1}^{i}\boldsymbol{x}_{\ell,m} = \frac{\boldsymbol{x}_{\ell,1}}{i} + \frac{1}{i}\sum_{m=2}^{i}\boldsymbol{x}_{\ell,m} \tag{2.26}$$

and noticing that the term on the LHS and the second term on the RHS both converge a.s. to $\mathbb{E}\,\boldsymbol{x}_{\ell,m}$ from the strong law of large numbers (SLLN), implying that the remaining term vanishes.

[3]From the definition of $\boldsymbol{x}_{\ell,i}$ in (2.17), we have that:

$$\mathbb{E}_f\boldsymbol{x}_{\ell,i} = D(f_\ell||L_\ell(\theta')) - D(f_\ell||L_\ell(\theta)). \tag{2.31}$$

In view of Assumption 1.1, it follows that $\mathbb{E}_f\boldsymbol{x}_{\ell,i}$ exists and takes a finite value. Therefore $\boldsymbol{x}_{\ell,i}$ is an integrable random variable and the convergence follows from the SLLN.

$$= \sum_{\ell=1}^{K} \pi_\ell \Big(D(f_\ell || L_\ell(\theta')) - D(f_\ell || L_\ell(\theta)) \Big). \tag{2.33}$$

Using definition (2.11) and expressions (2.17)–(2.18) we conclude that

$$\frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')} \xrightarrow{\text{a.s.}} D(\theta') - D(\theta). \tag{2.34}$$

Recall that, by definition, $\theta^\star$ is the unique minimizer of the network KL divergence $D(\theta)$. Let us replace θ by $\theta^\star$ in (2.34). Then we have, for all $\theta' \neq \theta^\star$, that:

$$\frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta^\star)}{\boldsymbol{\mu}_{k,i}(\theta')} \xrightarrow{\text{a.s.}} D(\theta') - D(\theta^\star) > 0, \tag{2.35}$$

which implies that, for all $\theta' \neq \theta^\star$,

$$\log \frac{\boldsymbol{\mu}_{k,i}(\theta^\star)}{\boldsymbol{\mu}_{k,i}(\theta')} \xrightarrow{\text{a.s.}} +\infty \Rightarrow \boldsymbol{\mu}_{k,i}(\theta') \xrightarrow{\text{a.s.}} 0. \tag{2.36}$$

The proof is complete since the belief components should add up to 1. □

We therefore find that all agents reach *consensus* asymptotically around $\theta^\star$. We can compare this outcome with the single-agent case. Thus, consider a single-agent scenario where agent k evolves in isolation. Agent k would decide for the hypothesis θ that yields the best match between $L_k(\xi|\theta)$ and $f_k(\xi)$. This can be achieved by picking the hypothesis that minimizes $D(f_k||L_k(\theta))$ over $\theta \in \Theta$. In the social learning setup, agents pick the hypothesis that minimizes the *network* KL divergence $D(\theta)$, which is a weighted average of the individual KL divergences.

The minimization of the network KL divergence introduces some *robustness* into the final choice of the agents. To see this, consider a malfunctioning agent m in the network, whose divergence $D(f_\mathsf{m}||L_\mathsf{m}(\theta))$ is minimized by multiple hypotheses in the subset $\Theta^\mathsf{m} \subset \Theta$. At any other functioning agent $k \neq \mathsf{m}$, $D(f_\mathsf{k}||L_\mathsf{k}(\theta))$ is minimized at a unique hypothesis $\theta^\mathsf{f} \in \Theta$. If agent m evolved in isolation, according to the discussion in Section 1.1.1, it would not be able to decide for a single hypothesis and would remain confused among the hypotheses in Θ^m. In social learning, this can be avoided. Since agents choose the hypothesis minimizing the *network* KL divergence, the majority of functioning agents will steer the whole network, including agent m, toward the unique minimizer hypothesis θ^f.

2.1.3 Convergence Behavior under Objective Evidence

Consider now the case in which the unknown true distribution f_k coincides with one of the likelihoods pertaining to agent k, namely, $L_k(\xi|\theta_0)$ for all agents $k = 1, 2, \ldots, K$, where θ_0 belongs to Θ. In this case, we refer to θ_0 as the true state of the world, as it explains perfectly the nature of all observations collected across the network. In the scenario under *objective evidence*, we say that:

$$\boldsymbol{\xi}_{k,i} \sim L_k(\xi|\theta_0), \tag{2.37}$$

For ease of notation, under objective evidence we denote the KL divergence between the true likelihood $L_k(\xi|\theta_0)$ and any other likelihood $L_k(\xi|\theta)$ by

$$d_k(\theta) \triangleq D(L_k(\theta_0)||L_k(\theta)). \tag{2.38}$$

In this scenario, we propose the following modification to Assumption 2.1 regarding the finiteness of KL divergences.

Assumption 2.4 (Finite KL divergences). We assume that, for all $k = 1, 2, \ldots, K$, and each pair of distinct hypotheses $\theta, \theta' \in \Theta$:

$$D(L_k(\theta)||L_k(\theta')) < \infty. \tag{2.39}$$

This condition ensures that for any choice of $\theta_0 \in \Theta$, the KL divergences between the true likelihood $L_k(\xi|\theta_0)$ and any other likelihood $L(\xi|\theta')$, $\theta' \neq \theta_0$, are well posed. Assumption 2.4 can be relaxed to require that for any $\theta \neq \theta_0$:

$$D(L_k(\theta_0)||L_k(\theta)) < \infty \tag{2.40}$$

in a similar way as in Assumption 2.1. While in the general formulation described in Section 2.1.2, agents sought to identify the hypothesis and its corresponding likelihood that best explained the observations, under objective evidence, social learning is used as a means of learning the truth. The concept of truth learning corresponds to the capacity of each agent to concentrate their beliefs around the true hypothesis, as described in the following definition.

Definition 2.1.1 (Truth learning). We say that agent k running strategy (2.2)–(2.3) learns the truth when the following convergence behavior is observed:

$$\boldsymbol{\mu}_{k,i}(\theta_0) \xrightarrow{\text{a.s.}} 1. \tag{2.41}$$

In practical applications, it is possible that a certain agent cannot distinguish a hypothesis from the truth. In this case, when an agent k cannot distinguish θ from θ_0, it will hold that

$$L_k(\xi|\theta) = L_k(\xi|\theta_0), \text{ for all } \xi \in \mathcal{X}_k \Leftrightarrow d_k(\theta) = 0. \tag{2.42}$$

Consider the following example. Agent k is observing RGB images of animals and trying to detect whether these images correspond to a dog, a wolf, or a cat, i.e., $\Theta = \{\text{dog}, \text{wolf}, \text{cat}\}$. If the agent only sees a part of the full image, i.e., $\boldsymbol{\xi}_{k,i}$ corresponds to a small RGB patch, it is possible that this limited information is only helpful in distinguishing dogs from cats, but not dogs from wolves. In this case, the agent's likelihood models corresponding to a dog and a wolf can be identical $L_k(\xi|\text{dog}) = L_k(\xi|\text{wolf})$ for all ξ.

In isolation, if the agent k is not able to distinguish whether its observations are arising from hypothesis θ_0 or θ, it cannot learn the truth. To see that, consider the single-agent case—subscript k is dropped—with flat prior belief, i.e., $\mu_0(\theta) = 1/H$, where the recursive Bayesian

update in (1.11) yields:

$$\boldsymbol{\mu}_i(\theta) \propto L(\boldsymbol{\xi}_i|\theta)\boldsymbol{\mu}_{i-1}(\theta) = \prod_{m=1}^{i} L(\boldsymbol{\xi}_m|\theta)\frac{1}{H}. \tag{2.43}$$

If $d(\theta) = 0$ for two hypotheses $\theta_0 \neq \theta$, then

$$\frac{\boldsymbol{\mu}_i(\theta_0)}{\boldsymbol{\mu}_i(\theta)} = \prod_{m=1}^{i} \frac{L(\boldsymbol{\xi}_m|\theta_0)}{L(\boldsymbol{\xi}_m|\theta)} \frac{\frac{1}{H}}{\frac{1}{H}} = 1. \tag{2.44}$$

Therefore, the observations provide no information that allows the agent to distinguish hypotheses θ and θ_0, and thus $\boldsymbol{\mu}_i(\theta_0) = \boldsymbol{\mu}_i(\theta)$ for $i = 1, 2, \ldots$.

Motivated by this discussion, we associate with each agent k, a set of *locally indistinguishable hypotheses*, namely,

$$\Theta_k \triangleq \{\theta : d_k(\theta) = 0\}, \tag{2.45}$$

which is the set of hypotheses whose KL divergences with respect to the true hypothesis is zero. The complementary set made of *locally distinguishable* hypotheses by agent k is given by:

$$\bar{\Theta}_k \triangleq \Theta \setminus \Theta_k. \tag{2.46}$$

In many practical situations, the limited knowledge available *locally* at the *individual* agents precludes them from identifying the true state. When this happens, we say that the problem is not locally identifiable, which formally means that all local indistinguishable sets would have cardinality larger than one, i.e.,

$$|\Theta_k| > 1, \quad \forall k = 1, 2, \ldots, K. \tag{2.47}$$

In a collaborative setup, these local difficulties can be overcome by exchanging information with neighbors. Since the network is strongly connected, eventually after sufficient iterations, any piece of knowledge available at one agent will diffuse to all others. In social learning, instead of requiring agents to have locally identifiable likelihoods, it suffices for the models to be *globally identifiable*, as described in the next assumption.

Assumption 2.5 (Global identifiability). For each hypothesis $\theta \neq \theta_0$ there exists at least one agent k in the network for which

$$d_k(\theta) > 0. \tag{2.48}$$

Under global identifiability, it is clear that the network KL divergence $D(\theta)$ defined in (2.11) is zero for $\theta = \theta_0$ and strictly positive due to Assumption 2.5, since the entries of the Perron eigenvector are positive. Therefore, the true hypothesis becomes the unique minimizer of $D(\theta)$. We state the belief convergence result in Corollary 2.1

Corollary 2.1 (Belief convergence under objective truth). *Consider a strongly connected network with a left-stochastic combination matrix, and assume all observations are generated*

by the same model θ_0, i.e., $\boldsymbol{\xi}_{k,i} \sim L_k(\xi|\theta_0)$ for all $k = 1, 2, \ldots, K$. Under Assumptions 2.2, 2.4, and 2.5, the beliefs of strategy (2.2)–(2.3) converge almost surely to the truth, namely,

$$\boldsymbol{\mu}_{k,i}(\theta_0) \xrightarrow{\text{a.s.}} 1. \tag{2.49}$$

Proof. Corollary 2.1 is proven by noticing that Assumption 2.5 implies that $\theta^\star$, from Theorem 2.2, coincides with θ_0. When $f_\ell(\xi) = L_\ell(\xi|\theta_0)$, the network divergence is written as

$$D(\theta) = \sum_{\ell=1}^{K} \pi_\ell d_\ell(\theta). \tag{2.50}$$

Clearly, $D(\theta_0)$ is equal to 0. From Assumption 2.5, for each $\theta \neq \theta_0$, there exists at least one agent for which $d_\ell(\theta) > 0$. From this assumption and the fact that $\pi \succ 0$, we have that

$$D(\theta) > 0, \quad \theta \neq \theta_0. \tag{2.51}$$

Hence, $\theta^\star$ is unique and equal to θ_0. □

In the next example, we illustrate the phenomenon of truth learning in strongly connected networks corresponding to the result in Corollary 2.1.

Example 2.1 (Truth learning in strongly connected networks). Consider a strongly connected network with 10 agents, whose topology is illustrated in the left panel of Figure 2.2. The combination matrix A follows the Metropolis rule [37], which yields a doubly-stochastic matrix.

Agents consider a set of 3 hypotheses, i.e., $\Theta = \{1, 2, 3\}$, which explains the nature of the world they observe. For simplicity, agents share the same set of likelihood models, i.e., for each k, $L_k(\xi|\theta) = L(\xi|\theta)$ for all $\theta \in \Theta$, which are Gaussian models with variance 1 and different means given by the hypotheses θ:

$$L(\xi|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\xi - \theta)^2}{2}\right\}, \quad \theta \in \Theta. \tag{2.52}$$

We assume the observations $\boldsymbol{\xi}_{k,i}$ are sampled independently from the likelihood $L(\xi|1)$ for all agents, i.e., $\theta_0 = 1$. The likelihood models can be seen in the middle panel of Figure 2.2.

We illustrate the result of Corollary 2.1 by simulating the social learning algorithm (??)–(??) for 30 iterations. The belief convergence is shown in the right panel of Figure 2.2 for agent 1. All other agents present similar convergence, in the sense that they all concentrate their beliefs around the true hypothesis $\theta_0 = 1$. □

2.2 Influence and Disagreement

In real-life social networks, instead of a strongly connected graph where information flows in both directions between every two agents, we often encounter situations where communication

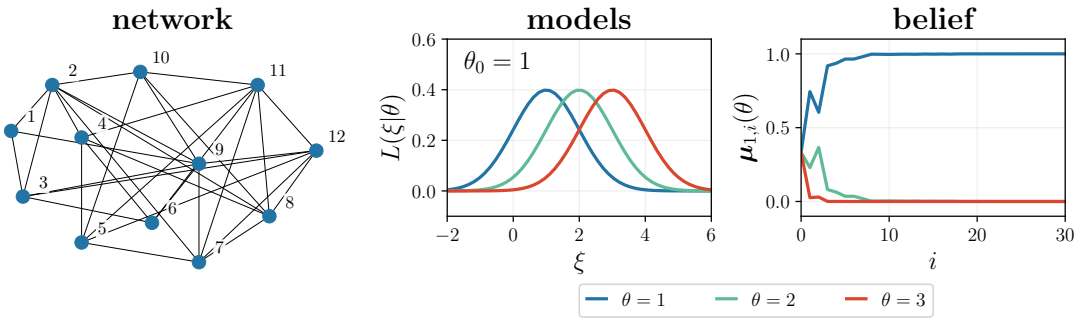


Figure 2.2: (*Left*) Strongly connected network topology. (*Middle*) Likelihood models. (*Right*) Belief convergence for agent 1, showing asymptotic truth learning.

happens in a unidirectional manner over some parts of the network. For example, some influential users produce content that reaches simultaneously millions of followers, without receiving any feedback from all these followers. This asymmetric communication dynamics can be represented mathematically in terms of *weakly connected graphs*. The concept has been first explored in [47], in the context of distributed learning, where authors consider that information can flow from some graph components to the others in *only one* direction. The same concept is explored in social learning under arithmetic [48] and geometric [49] averaging. In all applications, the weakly connected graph introduces the effect of *influence* between different network components and the possibility of *disagreement* between agents. We report here the results from [49].

2.2.1 Weakly Connected Networks

A weakly connected network is generally defined as a network in which there exists a path linking every two agents in *at least one* direction. A particular case is the definition of a *connected network*, that is, a network where there exists a path linking every two agents in *both* directions. Note that the strongly connected network presented in Section 2.1.1 is a particular case of this definition. In this section, when treating weakly connected networks we are however interested in a particular structure that we describe as follows.

A $K-$agent weakly connected network can be divided into $S+R$ disjoint subnetworks: S *sending (sub)networks*, denoted by the sets $\mathcal{S}_s$ with $s = 1, 2, \ldots, S$, and R *receiving (sub)networks*, denoted by the sets $\mathcal{R}_r$ for $r = 1, 2, \ldots, R$. Each set $\mathcal{S}_s$ or $\mathcal{R}_r$ consists of the nodes in the respective subnetwork. We therefore have:

$$\mathcal{S} \triangleq \bigcup_{s=1}^{S} \mathcal{S}_s, \quad \mathcal{R} \triangleq \bigcup_{r=1}^{R} \mathcal{R}_r, \quad \mathcal{S} \cup \mathcal{R} = \{1, 2, \ldots, K\}, \tag{2.53}$$

where $\mathcal{S}$ and $\mathcal{R}$ are the union of sending and receiving networks, respectively. The connectivity within and between different (sub)networks is characterized according to the following rules:

1. Each sending network is strongly connected.

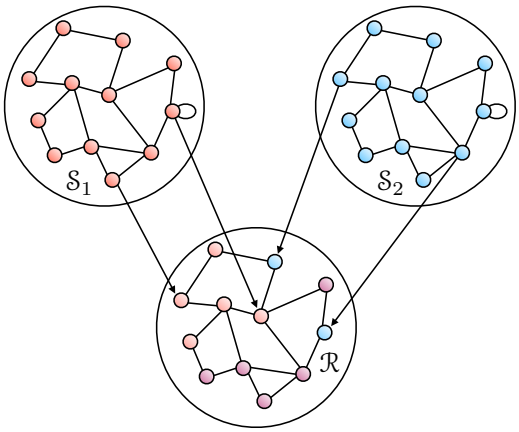


Figure 2.3: Weakly connected network with two sending networks, namely $\mathcal{S}_1$ and $\mathcal{S}_2$, and one receiving network $\mathcal{R}$.

2. Sending networks do not communicate with each other.

3. Each receiving network is assumed to be connected.

4. Communication occurs from a sending network to a receiving network, but not the other way around.

With each sending network $\mathcal{S}_s$ we associate a combination matrix $A_{\mathcal{S}_s}$, which, in view of rule 1., implies that we can associate with it a Perron eigenvector $\pi^{(s)}$ of dimension $|\mathcal{S}_s| \times 1$. Similarly, we associate a combination matrix $A_{\mathcal{R}_r}$ with each receiving network $\mathcal{R}_r$. An illustration of a weakly connected network can be found in Figure 2.3 with $S = 2$ and $R = 1$.

Without loss of generality, we assume agents are numbered starting from agents belonging to sending networks $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_S$, i.e., sending agents, followed by agents from the receiving networks $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_R$, i.e., receiving agents. In this manner, the overall combination matrix will have the upper block triangular form:

$$A = \left[\begin{array}{c|c} A_{\mathcal{S}} & A_{\mathcal{S}\mathcal{R}} \\ \hline 0 & A_{\mathcal{R}} \end{array}\right], \tag{2.54}$$

where the upper left block contains the combination matrices pertaining to the sending networks, i.e.,

$$A_{\mathcal{S}} = \text{blkdiag}\Big\{A_{\mathcal{S}_1}, A_{\mathcal{S}_2}, \ldots, A_{\mathcal{S}_S}\Big\}. \tag{2.55}$$

The upper right block $A_{\mathcal{S}\mathcal{R}}$ contains the combination weights associated to the links from sending agents to receiving agents. The lower left block contains only zeros, since there are no directed edges from receiving to sending agents. Finally, the lower right block $A_{\mathcal{R}}$ collects the weights among receiving agents. The overall network is assumed to be left stochastic and an existing edge between two agents exists if its associated weight is positive, that is,

$$A^{\mathsf{T}}\mathbb{1} = \mathbb{1}, \quad a_{\ell k} > 0 \Rightarrow \ell \in \mathcal{N}_k. \tag{2.56}$$

Relation (2.56) immediately implies that $A_{\mathcal{S}_s}$ is left-stochastic for every $s = 1, 2, \ldots, S$. Another interesting property of weakly connected networks concerns the convergence of the matrix powers A^n for growing n. In [47], it was shown that the power of A converges according to the following limit:

$$A_\infty \triangleq \lim_{n\to\infty} A^n = \left[\begin{array}{c|c} E & EW \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} E & \Omega \\ \hline 0 & 0 \end{array}\right], \tag{2.57}$$

where E has dimension $|\mathcal{S}| \times |\mathcal{S}|$ and is given by

$$E = \text{blkdiag}\Big\{\pi^{(1)}\mathbb{1}_{|\mathcal{S}_1|}^{\mathsf{T}},\, \pi^{(2)}\mathbb{1}_{|\mathcal{S}_2|}^{\mathsf{T}}, \ldots,\, \pi^{(S)}\mathbb{1}_{|\mathcal{S}_S|}^{\mathsf{T}}\Big\}, \tag{2.58}$$

while Ω has dimension $|\mathcal{S}| \times |\mathcal{R}|$ and is equal to

$$\Omega = EW, \quad W = A_{\mathcal{S}\mathcal{R}}(I_{|\mathcal{R}|} - A_{\mathcal{R}})^{-1}. \tag{2.59}$$

Here, the notation $\mathbb{1}_{|\mathcal{X}|}$ denotes a vector of ones with dimension $|\mathcal{X}| \times 1$ and $I_{|\mathcal{X}|}$ is the identity matrix with dimension $|\mathcal{X}| \times |\mathcal{X}|$, where $|\mathcal{X}|$ is the cardinality of set $\mathcal{X}$. Matrix Ω can alternatively be expressed as

$$\Omega = EA_{\mathcal{S}\mathcal{R}}(I_{|\mathcal{R}|} + A_{\mathcal{R}} + A_{\mathcal{R}}^2 + \ldots). \tag{2.60}$$

We denote the elements of Ω by $\omega_{\ell k}$ with $\ell \in \mathcal{S}$ and $k \in \mathcal{R}$. Since A is left-stochastic, we have for any receiving agent k that

$$\sum_{\ell\in\mathcal{S}} \omega_{\ell k} = 1. \tag{2.61}$$

If we consider a sending agent $\ell \in \mathcal{S}_s$ and a receiving agent $k \in \mathcal{R}_r$, then the weight $\omega_{\ell k}$ can be zero only if there is no edge linking sending subnetwork $\mathcal{S}_s$ and receiving subnetwork $\mathcal{R}_r$.

2.2.2 Convergence Behavior

The convergence behavior depends on the nature of the agents. First, we note that sending agents, i.e., $k \in \mathcal{S}$, follow the convergence behavior of strongly connected networks established in Section 2.1.2 [42], [49]. For each sending network $\mathcal{S}_s$, the belief evolution for any agent $k \in \mathcal{S}_s$ is governed by Theorem 2.1, with its particular Perron eigenvector $\pi^{(s)}$.

More explicitly, from Theorem 2.1, under Assumptions 2.1, 2.2, and 2.3, for $k \in \mathcal{S}_s$:

$$\boldsymbol{\mu}_{k,i}(\theta_s^\star) \xrightarrow{\text{a.s.}} 1, \tag{2.62}$$

where $\theta_s^\star$ is the unique minimizer of the network KL divergence for sending network s:

$$\sum_{k\in\mathcal{S}_s} \pi_k^{(s)} D(f_k||L_k(\theta)), \tag{2.63}$$

for $s = 1, 2, \ldots, S$.

Second, the convergence behavior at receiving agents, i.e., $k \in \mathcal{R}$, needs to be established. To

do that, we define the *average KL divergence* at any receiving agent k as follows:

$$D_k(\theta) \triangleq \sum_{\ell \in \mathcal{S}} \omega_{\ell k} D(f_\ell || L_\ell(\theta)). \tag{2.64}$$

Observe that the sum is over all sending agents linked to k. Using this KL divergence, we replace Assumption 2.3 by the following identifiability condition for weakly connected graphs.

Assumption 2.6 (Unique minimizer in weakly connected graphs). For each $k \in \mathcal{R}$, the function $D_k(\theta)$ has a unique minimizer:

$$\theta_k^\star \triangleq \operatorname*{argmin}_{\theta \in \Theta} D_k(\theta). \tag{2.65}$$

We state next a theorem characterizing the convergence of beliefs within weakly connected networks for the Adapt-Then-Combine social learning strategy in (2.2)–(2.3). The result was stated in [49], and its proof appears in [42].

Theorem 2.2 (Belief convergence of receiving agents in weakly connected networks [42], [49]**).** *Consider a weakly connected network for which* (2.57) *holds. Under Assumptions 2.1, 2.2 and 2.6, the receiving agents under the social learning strategy* (2.2)–(2.3) *will have their beliefs converge to the minimizers of* (2.65) *almost surely, i.e., for $k \in \mathcal{R}$:*

$$\boldsymbol{\mu}_{k,i}(\theta_k^\star) \xrightarrow{\text{a.s.}} 1. \tag{2.66}$$

Moreover, for all $\theta \neq \theta_k^\star$, the convergence of the belief to zero takes place at an exponential rate as:

$$\lim_{i \to \infty} \frac{\log \boldsymbol{\mu}_{k,i}(\theta)}{i} \overset{\text{a.s.}}{=} D_k(\theta_k^\star) - D_k(\theta). \tag{2.67}$$

Two important phenomena can be observed from the result in Theorem 2.2. First, weakly connected networks introduce a hierarchy of *influence* between different subnetworks, where the sending agents influence the beliefs of receiving agents. From (2.64), we see that the accepted hypothesis $\theta_k^\star$ is solely determined by the statistical models pertaining to the sending agents. Second, receiving agents can end up in a state of *disagreement*. As seen in Theorem 2.2, the hypothesis $\theta_k^\star$ around which receiving agents concetrate their beliefs varies with k and depends on the graph topology through the weights $\omega_{\ell k}$. From (2.60), we see that the elements of the matrix Ω take into account the *cumulative influence* over all paths from sending agents to receiving agents. Differences in this cumulative influence across different agents can result in them converging to distinct hypotheses.

Example 2.2 (Disagreement in weakly connected networks). Consider a weakly connected network with 12 agents divided into two sending networks and one receiving network as follows:

$$\mathcal{S}_1 \triangleq \{1, 2, 3, 4\}, \tag{2.68}$$
$$\mathcal{S}_2 \triangleq \{5, 6, 7, 8\}, \tag{2.69}$$
$$\mathcal{R} \triangleq \{9, 10, 11, 12\}. \tag{2.70}$$

A diagram showing the topology of the network can be seen in the left panel of Figure 2.4.

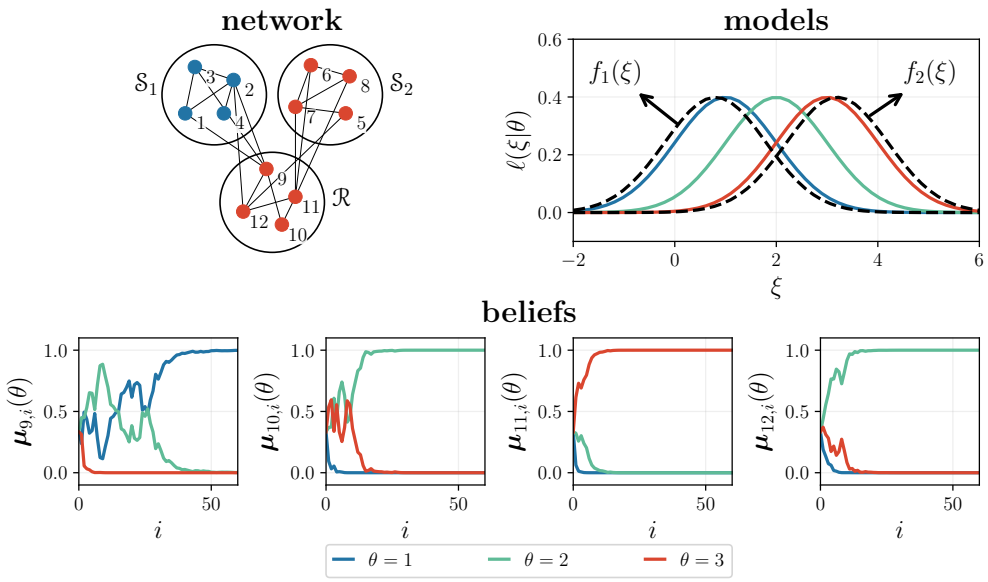


Figure 2.4: (*Top left*) Weakly connected network topology. (*Top right*) Likelihood models and true models for sending networks $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively $f_1(\xi)$ and $f_2(\xi)$. (*Bottom*) Belief convergence for receiving agents over time.

We assume that all agents share the same set of Gaussian likelihoods, introduced in Example 2.1, more precisely in (2.52). The true models $f_\ell(\xi)$ do not belong to this set of likelihoods, instead they are given by the following Gaussian models, which differ across different subnetworks:

$$f_\ell(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\xi-0.8)^2}{2}\right\} \triangleq f_1(\xi), \quad \ell \in \mathcal{S}_1, \tag{2.71}$$

$$f_\ell(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\xi-3.2)^2}{2}\right\} \triangleq f_2(\xi), \quad \ell \in \mathcal{S}_2 \cup \mathcal{R}. \tag{2.72}$$

The likelihood models and true models $f_1(\xi)$ and $f_2(\xi)$ can be seen in the top right panel of Figure 2.4. We compute the KL divergences between the true model $f_s(\xi)$ for $s = 1, 2$ and the likelihoods $L(\xi|\theta)$ for $\theta \in \Theta$ in the following way:

$$D(f_s||L(\theta)) = \mathbb{E}_{f_s}\left(\log \frac{f_s(\boldsymbol{\xi})}{L(\boldsymbol{\xi}|\theta)}\right) = \frac{1}{2}(\theta - \mathbb{E}_{f_s}(\boldsymbol{\xi}))^2, \tag{2.73}$$

from which we can write the average KL divergence for each receiving agent k as:

$$D_k(\theta) = \frac{1}{2}(\theta-0.8)^2 \sum_{\ell\in\mathcal{S}_1} \omega_{\ell k} + \frac{1}{2}(\theta-3.2)^2 \sum_{\ell\in\mathcal{S}_2} \omega_{\ell k}. \tag{2.74}$$

In this example, the hypothesis that minimizes (2.74), namely, $\theta_k^\star$, depends on the *cumulative weights* for each subnetwork, that is, on

$$\omega_k^{(1)} \triangleq \sum_{\ell\in\mathcal{S}_1} \omega_{\ell k}, \quad \omega_k^{(2)} \triangleq \sum_{\ell\in\mathcal{S}_2} \omega_{\ell k}. \tag{2.75}$$

Table 2.1: Cumulative weights given to the information received from the two sending networks $\mathcal{S}_1$ and $\mathcal{S}_2$ for each receiving agent k.

Agent k	$\omega_k^{(1)}$	$\omega_k^{(2)}$
9	0.8	0.2
10	0.5	0.5
11	0.2	0.8
12	0.5	0.5

The larger the ratio $\omega_k^{(1)}/\omega_k^{(2)}$ is, the more influence $\mathcal{S}_1$ bears on the minimizer of (2.74), driving it closer to hypothesis 1. Whereas for smaller values of the same ratio, $\mathcal{S}_2$ has more influence and thus steers the minimizer toward hypothesis 3. We can see in Table 2.1 the list of cumulative weights for each receiving agent with respect to the two sending networks. The cumulative weights quantify the influence of each sending network onto each receiving agent. For example, we see that agent 9 is mostly influenced by sending network $\mathcal{S}_1$, while agent 11 is mostly influenced by sending network $\mathcal{S}_2$. Agents 10 and 12 are influenced equally by both sending components. The belief convergence shown in Figure 2.4 confirms our intuition and shows disagreement between receiving agents. □

Stationary World Part I

3 Recovering Influences in Weak Graphs

3.1 Introduction[1]

In the social learning setting described in Chapter 1, several agents linked through a network topology form their individual opinions about a phenomenon of interest by exchanging beliefs with their neighbors. One relevant network topology for social learning takes the form of *weakly connected* networks, which we described in Chapter 2. Under this model, there are two categories of subnetworks: *sending* and *receving* subnetworks. Sending agents feed information to receiving agents without getting any information back from them [47], [48], [60]. This scenario is common over social networks. For example, a celebrity may have a large number of followers, whose individual opinions are not necessarily followed by the celebrity. Another example is that of media channels, which promote the emergence of opinions by feeding data to users without taking into account users' feedback.

One fundamental challenge arising in the study of social learning problems is to understand the mechanism of opinion formation. This was explained in Section 2.2, where we showed how the receiving agents are *completely influenced* by the sending subnetworks. Naturally, the network topology plays an important role in determining the asymptotic opinion formation. This observation motivates the question that is addressed in the current chapter, and which can be seen as a *dual* learning problem. Given the observation of the receiving agents' behavior, we want to establish whether it is possible to learn *topological influences* from the sending components to the receiving agents.

This question is interesting because it allows us to identify the main sources of information in a network and how they influence opinion formation. This problem is challenging because we assume that we can only observe the beliefs evolving at the receiving agents. In particular, we will only be able to recover topological influences in terms of the limiting weights that each receiving agent experiences from each sending component. We refer to this as macroscopic information since these weights incorporate: **i)** The global effect coming from *all* agents belonging to a sending component, and **ii)** the effect of intermediate receiving agents linked to the receiving agent under consideration. The relevance in estimating these global weights relies on the fact that the limiting beliefs of the receiving agents depend solely on this aggregate

[1] This chapter is adapted from [42], [51].

information.

We will establish conditions under which the recovery of influence weights becomes feasible. More specifically, given H hypotheses and S sending components, under the assumption of homogeneous statistical models within each sending component, we will ascertain that a necessary condition to achieve consistent influence recovery is (Lemma 3.1):

$$H \geq S. \tag{3.1}$$

Once the necessary condition is established, we will examine some useful models to see whether influence recovery *can be* in fact achieved. We consider first a *structured* Gaussian model where: **i)** the true underlying (Gaussian) distributions are distinct across the sending subnetworks; and **ii)** the (Gaussian) likelihoods are equal across the sending subnetworks, and contain the true distributions. For this setting, we will show in Theorem 3.1 that influence recovery is feasible only when $S = 2$. We then recognize that one fundamental element for influence recovery is the *diversity* between the sending subnetworks. Adding this further element, we will establish in Theorem 3.2 that the problem is feasible for *any* S provided that (3.1) holds, and even under more general (e.g., non-Gaussian) models.

In summary, we remark that there are two learning problems coexisting in our work: A social learning problem and an influence recovery problem. The former is the direct inferential problem studied in Chapter 2, and for which the agents are deployed. The latter is the *reverse* problem, which is in fact based on observation of the output (the beliefs) of the direct learning problem. One useful conclusion of our analysis is to reveal an interplay between these two coexisting learning problems—see Section 3.6 further ahead.

3.2 Problem Setting

In this chapter, we consider the weakly connected network setting described in Section 2.2, where the network consists of S sending subnetworks and R receiving subnetworks. In particular, we assume here that each receiving subnetwork is connected to at least one agent in each sending subnetwork.

The learning procedure used is the social learning strategy found in (2.2) and (2.3). We describe it here in some detail: For each admissible hypothesis $\theta \in \Theta$ at time i, each agent k uses its own fresh *private* observation, $\boldsymbol{\xi}_{k,i}$, to compute the local likelihood $L_k(\boldsymbol{\xi}_{k,i}|\theta)$. Using this likelihood, agent k updates its local belief, $\boldsymbol{\mu}_{k,i-1}(\theta)$, obtaining an intermediate belief $\boldsymbol{\psi}_{k,i}(\theta)$ through a Bayesian update:

$$\boldsymbol{\psi}_{k,i}(\theta) = \frac{\boldsymbol{\mu}_{k,i-1}(\theta) L_k(\boldsymbol{\xi}_{k,i}|\theta)}{\sum_{\theta' \in \Theta} \boldsymbol{\mu}_{k,i-1}(\theta') L_k(\boldsymbol{\xi}_{k,i}|\theta')}. \tag{3.2}$$

Then, agent k aggregates the intermediate beliefs received from its neighbors through the following combination rule, which is equivalent to the geometric-average combination seen in

(2.3):

$$\boldsymbol{\mu}_{k,i}(\theta) = \frac{\exp\left\{\sum_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\theta)\right\}}{\sum_{\theta' \in \Theta} \exp\left\{\sum_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\theta')\right\}}, \tag{3.3}$$

where $a_{\ell k}$ is the *nonnegative combination weight* that agent k uses to scale the intermediate log-belief received from agent ℓ. The $\{a_{\ell k}\}$ are the elements of the combination matrix A—see (2.54) and the discussion surrounding it for an overview of the properties of A.

3.2.1 Limiting Beliefs of Receiving Agents

Let us consider a single-agent scenario where agent ℓ operates alone. A natural way for agent ℓ to choose a hypothesis would be to choose the θ that gives the best match between a model $L_\ell(\xi|\theta)$ and the distribution of the observed data, $f_\ell(\xi)$. One measure of the match between $f_\ell(\xi)$ and $L_\ell(\xi|\theta)$ is the KL divergence $D(f_\ell||L_\ell(\theta))$. The smaller the value of this divergence is, the greater the match between the data and the model will be. As seen in Chapter 1, the single-agent recursive Bayesian update yields a belief concentrated on the hypothesis θ that minimizes the divergence $D(f_\ell||L_\ell(\theta))$.

In the social learning context, this optimization problem turns into a *distributed* optimization problem. In particular, under our social learning setting over weak graphs, we have seen in Theorem 2.2 that the social learning strategy in (3.2)–(3.3) ends up minimizing (without knowing the true distributions) the following *average* divergence at *receiving* agent $k \in \mathcal{R}$:

$$D_k(\theta) \triangleq \sum_{\ell \in \mathcal{S}} \omega_{\ell k} D(f_\ell||L_\ell(\theta)), \tag{3.4}$$

which is a weighted combination, through the limiting combination weights $\{\omega_{\ell k}\}$, of the KL divergences of the *sending* agents reaching k. We recall that the weights $\{\omega_{\ell k}\}$ correspond to the elements of the matrix Ω defined in (2.59). The role of average divergence measures like the one in (3.4) already appeared in the case of strongly connected networks. For example, it was shown in Theorem 2.1 that with the geometric-average strategy in (3.2) and (3.3), each agent ends up minimizing the *same* weighted combination of divergences in (2.11). Under objective evidence—see Section 2.1.3, we saw in Corollary 2.1 that such minimization leads each individual agent to discover the true underlying hypothesis.

In our weak-graph setting, however, the effect of minimizing $D_k(\theta)$ (which depends on the particular receiving agent k) will be less obvious. We already see from (3.4) that the average divergence combines *topological* attributes, encoded in the limiting combination weights, with *inferential attributes*, encoded in the local KL divergences. The interplay arising between the network topology and social learning will be critical in determining the belief convergence of the receiving agents. We report here the result of Theorem 2.2 for ease of reference. Under Assumptions 2.1, 2.2 and 2.6, for $k \in \mathcal{R}$, we have that:

$$\boldsymbol{\mu}_{k,i}(\theta_k^\star) \xrightarrow{\text{a.s.}} 1. \tag{3.5}$$

Moreover, for all $\theta \neq \theta_k^\star$, the convergence of the belief to zero takes place at an exponential rate:

$$\frac{\log \boldsymbol{\mu}_{k,i}(\theta)}{i} \xrightarrow{\text{a.s.}} D_k(\theta_k^\star) - D_k(\theta). \tag{3.6}$$

The proof combines the techniques used to establish the convergence of the social learning algorithm, e.g., in [40], [41] for strongly connected graphs, with the convergence results of the combination matrix over weak graphs used in [5], [48].

Several insightful conclusions arise from the result above. First, the limiting belief *of each receiving agent* is degenerate, meaning that it collapses to a single hypothesis, when sufficient time for learning is allowed. Second, different agents can in principle *disagree*, since they can converge to different hypotheses. The hypothesis around which the belief concentrates will depend on a weighted combination of KL divergences. Third, we see from (3.4) that only the local divergences corresponding to the sending agents, $\ell \in \mathcal{S}$, determine the value of $D_k(\theta)$ and, hence, of $\theta_k^\star$. Therefore, the limiting hypothesis $\theta_k^\star$ at agent k is determined by the KL divergences of statistical models pertaining to *sending subnetworks*, and, hence, *it does not depend on the data sensed at receiving agent* k.

In a nutshell, we see the emergence of two effects: **i)** Influence effect, i.e., the final states of the receiving agents are dependent only upon the properties of the detection problems at the *sending* agents; **ii)** Disagreement effect, i.e., different network topologies allow the sending agents to drive the receiving agents to potentially different decisions.

3.2.2 Canonical Examples

In order to examine in more detail the role the network topology plays in determining the limiting beliefs of receiving agents, we consider a simple yet insightful example. The sending and receiving components are:

$$\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \quad \text{and} \quad \mathcal{R}, \tag{3.7}$$

namely, we have two sending subnetworks, $\mathcal{S}_1$ and $\mathcal{S}_2$, and one receiving subnetwork $\mathcal{R}$.

For what concerns the inferential problem, we assume there are three possible hypotheses, $\theta \in \{1, 2, 3\}$. The likelihood functions are the same across *all* agents. In particular, we assume that, for all $\xi \in \mathbb{R}$, and for $\theta \in \{1, 2, 3\}$:

$$L(\xi|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\xi - \mathsf{m}_\theta)^2}{2}\right\}, \tag{3.8}$$

where the means corresponding to the different hypotheses are chosen as, for some $\Delta > 0$:

$$\mathsf{m}_1 = -\Delta, \;\; \mathsf{m}_2 = 0, \;\; \mathsf{m}_3 = +\Delta. \tag{3.9}$$

We further assume that the *true* distributions of the sending subnetworks are Gaussian distributions, with expectations chosen among the expectations in (3.9). In particular, we assume that agents belonging to subnetwork $\mathcal{S}_1$ generate data according to model $\theta = 1$, i.e., with

expectation equal to $-\Delta$, whereas agents belonging to subnetwork $\mathcal{S}_2$ generate data according to model $\theta = 3$, i.e., with expectation equal to $+\Delta$. Formally we write:

$$f_\ell(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\xi+\Delta)^2}{2}\right\}, \quad \forall \ell \in \mathcal{S}_1, \tag{3.10}$$

$$f_\ell(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\xi-\Delta)^2}{2}\right\}, \quad \forall \ell \in \mathcal{S}_2. \tag{3.11}$$

Recalling that the KL divergence between two unit-variance Gaussian distributions of expectations a and b is given by $0.5(a-b)^2$, under the setting described above we can write, for all $k \in \mathcal{R}$:

$$\begin{aligned} D_k(\theta) &= \sum_{\ell \in \mathcal{S}} \omega_{\ell k} D(f_\ell || L_\ell(\theta)) \\ &= \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} D(f_\ell || L(\theta)) + \sum_{\ell \in \mathcal{S}_2} \omega_{\ell k} D(f_\ell || L(\theta)) \\ &= \frac{(-\Delta - \mathsf{m}_\theta)^2}{2} \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} + \frac{(\Delta - \mathsf{m}_\theta)^2}{2} \sum_{\ell \in \mathcal{S}_2} \omega_{\ell k}, \end{aligned} \tag{3.12}$$

which further implies:

$$D_k(1) = 2\Delta^2 \sum_{\ell \in \mathcal{S}_2} \omega_{\ell k}, \quad D_k(2) = \frac{\Delta^2}{2}, \quad D_k(3) = 2\Delta^2 \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k}, \tag{3.13}$$

where, in the intermediate equality, we used (2.61). As a result, we can compute the limiting hypothesis, for each $k \in \mathcal{R}$, as:

$$\theta_k^\star = \text{argmin} \left\{ 4 \sum_{\ell \in \mathcal{S}_2} \omega_{\ell k}, \;\; 1, \;\; 4 \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} \right\} \tag{3.14}$$

From (2.60), one can argue that $\sum_{\ell \in \mathcal{S}_s} \omega_{\ell k}$ reflects the sum of influences over *all* paths connecting *all* sending agents in subnetwork s to receiving agent k.

In order to find the minimizer in (3.14), we start by using (2.61) in (3.14), which yields:

$$\theta_k^\star = \text{argmin} \left\{ 1 - \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k}, \;\; 0.25, \;\; \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} \right\}. \tag{3.15}$$

In view of Theorem 2.2, the belief of the k-th receiving agent will converge to $\theta_k^\star = 1$ if the following two conditions are simultaneously verified:

$$\begin{aligned} 1 - \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} < 0.25 \Leftrightarrow \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} > 0.75, \\ 1 - \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} < \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} \Leftrightarrow \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} > 0.5. \end{aligned} \tag{3.16}$$

Taking the most stringent condition in (3.16) reveals that:

$$\theta_k^\star = 1 \;\Leftrightarrow\; \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} > 0.75. \tag{3.17}$$

In summary, we conclude that agent k follows the opinion promoted by sending subnetwork $\mathcal{S}_1$ if the influence of subnetwork $\mathcal{S}_1$ on agent k is "sufficiently large".

The situation is reversed if the influence of subnetwork $\mathcal{S}_2$ is sufficiently large, namely,

$$\theta_k^\star = 3 \;\Leftrightarrow\; \sum_{\ell \in \mathcal{S}_2} \omega_{\ell k} > 0.75, \tag{3.18}$$

where we recall that hypothesis $\theta = 3$ is promoted by subnetwork $\mathcal{S}_2$. However, there is another possibility. It occurs when:

$$\sum_{\ell \in \mathcal{S}_1} \omega_{\ell k} < 0.75 \quad \text{and} \quad \sum_{\ell \in \mathcal{S}_2} \omega_{\ell k} < 0.75. \tag{3.19}$$

In this case, no clear dominance from one subnetwork can be ascertained, and each receiving agent will choose $\theta_k^\star = 2$, i.e., an *opinion that does not coincide with any of the opinions promoted by the sending subnetworks.*

From (3.17) and (3.18), we see that the dominance of one of the sending subnetworks is determined by the aggregate influence $\sum_{\ell \in \mathcal{S}_1} \omega_{\ell k}$, with the complementary aggregate influence being $\sum_{\ell \in \mathcal{S}_2} \omega_{\ell k} = 1 - \sum_{\ell \in \mathcal{S}_1} \omega_{\ell k}$. The main way to manipulate these factors consists in varying the *sizes* of the sending subnetworks or their connections with the receiving agents.

In order to illustrate more carefully the possible scenarios, we consider the following simulation framework:

- The strongly connected sending components $\mathcal{S}_1$ and $\mathcal{S}_2$ are generated as Erdős-Rényi random graphs with connection probability q, and the entries of the corresponding combination matrix are determined by the averaging rule, namely,[2]

$$a_{\ell k} = \begin{cases} 1/n_k, & \text{if } k \neq \ell \text{ are neighbors or } k = \ell, \\ 0, & \text{otherwise,} \end{cases} \tag{3.20}$$

  where n_k is the number of neighbors of node k (including node k itself). In our experiments we set $q = 0.7$.

- An agent k is connected to a sending agent through a Bernoulli distribution with parameter π_s, which depends on the sending subnetwork s. Given the total number d_k, of directed edges from sending agents to agent k, we initially set $a_{\ell k} = 1/d_k$. The combination matrix A of the overall network $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$ is normalized so that it is left-stochastic.

It is now possible to examine different scenarios by manipulating the size of the sending subnetworks as well as the send-receive connection probabilities π_s.

[2]When drawing the random graph, we have verified that there exists at least one self-loop.

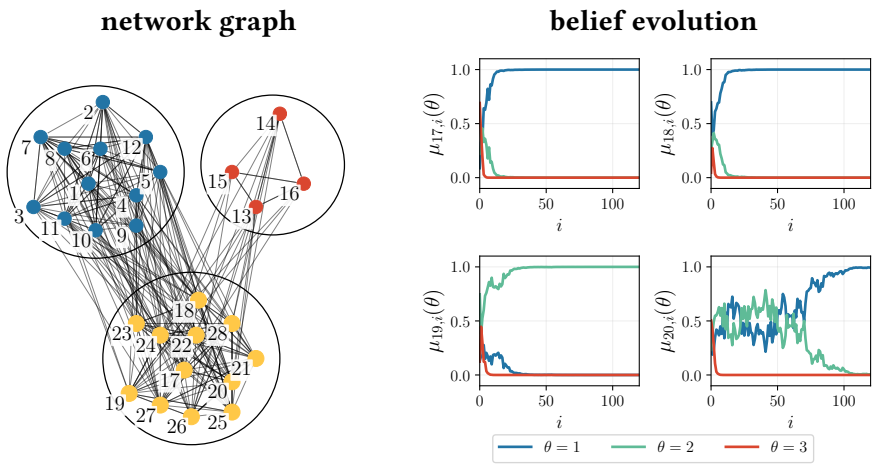


Figure 3.1: *How majorities build a majority.* (*Left*) Weakly connected network, where the size of sending subnetwork $\mathcal{S}_1$ is dominant. (*Right*) Convergence of beliefs at receiving gents.

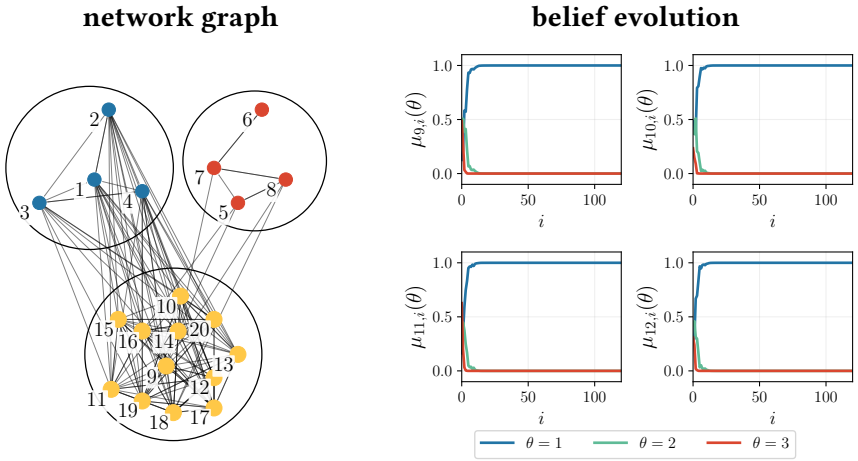


Figure 3.2: *How filter bubbles build a majority.* (*Left*) Weakly connected network, where the connectivity from sending subnetwork $\mathcal{S}_1$ is dominant. (*Right*) Convergence of beliefs at receiving agents.

– Setup 1 or **How majorities build a majority**. In Figure 3.1, we set $\pi_1 = \pi_2 = 0.5$, i.e. it is equally probable that a receiving agent connects to any sending agent, irrespective of the sending subnetwork. In view of this uniformity, we can expect that the limiting weights $\omega_{\ell k}$ are sufficiently uniform across the two sending subnetworks and, hence, that the value of $\sum_{\ell \in \mathcal{S}_1} \omega_{\ell k}$ is primarily determined by the subnetwork size $|\mathcal{S}_1|$. In the example we are going to illustrate, we assume that the number of agents in subnetwork $\mathcal{S}_1$ is three times larger than the size of subnetwork $\mathcal{S}_2$. For clarity of visualization, we display only the belief of four receiving agents. From the lowermost panel in Figure 3.1, we observe that receiving agents $17, 18, 20$ converge to $\theta = 1$, i.e., to the opinion promoted by $\mathcal{S}_1$. We see also that agent 19 takes a minority position and opts for $\theta = 2$, i.e., it does follow neither the opinion promoted by $\mathcal{S}_1$ nor by $\mathcal{S}_2$. This shows the following interesting effect. Even if subnetwork $\mathcal{S}_1$ is bigger, for the specific topology shown in the example (see uppermost panel of Figure 3.1), the aggregate weight of agent 19 is $\sum_{\ell \in \mathcal{S}_1} \omega_{\ell\, 19} = 0.6885$. This means that condition (3.19) is actually verified, which explains why

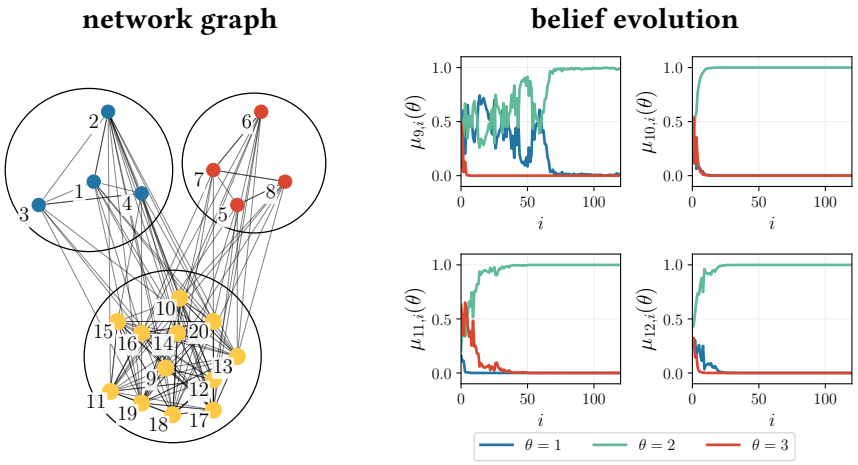


Figure 3.3: *Truth is somewhere in between.* (*Left*) Weakly connected network with balanced influences. (*Right*) Convergence of beliefs at receiving agents.

agent 19 opts for $\theta = 2$. Including in the analysis also the agents that are not displayed, in this example we have that $2/3$ of the receiving agents in the network opt for $\theta = 1$. In summary, we observed that building a majority of agents in $\mathcal{S}_1$ relative to $\mathcal{S}_2$ yields a majority of receiving agents opting for the hypothesis promoted by $\mathcal{S}_1$.

– Setup 2 or **How filter bubbles build a majority**. Under this setup, we assume that both sending components have the same size, however π_s is different for each of the two components. We set $\pi_1 = 0.9$ and $\pi_2 = 0.1$ in order to motivate agent k to have more connections with subnetwork $\mathcal{S}_1$ than with $\mathcal{S}_2$. This scenario is considered in Figure 3.2, where we see that the displayed receiving agents end up agreeing with opinion $\theta = 1$, i.e., with the opinion promoted by the sending component $\mathcal{S}_1$. Including in the analysis also the receiving agents that are not displayed, in this example we have that *all* the receiving agents in the network opt for $\theta = 1$. Therefore, closing a receiving agent into the "*filter bubble*" determined by the overwhelming flow of data coming from $\mathcal{S}_1$ essentially makes these agents blind to the solicitations coming from $\mathcal{S}_2$.

– Setup 3 or **Truth is somewhere in between**. We now address the balanced case where the sending subnetworks have the same size and similar number of connections to the receiving subnetwork ($\pi_1 = \pi_2 = 0.5$). Under this setting, it is expected that no dominant behavior emerges, and (3.19) holds. We see in Figure 3.3 that the opinions of receiving agents $9, 10, 11, 12$ tend to converge with full confidence to hypothesis $\theta = 2$ ($\mathsf{m}_\theta = 0$), which is an opinion pushed by none of the sending agents. How can we explain this effect? One interpretation is that, in the presence of *conflicting suggestions* coming from the two subnetworks, the receiving agent opts for a conservative choice. If sending subnetwork $\mathcal{S}_1$ says "choose $-\Delta$", while sending subnetwork $\mathcal{S}_2$ says "choose $+\Delta$", then the receiving agent prefers to be agnostic and stays in the middle, i.e., it chooses 0. Referring to real-life situations, we can think of one person betting on a soccer match between teams A and B. Assuming that discordant solicitations come from the environment, i.e., the person receives data suggesting to bet on the victory of team A, as well as data suggesting to bet on the victory of team B. If there is no sufficient evidence to let one suggestion prevail, then the most probable choice would be betting on a

draw! This "*truth-is-somewhere-in-between*" effect is a remarkable effect that is peculiar to the weakly connected setting, and that has been not observed before, e.g., it was not present in [48].

In summary, it is the cumulative influence of a sending group over a receiving agent that determines whether it will follow the group's opinion or not. This situation emulates the social phenomenon of herd behavior: agents choose to ignore their private signal in order to follow the most influencing group of agents. When none of the above dominance situations occurs, the receiving agent can opt for an opinion that is not promoted by any of the sending agents.

3.3 Influence Recovery

In the previous section we examined the effect of the network topology on the social learning of the agents. In particular, we discovered how the topology and the states of the sending agents determine the opinion formation by the receiving agents. The way the information is delivered across the network ultimately determines the minimizers in (2.65), i.e., the hypothesis around which each receiving agent's belief will concentrate. We now examine the reverse problem. Assume we observe the belief evolution of part of the network. We would like to use this information to infer the underlying topological influences. This is a useful question to consider because understanding the topology can help us understand why a particular agent adopts a certain opinion. The main question we consider now is this: given some measurements collected at the receiving agents, can we recover the influence exerted from sending subnetworks?

We answer this question under the following assumption of homogeneity of likelihoods and true distributions inside the individual sending subnetworks.

Assumption 3.1 (Homogeneity within sending subnetworks). For $s = 1, 2, \ldots, S$, we assume that the distribution and the likelihood functions within the s-th sending subnetwork are equal across all agents in that subnetwork, namely, for all $\ell \in \mathcal{S}_s$:

$$f_\ell = f^{(s)}, \qquad L_\ell(\theta) = L^{(s)}(\theta). \tag{3.21}$$

One main consequence of Assumption 3.1 is that (3.4) becomes:

$$\begin{aligned} D_k(\theta) &= \sum_{\ell \in \mathcal{S}} \omega_{\ell k} D(f_\ell || L_\ell(\theta)) \\ &= \sum_{s=1}^{S} \left(D(f^{(s)} || L^{(s)}(\theta)) \sum_{\ell \in \mathcal{S}_s} \omega_{\ell k} \right), \end{aligned} \tag{3.22}$$

where $\mathcal{S}_s$ denotes the collection of agents in the s-th sending subnetwork. Equation (3.22) has the following relevant implication. Under Assumption 3.1, the network topology influences the average divergence $D_k(\theta)$ through an *aggregate* weight:

$$x_{sk} \triangleq \sum_{\ell \in \mathcal{S}_s} \omega_{\ell k} = \sum_{\ell \in \mathcal{S}_s} w_{\ell k}. \tag{3.23}$$

The latter equality, using $w_{\ell k}$ instead of $\omega_{\ell k}$, comes straightforwardly from (2.58) and (2.59).

This equality reveals that the aggregate weights depend solely on the matrix W, and not on the matrix E of Perron eigenvectors. In other words, the inner structure of the pertinent sending subnetwork s does not influence the aggregate weight x_{sk}. We notice that, while a combination weight $a_{\ell k}$ accounts for a *local, small-scale* pairwise interaction between agent ℓ and agent k, the aggregate weight x_{sk} accounts for *macroscopic* topology effects, for two reasons. First of all, x_{sk} is determined by the *limiting* weights $\omega_{\ell k}$, which embody not only direct connection effects between ℓ and k, but also effects *mediated* by multi-hop paths connecting ℓ and k. Second, from (3.23) we see that x_{sk} embodies the *global* effect coming from all agents belonging to the s-th sending component. In other words, x_{sk} is a measure of the effect from all agents in sending subnetwork s on agent k. Since, in view of Theorem 2.2, the average divergence determines the behavior of the limiting belief, we conclude from (3.22) that the network topology ultimately determines the particular hypothesis chosen by a receiving agent only through these *global* influence weights $\{x_{sk}\}$.

We assume that the data available for estimating x_{sk} are the shared (intermediate) beliefs, $\boldsymbol{\psi}_{k,i}(\theta)$. We will say that *consistent* influence recovery is achievable if the x_{sk} can be correctly guessed when sufficient time is given for learning, i.e., we will focus on the *limiting* data, for all $\theta \neq \theta_k^\star$:[3]

$$y_k(\theta) \triangleq \lim_{i\to\infty} \frac{\log \boldsymbol{\psi}_{k,i}(\theta)}{i} \overset{\text{a.s.}}{=} D_k(\theta_k^\star) - D_k(\theta). \tag{3.25}$$

Accordingly, the influence recovery problem we are interested in can be formally stated as follows. For any receiving agent k, introduce its influence-weight vector:

$$x_k \triangleq [x_{1k}, x_{2k}, \ldots, x_{Sk}]^\mathsf{T}, \tag{3.26}$$

and consider the vector stacking the H limiting beliefs $y_k(\theta)$ (i.e., the data):

$$y_k \triangleq [y_k(1), y_k(2), \ldots, y_k(H)]^\mathsf{T}, \tag{3.27}$$

The main question is whether we can estimate x_k consistently from observation of y_k. In the sequel we will sometimes refer to this problem as a *macroscopic* topology learning problem—see Figure 3.4 for an illustration. In order to avoid confusion, we remark that the method proposed in this work does not allow retrieving the topology of the network (for that purpose, we refer the reader instead to [61], [62]), but the influence quantified by the aggregate weights x_{sk} that each sending subnetwork exerts on each receiving agent. While this information has the real topology of the network embedded in it, some other information is missing. For instance, topology inside the sending subnetworks and inside the receiving subnetworks is not considered.

As compared to topology inference problems, we are faced here with one critical element of novelty. We have no data coming from the sending agents. This means that correlation between sending and receiving agent pairs cannot be performed. This is in sharp contrast with traditional topology inference problems, where the estimation of connections between pairs of agents is

[3] In view of (3.2), we can write for $\theta, \theta' \in \Theta$:

$$\log \frac{\boldsymbol{\psi}_{k,i}(\theta)}{\boldsymbol{\psi}_{k,i}(\theta')} = \log \frac{\boldsymbol{\mu}_{k,i-1}(\theta)}{\boldsymbol{\mu}_{k,i-1}(\theta')} + \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta)}{L_k(\boldsymbol{\xi}_{k,i}|\theta')}. \tag{3.24}$$

Thus the asymptotic properties of $\boldsymbol{\psi}_{k,i}(\cdot)$ are the same as $\boldsymbol{\mu}_{k,i}(\cdot)$.

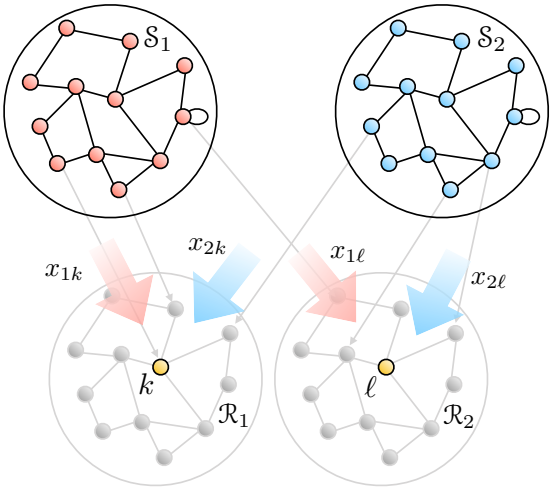


Figure 3.4: Influence recovery (or macroscopic topology learning) problem. The problem of influence recovery aims at finding the *global* influence weights x_{sk} from sending subnetwork s to receiving agent k. For example, consider a weakly connected network with two sending subnetworks, $\mathcal{S}_1$ and $\mathcal{S}_2$, and two receiving subnetworks $\mathcal{R}_1$ and $\mathcal{R}_2$. The weight x_{1k} in the figure embodies the influence of *all* sending agents in $\mathcal{S}_1$, from *all* paths (possibly including intermediate receiving agents) leading to receiving agent $k \in \mathcal{R}_1$.

heavily based on comparison (e.g., correlation) between data streams coming from these *pairs* of agents [61]–[63]. In contrast, we focus here on the asymmetrical case that, when estimating the weights x_{sk} from sending to receiving agents, no data are available from the sending agents. For this reason, the influence recovery problem addressed in this work is significantly different from traditional topology problems studied in the literature.

3.4 Is Influence Recovery Feasible?

We now examine the feasibility of the influence recovery problem illustrated in the previous section.

Let us preliminarily introduce a matrix $D = [d_{\theta s}]$, which collects the $H \times S$ divergences between any true distribution in the sending subnetworks and any likelihood, and whose (θ, s)-th entry is:

$$[D]_{\theta s} = d_{\theta s} = D(f^{(s)} || L^{(s)}(\theta)). \tag{3.28}$$

Using (3.26) and (3.28) in (3.22), the network divergence of receiving agent k, evaluated at θ, can be written as:

$$D_k(\theta) = \sum_{s=1}^{S} d_{\theta s} x_{sk}. \tag{3.29}$$

Through (3.27) we can rewrite the limiting data in (3.25) as:

$$y_k(\theta) = D(\theta_k^\star) - D(\theta) = \sum_{s=1}^{S} (d_{\theta_k^\star s} - d_{\theta s})\, x_{sk}. \tag{3.30}$$

It is useful to introduce the matrix:

$$B_k \triangleq \left(\mathbb{1}_H e_{\theta_k^\star}^{\mathsf{T}} - I_H\right) D, \tag{3.31}$$

where e_m is an $H \times 1$ vector with all zeros and a one in the m-th position. It is important to note that B_k has its $\theta_k^\star$-th row equal to zero. We can now formulate the influence recovery problem in terms of the following constrained system:

$$\text{Find } \widetilde{x}_k \in \mathbb{R}^S, \quad \text{such that } \begin{cases} y_k = B_k\,\widetilde{x}_k, \\ \sum_{s=1}^S \widetilde{x}_{sk} = 1, \\ \widetilde{x}_k > 0, \end{cases} \tag{3.32}$$

where we remark that the notation $\widetilde{x}_k > 0$ signifies that all entries in the solution vector $\widetilde{x}_k$ must be strictly positive. This positivity constraint is enforced because by assumption, each receiving subnetwork is connected to at least one agent from each sending subnetwork, which implies that the true vector we are looking for, x_k, has all positive entries. The equality constraint in (3.32) can be readily included in matrix form by introducing the augmented matrix and vector:

$$C_k \triangleq \begin{bmatrix} B_k \\ \mathbb{1}_S^{\mathsf{T}} \end{bmatrix}, \qquad \widetilde{y}_k \triangleq \begin{bmatrix} y_k \\ 1 \end{bmatrix}, \tag{3.33}$$

which allow rewriting (3.32) as:

$$\text{Find } \widetilde{x}_k \in \mathbb{R}^S : \;\; \widetilde{y}_k = C_k\,\widetilde{x}_k, \quad \widetilde{x}_k > 0. \tag{3.34}$$

We are now ready to state formally the concept of feasibility for the influence recovery problem. First, we want to solve the problem under the assumption that the matrix of divergences, D, is known, i.e., that sufficient knowledge is available about the underlying statistical models (likelihoods and true distributions). In this respect, we remark that the matrix B_k in (3.31) depends on $\theta_k^\star$, which in turn depends on the unknowns x_{sk} as well through (2.65). However, from (3.5) we know that the beliefs (and also the intermediate beliefs) converge to 1 at $\theta_k^\star$. Therefore, we can safely estimate $\theta_k^\star$ from the limiting data $y_k(\theta)$, which is tantamount to assuming that the matrix B_k is known.

Therefore, achievability of a consistent solution for the influence recovery problem translates into the condition that the linear system in (3.34) should admit a unique solution. We will now prove the following result.

Lemma 3.1 (Necessary condition for influence recovery). *The influence recovery problem described by the system in (3.34) admits a unique solution if, and only if:*

$$\text{rank}(C_k) = S. \tag{3.35}$$

Thus, a necessary condition for influence recovery is that the number of hypotheses is at least equal to the number of sending subnetworks, namely, that:

$$H \geq S. \tag{3.36}$$

Proof. We remark that we are not concerned with the existence of a solution for the constrained linear system (3.34). In fact, this system admits at least a solution, namely, the true weight vector, $x_k \in \mathbb{R}_+^S$, which by assumption fulfills the equation $\widetilde{y}_k = C_k\, x_k$.

Let us now focus on the unconstrained system (i.e., the system in (3.34) *without the inequality constraints*), whose set of solutions is given by [59]:

$$\widetilde{x}_k = C_k^{\dagger}\, \widetilde{y}_k + (I_S - C_k^{\dagger} C_k) z, \tag{3.37}$$

where $z \in \mathbb{R}^S$ is an arbitrary vector, and $C_k^{\dagger}$ is the Moore-Penrose pseudoinverse of C_k. If $\mathrm{rank}(C_k) = S$, it is well known [59] that $C_k^{\dagger} = (C_k^{\mathsf{T}} C_k)^{-1} C_k^{\mathsf{T}}$, which implies that the second term on the RHS in (3.37) is zero, which in turn implies that the unconstrained system has the unique solution:

$$\widetilde{x}_k = C_k^{\dagger} \widetilde{y}_k = (C_k^{\mathsf{T}} C_k)^{-1} C_k^{\mathsf{T}} \widetilde{y}_k = x_k. \tag{3.38}$$

The latter equality holds because, if the unconstrained system has a *unique* solution, this is also the unique solution for the constrained system, i.e., it coincides with x_k and satisfies the positivity constraints. Accordingly, we have proved that whenever $\mathrm{rank}(C_k) = S$, the *constrained* system has the unique solution corresponding to the true vector x_k.

We now show that when $\mathrm{rank}(C_k) < S$ the constrained system has infinite solutions. Since any solution of the unconstrained system takes on the form (3.37), and since x_k *is* a particular solution, there will exist a certain vector z_0 such that the x_k can be written as:

$$x_k = C_k^{\dagger}\, \widetilde{y}_k + (I_S - C_k^{\dagger} C_k) z_0. \tag{3.39}$$

Consider a solution $\widetilde{x}_k$ in (3.37) that corresponds to another vector, $z = z_0 + \epsilon$, where ϵ is a perturbation vector:

$$\widetilde{x}_k = C_k^{\dagger}\, \widetilde{y}_k + (I_S - C_k^{\dagger} C_k)(z_0 + \epsilon) = x_k + (I_S - C_k^{\dagger} C_k)\epsilon. \tag{3.40}$$

Since by assumption $x_k > 0$, we conclude from (3.40) that for sufficiently small perturbations it is always possible to obtain a distinct $\widetilde{x}_k > 0$, which implies that the *constrained* system in (3.34) has infinite solutions.

In summary, we conclude that the influence recovery problem is feasible if, and only if, $\mathrm{rank}(C_k) = S$. Finally, by observing that the augmented matrix C_k is an $(H+1) \times S$ matrix with an all-zeros row, we have in fact proved the claim of the lemma. □

Lemma 3.1 has at least three useful implications. First, it reveals a fundamental interplay between social learning and influence recovery: the possibility of estimating x_k depends on the comparison between two seemingly unrelated quantities, the number of hypotheses H (an attribute of the social inferential problem) and the number of sending subnetworks S (an attribute of the network topology).

Second, the necessary condition in (3.36) highlights that influence recovery over social networks is challenging. For example, if the agents of the social network want to solve a binary detection problem ($H = 2$), then the maximum number of sending subnetworks that could allow faithful

influence estimation is $S = 2$. Increasing the complexity of the social learning problem (i.e., increasing H) is beneficial to influence estimation, since it allows to increase also S.

Third, we see that having more sending subnetworks makes influence recovery more complicated. This is because increasing the number of sending subnetworks increases the number of unknowns (i.e., the dimension of x_k), while not adding information since in our setting we are not allowed to probe the sending nodes. Remarkably, when examining jointly the social learning and the influence recovery problems, *the role of the data and of the unknowns is exchanged.* In the social learning problem, more hypotheses means more unknowns and more sending subnetworks means more data; in the influence recovery problem, the situation is exactly reversed.

3.4.1 Structured Gaussian Models

In this section we consider the practical case of a Gaussian model, defined as follows.

- All agents use the same family of likelihood functions $\{L(\theta)\}$, for $\theta = 1, 2, \ldots, H$.
- These likelihoods are unit-variance Gaussian likelihoods with different means $\{\mathsf{m}_\theta\}$.
- Each true distribution coincides with one of the likelihoods. This implies that the distribution of the s-th sending subnetwork, $f^{(s)}$, is a unit-variance Gaussian distribution with mean ν_s that is chosen among the means $\{\mathsf{m}_\theta\}$, namely, for $s = 1, 2, \ldots, S$:

$$\nu_s \in \{\mathsf{m}_1, \mathsf{m}_2, \ldots, \mathsf{m}_H\}. \tag{3.41}$$

- The sending subnetworks have different means.

Using (3.28) and the definition of KL divergence between Gaussian distributions, the matrix D is given by:

$$D = \frac{1}{2} \begin{bmatrix} (\mathsf{m}_1 - \nu_1)^2 & (\mathsf{m}_1 - \nu_2)^2 & \ldots & (\mathsf{m}_1 - \nu_S)^2 \\ (\mathsf{m}_2 - \nu_1)^2 & (\mathsf{m}_2 - \nu_2)^2 & \ldots & (\mathsf{m}_2 - \nu_S)^2 \\ \vdots & & & \vdots \\ (\mathsf{m}_H - \nu_1)^2 & (\mathsf{m}_H - \nu_2)^2 & \ldots & (\mathsf{m}_H - \nu_S)^2 \end{bmatrix}. \tag{3.42}$$

From (3.42) it is readily seen that, if the sending subnetworks share the same true distribution (i.e., if $\nu_1 = \nu_2 = \cdots = \nu_S$), then the matrix D has rank 1, and, hence, the influence recovery problem is obviously not feasible. As said, we will instead focus on the opposite case where the true expectations are all distinct.

For ease of presentation, and without loss of generality we can assume that the sending subnetworks are numbered so that the expectations of the true distributions are:

$$\nu_1 = \mathsf{m}_1,\ \nu_2 = \mathsf{m}_2, \ldots, \nu_S = \mathsf{m}_S, \tag{3.43}$$

which implies that (3.42) takes on the form:

$$D = \frac{1}{2}\begin{bmatrix} 0 & (\mathsf{m}_1 - \mathsf{m}_2)^2 & \dots & (\mathsf{m}_1 - \mathsf{m}_S)^2 \\ (\mathsf{m}_2 - \mathsf{m}_1)^2 & 0 & \dots & (\mathsf{m}_2 - \mathsf{m}_S)^2 \\ \vdots & & & \vdots \\ (\mathsf{m}_H - \mathsf{m}_1)^2 & (\mathsf{m}_H - \mathsf{m}_2)^2 & \dots & (\mathsf{m}_H - \mathsf{m}_S)^2 \end{bmatrix}. \tag{3.44}$$

The structure in (3.44) implies that, for $H = S$, the matrix D is a Euclidean distance matrix (but for the constant $1/2$) [64]. These matrices are constructed as follows. Given points $r_1, r_2, \dots, r_L$, belonging to $\mathbb{R}^{\mathsf{dim}}$, the (i, j)-th entry of the matrix $\mathsf{EDM}(r_1, r_2, \dots, r_L)$ is given by the squared Euclidean distance between points r_i and r_j. Accordingly, we see from (3.44) that, for $H = S$:

$$D = \frac{1}{2}\mathsf{EDM}(\mathsf{m}_1, \mathsf{m}_2, \dots, \mathsf{m}_H). \tag{3.45}$$

For $H > S$, the matrix D can be described as an *extended* Euclidean distance matrix, constructed as follows. Let:

$$\begin{aligned} E_S &\triangleq \frac{1}{2}\mathsf{EDM}(\mathsf{m}_1, \mathsf{m}_2, \dots, \mathsf{m}_S), \\ E_H &\triangleq \frac{1}{2}\mathsf{EDM}(\mathsf{m}_1, \mathsf{m}_2, \dots, \mathsf{m}_H), \\ E_{H-S} &\triangleq \frac{1}{2}\mathsf{EDM}(\mathsf{m}_{S+1}, \mathsf{m}_{S+2}, \dots, \mathsf{m}_H), \end{aligned} \tag{3.46}$$

and let F be the $(H-S)\times S$ matrix with entries, for $\theta = S+1, S+2, \dots, H$ and $s = 1, 2, \dots, S$:

$$[F]_{\theta s} = \frac{1}{2}(\mathsf{m}_\theta - \mathsf{m}_s)^2. \tag{3.47}$$

Then, we have the following representation:

$$D = \begin{bmatrix} E_S \\ F \end{bmatrix}, \qquad E_H = \begin{bmatrix} E_S & F^{\mathsf{T}} \\ F & E_{H-S} \end{bmatrix}. \tag{3.48}$$

The following theorem, which establishes the feasibility of the influence recovery problem for the considered Gaussian model, relies heavily on some fundamental properties of Euclidean distance matrices.

Theorem 3.1 (Influence recovery under structured Gaussian models). *Let $S \geq 2$ and $H \geq S$. Assume that all sending subnetworks have the same family of unit-variance Gaussian likelihood functions $L(\theta)$ with distinct means $\{\mathsf{m}_\theta\}$, for $\theta = 1, 2, \dots, H$. Assume that the true distributions $f^{(s)}$, within the sending subnetworks $s = 1, 2, \dots, S$, are unit-variance Gaussian with distinct means ν_s, chosen from the collection $\{\mathsf{m}_\theta\}$. Then, under Assumption 2.6 (so that the matrix B_k in (3.31) is well defined), for all receiving agents $k \in \mathcal{R}$ we have that:*

$$\mathrm{rank}(C_k) = 2. \tag{3.49}$$

Proof. The proof is reported in Appendix 3.A. □

In view of Lemma 3.1, Eq. (3.49) has the following implication. Under the considered Gaussian model, influence recovery is feasible only when $S = 2$. We remark also that, when $S = 2$, condition (3.36) plays no role, since any meaningful classification problem has at least $H = 2$. In summary, Theorem 3.1 reveals that the structure of the Gaussian model makes influence recovery very challenging, as this problem is not solvable for networks with more than 2 sending subnetworks. Thus, the theorem reveals that $H \geq S$ is *not* a sufficient condition for consistent influence recovery.

3.4.2 Diversity Models

We can now examine the effect that diversity in the models of the sending subnetworks can have on influence recovery. Since the limiting beliefs are essentially determined by the divergence matrix D, it is meaningful to impose a form of diversity in terms of the divergences between distributions and likelihoods. In other words, differently from the Gaussian case illustrated in the previous section, we now require that the entries of D are not tightly related to each other, namely, we allow them to assume values in $\mathbb{R}_+^{H\times S}$ (where we denote by $\mathbb{R}_+$ the nonnegative reals) with no strong structure linking them.

One typical model for this type of diversity is that the divergences perceived by the different agents (i.e., across index s), and corresponding to different hypotheses (i.e., across index h), are modeled as absolutely continuous random variables. This randomness is a formal way to embody some degree of variability in how the agents "see" the world. For example, this is a useful model to consider when the agents, due to imperfect knowledge, have likelihoods that are slightly *perturbed* versions of some nominal model. Examples of this type are illustrated in the next section.

In order to avoid confusion, it is important to remark one fundamental property. Under the diversity setting, the matrix $\boldsymbol{D}$ is random[4] with entries modeled as absolutely continuous random variables. The full-rank property for this type of matrices is a classical result. However, we observe from (3.31) that the matrix $\boldsymbol{B}_k$ is obtained from $\boldsymbol{D}$ by multiplying a matrix that depends on a random variable $\boldsymbol{\theta}_k^\star$, which in turn depends statistically upon the entries of $\boldsymbol{D}$. Finally, we know from (3.33) that $\boldsymbol{C}_k$ is obtained from $\boldsymbol{B}_k$ by adding an all-ones row. Accordingly, to determine the rank of $\boldsymbol{C}_k$ we need to address carefully these intricate dependencies. This is accomplished in the proof of the forthcoming Theorem 3.2.

Theorem 3.2 (Influence recovery under general models with diversity). *Let $H \geq S$, and assume that the array $\{\boldsymbol{d}_{\theta s}\}$, with $\theta = 1, 2, \ldots, H$ and $s = 1, 2, \ldots, S$, is made of random variables that are jointly absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}_+^{H\times S}$. Then, for all receiving agents $k \in \mathcal{R}$ we have that, with probability 1, Assumption 2.6 is verified and the matrix $\boldsymbol{C}_k$ is full column rank, namely,*

$$\mathbb{P}\left(\boldsymbol{\theta}_k^\star \text{ is unique and } \operatorname{rank}(\boldsymbol{C}_k) = S\right) = 1. \tag{3.50}$$

Proof. The proof is reported in Appendix 3.B. □

[4] Accordingly, we will now use the bold notation for the matrix entries, $\boldsymbol{d}_{\theta s}$, as well as for other related quantities.

The meaning of Theorem 3.2 is that *configurations of KL divergence that lead to a rank-deficient matrix C_k are rare.* In other words, if some diversity exists in the statistical models of the sending components, then the influence recovery problem is feasible for almost all configurations.

3.5 Simulation Results

We now present some illustrative examples. The first example refers to the Gaussian model presented in Section 3.4.1. The other two examples refer to the setting with diversity presented in Section 3.4.2.

a) Gaussian with $H = S = 2$. We consider the topology shown in the left panel of Figure 3.5. The likelihoods and true distributions for the sending subnetworks are unit-variance Gaussian with means $\nu_1 = \mathsf{m}_1 = 1$, $\nu_2 = \mathsf{m}_2 = 2$. The receiving agents[5] employ the same likelihoods of the sending agents, and their true distributions are unit-variance Gaussian with mean equal to 1. In Figure 3.5 (right) we show the belief convergence for four receiving agents.

Next, we address the influence recovery problem. First, for an observation time i, we construct the empirical data $\widehat{y}_k(\theta) = (1/i) \log \boldsymbol{\psi}_{k,i}(\theta)$, and construct an estimate $\widehat{\theta}_k^\star$ as the value of θ that maximizes $\widehat{y}_k(\theta)$ (i.e., the hypothesis where $\widehat{y}_k(\theta)$ will collapse to 1). We can then construct an estimate for B_k as:

$$\widehat{B}_k = \left(\mathbb{1}_H e_{\widehat{\theta}_k^\star}^{\mathsf{T}} - I_H \right) D, \tag{3.51}$$

from which we obtain $\widehat{C}_k$ by adding an all-ones row, according to (3.33). At this point, we have verified on the simulated data that, for any receiving agent $k \in \{9, 10, 11, 12\}$, the matrices $\widehat{C}_k$ are full column rank. Then, we used (3.38) with empirical matrices replacing the exact ones to estimate the connection-weight vector x_k as:[6]

$$\widehat{x}_k = \widehat{C}_k^\dagger \begin{bmatrix} \widehat{y}_k \\ 1 \end{bmatrix} = (\widehat{C}_k^{\mathsf{T}} \widehat{C}_k)^{-1} \widehat{C}_k^{\mathsf{T}} \begin{bmatrix} \widehat{y}_k \\ 1 \end{bmatrix}. \tag{3.52}$$

We see from Figure 3.6 that this procedure allows us to retrieve the influence weights $\{x_{sk}\}$, provided that the system evolves for a sufficiently long time.

b) Randomly perturbed Gaussian with $H = S = 3$. The network topology has three sending subnetworks and one receiving subnetwork as shown in the left panel of Figure 3.7. When $S > 2$, we know from Theorem 3.1 that for the *structured* Gaussian model, diversity in the sending components is not enough to ensure the full column rank of the matrix C_k. In order to increase diversity, we consider a *randomly perturbed* model for the likelihood functions, where the likelihood of the s-th sending subnetwork, evaluated at hypothesis θ, is unit-variance Gaussian with mean $\theta + \boldsymbol{\epsilon}_{\theta s}$. The random variables $\{\boldsymbol{\epsilon}_{\theta s}\}$ are equally correlated zero-mean Gaussian with variance equal to 0.02 and Pearson correlation coefficient equal to 0.5. For the receiving subnetwork we use the same type of random perturbation of the likelihoods. The true distributions for *all* sending and receiving agents are unit-variance Gaussian with mean equal to

[5] We recall that the models of the receiving agents will be ultimately immaterial as regards their limiting beliefs.

[6] The symbol $\widehat{\ }$ is used for quantities *estimated* from the data, to be not confused with the symbol $\widetilde{\ }$ used for the *exact* quantities appearing in (3.34).

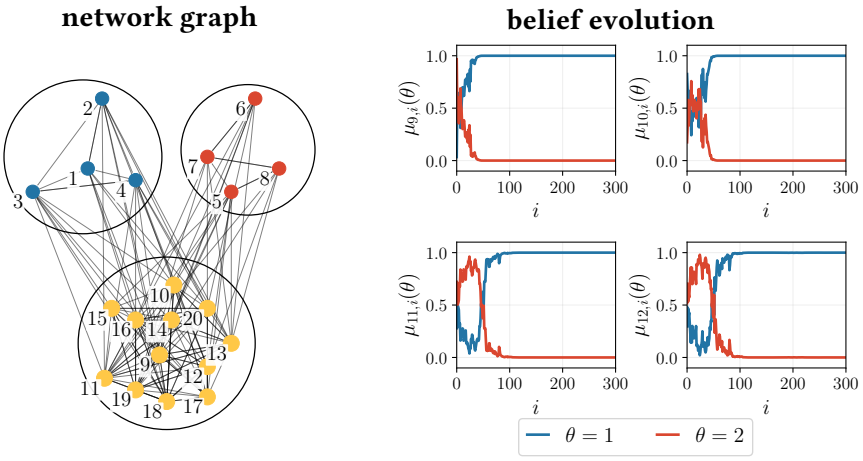


Figure 3.5: Unperturbed Gaussian model: (*Left*) Network topology. (*Right*) Belief convergence at the receiving agents.

1. The belief convergence for four receiving agents can be seen in the right panel of Figure 3.7. In Figure 3.8 we see how the estimates $\{\widehat{x}_{sk}\}$ of the influence weights converge to the true values $\{x_{sk}\}$. In contrast with the *structured* Gaussian case, influence recovery is now feasible for $S > 2$ and even if the true distributions are equal across all sending components. This change in behavior is due to the *diversity* in the models of the sending subnetworks, represented by the different means of the likelihoods. Moreover, we see from the parameters of the random variables $\{\boldsymbol{\epsilon}_{\theta s}\}$ that a relatively small perturbation is already sufficient to enable consistent influence recovery.

c) Beta with $H = S = 3$. Finally, we consider a non-Gaussian example. Moreover, since in the previous examples (motivated by what is typically observed in many networks) we have considered a number of receiving agents fairly larger than the size of the sending subnetworks, we now explore a case where the size of the receiving subnetwork is equal to the size of the sending subnetworks.

The non-Gaussian setting used in Figure 3.9 considers likelihood functions following a Beta distribution with scale parameter equal to 2 and with shape parameters given by $\theta + 1 + \boldsymbol{u}_{\theta s}$, where $\{\boldsymbol{u}_{\theta s}\}$, for $\theta \in \{1, 2, 3\}$ and $s \in \{1, 2, 3\}$, are independent random variables sampled from a uniform distribution with support $[-0.1, 0.1]$. The true distributions coincide with the unperturbed likelihoods, i.e., the true distribution of the s-th sending subnetwork is a Beta distribution with scale parameter equal to 2 and shape parameter equal to $s + 1$. For the receiving subnetwork we apply the same type of random perturbation of the likelihoods, whereas the true distributions are Beta with scale and shape parameters equal to 2. The belief convergence for the receiving agents can be seen in the right panel of Figure 3.9. In Figure 3.10, we see the convergence of the estimated influence weights.

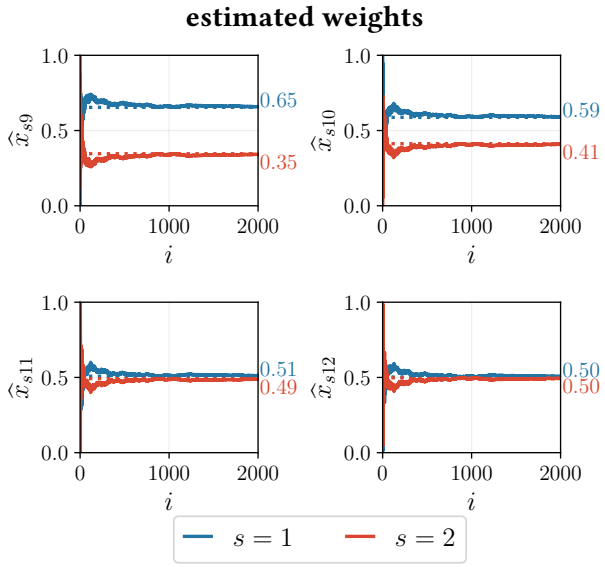


Figure 3.6: Unperturbed Gaussian model: Estimated influence weights. For each of the four panels, the numbers on the right denote the true values $\{x_{sk}\}$, with different colors denoting different s, according to the legend.

3.5.1 An Example of Noisy Influence Recovery

Let us consider an influence recovery problem that is feasible according to our previous models and results. In practice, different sources of error can alter these models (and possibly the results). In this section we focus on a relevant source of error and show that the proposed strategy is stable with respect to it.

In the previous treatment, the divergence matrix D was assumed known. However, in some applications this knowledge can be approximate, and D can be known up to a certain error $\delta D \in \mathbb{R}^{H\times S}$. Under this assumption, the solution in (3.38) is replaced by the following noisy version (agent index k suppressed for ease of notation):

$$x + \delta x = (C + \delta C)^{\dagger}\, \widetilde{y}, \tag{3.53}$$

where $\delta C \in \mathbb{R}^{(H+1)\times S}$ is the error induced by δD on C—see (3.31) and (3.33)—and $\delta x \in \mathbb{R}^{S}$ is the error induced by δC on the true solution x. We now quantify the error δx.

Since we are considering a feasible influence recovery problem, we have $H \geq S$ and $\mathrm{rank}(C) = S$. We also know that a noisy matrix δD would typically preserve the rank[7] of C, and, hence, we assume that $\mathrm{rank}(C + \delta C) = S$. Finally, we introduce the condition number $\kappa \triangleq \|C\|_2\|C^{\dagger}\|_2$ (where $\|\cdot\|_2$ is the spectral norm), and assume that the matrix C is well-conditioned and the noise is small such that $\|\delta C\|_2\|C^{\dagger}\|_2 < 1$. Under these assumptions, Theorem 5.1 in [65]

[7]For example, if the entries of δD are modeled as jointly absolutely continuous random variables, reasoning as in Theorem 3.2 we have $\mathrm{rank}(C + \delta C) = S$ with probability 1.

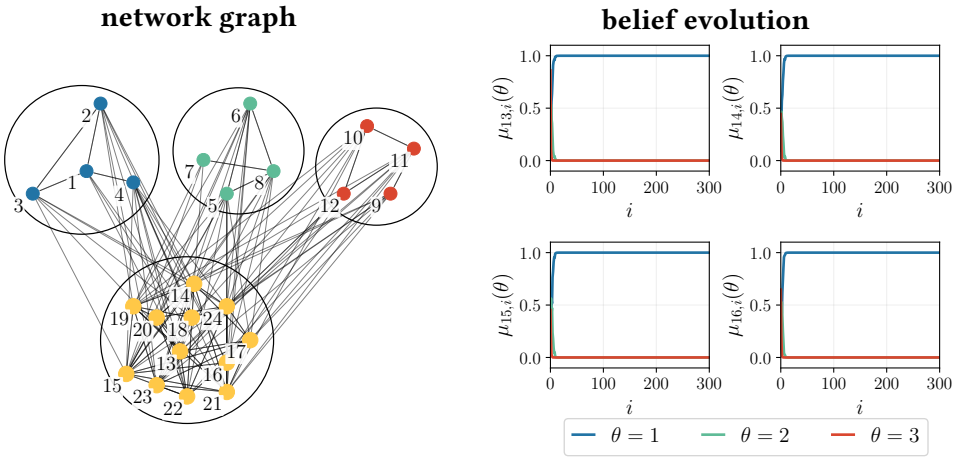


Figure 3.7: Perturbed Gaussian model: (*Left*) Network topology. (*Right*) Belief convergence at the receiving agents.

provides the following bound on the relative error:

$$\frac{\|\delta x\|_2}{\|x\|_2} \leq \frac{\kappa}{1 - \kappa \frac{\|\delta C\|_2}{\|C\|_2}} \frac{\|\delta C\|_2}{\|C\|_2}, \tag{3.54}$$

which reveals that, for sufficiently small deviations δC, the relative error $\|\delta x\|_2/\|x\|_2$ is on the same order as the relative error $\|\delta C\|_2/\|C\|_2$ [59].

Let us now provide a numerical example to illustrate how (3.54) works in practice. We consider the same setting of Figure 3.5, focusing on receiving agent 10, for which the exact weight vector is given by $x = [0.59,\ 0.41]^{\mathsf{T}}$. Now, in the considered example we have:

$$D = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}, \tag{3.55}$$

We generate the matrix δD as follows. The off-diagonal entries of δD are independent random variables following a zero-mean Gaussian distribution with standard deviation σ (when the resulting off-diagonal entries of $D + \delta D$ are negative we resample until nonnegative entries are obtained). The main-diagonal entries of δD are independent random variables distributed as the absolute value of zero-mean Gaussian random variables with standard deviation σ. Then we apply the influence estimation procedure described in Section 3.5. In Table 3.1 we report, for several values of σ, the root-mean-square error, $\|\delta x\|_2^{\mathsf{rms}}$, computed over 10^3 Monte Carlo iterations for each value of σ. Examining Table 3.1, we see that the influence recovery strategy is in fact stable w.r.t. to the noise introduced on the divergence matrix.[8]

[8] In the considered example, it is straightforward to relate the error $\|\delta x\|_2^{\mathsf{rms}}$ to the errors relative to the individual entries of the true solution, $x = [0.59, 0.41]^{\mathsf{T}}$. Since both the perturbed and true solution have sum equal to 1, the entries of δx have sum equal to 0, which implies that their (common) root-mean-square value is $\|\delta x\|_2^{\mathsf{rms}}/\sqrt{2}$.

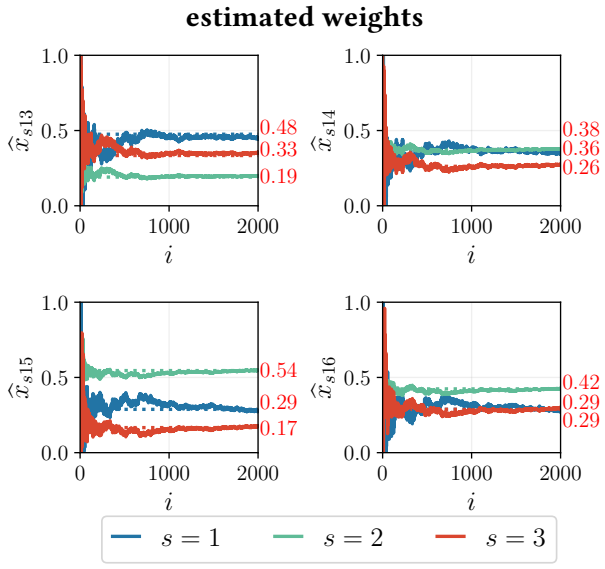


Figure 3.8: Perturbed Gaussian model: Estimated influence weights. For each of the four panels, the numbers on the right denote the true values $\{x_{sk}\}$, with different colors denoting different s, according to the legend.

Table 3.1: Root-mean-square error for different values of σ. The true solution is $x = [0.59,\ 0.41]^{\mathsf{T}}$.

σ	0.001	0.026	0.051	0.075	0.100
$\|\delta x\|_2^{\mathsf{rms}}$	0.001	0.033	0.069	0.105	0.156

3.6 Concluding Remarks

We considered a network where agents solve the Social Learning (SL) problem. These agents aim at forming their opinions after consulting the beliefs of their neighbors through an iterative update-and-combine SL algorithm. In this chapter we addressed the Influence Recovery (IR) problem, where a receiving agent (or some entity monitoring its behavior) attempts to get knowledge about the influence, in the form of network connections, from each sending subnetwork upon that receiving agent. We can refer to the SL problem as the *direct* learning problem, in the sense that it is the original inferential problem the network is deployed for. Likewise, we can refer to the IR problem as the *dual* learning problem, since it is an inferential procedure that takes as input data *the output of the direct SL problem.*

The analysis conducted in this chapter has revealed an interesting interplay between SL and IR problems. First, we established in Lemma 3.1 that $H \geq S$ is a necessary condition to achieve consistent IR, where S denotes the number of sending subnetworks, and H the number of hypotheses. In a sense, the number of hypotheses is an index (even if not the only one) of complexity associated to the SL problem since, other conditions being equal, more hypotheses make the SL problem more complicated. Likewise, the number of sending components represents an index of complexity of the IR problem, since, other conditions being equal, estimating more links is more complicated. According to these remarks, the condition $H \geq S$ *implies that the IR problem can be feasible when its complexity is not greater than the complexity of the SL*

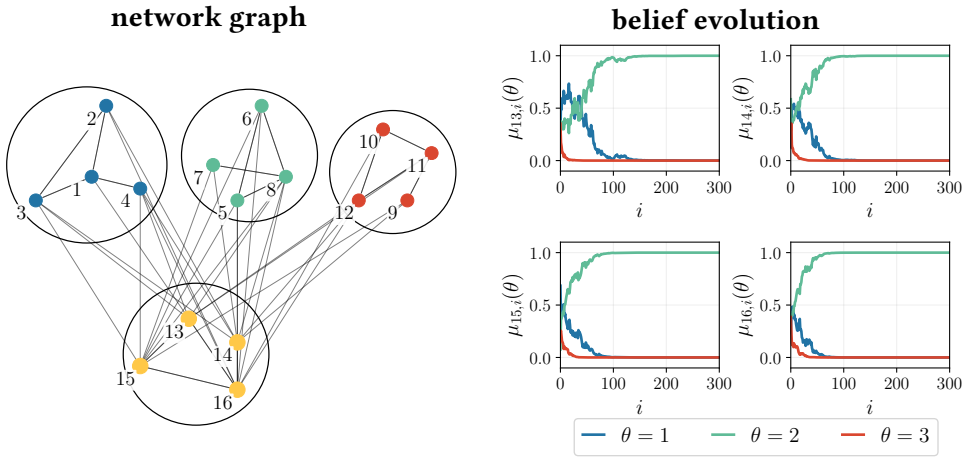


Figure 3.9: Perturbed Beta model: (*Left*) Network topology. (*Right*) Belief convergence at the receiving agents.

problem. In contrast, in topology inference problems, the connections between agents are inferred from some kind of pairwise measure of their dependence. In our setting, since we cannot measure the output of the sending subnetwork, we do not have data quantifying the direct topological dependence between a receiving and a sending agent. Our IR inference is based instead on beliefs at receiving agents. The belief contains some richness of information, i.e., its H components, which is critical to enable feasibility of the IR problem. In particular, $H \geq S$ means that the richness of information in the belief function should be greater than or equal to the number of unknown influence weights to be estimated, S.

Having established a necessary condition for consistent IR, we moved on to examine some useful models to see whether and when consistent IR is in fact achievable. First, we have considered a structured Gaussian model where *all* sending subnetworks use the same family of Gaussian likelihoods, and the sending subnetworks have distinct true distributions, each one coinciding with one of the likelihoods. We have shown in Theorem 3.1 that the IR problem is feasible only if $S = 2$, for any $H \geq 2$. The limited possibility of achieving consistent IR can be explained by the limited diversity existing between the different subnetworks, i.e., they all use the same family of likelihoods. This observation motivated the analysis of more general models with a certain degree of *diversity*, a condition formalized by saying that the KL divergences between true distributions and likelihoods are not structured, i.e., they are nonnegative real numbers with no particular relationship among them. Under this setting we have showed that, if $H \geq S$, the IR problem becomes feasible for almost all configurations, in a precise mathematical sense as stated in Theorem 3.2. In summary, two critical features that enable consistent IR are: *More hypotheses than sending components* and *a sufficient degree of diversity.*

3.A Proof of Theorem 3.1

Preliminarily, it is useful to introduce some auxiliary matrices. We let, for all $\theta = 1, 2, \ldots, H$:

$$\mathcal{I}(\theta) \triangleq \mathbb{1}_H e_\theta^\mathsf{T} - I_H \tag{3.56}$$

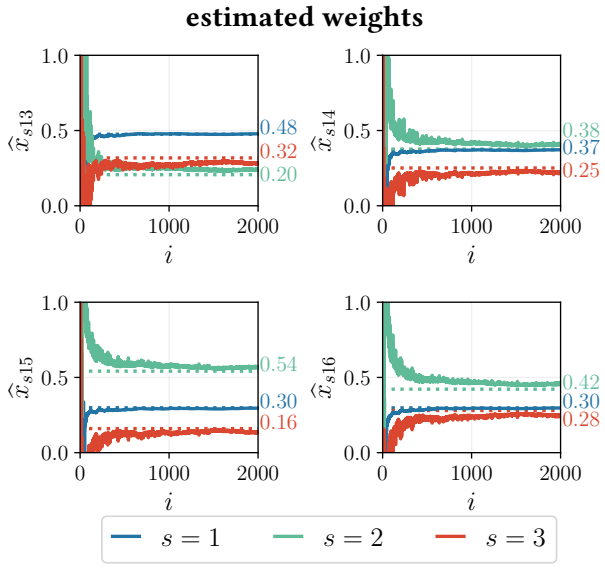


Figure 3.10: Perturbed Beta model: Estimated influence weights. For each of the four panels, the numbers on the right denote the true values $\{x_{sk}\}$, with different colors denoting different s, according to the legend.

and

$$B(\theta) \triangleq \mathfrak{I}(\theta)D, \quad C(\theta) = \begin{bmatrix} B(\theta) \\ \mathbb{1}_S^\mathsf{T} \end{bmatrix}. \tag{3.57}$$

In view of Eqs. (3.31) and (3.33), the definitions in (3.56) and (3.57) imply:

$$B_k = B(\theta_k^\star), \quad C_k = C(\theta_k^\star). \tag{3.58}$$

We continue by showing some useful properties of the matrix D under the considered Gaussian model. Let us focus on the representation in (3.48). It is a known result that the rank of a Euclidean distance matrix with n points in $\mathbb{R}^{\mathsf{dim}}$ is at most $\mathsf{dim} + 2$ [64]. Since in our case $\mathsf{dim} = 1$, we, can write:

$$\mathrm{rank}(E_S) \leq 3 \tag{3.59}$$

Moreover, for the cases $S = 2$ and $S = 3$ we have that:

$$E_2 = \frac{1}{2}\begin{bmatrix} 0 & (\mathsf{m}_1 - \mathsf{m}_2)^2 \\ (\mathsf{m}_2 - \mathsf{m}_1)^2 & 0 \end{bmatrix}, \tag{3.60}$$

$$E_3 = \frac{1}{2}\begin{bmatrix} 0 & (\mathsf{m}_1 - \mathsf{m}_2)^2 & (\mathsf{m}_1 - \mathsf{m}_3)^2 \\ (\mathsf{m}_2 - \mathsf{m}_1)^2 & 0 & (\mathsf{m}_2 - \mathsf{m}_3)^2 \\ (\mathsf{m}_3 - \mathsf{m}_1)^2 & (\mathsf{m}_3 - \mathsf{m}_2)^2 & 0 \end{bmatrix} \tag{3.61}$$

and, hence:

$$\det(E_2) = -\frac{1}{4}(\mathsf{m}_1 - \mathsf{m}_2)^2,$$

$$\det(E_3) = \frac{1}{4}(\mathsf{m}_1 - \mathsf{m}_2)^2(\mathsf{m}_1 - \mathsf{m}_3)^2(\mathsf{m}_2 - \mathsf{m}_3)^2. \tag{3.62}$$

Therefore, when the points that determine the Euclidean distance matrix are all distinct, both the above matrices are full rank. Thus, when $S = 2$, we have that $\mathrm{rank}(E_S) = 2$. When $S > 2$, since E_3 is full rank, and in view of (3.59), we have instead $\mathrm{rank}(E_S) = 3$. From the representation of D in (3.48), we then conclude that:

$$\mathrm{rank}(D) = \begin{cases} 2, & \text{if } S = 2, \\ 3, & \text{if } S > 2. \end{cases} \tag{3.63}$$

Next we state and prove a useful lemma.

Lemma 3.2. *Let $\mathcal{I}(\theta)$ be defined as in (3.56). Then, for all $\theta = 1, 2, \ldots, H$ we have that:*

$$I_H - \mathcal{I}^\dagger(\theta)\mathcal{I}(\theta) = \frac{1}{H}\mathbb{1}\mathbb{1}^\mathsf{T}. \tag{3.64}$$

Proof. For ease of notation, in the following proof the explicit dependence on θ is suppressed, and we write $\mathcal{I}$ in place of $\mathcal{I}(\theta)$. By definition of the Moore-Penrose inverse, matrix $\mathcal{I}^\dagger$ satisfies:

$$\mathcal{I}\mathcal{I}^\dagger\mathcal{I} = \mathcal{I}, \qquad (\mathcal{I}^\dagger\mathcal{I})^\mathsf{T} = \mathcal{I}^\dagger\mathcal{I}. \tag{3.65}$$

Then we note that:

$$\mathcal{I}(I_H - \mathcal{I}^\dagger\mathcal{I}) = \mathcal{I} - \mathcal{I}\mathcal{I}^\dagger\mathcal{I} = \mathcal{I} - \mathcal{I} = 0, \tag{3.66}$$

where in the second equality we used the first identity in (3.65). Equation (3.66) implies that the columns of $(I_H - \mathcal{I}^\dagger\mathcal{I})$ belong to the null space of $\mathcal{I}$, denoted by $\mathscr{N}(\mathcal{I}) = \{v : \mathcal{I}v = 0\}$. On the other hand, in view of (3.56) we can write:

$$\mathcal{I}v = \mathbb{1}_H e_\theta^\mathsf{T} v - v = \mathbb{1}_H v_\theta - v = 0, \tag{3.67}$$

with v_θ the θ-th element of v. As a result, Eq. (3.67) will be satisfied only if $v_h = v_\theta$ for all $h = 1, \ldots, H$. Therefore, we obtain:

$$\mathscr{N}(\mathcal{I}) = \{\alpha\mathbb{1}_H : \alpha \in \mathbb{R}\}, \tag{3.68}$$

further implying, in light of (3.66), that, for each $h = 1, 2, \ldots, H$, the h-th column of $I_H - \mathcal{I}^\dagger\mathcal{I}$ is of form $\alpha_h\mathbb{1}_H$ for some $\{\alpha_h\}$. On the other hand, since $I_H - \mathcal{I}^\dagger\mathcal{I}$ is symmetric in view of the second identity in (3.65), we conclude that $\alpha_h = \overline{\alpha}$ for all h, namely,

$$I_H - \mathcal{I}^\dagger\mathcal{I} = \overline{\alpha}\mathbb{1}_H\mathbb{1}_H^\mathsf{T}, \tag{3.69}$$

for some $\overline{\alpha} \in \mathbb{R}$. Finally, since in particular $\mathbb{1}_H \in \mathscr{N}(\mathcal{I})$, we can write:

$$(I_H - \mathcal{I}^\dagger\mathcal{I})\mathbb{1}_H = \mathbb{1}_H - \mathcal{I}^\dagger\mathcal{I}\mathbb{1}_H = \mathbb{1}_H, \tag{3.70}$$

which, in view of (3.69), yields:

$$\overline{\alpha}\,\mathbb{1}_H\mathbb{1}_H^{\mathsf{T}}\mathbb{1}_H = \overline{\alpha}H\mathbb{1}_H = \mathbb{1}_H \Rightarrow \overline{\alpha} = \frac{1}{H} \tag{3.71}$$

and we have in fact proved (3.64). □

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. We will now show that

$$\operatorname{rank}(C(\theta)) = 2 \quad \text{for all } \theta = 1, 2, \ldots, H, \tag{3.72}$$

which clearly implies the claim of the theorem in view of the second equation in (3.58).

For the case $H = S = 2$, it is immediately seen that the matrix $C(\theta)$ (assuming, e.g., $\theta = 1$) takes on the form:

$$C(1) = \begin{bmatrix} 0 & 0 \\ -\dfrac{(\mathsf{m}_1 - \mathsf{m}_2)^2}{2} & \dfrac{(\mathsf{m}_1 - \mathsf{m}_2)^2}{2} \\ 1 & 1 \end{bmatrix}, \tag{3.73}$$

which reveals that $\operatorname{rank}(C(\theta)) = 2$.

Let us move on to examine the other cases where $H \geq S$ (excluding $H = S = 2$). We will examine first the properties of the matrix $B(\theta)$ in (3.57). As done before, the dependence on θ is suppressed for ease of notation, and, in particular, we write B, C, and $\mathcal{J}$ in place of $B(\theta)$, $C(\theta)$, and $\mathcal{J}(\theta)$, respectively. Applying Sylvester's inequality to the first equation in (3.57) we can write [59]:

$$\operatorname{rank}(B) \geq \operatorname{rank}(D) + \operatorname{rank}(\mathcal{J}) - H = \operatorname{rank}(D) - 1, \tag{3.74}$$

where in the latter equality we used the fact that $\operatorname{rank}(\mathcal{J}) = H - 1$. Therefore, from (3.63) and (3.74) we conclude that:

$$\operatorname{rank}(B) \geq 1, \quad \text{if } S = 2, \tag{3.75}$$

$$\operatorname{rank}(B) \geq 2, \quad \text{if } S > 2. \tag{3.76}$$

Now we would like to see if equality is satisfied for the cases $S = 2$ (with $H > 2$) and $S > 2$ (with $H \geq S$).

To this end, we start by noticing that equality in Sylvester's inequality holds if, and only if, there exist matrices X and Y that solve [59]:

$$DX + Y\mathcal{J} = I_H, \tag{3.77}$$

which in turn admits a solution if, and only if, [66]:

$$(I_H - DD^{\dagger})(I_H - \mathcal{J}^{\dagger}\mathcal{J}) = 0. \tag{3.78}$$

Applying Lemma 3.2, from (3.78) we get:

$$(I_H - DD^\dagger)\frac{1}{H}\mathbb{1}_H\mathbb{1}_H^\mathsf{T} = 0, \tag{3.79}$$

which means that the equality sign in (3.75) or (3.76) holds if, and only if:

$$DD^\dagger\mathbb{1}_H = \mathbb{1}_H. \tag{3.80}$$

In particular, we will now show that (3.80) does not hold for $S = 2$, while it holds for $S > 2$.

Let us start with the case $S = 2$ (and $H > 2$). We will appeal to the representation of D in (3.48), which for the case $S = 2$ can be written as:

$$D = \frac{1}{2}\begin{bmatrix} 0 & (\mathsf{m}_1 - \mathsf{m}_2)^2 \\ (\mathsf{m}_2 - \mathsf{m}_1)^2 & 0 \\ (\mathsf{m}_3 - \mathsf{m}_1)^2 & (\mathsf{m}_3 - \mathsf{m}_2)^2 \\ \vdots & \vdots \\ (\mathsf{m}_H - \mathsf{m}_1)^2 & (\mathsf{m}_H - \mathsf{m}_2)^2 \end{bmatrix}. \tag{3.81}$$

Let us now consider the linear system $Dv = \mathbb{1}_H$. From the first two rows of D, we get the unique solution: $v = 2(\mathsf{m}_1 - \mathsf{m}_2)^{-2}\mathbb{1}_2$. Considering now the third row, we get the identity $(\mathsf{m}_3 - \mathsf{m}_1)^2 + (\mathsf{m}_3 - \mathsf{m}_2)^2 = (\mathsf{m}_1 - \mathsf{m}_2)^2$, which is true only if the third point, m_3, is equal to one of the previous points. We conclude that there exist no v such that $Dv = \mathbb{1}_H$, which further implies that $DD^\dagger\mathbb{1}_H \neq \mathbb{1}_H$ Therefore, for $S = 2$ Eq. (3.75) gives $\mathrm{rank}(B) > 1$, which since B is of dimension $H \times 2$, with $H > 2$, implies that $\mathrm{rank}(B) = 2$.

Let us move on to examine the case $S > 2$ and $H \geq 2$. It is known that, for an $L \times L$ Euclidean distance matrix M, one has $MM^\dagger\mathbb{1}_L = \mathbb{1}_L$, implying that $\mathbb{1}_L$ belongs to the range space of M [67]. We can apply this result to the matrices E_S and E_H in (3.48), since they are proportional to Euclidean distance matrices. In particular, we can say that there exist vectors u_S and u_H such that $E_S u_S = \mathbb{1}_S$ and $E_H u_H = \mathbb{1}_H$. In particular, one of the (infinite) solutions is given by

$$u_H^\star = \begin{bmatrix} u_S \\ 0 \end{bmatrix}. \tag{3.82}$$

Applying now (3.82) into (3.48), we can write:

$$\mathbb{1}_H = E_H u_H^\star = \begin{bmatrix} E_S & F^\mathsf{T} \\ F & E_{H-S} \end{bmatrix}\begin{bmatrix} u_S \\ 0 \end{bmatrix} = \begin{bmatrix} E_S \\ F \end{bmatrix} u_S = Du_S. \tag{3.83}$$

Equation (3.80) now follows by observing that:

$$DD^\dagger \underbrace{\mathbb{1}_H}_{Du_S} = \underbrace{DD^\dagger D}_{D} u_S = Du_S = \mathbb{1}_H. \tag{3.84}$$

We have in fact shown that (3.80) holds true for $S > 2$, which implies that (3.76) becomes an equality for $S > 2$.

In summary, we have shown so far that $\mathrm{rank}(B) = 2$ for all $H \geq S$ (but for the case $H = S = 2$, which has been examined separately). We will now use this result to prove the claim of the theorem, namely, that $\mathrm{rank}(C) = 2$. Since C is obtained from B by adding an all-ones row, determining the rank of C from that of B amounts to check whether the row vector $\mathbb{1}_S^\mathsf{T}$ lies in the row space of B, which is tantamount to ascertaining whether there exists z such that:

$$z^\mathsf{T} \mathcal{J} D = \mathbb{1}_S^\mathsf{T}. \tag{3.85}$$

Since we exclude the case $H = S = 2$, we have always $H \geq 3$. Now, let us consider an EDM E_3 defined on 3 distinct points p_1, p_2, p_3. Since in this case E_3 is full rank, the system $v_3^\mathsf{T} E_3 = \mathbb{1}_3^\mathsf{T}$ has the following (unique) solution:

$$v_3^\mathsf{T} = \begin{bmatrix} \frac{e_{13}+e_{12}-e_{23}}{e_{13}e_{12}} & \frac{e_{12}+e_{23}-e_{13}}{e_{12}e_{23}} & \frac{e_{13}+e_{23}-e_{12}}{e_{13}e_{23}} \end{bmatrix}, \tag{3.86}$$

where we denoted by $e_{ij} = 1/2(p_i - p_j)^2$ the (i,j)-th entry of E_3. Let us now introduce the vector:

$$v_H^\mathsf{T} = [v_3^\mathsf{T} \;\; 0_{H-3}^\mathsf{T}]. \tag{3.87}$$

Since, for $H \geq 3$, we know that $\mathrm{rank}(E_H) = 3$, we conclude that:

$$v_3^\mathsf{T} E_3 = \mathbb{1}_3^\mathsf{T} \Rightarrow v_H^\mathsf{T} E_H = \mathbb{1}_H^\mathsf{T}, \tag{3.88}$$

which, using the block representation of D in (3.48), yields:

$$v_H^\mathsf{T} D = \mathbb{1}_S^\mathsf{T}. \tag{3.89}$$

In view of (3.89), one solution z to (3.85) exists if $z^\mathsf{T} \mathcal{J} = v_H^\mathsf{T}$, that is, if v_H^T lies in the row space of $\mathcal{J}$.

On the other hand, from the definition in (3.56), we see that the matrix $\mathcal{J}$ can be represented as:

$$\mathcal{J} = \begin{bmatrix} -1 & 0 & \dots & 0 & \mathbf{1} & 0 \dots 0 \\ 0 & -1 & \dots & 0 & \mathbf{1} & 0 \dots 0 \\ \vdots & \vdots & & \vdots & & \\ 0 & 0 & \dots & -1 & \mathbf{1} & 0 \dots 0 \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \dots \mathbf{0} \\ 0 & 0 & \dots & 0 & \mathbf{1} & -1 \dots 0 \\ \vdots & \vdots & & \vdots & & \\ 0 & 0 & \dots & 0 & \mathbf{1} & 0 \cdots -1 \end{bmatrix}, \tag{3.90}$$

where the bold notation highlights the θ-th row and column. According to (3.90), the row space

of $\mathcal{J}$ is:

$$\mathrm{Row}(\mathcal{J}) = \left\{ [\alpha_1 \; \alpha_2 \ldots \alpha_H] : \alpha_\theta = -\sum_{h \neq \theta} \alpha_h \right\}, \tag{3.91}$$

which is equivalent to:

$$\mathrm{Row}(\mathcal{J}) = \left\{ [\alpha_1 \; \alpha_2 \ldots \alpha_H] : \; \alpha^\mathsf{T} \mathbb{1}_H = 0 \right\}. \tag{3.92}$$

Examining (3.86), from straightforward algebra it can be shown that $v_3^\mathsf{T} \mathbb{1}_3 = 0$, which, in light of (3.87), implies that $v_H^\mathsf{T} \mathbb{1}_H = 0$. Using (3.92), we conclude that v_H^T lies in fact in the row space of $\mathcal{J}$, which finally implies, for $H \geq S$ (excluding the case $H = S = 2$) that $\mathrm{rank}(C) = 2$. □

3.B Proof of Theorem 3.2

We remark that in our setting the divergences are modeled as random variables, which implies that the value of $\boldsymbol{\theta}_k^\star$ is random as well. We should take this into account when proving the claim of the theorem. First, we observe that:

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{\theta}_k^\star \text{ is unique and } \mathrm{rank}(\boldsymbol{C}_k) = S\right) \\
&= \mathbb{P}\left(\boldsymbol{\theta}_k^\star \text{ is unique and } \mathrm{rank}(\boldsymbol{C}(\boldsymbol{\theta}_k^\star)) = S\right) \\
&= \sum_{\theta=1}^{H} \mathbb{P}\left(\boldsymbol{\theta}_k^\star = \theta, \mathrm{rank}(\boldsymbol{C}(\theta)) = S\right).
\end{aligned} \tag{3.93}$$

We now show that, for all $\theta = 1, 2, \ldots, H$:

$$\mathbb{P}\left(\mathrm{rank}(\boldsymbol{C}(\theta)) = S\right) = 1. \tag{3.94}$$

It is useful to visualize the matrix $\boldsymbol{C}(\theta)$ as follows:

$$\begin{bmatrix} \boldsymbol{d_{\theta 1}} - \boldsymbol{d}_{11} & \boldsymbol{d_{\theta 2}} - \boldsymbol{d}_{12} & \dots & \boldsymbol{d_{\theta S}} - \boldsymbol{d}_{1S} \\ \boldsymbol{d_{\theta 1}} - \boldsymbol{d}_{21} & \boldsymbol{d_{\theta 2}} - \boldsymbol{d}_{22} & \dots & \boldsymbol{d_{\theta S}} - \boldsymbol{d}_{2S} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{d_{\theta 1}} - \boldsymbol{d}_{(\theta-1)1} & \boldsymbol{d_{\theta 2}} - \boldsymbol{d}_{(\theta-1)2} & \dots & \boldsymbol{d_{\theta S}} - \boldsymbol{d}_{(\theta-1)S} \\ 0 & 0 & \dots & 0 \\ \boldsymbol{d_{\theta 1}} - \boldsymbol{d}_{(\theta+1)1} & \boldsymbol{d_{\theta 2}} - \boldsymbol{d}_{(\theta+1)2} & \dots & \boldsymbol{d_{\theta S}} - \boldsymbol{d}_{(\theta+1)S} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{d_{\theta 1}} - \boldsymbol{d}_{H1} & \boldsymbol{d_{\theta 2}} - \boldsymbol{d}_{H2} & \dots & \boldsymbol{d_{\theta S}} - \boldsymbol{d}_{HS} \\ 1 & 1 & \dots & 1 \end{bmatrix}. \tag{3.95}$$

The matrix $\boldsymbol{C}(\theta)$ has $H-1$ *random* rows (i.e., excluding the all-zeros and all-ones rows). Thus, when $H > S$ there are at least S rows with random entries. These random entries are jointly absolutely continuous since **i)** so are the entries of $\boldsymbol{D}$; and **ii)** the mapping from $\boldsymbol{D}$ to (the random entries of) $\boldsymbol{C}(\theta)$ is non-singular.[9] This implies that, for $H > S$:

$$\mathbb{P}\left(\operatorname{rank}(\boldsymbol{C}(\theta)) = S\right) = 1, \tag{3.96}$$

which proves (3.94) for the case $H > S$.

We switch to the case $H = S$. Let us denote by $\boldsymbol{B}_{S-1}(\theta)$ the sub-matrix of $\boldsymbol{B}(\theta)$ obtained by deleting its last column, and with $\boldsymbol{b}_S(\theta)$ the last column of $\boldsymbol{B}(\theta)$. We can write:

$$\boldsymbol{C}(\theta) = \begin{bmatrix} \boldsymbol{B}_{S-1}(\theta) & \boldsymbol{b}_S(\theta) \\ \mathbb{1}_{S-1}^{\mathsf{T}} & 1 \end{bmatrix}. \tag{3.97}$$

We notice that $\boldsymbol{B}_{S-1}(\theta)$ depends only on the sub-matrix $\boldsymbol{D}_{S-1}$ that is obtained by deleting from $\boldsymbol{D}$ the last column. It is thus meaningful to introduce the set of matrices:

$$\mathcal{E} \triangleq \left\{D_{S-1} : \operatorname{rank}(B_{S-1}(\theta)) = S-1\right\}. \tag{3.98}$$

Recalling that $B_{S-1}(\theta)$ contains an all-zeros row, we see that, given a matrix $D_{S-1} \in \mathcal{E}$, there exists a unique sequence of weights:

$$w_1, w_2, \dots, w_{\theta-1}, w_{\theta+1}, \dots, w_S \tag{3.99}$$

[9] For example, property **ii)** can be grasped by noting that, conditioned on $d_{\theta 1}, \dots, d_{\theta S}$, the random entries in (3.95) are jointly absolutely continuous.

to obtain the row vector $\mathbb{1}_{S-1}^{\mathsf{T}}$ as a weighted linear combination of the rows of $B_{S-1}(\theta)$. Accordingly, given a matrix $D_{S-1} \in \mathcal{E}$, the rank of $\boldsymbol{C}(\theta)$ will be equal to S if the last row in $\boldsymbol{C}(\theta)$ cannot be obtained as a linear combination of the rows of $\boldsymbol{B}(\theta)$. In view of (3.97), this corresponds to check whether the linear combination of the elements in $\boldsymbol{b}_S$ with the same weights is equal to 1, namely, if:

$$\sum_{h \neq \theta} w_h (\boldsymbol{d}_{\theta S} - \boldsymbol{d}_{hS}) = 1. \tag{3.100}$$

Consider now a matrix $D_{S-1} \in \mathcal{E}$ We have that:

$$\mathbb{P}\left[\sum_{h \neq \theta} w_h (\boldsymbol{d}_{\theta S} - \boldsymbol{d}_{hS}) = 1 \middle| D_{S-1} \right] = 0, \tag{3.101}$$

since (also conditioned on $\boldsymbol{D}_{S-1}$) the random variables $\{\boldsymbol{d}_{hS}\}$, with $h = 1, 2, \ldots, H$, are jointly absolutely continuous. We then conclude that:

$$\mathbb{P}\left(\operatorname{rank}(\boldsymbol{C}(\theta)) = S | D_{S-1}\right) = 1, \tag{3.102}$$

which implies (3.94) since, in view of the joint absolute continuity of the entries in $\boldsymbol{D}$, we have that:

$$\mathbb{P}\left(\operatorname{rank}(\boldsymbol{B}_{S-1}(\theta)) = S - 1\right) = 1 \Rightarrow \mathbb{P}\left(\boldsymbol{D}_{S-1} \in \mathcal{E}\right) = 1. \tag{3.103}$$

If we now apply (3.94) in (3.93), we conclude that:

$$\begin{aligned} &\mathbb{P}\left(\boldsymbol{\theta}_k^\star \text{ is unique and } \operatorname{rank}(\boldsymbol{C}_k) = S\right) \\ &= \sum_{\theta=1}^{H} \mathbb{P}\left(\boldsymbol{\theta}_k^\star = \theta\right) = \mathbb{P}\left(\boldsymbol{\theta}_k^\star \text{ is unique}\right). \end{aligned} \tag{3.104}$$

The proof of the theorem will be now complete if we show that the probability of having a unique $\boldsymbol{\theta}_k^\star$ is equal to 1. To this aim, by using (2.65) and (3.29), we see that:

$$\boldsymbol{\theta}_k^\star = \operatorname*{argmin}_{\theta \in \Theta} \sum_{s=1}^{S} x_{sk} \boldsymbol{d}_{\theta s}. \tag{3.105}$$

Let us consider the summations in (3.105) corresponding to different values of θ. Since the random variables $\{\boldsymbol{d}_{\theta s}\}$ are jointly absolutely continuous (and since x_k is not an all-zeros vector), the probability that two or more summations are equal is zero, which finally implies that $\boldsymbol{\theta}_k^\star$ is unique.

4 Exchange of Partial Information

4.1 Introduction[1]

In this chapter, we consider that agents observe the world under the objective evidence scenario. That is, each agent observes world measurements, describing some phenomenon or event of interest, which are explained by one of the hypotheses, namely the *true state of the world*, denoted by $\theta_0 \in \Theta$. In this case, the observed signal at each instant i for each agent k is generated according to the marginal likelihood function $L_k(\xi|\theta_0)$:

$$\boldsymbol{\xi}_{k,i} \sim L_k(\xi|\theta_0) \tag{4.1}$$

The purpose of social learning in this context is to allow agents to *learn the truth* by leveraging collaboration and the exchange of opinions. In other words, agents will share their beliefs with neighbors, and information will thus diffuse across the network, enabling truth learning—see Section 2.1.3 on social learning under objective evidence. For example, consider a network of meteorological stations at different locations monitoring weather conditions. Each of the stations (or agents) observes measurements such as temperature, humidity, atmospheric pressure and wind speed, which are functions of the underlying weather condition (or state of the world). Agents then try to infer the underlying weather state such as declaring that it is sunny, rainy, cloudy, or snowing.

As described in Chapters 1 and 2, several existing social learning implementations successfully drive the agents to identify the true state of nature with full confidence. Under objective evidence, global identifiability plays an important role in enabling truth learning—see Assumption 2.5. In this case, for each $\theta \neq \theta_0$, there exists at least one agent $k^\star(\theta)$ that is able to distinguish hypothesis θ from the truth θ_0. We refer to each of these agents as a *clear-sighted agent*.

In this chapter, we examine the scenario in which agents within a strongly connected network do not share their full belief vectors but only the confidence they have in a particular *hypothesis of interest* (such as their opinion about whether the weather conditions are rainy or not). In this setting, agents only share *partial information*. The best learning outcome agents could hope for with the sharing of such minimal information is to infer whether the hypothesis of

[1] This chapter is adapted from [68], [69].

interest corresponds to the truth or not. We will see that this process gives rise to a rich set of convergence regimes.

The main contributions of this chapter consist in the characterization of the learning and mislearning regimes, under the social learning process with partial information. We will propose two approaches for diffusing partial information: **i)** An approach without self-awareness; **ii)** An approach with self-awareness, in which each agent can combine neighbors' *partial* information to its own *full* belief vector. The theoretical results highlight some interesting phenomena. One of them being that truth sharing preserves truth learning, but also that, when the hypothesis of interest is false, a sufficient distance between this hypothesis and the truth must exist in order for agents to make a clear distinction between both and to correctly discard the presumed hypothesis.

4.2 Problem Setting

In traditional social learning, in order to learn the true state of the world out of a set of H hypotheses, agents share the full extension of their intermediate belief vector with their respective neighbors. We consider now that agents are interested in answering a different question. For instance, in our example of the network of meteorological stations, consider that these stations want to answer the question "is it sunny?". Do the agents still need to share their entire belief vectors repeatedly to find out whether it is sunny or not? If we devise a cooperation scheme where agents share only, at every iteration, the confidence they have regarding the "sunny" condition, can agents still learn?

In this chapter, we adapt the social learning framework by incorporating the following communication constraint (due, for example, to communication or regulation requirements): Agents share a *single* belief component, namely $\psi_{\ell,i}(\theta_{\mathsf{TX}})$, where θ_{TX} denotes a *hypothesis of interest* or the *transmitted hypothesis*. This constraint reflects a situation in which agents possess a certain level of private knowledge, but, for reasons such as social dynamics, limited bandwidth, regulation, diffuse only certain aspects of it. For example, consider the following situation. A group of agents exchanges reviews concerning a product from brand θ_1 that was recently released in the market. The information contained in these reviews is limited to the product of interest, i.e., the hypothesis of interest. The content of these reviews can be positive or negative according to the agent's perception of the product, i.e., the review conveys a soft decision. From these repeated interactions, agents would like to reach a conclusion on which product brand is best among brands $\{\theta_1, \theta_2, \theta_3\}$. In their reviews, agents do not share opinions on brands θ_2, θ_3, which correspond to the non-transmitted hypotheses.

Besides the appeal of the partial information strategy from a behavioral standpoint, a second relevant aspect consists in taking into account *compressed information*. While technological advances allow for improved communication bandwidth capacity, therefore enabling many edge-intelligent solutions such as distributed and federated learning, we see a growing interest for communication efficiency [70].

In Figure 4.1, we see how the output of the Bayesian update step is limited to the transmitted component $\psi_{\ell,i}(\theta_{\mathsf{TX}})$. We see also that an intermediate step is included between the Bayesian

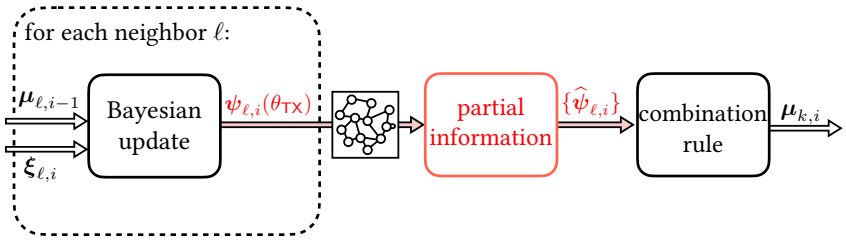


Figure 4.1: Diagram of the social learning strategy with a partial information mechanism.

update and combination steps. This additional step is referred to as a *partial information* mechanism and its role consists in transforming the received transmitted components into a valid belief vector to be used in the combination step. To this end, the red block in Figure 4.1 implements some transformation

$$\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}}) \mapsto \widehat{\boldsymbol{\psi}}_{\ell,i} \tag{4.2}$$

to incorporate the information contained in the transmitted component into an estimate $\widehat{\boldsymbol{\psi}}_{\ell,i}$ of the locally-updated belief vector $\boldsymbol{\psi}_{\ell,i}$ for each neighbor $\ell \in \mathcal{N}_k$. Different transformations correspond to different application scenarios and represent various types of behavior of the learning agents.

Given that the only information shared by neighbors is the transmitted component of the intermediate beliefs, the transformation $\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}}) \mapsto \widehat{\boldsymbol{\psi}}_{\ell,i}$, performed in the partial information mechanism, can be designed according to:

$$\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta) = \begin{cases} \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}}), & \theta = \theta_{\mathsf{TX}}, \\ \frac{1}{H-1}(1 - \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})), & \theta \neq \theta_{\mathsf{TX}}. \end{cases} \tag{4.3}$$

Intuitively the partial information mechanism in (4.3) preserves the component of interest shared by agent ℓ, i.e., $\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})$, and redistributes the excess mass, i.e., $1 - \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})$, over the remaining hypotheses $\theta \neq \theta_{\mathsf{TX}}$ uniformly following a maximum entropy principle. We say that (4.3) implements partial information using a *memoryless* approach, that is, disregarding prior knowledge that might bias the agents to give more or less importance to the non-transmitted components. From a behavioral perspective, this choice reflects well situations where the agents focus on the transmitted hypothesis (for example, they are discussing/sharing opinions on a particular candidate in an election process), and their learning mechanism does not allow them to care about the detailed update of the other components.

We propose two algorithms that include partial information sharing regarding one hypothesis of interest θ_{TX} and examine how the constrained communication affects truth learning in this setup. The objective of the agents, in both approaches, is to verify whether the state of nature agrees with θ_{TX} or not. We consider that agent k succeeds in doing so whenever it learns the truth according to Definition 4.2.1.

Definition 4.2.1 (Truth Learning with Partial Information). Within the partial information framework, the definition of truth learning depends on the choice of θ_{TX}:

- If $\theta_{\mathsf{TX}} = \theta_0$, agent k learns the truth when

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 1. \tag{4.4}$$

- If $\theta_{\mathsf{TX}} \neq \theta_0$, agent k learns the truth when

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 0. \tag{4.5}$$

Any other case is classified as a mislearning outcome.

4.2.1 Social Learning under Partial Information

In the first partial information approach, we propose the following modified version of the social learning algorithm (2.2)–(2.3), where at each instant $i = 1, 2, \ldots$ each agent k performs the following operations:

$$\boldsymbol{\psi}_{k,i}(\theta) = \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta)\boldsymbol{\mu}_{k,i-1}(\theta)}{\sum_{\theta' \in \Theta} L_k(\boldsymbol{\xi}_{k,i}|\theta')\boldsymbol{\mu}_{k,i-1}(\theta')}, \tag{4.6}$$

$$\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta) = \begin{cases} \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}}), & \theta = \theta_{\mathsf{TX}}, \\ \frac{1}{H-1}(1 - \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})), & \theta \neq \theta_{\mathsf{TX}}, \end{cases} \tag{4.7}$$

$$\boldsymbol{\mu}_{k,i}(\theta) = \frac{\exp\left\{\sum_{\ell=1}^{K} a_{\ell k} \log \widehat{\boldsymbol{\psi}}_{\ell,i}(\theta)\right\}}{\sum_{\theta' \in \Theta} \exp\left\{\sum_{\ell=1}^{K} a_{\ell k} \log \widehat{\boldsymbol{\psi}}_{\ell,i}(\theta')\right\}}. \tag{4.8}$$

In (4.6), agent k performs a local Bayesian update to incorporate its new private observation $\boldsymbol{\xi}_{k,i}$. By doing so, the agent builds its intermediate belief vector $\boldsymbol{\psi}_{k,i}$, which in the traditional social learning implementation would have been the variable shared with the neighbors of k. In the partial information setting, however, agent k will only share the component $\boldsymbol{\psi}_{k,i}(\theta_{\mathsf{TX}})$ with its neighbors, which will then split the remaining mass $1 - \boldsymbol{\psi}_{k,i}(\theta_{\mathsf{TX}})$ uniformly across the hypotheses $\theta \neq \theta_{\mathsf{TX}}$. This process is shown in (4.7), which gives origin to the *modified belief vector* $\widehat{\boldsymbol{\psi}}_{k,i}$. The final belief vector $\boldsymbol{\mu}_{k,i}$ is obtained by locally aggregating the neighbors' modified belief vectors using the same log-linear combination rule as shown in (4.8).

Note that the aggregation step in (4.8) implies for every pair $\theta, \theta' \in \Theta$ that

$$\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')} = \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta)}{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta')}. \tag{4.9}$$

The social learning strategy with partial information hereby proposed admits a useful relation

to traditional social learning, which is described in Proposition 4.1.

Proposition 4.1 (Binary hypothesis test). *The social learning algorithm with partial information presented in* (4.6)-(4.8) *can be interpreted as solving a binary hypothesis test problem for the set* $\Theta_b = \{\theta_{\mathsf{TX}}, \bar{\theta}_{\mathsf{TX}}\}$*, with the likelihood* $L_k(\xi|\bar{\theta}_{\mathsf{TX}})$ *associated to* $\bar{\theta}_{\mathsf{TX}}$ *defined as:*

$$L_k(\xi|\bar{\theta}_{\mathsf{TX}}) \triangleq \sum_{\tau \neq \theta_{\mathsf{TX}}} \frac{L_k(\xi|\tau)}{H-1}. \tag{4.10}$$

In other words, the original problem with H *hypotheses considered by the agents can be reformulated as a binary hypothesis test problem over* Θ_b*, with a fictitious likelihood for the "aggregate" fictitious hypothesis* $\bar{\theta}_{\mathsf{TX}}$*, namely,* $L_k(\xi|\bar{\theta}_{\mathsf{TX}})$*.*

Proof. See Appendix 4.A. □

The proposition shows that the algorithm under partial information in (4.6)–(4.8) can be reinterpreted in terms of a traditional social learning algorithm with a binary set of hypotheses $\Theta_b = \{\theta_{\mathsf{TX}}, \bar{\theta}_{\mathsf{TX}}\}$, and with likelihoods $L_k(\xi|\theta_{\mathsf{TX}})$ and $L_k(\xi|\bar{\theta}_{\mathsf{TX}})$. Intuitively, hypothesis $\bar{\theta}_{\mathsf{TX}}$ corresponds to an artificial hypothesis that representing any hypothesis distinct from θ_{TX}.

When $\theta_{\mathsf{TX}} \neq \theta_0$, the algorithm is equivalent to a traditional (binary) social learning algorithm with mismatched distribution, i.e., with a distribution of the data that does not match the assumed likelihood. Under these conditions, the evolution of beliefs, particularly its asymptotic learning behavior, is known to depend on the KL divergence between the true likelihood $L_k(\xi|\theta_0)$ and likelihoods $L_k(\xi|\bar{\theta}_{\mathsf{TX}})$ and $L_k(\xi|\theta_{\mathsf{TX}})$, an can be characterized using the theoretical results in [40]. In order to keep this chapter self-contained, we establish the convergence behavior of beliefs in Lemma 4.1 (see Appendix 4.B), Theorem 4.1 and Theorem 4.3.

In the second approach, we take into account the fact that each agent k has full knowledge about its own intermediate belief vector $\boldsymbol{\psi}_{k,i}$. Agent k will still perform the same Bayesian update seen in (4.6) and share only its belief component corresponding to the hypothesis of interest θ_{TX}, reflected in (4.7). However, now, we rewrite the combination step of the algorithm in such a way that agent k combines its neighbors' modified beliefs $\{\widehat{\boldsymbol{\psi}}_{\ell,i}\}_{\ell \in \mathcal{N}_k \backslash k}$ with its own *true* belief $\boldsymbol{\psi}_{k,i}$, using the following log-linear combination rule:

$$\boldsymbol{\mu}_{k,i}(\theta) = \frac{\exp\left\{ a_{kk} \log \boldsymbol{\psi}_{k,i}(\theta) + \sum\limits_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \widehat{\boldsymbol{\psi}}_{\ell,i}(\theta) \right\}}{\sum\limits_{\theta' \in \Theta} \exp\left\{ a_{kk} \log \boldsymbol{\psi}_{k,i}(\theta') + \sum\limits_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \widehat{\boldsymbol{\psi}}_{\ell,i}(\theta') \right\}}. \tag{4.11}$$

Note that this combination step leads to:

$$\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')} = a_{kk} \log \frac{\boldsymbol{\psi}_{k,i}(\theta)}{\boldsymbol{\psi}_{k,i}(\theta')} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \frac{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta)}{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta')}, \tag{4.12}$$

where we can distinguish two terms on the RHS of (4.12): A first term representing the *self-awareness* of agent k and a second term, which combines the neighbors' partial information contribution. In this formulation, it is necessary that $a_{kk} > 0$ in order for the self-awareness of agent k to count in the combination step. We will refer to it as *self-awareness coefficient* and we will assume that $a_{kk} > 0$ for all $k = 1, 2, \ldots, K$ in this setup.

4.2.2 Non-Transmitted Components

Before presenting the theoretical results, it is useful to make a parallel between the evolution of non-transmitted belief components for both partial information approaches. For the algorithm without self-awareness, all non-transmitted components of the belief vector *evolve equally* over time. To see that, replace (4.7) into (4.9) for any two non-transmitted components $\tau, \tau' \neq \theta_{\mathsf{TX}}$:

$$\begin{aligned}\log \frac{\boldsymbol{\mu}_{k,i}(\tau)}{\boldsymbol{\mu}_{k,i}(\tau')} &= \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\widehat{\boldsymbol{\psi}}_{\ell,i}(\tau)}{\widehat{\boldsymbol{\psi}}_{\ell,i}(\tau')} \\ &= \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\frac{1-\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}{H-1}}{\frac{1-\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}{H-1}} = 0 \\ &\Rightarrow \boldsymbol{\mu}_{k,i}(\tau) = \boldsymbol{\mu}_{k,i}(\tau'). \end{aligned} \tag{4.13}$$

Since the entries of the vector $\boldsymbol{\mu}_{k,i}$ sum up to one, it follows that we can write, for any non-transmitted hypothesis $\tau \neq \theta_{\mathsf{TX}}$:

$$\begin{aligned}\sum_{\tau \neq \theta_{\mathsf{TX}}} \boldsymbol{\mu}_{k,i}(\tau) &= 1 - \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \\ \Rightarrow \boldsymbol{\mu}_{k,i}(\tau) &= \frac{1 - \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})}{H-1}. \end{aligned} \tag{4.14}$$

This equal evolution for all $\tau \neq \theta_{\mathsf{TX}}$ will have the following important effect on the learning behavior: If one non-transmitted hypothesis is rejected, then so are all the non-transmitted hypotheses.

For the approach with self-awareness, the non-transmitted belief components no longer evolve equally as is the case for the first partial information strategy. More precisely, for two non-transmitted hypotheses $\tau, \tau' \neq \theta_{\mathsf{TX}}$, considering (4.7), the combination step in (4.12) yields:

$$\begin{aligned}\log \frac{\boldsymbol{\mu}_{k,i}(\tau)}{\boldsymbol{\mu}_{k,i}(\tau')} &= a_{kk} \log \frac{\boldsymbol{\psi}_{k,i}(\tau)}{\boldsymbol{\psi}_{k,i}(\tau')} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \frac{\frac{1-\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}{H-1}}{\frac{1-\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}{H-1}} \\ &\overset{\text{(a)}}{=} a_{kk} \log \frac{\boldsymbol{\mu}_{k,i-1}(\tau)}{\boldsymbol{\mu}_{k,i-1}(\tau')} + a_{kk} \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\tau)}{L_k(\boldsymbol{\xi}_{k,i}|\tau')}, \end{aligned} \tag{4.15}$$

where in (a) we used (4.6). In Lemma 4.5 (Appendix 4.D), we show that the log-ratio of beliefs for non-transmitted hypotheses in the LHS of (4.15) converges in distribution to some asymptotic random variable. In practice, this implies that the non-transmitted components will exhibit an oscillatory behavior over time. Lemma 4.6 (Appendix 4.D) will nevertheless ensure the following

stronger result: for the algorithm with self-awareness all non-transmitted components are *rejected in parallel.* Although not as strong as the equal evolution seen in (4.14), this property will be essential to enable learning in the self-aware case.

4.3 Performance Analysis

Before delving on the analysis of the learning performance, it is useful to introduce some auxiliary quantities. First, we recall from (2.38) that, under objective evidence, the KL divergence between the true likelihood and the likelihood corresponding to some hypothesis θ at agent k is denoted by:

$$d_k(\theta) = D(L_k(\theta_0)||L_k(\theta)) \tag{4.16}$$

Using the Perron eigenvector entries, the *network KL divergence* is denoted for all $\theta \neq \theta_0$ as

$$D(\theta) \triangleq \sum_{\ell=1}^{K} \pi_\ell d_\ell(\theta), \tag{4.17}$$

which will play an important role in the results that follow.

Second, recall the definition of $\bar{\theta}_{\mathsf{TX}}$, which corresponds to a "fictitious" hypothesis that represents occurrence of any hypothesis distinct from θ_{TX}. This fictitious hypothesis does not explicitly belong to Θ, and therefore is not associated to any of the likelihood functions. To this end, we use instead the definition of the fictitious likelihood, seen in (4.10), which embodies compressed information on all likelihoods relative to $\theta \neq \theta_{\mathsf{TX}}$. These two concepts allow us to extend the notation of the KL divergence introduced in (4.16) to likelihood $L_k(\xi|\bar{\theta}_{\mathsf{TX}})$:

$$d_k(\bar{\theta}_{\mathsf{TX}}) \triangleq \mathbb{E}\left(\log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta_0)}{L_k(\boldsymbol{\xi}_{k,i}|\bar{\theta}_{\mathsf{TX}})}\right). \tag{4.18}$$

We also introduce the corresponding network divergence:[2]

$$D(\bar{\theta}_{\mathsf{TX}}) \triangleq \sum_{\ell=1}^{K} \pi_\ell d_\ell(\bar{\theta}_{\mathsf{TX}}). \tag{4.21}$$

In the following sections, we are interested in determining for each of the algorithms, and for different choices of the transmitted hypothesis, the conditions for learning and mislearning.

[2]From the convexity of $-\log(\cdot)$ and using Jensen's inequality, we have that:

$$\log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta_0)}{\frac{1}{H-1}\sum\limits_{\tau\neq\theta_{\mathsf{TX}}} L_k(\boldsymbol{\xi}_{k,i}|\tau)} \leq \frac{1}{H-1}\sum_{\tau\neq\theta_{\mathsf{TX}}}\left(\log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta_0)}{L_k(\boldsymbol{\xi}_{k,i}|\tau)}\right). \tag{4.19}$$

Taking expectation of both sides in (4.19) allows us to relate $d_k(\bar{\theta}_{\mathsf{TX}})$ to the KL divergences relative to the non-transmitted hypotheses according to:

$$d_k(\bar{\theta}_{\mathsf{TX}}) \leq \sum_{\tau\neq\theta_{\mathsf{TX}}} \frac{d_k(\tau)}{H-1}. \tag{4.20}$$

From (4.20), we see that the finite KL divergence assumption (Assumption 2.4) extends naturally to $d_k(\bar{\theta}_{\mathsf{TX}})$ for all $k = 1, 2, \ldots, K$.

The convergence analysis will be split in two complementary cases: **i)** when $\theta_{\mathsf{TX}} = \theta_0$; and **ii)** when $\theta_{\mathsf{TX}} \neq \theta_0$.

4.3.1 Truth Learning when $\theta_{\mathsf{TX}} = \theta_0$

For both partial information strategies, we will show that truth sharing, i.e., choosing $\theta_{\mathsf{TX}} = \theta_0$, results in truth learning. Consider first the approach without self-awareness, namely algorithm (4.6)–(4.8). Truth learning under truth sharing is guaranteed conditioned on the existence of at least one agent that is clear-sighted in the following sense (we use the notation $\bar{\theta}_0$ in place of $\bar{\theta}_{\mathsf{TX}}$ since we are focusing on the case $\theta_{\mathsf{TX}} = \theta_0$).

Assumption 4.1 (Existence of a clear-sighted agent: Approach without self-awareness). There exists at least one agent $k^\star$ that satisfies the following condition:

$$d_{k^\star}(\bar{\theta}_0) > 0. \tag{4.22}$$

From (4.22), we require that this clear-sighted agent is endowed with the ability of distinguishing the true likelihood $L_{k^\star}(\xi|\theta_0)$ from the fictitious likelihood $L_{k^\star}(\xi|\bar{\theta}_0)$ defined in (4.10). Note that Assumption 4.1 implies that $|\bar{\Theta}_{k^\star}| > 0$. Actually, requiring that the true likelihood is not a combination, with weights $1/(H-1)$, of all the likelihoods for $\theta \neq \theta_0$, is tantamount to requiring that the true likelihood is not a combination, with weights $1/|\bar{\Theta}_{k^\star}|$, of the *distinguishable* hypotheses. This is not a strong assumption, since the case in which the true likelihood matches *exactly* a mixture of the likelihoods relative to the distinguishable hypotheses with uniform weights is deemed to be an unlucky coincidence.

Theorem 4.1 (Truth sharing implies truth learning: Approach without self-awareness). *Under Assumptions 2.4, 2.2 and 4.1, if $\theta_{\mathsf{TX}} = \theta_0$, then every agent k learns the truth, i.e.,*

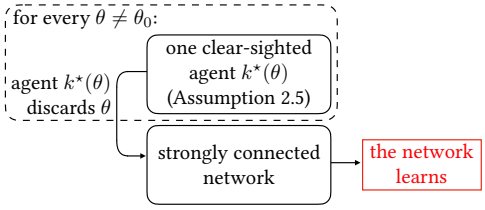
$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 1. \tag{4.23}$$

Proof. See Appendix 4.B. □

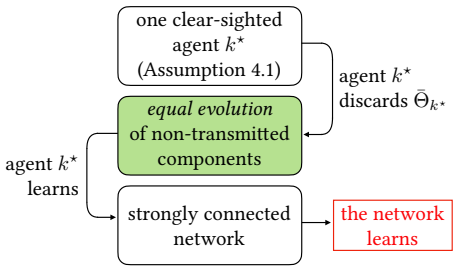
Next, consider the partial information approach with self-awareness, whose algorithm can be seen in (4.6), (4.7) and (4.11). For this algorithm, truth learning under truth sharing requires another notion of clear-sighted agent.

Assumption 4.2 (Existence of a clear-sighted agent: Approach with self-awareness). There exists at least one agent $k^\star$ that satisfies the following condition. For any convex combination vector $\alpha \in \Delta^{|\bar{\Theta}_{k^\star}|}$,

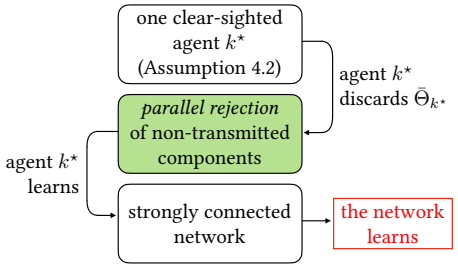
$$\mathbb{E}\left(\log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta_0)}{\sum\limits_{\tau\in\bar{\Theta}_{k^\star}} \alpha(\tau) L_k(\boldsymbol{\xi}_{k,i}|\tau)}\right) \geq c > 0. \tag{4.24}$$



(a) Traditional social learning.



(b) Partial information approach without self-awareness.



(c) Partial information approach with self-awareness.

Figure 4.2: Learning mechanism for the traditional social learning and the two partial information approaches under the truth-sharing regime.

First note that Assumption 4.2 implies that $|\bar{\Theta}_{k^\star}| > 0$. This assumption is stronger than Assumption 4.1 since it requires that the true likelihood $L_{k^\star}(\xi|\theta_0)$ is not an *arbitrary* mixture of the likelihoods relative to distinguishable hypotheses. We will discuss these differences in due detail in Section 4.3.3.

Theorem 4.2 (Truth sharing implies truth learning: Approach with self-awareness). *Under Assumptions 2.4, 2.2 and 4.2, when $\theta_{\mathsf{TX}} = \theta_0$ we have:*

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 1. \tag{4.25}$$

Proof. See Appendix 4.C. □

Theorems 4.1 and 4.2 ensure, under some technical assumptions, that both partial information approaches drive agents to learn the truth when they share information relative only to the true hypothesis. These results motivate us to draw a parallel between the learning behaviors of the partial information and the traditional social learning strategies.

Learning with Traditional Social Learning (Figure 4.2a)

Consider a fixed hypothesis $\theta \neq \theta_0$. In view of the global identifiability assumption (Assumption 2.5), there is (at least) one *clear-sighted* agent whose data and likelihoods allow it to

distinguish θ from θ_0. Due to the propagation of information across the strongly connected network, the other agents are eventually endowed with the same ability. As a result, all agents are able to discard θ from θ_0. Repeating the above process for every $\theta \neq \theta_0$ leads the agents to choose finally θ_0.

Learning without Self-Awareness (Figure 4.2b)

In this case, the meaning of the qualification "clear-sighted" changes. Let $k^\star$ be the index of a clear-sighted agent, and recall that the distinguishable set of this agent is denoted by $\bar{\Theta}_{k^\star}$. In the context of partial information without self-awareness, the clear-sighted agent is required to distinguish θ_0 from some fictitious hypothesis "aggregating" the hypotheses in $\bar{\Theta}_{k^\star}$—see Assumption 4.1. We showed that if this condition is verified, then all agents decide correctly. This result admits a useful interpretation. Assumption 4.1 implies that the clear-sighted agent has some capability of discounting $\bar{\Theta}_{k^\star}$. Now, since we have shown in (3.74) that under partial information without self-awareness the beliefs evaluated at $\theta \neq \theta_{\mathsf{TX}}$ *evolve equally in parallel* (i.e., $\boldsymbol{\mu}_{k,i}(\theta)$ takes the same value for all $\theta \neq \theta_{\mathsf{TX}}$ during the algorithm evolution), once the clear-sighted is able to discount the hypotheses in $\bar{\Theta}_{k^\star}$, it is also able to discount *all* $\theta \neq \theta_{\mathsf{TX}}$. Finally, this possibility is extended to all the other agents by propagation of information across the strongly connected network.

Learning with Self-Awareness (Figure 4.2c)

As happens in the case without self-awareness, we need a clear-sighted agent, say agent $k^\star$, that is required to distinguish θ_0 from some aggregate hypothesis involving $\bar{\Theta}_{k^\star}$, but now in a different sense. In the self-aware strategy, the clear-sighted agent must be able to discern the likelihood at θ_0 from *any convex combination* of likelihoods of the distinguishable hypotheses, as detailed in condition (4.24). This condition is stronger than (4.22) since (4.22) requires discriminability for a particular (uniform) combination. The reason for this stronger requirement is as follows. In the Bayesian update rule, the social learning algorithms evaluate convex combinations of the likelihoods that use as weights the beliefs $\boldsymbol{\mu}_{k,i-1}(\theta)$—see the denominator in (4.6). In the social learning strategy without self-awareness, due to the equal evolution of the beliefs at the non-transmitted hypotheses, this convex combination ends up being a uniformly weighted convex combination of the likelihoods. In contrast, as discussed in Section 4.2.2, in the strategy with self-awareness the beliefs at $\theta \neq \theta_{\mathsf{TX}}$ experience unpredictable mutual oscillations, and due to this unpredictability we require discriminability with respect to any convex combination.

Now, as we show in Lemma 4.3 (Appendix 4.C), if condition (4.24) is satisfied, the clear-sighted agent is able to discount the hypotheses in $\bar{\Theta}_{k^\star}$. We will be able to show that also in this case the correct choice of the clear-sighted agent propagates across the other agents, albeit with a different learning mechanism, due to the self-awareness term. Comparing (4.8) against (4.11), we see that the self-awareness term introduces a slight asymmetry in the social learning algorithm, since the self-loop term is treated differently from all the other terms. On the theoretical side, this slight asymmetry entails a significant complication in the technical proofs required to examine the learning performance. On the practical side, the beliefs at the non-transmitted

hypotheses do not evolve equally in parallel as happens in the case without self-awareness. Instead, as already mentioned, the beliefs will feature mutual oscillations among different entries $\theta \neq \theta_{\mathsf{TX}}$. Lemma 4.6 (Appendix 4.D) is used to show that the oscillatory behavior of the beliefs does not impair the extension of this knowledge to the remaining $\theta \neq \theta_{\mathsf{TX}}$. As a result, despite the oscillatory behavior, the clear-sighted agent is able to discount all the hypotheses $\theta \neq \theta_{\mathsf{TX}}$. Finally, this possibility is extended to all the other agents by propagation of information across the network (see Lemma 4.4 in Appendix 4.C).

4.3.2 Truth Learning/Mislearning when $\theta_{\mathsf{TX}} \neq \theta_0$

For both partial information approaches, we will establish conditions for obtaining truth learning and mislearning as an outcome of choosing $\theta_{\mathsf{TX}} \neq \theta_0$. First, we introduce these results for the strategy without self-awareness.

Theorem 4.3 (Learning/mislearning regimes: Approach without self-awareness). *Under Assumptions 2.4 and 2.2, for every agent $k = 1, 2, \ldots, K$, we observe two convergence behaviors:*[3]

1. *If $D(\theta_{\mathsf{TX}}) > D(\bar{\theta}_{\mathsf{TX}})$,*

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 0 \text{ and then } \boldsymbol{\mu}_{k,i}(\theta) \xrightarrow{\text{a.s.}} \frac{1}{H-1}, \tag{4.26}$$

   for all $\theta \neq \theta_{\mathsf{TX}}$.

2. *If $D(\theta_{\mathsf{TX}}) < D(\bar{\theta}_{\mathsf{TX}})$,*

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 1. \tag{4.27}$$

Proof. See Appendix 4.E. □

Theorem 4.3 shows two possible convergence behaviors for the beliefs across the network: Asymptotically, either agents correctly discard θ_{TX} or they mistakenly believe that θ_{TX} is the true hypothesis. The former case takes place whenever the transmitted hypothesis is sufficiently distinct from the true hypothesis. The latter case happens whenever the transmitted hypothesis is more easily confounded with the true one than the fictitious complementary hypothesis $\bar{\theta}_{\mathsf{TX}}$.

Before presenting similar results for the strategy with self-awareness, we introduce an extra assumption on the boundedness of the likelihood functions.

Assumption 4.3 (Bounded likelihoods). Let there be a finite constant $B > 0$ such that, for all k:

$$\left| \log \frac{L_k(\xi|\tau)}{L_k(\xi|\tau')} \right| \leq B, \tag{4.28}$$

[3] We rule out the pathological case in which $D(\theta_{\mathsf{TX}}) = D(\bar{\theta}_{\mathsf{TX}})$, which typically results in a (non-convergent) asymptotic oscillatory behavior of the belief components.

for all $\tau, \tau' \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$ and for all $\xi \in \mathcal{X}_k$.

Theorem 4.4 (Learning/mislearning regimes: Approach with self-awareness). *Under Assumptions 2.4 and 2.2, when $\theta_{\mathsf{TX}} \neq \theta_0$, for any agent k, we have:*

1. *If $D(\theta_{\mathsf{TX}}) > \frac{1}{H-1} \sum_{\tau \neq \theta_{\mathsf{TX}}} D(\tau)$,*

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 0. \tag{4.29}$$

2. *Under Assumption 4.3, if $D(\theta_{\mathsf{TX}}) < D(\bar{\theta}_{\mathsf{TX}}) - \sum_{k=1}^{K} a_{kk}(d_k(\bar{\theta}_{\mathsf{TX}}) + \pi_k B)$,*

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 1. \tag{4.30}$$

Proof. See Appendix 4.F. □

Comparing the conditions for truth-learning when $\theta_{\mathsf{TX}} = \theta_0$ (Theorems 4.1 and 4.2) against the conditions for truth-learning/mislearning when $\theta_{\mathsf{TX}} \neq \theta_0$ (Theorems 4.3 and 4.4), we see that a fundamental difference arises. The conditions relative to the case $\theta_{\mathsf{TX}} = \theta_0$ are formulated *at an individual agent level*, i.e., they depend on local characteristics of a clear-sighted agent. In contrast, the conditions relative to the case $\theta_{\mathsf{TX}} \neq \theta_0$ are formulated at a *network* level, since they depend on average KL divergences and network parameters in a way that does not allow disentangling the individual agent contributions.

Let us now provide some interpretation of the results in Theorems 4.3 and 4.4. We will examine the two theorems separately.

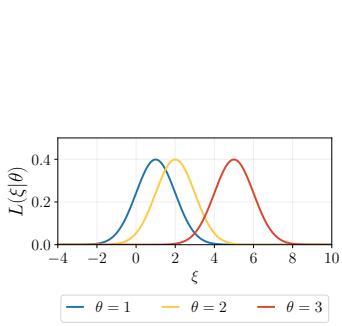
Learning and Mislearning without Self-Awareness

To explain the intuition behind Theorem 4.3, we will introduce a numerical example. Let there be a strongly connected network of $K = 10$ agents solving a social learning problem under the partial information regime without self-awareness, i.e., under (4.6)-(4.8). The set of hypotheses is $\Theta = \{1, 2, 3\}$, where we assume the true hypothesis is $\theta_0 = 1$. We consider that all agents possess the same family of Gaussian likelihood functions with same variance and distinct means, denoted by $L(\xi|\theta)$ for $\theta \in \Theta$, which are illustrated in Figure 4.3a.

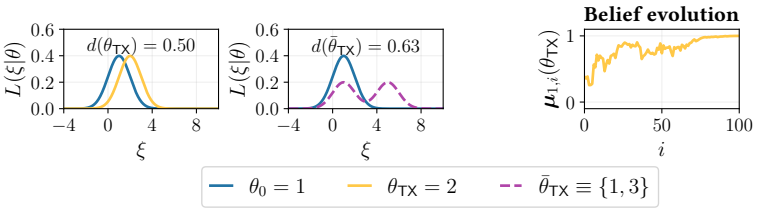
Since the likelihood functions are the same across all agents, the Perron eigenvector does not play a role in the convergence behavior, and only the following two quantities of interest will determine the behavior of all agents—subscript k is dropped:

$$d(\theta_{\mathsf{TX}}) \quad \text{and} \quad d(\bar{\theta}_{\mathsf{TX}}), \tag{4.31}$$

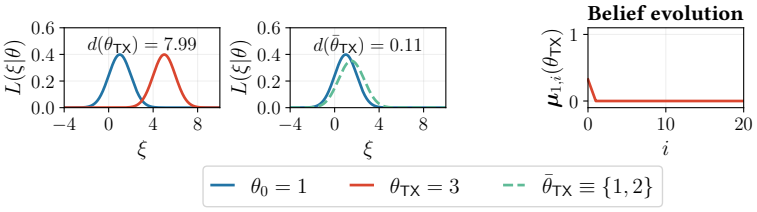
which, in the considered example, are the same across all agents, and which quantify, respectively, the KL divergence between the likelihood of the true hypothesis and hypothesis of



(a) Family of Gaussian likelihoods.



(b) Transmitted hypothesis $\theta_{\mathsf{TX}} = 2$.



(c) Transmitted hypothesis $\theta_{\mathsf{TX}} = 3$.

Figure 4.3: Example of family of likelihood functions with $\theta_0 = 1$. In the middle panels of (b) and (c), *solid lines* represent the actual likelihood functions and *dashed lines* depict the "fictitious" likelihood functions associated with the complementary hypothesis $\bar{\theta}_{\mathsf{TX}}$, defined in (4.10).

interest θ_{TX}, and the KL divergence between the likelihood of the true hypothesis and the fictitious likelihood of the complementary hypothesis $\bar{\theta}_{\mathsf{TX}}$—see (4.10).

Consider first that the hypothesis of interest is chosen to be $\theta_{\mathsf{TX}} = 2$. We see in Figure 4.3b that the likelihood relative to the transmitted hypothesis is closer to the true likelihood in comparison with the likelihood relative to the non-transmitted hypothesis, i.e.,

$$d(\theta_{\mathsf{TX}}) < d(\bar{\theta}_{\mathsf{TX}}), \tag{4.32}$$

which implies that condition 2) of Theorem 4.3 is satisfied, and all agents are fooled into believing that θ_{TX} is the true state. This behavior is confirmed by the experiment shown in the righmost panel of Figure 4.3b for agent 1.

When the hypothesis of interest is chosen as $\theta_{\mathsf{TX}} = 3$, Figure 4.3c shows that the likelihood relative to the transmitted hypothesis is farther from the true likelihood in comparison with the likelihood relative to the non-transmitted hypothesis, i.e.,

$$d(\theta_{\mathsf{TX}}) > d(\bar{\theta}_{\mathsf{TX}}), \tag{4.33}$$

and agents can properly distinguish the transmitted hypothesis as being false, as seen in case 1) of Theorem 4.3. Therefore agents are able to discount hypothesis θ_{TX}, as shown in the belief evolution in the rightmost panel of Figure 4.3c.

Learning and Mislearning with Self-Awareness

Let us comment on the result for the algorithm with self-awareness in the case $\theta_{\mathsf{TX}} \neq \theta_0$. The addition of a self-awareness term is expected to improve the learning performance, and this behavior will be examined in the forthcoming section. However, as already noticed, the

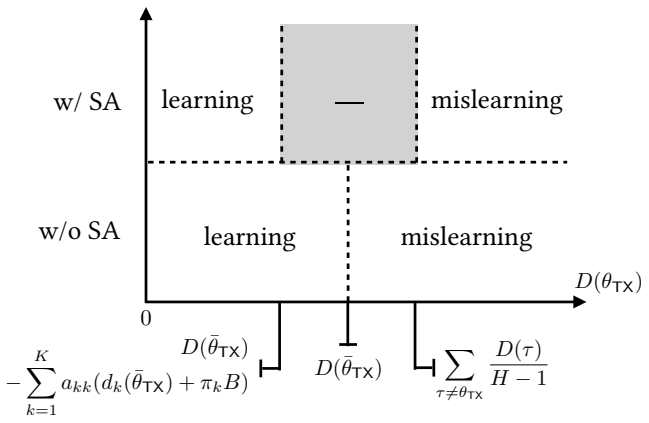


Figure 4.4: Learning regions for $D(\theta_{\mathsf{TX}})$ for the partial information algorithm without and with self-awareness (denoted by "w/o SA" and "w/ SA" respectively) when $\theta_{\mathsf{TX}} \neq \theta_0$.

asymmetry introduced in the learning algorithm by the self-awareness term makes the theoretical analysis more complicated. For example, different from the case without self-awareness, the learning/mislearning bounds are not tight, and, as far as we can tell, do not suggest a neat physical interpretation of the learning/mislearning behavior. We notice furthermore that the RHS of condition 2) in Theorem 4.4, can in principle be negative, particularly when the self-awareness coefficients approach 1. In this case, the mislearning condition is never satisfied, and simulation results, detailed in the next section, suggest that higher self-weights can mitigate mislearning.

In summary, Theorems 4.3 and 4.4 show, for both partial information approaches, that when $\theta_{\mathsf{TX}} \neq \theta_0$ there exist regions of $D(\theta_{\mathsf{TX}})$ for which respectively truth learning and mislearning occur. These regions are illustrated in Figure 4.4. As a general comment applying to both algorithms, we see that if the transmitted hypothesis is more easily confounded with the true one (small $D(\theta_{\mathsf{TX}})$) we have mislearning, while the converse behavior occurs for relatively high values of $D(\theta_{\mathsf{TX}})$. However, a difference emerges between the results available from the two theorems. For the algorithm without self-awareness, we can determine the learning/mislearning behavior for any value of $D(\theta_{\mathsf{TX}})$, whereas for the algorithm with self-awareness, we cannot determine the behavior whenever $D(\theta_{\mathsf{TX}})$ is found in the gray area of Figure 4.4.

Exploiting the structure of the lower boundary in Figure 4.4, we can examine how this boundary is influenced by the self-weights a_{kk}. If the value of one or more self-terms decreases (i.e., if self-awareness decreases) the lower boundary moves upward, and the region where mislearning occurs becomes wider, eventually approaching the threshold $D(\bar{\theta}_{\mathsf{TX}})$ pertaining to the algorithm without self-awareness when the self-terms vanish. Conversely, if a_{kk} increases the lower boundary moves downward. This implies that the gray area becomes wider, i.e., the region where we are sure to mislearn reduces. On the other hand, a wider gray area leaves open the possibility that correct learning occurs over an ampler range of cases. We will get confirmation of this behavior in the forthcoming section.

4.3.3 Discussion and Overview of Results

Comparative Discussion on Main Assumptions

As seen in Section 2.1.3, traditional social learning requires global identifiability, i.e., for every $\theta \neq \theta_0$, at least one agent should be able to distinguish θ from θ_0. In comparison, Assumptions 4.1 and 4.2 require the existence of one agent whose true likelihood is not equal to convex combinations of the other likelihoods, which are in some sense representative of the "alternative" w.r.t. the transmitted hypothesis. The situation that one likelihood is a convex combination of the other likelihoods is often an unlikely situation (for example, if we have Gaussian or exponential likelihoods, a convex combination thereof is not Gaussian or exponential). In summary, Assumptions 4.1 and 4.2 can be verified even if global identifiability is violated. For example, if $\theta^\star$ is indistinguishable from θ_0 at all agents, our results imply that when $\theta_{\mathsf{TX}} = \theta_0$ we can still guess the right hypothesis. This might appear strange in view of traditional social learning, however we must not forget that the problem of truth learning contemplates also the case $\theta_{\mathsf{TX}} \neq \theta_0$. In the latter case, the impact of indistinguishability becomes more relevant, since the partial information strategies learn well provided that condition 1) in Theorem 4.3 (without self-awareness) or condition 1) in Theorem 4.4 (with self-awareness) holds. Examining (4.26) and (4.29), we see that one necessary condition for them to hold is that $D(\theta_{\mathsf{TX}}) > 0$, which implies that some agent must be able to distinguish θ_{TX} from θ_0. In particular, if we require truth learning for all $\theta_{\mathsf{TX}} \neq \theta_0$, we need $D(\theta_{\mathsf{TX}}) > 0$ for all $\theta_{\mathsf{TX}} \neq \theta_0$, i.e., at least global identifiability is required. In summary, in the truth sharing regime, conditions for learning are weaker than in traditional social learning, whereas in the regime with $\theta_{\mathsf{TX}} \neq \theta_0$, global identifiability is necessary but not sufficient.

Main Questions in Social Learning with Partial Information

In summary, the main questions to be answered in social learning with partial information sharing are overall ones like:

1. In which instances the agents learn regardless of the true state?
2. When agents mislearn, how does this happen?

The answer to question 1 is provided by Theorems 4.1–4.4. In particular, since Theorems 4.1 and 4.2 reveal that, when $\theta_{\mathsf{TX}} = \theta_0$, truth learning is guaranteed, both with and without self-awareness, the answer to question 1 is contained in Eqs. (4.26) and (4.29), which provide conditions under which truth learning takes place regardless of the transmitted hypothesis. Likewise, the answer to question 2 is provided by (4.27) and (4.30), which in particular specify that when an agent mislearns, it gives full credit to the transmitted (wrong) hypothesis.

4.4 Simulation Results

In this section, we illustrate the results seen in Theorems 4.1–4.4. To do so, we set up an inference problem with ten hypotheses, i.e., $\Theta = \{1, 2, \dots, 10\}$, from which $\theta_0 = 1$ is the true

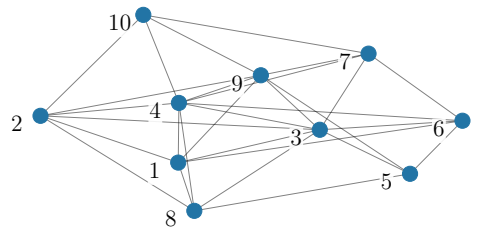


Figure 4.5: Strongly connected network topology with $K = 10$ agents.

state of nature. We consider a strongly connected network of 10 agents, whose topology can be seen in Figure 4.5, designed so that all agents have self-loops.

Besides, the adjacency matrix is designed to be left-stochastic using a parametrized averaging rule [37]:

$$a_{\ell k} = \begin{cases} \lambda, & \text{if } \ell = k, \\ (1-\lambda)/n_k, & \text{if } \ell \neq k \text{ and } \ell \in \mathcal{N}_k, \\ 0, & \text{otherwise.} \end{cases} \tag{4.34}$$

where n_ℓ is the degree of node (agent) ℓ, excluding node ℓ itself. Each agent is trying to determine whether some hypothesis $\theta_{\mathsf{TX}} \in \Theta$ corresponds to the true state of nature, by exchanging among neighbors partial information regarding the hypothesis of interest. In the following we consider two inference problems, one with continuous observations, the other with discrete observations.

4.4.1 Continuous Observations

The first example considers a family of unit-variance Gaussian likelihood functions given by:

$$f_n(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(\xi - 0.5(n-1))^2}{2} \right\}, \tag{4.35}$$

for $n = 1, 2, \ldots, 10$.

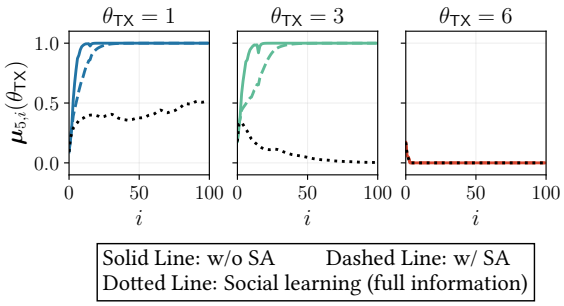
We assume that the inference problem is globally identifiable (see Assumption 2.5). More particularly, we consider the following identifiability limitations: For each agent $k = 1, 2, \ldots, 10$,

$$L_k(\xi|\theta) = \begin{cases} f_1(\xi), & \text{for } \theta \leq k, \\ f_\theta(\xi), & \text{for } \theta > k. \end{cases} \tag{4.36}$$

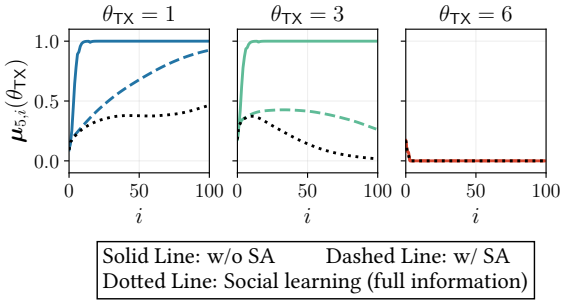
In this case, only agent 1 is able to solve the inference problem alone, that is, the indistinguishable set of hypotheses satisfies:

$$|\Theta_k| > 1, \text{ for } k = 2, \ldots, 10. \tag{4.37}$$

Under the aforementioned setup, we now examine both the partial information algorithm proposed in (4.6)–(4.8) and the algorithm with self-awareness in (4.6), (4.7) and (4.11). We also



(a) Self-awareness parameter $\lambda = 0.7$.



(b) Self-awareness parameter $\lambda = 0.95$.

Figure 4.6: Convergence of the belief component regarding different transmitted hypotheses for agent 5, where $\theta_0 = 1$.

wish to compare the performance of both algorithms with the performance of the traditional social learning algorithm (seen in (2.2)–(2.3)), in which the agents share all elements of the belief vector.

At first, we consider that the combination matrix is parameterized according to (4.34) with $\lambda = 0.7$. In Figure 4.6a we can see the evolution of belief at agent 5 (similar behavior is observed for the other agents) for each different hypothesis of interest θ_{TX}. Colorful solid and dashed lines refer to the partial information algorithm without and with self-awareness, respectively. Black dotted lines refer to traditional social learning.

We start by examining the behavior of the algorithm under truth sharing, i.e., when $\theta_{\mathsf{TX}} = \theta_0 = 1$ (leftmost panel in Figure 4.6a). We see that all social learning algorithms are able to learn the true hypothesis, as predicted by Theorems 4.1 and 4.2 for the partial information algorithms, and by the existing results on traditional social learning. We switch to the case $\theta_{\mathsf{TX}} \neq \theta_0$ (middle and rightmost panel in Figure 4.6a). As expected, traditional social learning learns well. The partial information algorithms behave instead in accordance with Theorems 4.3 and 4.4: When the hypothesis of interest is sufficiently "close" to the true one, which is the case for $\theta_{\mathsf{TX}} = 3$ (middle panel), the agent mistakenly learns that θ_{TX} is the true hypothesis. Conversely, when the hypothesis of interest is far enough from the true one, which is the case for $\theta_{\mathsf{TX}} = 6$ (rightmost panel), the agent learns well.

It is interesting to see what happens when all agents give more weight to their individual information by increasing the self-awareness parameter, setting it to parameter λ. In Figure 4.6b

we consider the case $\lambda = 0.95$. The algorithm with self-awareness is now able to learn the truth for any of the three transmitted hypothesis, and its convergence curve is now closer to the curve of the traditional social learning algorithm. In a nutshell, concentrating the weights of the combination matrix A around the self-loops entails a decrease in cooperation and hence a slower convergence. It also mitigates the effect of partial information received from neighbors, allowing for truth learning in all three cases.

Another interesting phenomenon emerging from the simulations pertains to the learning rate. In the considered example, the algorithm without self-awareness can be faster[4] than that with self-awareness, which can, in turn, be faster than traditional social learning. This can be counterintuitive, since one could expect that traditional social learning is the best one. However, in making this observation one should not forget the inherent trade-off of decision systems. Think of a decision system that always chooses θ_{TX}. This system learns *instantaneously* when $\theta_{\mathsf{TX}} = \theta_0$, but fails invariably in the other cases. In other words, the superiority of traditional social learning resides in the fact that it *always* allows correct learning. In contrast, the algorithms with partial information can learn faster *when they learn well*, but they can fail. Likewise, the fact that the algorithm without self-awareness can be faster than the algorithm with self-awareness when $\theta_{\mathsf{TX}} = \theta_0$, is justified by the fact that the latter algorithm can perform better when $\theta_{\mathsf{TX}} \neq \theta_0$.

4.4.2 Discrete Observations

Consider the same network topology seen in Figure 4.5 and combination matrix in (4.34). Under $\theta_0 = 1$, and the same identifiability constraints enunciated in the previous example, we now consider a family of discrete likelihood functions given by $f_n(\xi)$ for $n = 1, 2, \ldots, 10$, defined over a discrete space of signals $\mathcal{X} \triangleq \{0, 1, 2\}$, which can be seen in Figure 4.7. We highlight in blue the distribution f_1, which we associated with the true likelihoods $L_k(\xi|\theta_0)$ for all agents.

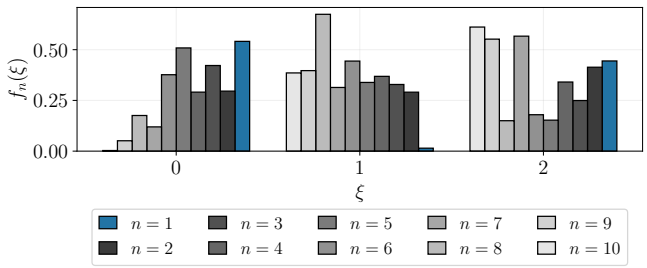
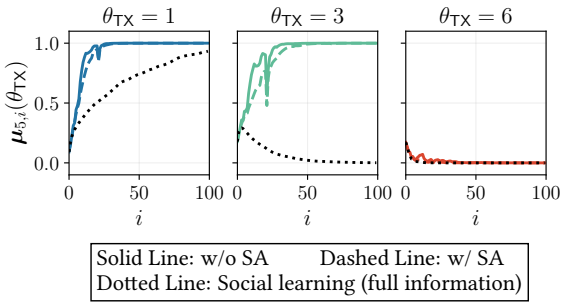


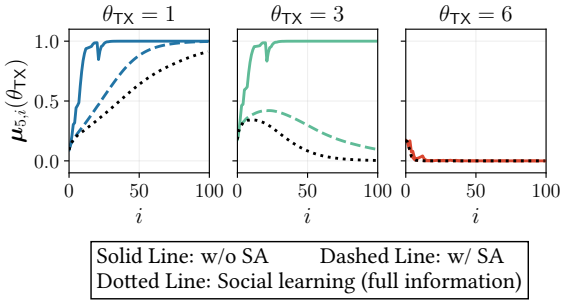
Figure 4.7: Family of discrete likelihood functions.

At first, we consider a self-awareness parameter $\lambda = 0.7$. We wish to compare the two partial information approaches for different transmitted hypotheses and the traditional social learning strategy with full information sharing. We can see in Figure 4.8a the evolution of belief at agent 5 for each transmitted hypothesis $\theta_{\mathsf{TX}} \in \Theta$ (similar behavior is observed for the other agents). As in the previous example, due to the likelihood functions setup, when $\theta_{\mathsf{TX}} = 3$, the true

[4]We have also noted that, for very small values of λ, it is possible for this convergence to be slightly slower instead.



(a) Self-awareness parameter $\lambda = 0.7$.



(b) Self-awareness parameter $\lambda = 0.95$.

Figure 4.8: Convergence of the belief component regarding different transmitted hypotheses for agent 5, where $\theta_0 = 1$.

hypothesis and $\theta_{TX} = 3$ are confounded by the algorithm with partial information, and the agent mislearns.

However, when the self-awareness parameter increases to $\lambda = 0.95$, again a switch in the convergence behavior happens as can be seen in Figure 4.8b for $\theta_{TX} = 3$. As the agents are more self-aware, they are able to make correct decisions for all scenarios of θ_{TX}.

Different from the previous example with Gaussian likelihoods, in the present example with discrete likelihoods Assumption 4.3 holds. This implies that we can exploit the lower boundary in Figure 4.4 corresponding to the algorithm with self-awareness. Examining this lower boundary, we see that as self-awareness grows (i.e., as λ grows), the mislearning region shrinks and gives place to a gray area, where either learning or mislearning could possibly occur. We recall that getting a wider gray region does not allow to conclude that the algorithm with self-awareness would learn inside this region. However, a wider gray area reduces the region where we would be *sure to observe mislearning*. As a matter of fact, in the specific example we are dealing with, self-awareness can be used to tune the learning behavior in the case $\theta_{TX} = 2$, and bring the network from mislearning to full learning.

4.5 Concluding Remarks

In this chapter, we introduced two approaches for taking into account partial information within the social learning framework, where agents communicate their belief about a single

hypothesis of interest. In the first approach, agents consider only partial beliefs. In the second, each individual agent becomes *self-aware*, in the sense that it exploits its own *full* belief (being still forced to use *partial* beliefs from its neighbors).

We established the following main trends. While the traditional social learning algorithms, which leverage full belief sharing, are always able to learn correctly the true hypothesis, a richer behavior characterizes social learning under partial information. Both social learning algorithms with partial information proposed in this work learn correctly when the hypothesis of interest is the true hypothesis. When the transmitted hypothesis is false, however, mislearning can occur. Moreover, we showed that there are cases where the algorithm without self-awareness mislearns, while the algorithm with self-awareness can be led to the right conclusion by increasing the self-weights in the combination matrix.

4.A Proof of Proposition 4.1

Let the belief vector $\boldsymbol{\mu}_{k,i}$ be split into two components for every agent k: $\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})$ and $\boldsymbol{\mu}_{k,i}(\bar{\theta}_{\mathsf{TX}})$, the latter defined as

$$\boldsymbol{\mu}_{k,i}(\bar{\theta}_{\mathsf{TX}}) = \sum_{\tau \neq \theta_{\mathsf{TX}}} \boldsymbol{\mu}_{k,i}(\tau). \tag{4.38}$$

Similarly, for the intermediate belief vector $\boldsymbol{\psi}_{k,i}$, we define:

$$\boldsymbol{\psi}_{k,i}(\bar{\theta}_{\mathsf{TX}}) = \sum_{\tau \neq \theta_{\mathsf{TX}}} \boldsymbol{\psi}_{k,i}(\tau). \tag{4.39}$$

Remember, from (3.74), that all non-transmitted components of $\boldsymbol{\mu}_{k,i}$ evolve equally according to:

$$\boldsymbol{\mu}_{k,i}(\tau) = \frac{\boldsymbol{\mu}_{k,i}(\bar{\theta}_{\mathsf{TX}})}{H-1} \tag{4.40}$$

for any $\tau \neq \theta_{\mathsf{TX}}$. Replace (4.6) into (4.39):

$$\begin{aligned}
\boldsymbol{\psi}_{k,i}(\bar{\theta}_{\mathsf{TX}}) &= \frac{\sum\limits_{\tau \neq \theta_{\mathsf{TX}}} \boldsymbol{\mu}_{k,i-1}(\tau) L_k(\boldsymbol{\xi}_{k,i}|\tau)}{\sum\limits_{\theta' \in \Theta} \boldsymbol{\mu}_{k,i-1}(\theta') L_k(\boldsymbol{\xi}_{k,i}|\theta')} \\
&\overset{\text{(a)}}{=} \frac{\sum\limits_{\tau \neq \theta_{\mathsf{TX}}} \boldsymbol{\mu}_{k,i-1}(\bar{\theta}_{\mathsf{TX}}) L_k(\boldsymbol{\xi}_{k,i}|\tau)/(H-1)}{\boldsymbol{\mu}_{k,i-1}(\theta_{\mathsf{TX}}) L_k(\boldsymbol{\xi}_{k,i}|\theta_{\mathsf{TX}}) + \sum\limits_{\theta' \neq \theta_{\mathsf{TX}}} \boldsymbol{\mu}_{k,i-1}(\bar{\theta}_{\mathsf{TX}}) L_k(\boldsymbol{\xi}_{k,i}|\theta')/(H-1)} \\
&\overset{\text{(b)}}{=} \frac{\boldsymbol{\mu}_{k,i-1}(\bar{\theta}_{\mathsf{TX}}) L_k(\boldsymbol{\xi}_{k,i}|\bar{\theta}_{\mathsf{TX}})}{\boldsymbol{\mu}_{k,i-1}(\theta_{\mathsf{TX}}) L_k(\boldsymbol{\xi}_{k,i}|\theta_{\mathsf{TX}}) + \boldsymbol{\mu}_{k,i-1}(\bar{\theta}_{\mathsf{TX}}) L_k(\boldsymbol{\xi}_{k,i}|\bar{\theta}_{\mathsf{TX}})},
\end{aligned} \tag{4.41}$$

where in (a) the non-transmitted components are replaced by (3.74), and in (b) the likelihood function corresponding to the complementary hypothesis $\bar{\theta}_{\mathsf{TX}}$ is replaced by (4.10). Note that (4.41) corresponds to the Bayesian update for the complementary hypothesis $\bar{\theta}_{\mathsf{TX}}$ under the set

of two hypotheses $\Theta_b = \{\theta_{\mathsf{TX}}, \bar{\theta}_{\mathsf{TX}}\}$, and with the fictitious likelihood in (4.10).

Similarly to the belief vectors $\boldsymbol{\mu}_{k,i}$, we now show that the non-transmitted components of the modified belief $\widehat{\boldsymbol{\psi}}_{k,i}$ evolve equally over time. From (4.7) and (4.39) we have, for any non-transmitted hypothesis $\tau \neq \theta_{\mathsf{TX}}$:

$$\widehat{\boldsymbol{\psi}}_{\ell,i}(\tau) = \frac{\boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})}{H-1}. \tag{4.42}$$

Consider now the combination step. Replacing (3.3) into (4.38) results in:

$$\begin{aligned}
\boldsymbol{\mu}_{k,i}(\bar{\theta}_{\mathsf{TX}}) &= \frac{\sum\limits_{\tau\neq\theta_{\mathsf{TX}}} \exp\left(\sum\limits_{\ell=1}^{K} a_{\ell k} \log \widehat{\boldsymbol{\psi}}_{\ell,i}(\tau)\right)}{\sum\limits_{\theta'\in\Theta} \exp\left(\sum\limits_{\ell=1}^{K} a_{\ell k} \log \widehat{\boldsymbol{\psi}}_{\ell,i}(\theta')\right)} \\
&\overset{\text{(a)}}{=} \frac{\sum\limits_{\tau\neq\theta_{\mathsf{TX}}} \exp\left(\sum\limits_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})}{H-1}\right)}{\exp\left(\sum\limits_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})\right) + \sum\limits_{\theta'\neq\theta_{\mathsf{TX}}} \exp\left(\sum\limits_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})}{H-1}\right)} \\
&= \frac{\exp\left(\sum\limits_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})\right)}{\exp\left(\sum\limits_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})\right) + \exp\left(\sum\limits_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})\right)},
\end{aligned} \tag{4.43}$$

where in (a), the non-transmitted components of the modified belief vector are replaced by (4.42). Note that (4.43) is equivalent to writing a (log-linear) combination step for the binary set of hypotheses $\Theta_b = \{\theta_{\mathsf{TX}}, \bar{\theta}_{\mathsf{TX}}\}$.

Since $\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})$ and $\boldsymbol{\mu}_{k,i}(\bar{\theta}_{\mathsf{TX}})$ (and similarly $\boldsymbol{\psi}_{k,i}(\theta_{\mathsf{TX}})$ and $\boldsymbol{\psi}_{k,i}(\bar{\theta}_{\mathsf{TX}})$) sum up to one, we have that for every $\theta \in \Theta_b$, the partial information algorithm enunciated in (2.2)–(2.3) behaves in the same manner as if each agent k performed the two steps in the traditional social learning algorithm seen in (2.2)–(2.3) for the two hypotheses in Θ_b, which agrees with the claim in Proposition 4.1.

4.B Proof of Theorem 4.1

We first introduce an intermediate result, where we show that for each agent the log-ratio between any non-transmitted and transmitted belief components will have an asymptotic exponential behavior. In order to avoid misunderstanding, we remark that this result is already known in social learning theory [40]. Nevertheless, we deem it useful to report here a proof for this result to make the chapter self-contained, and to make useful connections of this particular proof with other results that we prove relying on the recursive inequalities in Lemmas 4.8 and 4.9 further ahead.

Lemma 4.1 (Asymptotic rate of convergence). *Under Assumptions 2.4 and 2.2, for all $\theta \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$ and every agent $k = 1, 2, \ldots, K$, we have that:*

$$\frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} \xrightarrow{\text{a.s.}} D(\theta_{\mathsf{TX}}) - D(\bar{\theta}_{\mathsf{TX}}). \tag{4.44}$$

Proof. We know from (4.13) that for any non-transmitted hypotheses $\tau, \theta \neq \theta_{\mathsf{TX}}$:

$$\boldsymbol{\mu}_{k,i}(\tau) = \boldsymbol{\mu}_{k,i}(\theta). \tag{4.45}$$

Moreover, from (4.7):

$$\begin{aligned} \log \frac{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta)}{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta_{\mathsf{TX}})} &= \log \frac{\left(1 - \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})\right)/(H-1)}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})} \\ &= \log \frac{\sum_{\tau \neq \theta_{\mathsf{TX}}} \boldsymbol{\psi}_{\ell,i}(\tau)/(H-1)}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}. \end{aligned} \tag{4.46}$$

Substituting (4.6) into (4.46), we obtain:

$$\log \frac{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta)}{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta_{\mathsf{TX}})} = \log \frac{\sum_{\tau \neq \theta_{\mathsf{TX}}} \boldsymbol{\mu}_{\ell,i-1}(\tau) L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)/(H-1)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}}) L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}. \tag{4.47}$$

Using (4.45) in (4.47) yields:

$$\begin{aligned} \log \frac{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta)}{\widehat{\boldsymbol{\psi}}_{\ell,i}(\theta_{\mathsf{TX}})} &= \log \frac{\boldsymbol{\mu}_{\ell,i-1}(\theta) \sum_{\tau \neq \theta_{\mathsf{TX}}} L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)/(H-1)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}}) L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})} \\ &= \log \frac{\boldsymbol{\mu}_{\ell,i-1}(\theta)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})} + \log \frac{\sum_{\tau \neq \theta_{\mathsf{TX}}} L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)/(H-1)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})} \\ &\overset{\text{(a)}}{=} \log \frac{\boldsymbol{\mu}_{\ell,i-1}(\theta)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})} + \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\bar{\theta}_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}, \end{aligned} \tag{4.48}$$

where in (a) we used the definition of the likelihood for the non-transmitted hypotheses found in (4.10). Using (4.9), we obtain the following recursion:

$$\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} = \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\mu}_{\ell,i-1}(\theta)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})} + \sum_{\ell=1}^{K} a_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\bar{\theta}_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}. \tag{4.49}$$

Define the vectors:

$$\boldsymbol{y}_i(\theta) \triangleq \operatorname{col}\left\{\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})}\right\}_{k=1}^{K}, \quad \boldsymbol{x}_i \triangleq A^{\mathsf{T}} \operatorname{col}\left\{\log \frac{L_k(\boldsymbol{\xi}_{k,i}|\bar{\theta}_{\mathsf{TX}})}{L_k(\boldsymbol{\xi}_{k,i}|\theta_{\mathsf{TX}})}\right\}_{k=1}^{K}, \tag{4.50}$$

where the col operator concatenates a sequence of variables into a column vector. We can then

rewrite (4.49) in vector form for all $\theta \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$:

$$\boldsymbol{y}_i(\theta) = A^{\mathsf{T}} \boldsymbol{y}_{i-1}(\theta) + \boldsymbol{x}_i. \tag{4.51}$$

First, note that the recursion in (4.51) takes the form of the sequence of random vectors seen in auxiliary Lemma 4.8 (see Appendix 4.G). Since the random vectors $\boldsymbol{x}_i$ are i.i.d. across time and have finite expectation[5], Property 4.1 (also found in Appendix 4.G) can be applied and shows that $\boldsymbol{x}_i$ satisfies the three conditions (4.145)–(4.147) required by Lemma 4.8. Particularly, in view of Property 4.1, the vector $\bar{x}$ takes the form of the expectation vector $\mathbb{E}(\boldsymbol{x}_i)$. Since A is left-stochastic, all conditions in Lemma 4.8 are satisfied and we can therefore apply its result as follows. For each $\theta \neq \theta_{\mathsf{TX}}$, we have that

$$\begin{aligned}
\frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} &\xrightarrow{\text{a.s.}} \sum_{k=1}^{K} \pi_k \sum_{\ell=1}^{K} a_{\ell k} \mathbb{E}\left(\log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\bar{\theta}_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})} \right) \\
&= \sum_{\ell=1}^{K} \pi_\ell \mathbb{E}\left(\log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})} \right) - \sum_{\ell=1}^{K} \pi_\ell \mathbb{E}\left(\log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\bar{\theta}_{\mathsf{TX}})} \right) \\
&= D(\theta_{\mathsf{TX}}) - D(\bar{\theta}_{\mathsf{TX}}),
\end{aligned} \tag{4.53}$$

where we recall that $\sum_{k=1}^{K} \pi_k a_{\ell k} = \pi_\ell$ since π is the Perron eigenvector. □

Proof of Theorem 4.1. Note that the RHS of (4.44) represents a key quantity in the algorithm: conditionally on its sign, we have that the log-ratio of belief components on the LHS of (4.44) will increase or decrease indefinitely.

If $\theta_{\mathsf{TX}} = \theta_0$, we have that

$$D(\theta_0) = 0. \tag{4.54}$$

Under Assumption 4.1, there exists at least one clear-sighted agent in the network, say agent $k^\star$, for which

$$d_{k^\star}(\bar{\theta}_0) > 0. \tag{4.55}$$

From the positivity of the Perron eigenvector, we have that

$$D(\bar{\theta}_0) > 0. \tag{4.56}$$

Finally, from (4.53) with $\theta_{\mathsf{TX}} = \theta_0$, we obtain

$$\frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_0)} \xrightarrow{\text{a.s.}} D(\theta_0) - D(\bar{\theta}_0) < 0$$

[5]The i.i.d. property across time is inherited from variables $\boldsymbol{\xi}_{k,i}$ for all $k = 1, 2, \ldots, K$. Note that for each element of $\boldsymbol{x}_i$, $\mathbb{E}(\boldsymbol{x}_{k,i})$ can easily be rewritten as a function of two KL divergences:

$$\mathbb{E}(\boldsymbol{x}_{k,i}) = \sum_{\ell=1}^{K} a_{\ell k} \mathbb{E}\left(\log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\bar{\theta}_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})} \right) = \sum_{\ell=1}^{K} a_{\ell k} \left(d_\ell(\theta_{\mathsf{TX}}) - d_\ell(\bar{\theta}_{\mathsf{TX}}) \right). \tag{4.52}$$

The first term on the RHS is finite from Assumption 2.4, whereas the second term is finite from Assumption 2.4 and the inequality in (4.20).

$$\Rightarrow \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_0)} \xrightarrow{\text{a.s.}} -\infty$$
$$\Rightarrow \boldsymbol{\mu}_{k,i}(\theta) \xrightarrow{\text{a.s.}} 0, \tag{4.57}$$

which holds for all $\theta \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$. This, in turn, implies that

$$\boldsymbol{\mu}_{k,i}(\theta_0) \xrightarrow{\text{a.s.}} 1, \tag{4.58}$$

thus concluding the proof of Theorem 4.1. □

4.C Proof of Theorem 4.2

In order to reach the conclusion in Theorem 4.2, we need first to establish some intermediate results (see Lemmas 4.2, 4.3, and 4.4 enunciated next), which depend on auxiliary results found in Appendix 4.D (these results are stated in Lemmas 4.5, 4.6 and 4.7). We resort moreover to two auxiliary lemmas (see Lemmas 4.8 and 4.9 in Appendix 4.G) which refer to statistical properties of more general recursions.

Consider the truth sharing case, for which $\theta_{\mathsf{TX}} = \theta_0$. The first key result, which can be seen in Lemma 4.2 enunciated next, is that the random sequence

$$\boldsymbol{m}_i \triangleq \sum_{k=1}^{K} \pi_k \log \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \tag{4.59}$$

is a submartingale. To lighten the notation we will denote the KL divergence between the true likelihood function $L_k(\xi|\theta_0)$ and a convex combination of the likelihoods $L_k(\xi|\theta)$ by:

$$\delta_k(\alpha) \triangleq \mathbb{E}\left(\log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta_0)}{\sum\limits_{\theta\in\Theta} \alpha(\theta) L_k(\boldsymbol{\xi}_{k,i}|\theta)}\right), \tag{4.60}$$

where α is the convex combination vector, i.e., α is a vector belonging to the $H-$simplex Δ^H.

We define the signal profile at each instant i as $\boldsymbol{\xi}_i \triangleq \{\boldsymbol{\xi}_{1,i}, \boldsymbol{\xi}_{2,i}, \ldots, \boldsymbol{\xi}_{K,i}\}$. We also define the *filtration* over the past observations as $\mathcal{F}_j$ for $j = 1, 2, \ldots$, where $\mathcal{F}_j$ is the sub-σ-field generated by the observations up to instant j, namely,

$$\mathcal{F}_j \triangleq \sigma\left(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_j\right), \tag{4.61}$$

which satisfies $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_\infty \triangleq \sigma\left(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots\right)$.

Lemma 4.2 (Submartingale sequence). *Let $\theta_{\mathsf{TX}} = \theta_0$, and consider the random sequence $\{\boldsymbol{m}_i\}$ in (4.59). This sequence has the following properties.*

1. $$\mathbb{E}\left[\boldsymbol{m}_i|\mathcal{F}_{i-1}\right] \geq \boldsymbol{m}_{i-1} + \sum_{k=1}^{K} \pi_k \delta_k(\boldsymbol{\mu}_{k,i-1}). \tag{4.62}$$
2. *The sequence $\boldsymbol{m}_i$ is a nonpositive submartingale with respect to the process $\{\boldsymbol{\xi}_i\}$.*
3. *There exists a random variable $\boldsymbol{m}_\infty$ such that:*
$$\boldsymbol{m}_i \xrightarrow{\text{a.s.}} \boldsymbol{m}_\infty. \tag{4.63}$$
4. *The expectation $\mathbb{E}\left(\boldsymbol{m}_i\right)$ converges to a finite limit.*

Proof. Consider (4.11) with $\theta = \theta_{\mathsf{TX}}$. In view of (4.7), we have:

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) = \frac{\exp\left(\sum\limits_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})\right)}{\sum\limits_{\theta' \in \Theta} \exp\left(a_{kk} \log \boldsymbol{\psi}_{k,i}(\theta') + \sum\limits_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \widehat{\boldsymbol{\psi}}_{\ell,i}(\theta')\right)}. \tag{4.64}$$

To simplify the notation, we define the following auxiliary variables:

$$\mathcal{Y}_{k,i}(\theta_{\mathsf{TX}}) \triangleq \exp\left(\sum_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})\right), \tag{4.65}$$

$$\mathcal{Z}_{k,i}(\bar{\theta}_{\mathsf{TX}}) \triangleq \sum_{\tau \neq \theta_{\mathsf{TX}}} \exp\left(a_{kk} \log \boldsymbol{\psi}_{k,i}(\tau) + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \widehat{\boldsymbol{\psi}}_{\ell,i}(\tau)\right), \tag{4.66}$$

from which we can rewrite the combination step in (4.64) as:

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) = \frac{\mathcal{Y}_{k,i}(\theta_{\mathsf{TX}})}{\mathcal{Y}_{k,i}(\theta_{\mathsf{TX}}) + \mathcal{Z}_{k,i}(\bar{\theta}_{\mathsf{TX}})}. \tag{4.67}$$

Using (4.7), we can develop the expression for $\mathcal{Z}_{k,i}(\bar{\theta}_{\mathsf{TX}})$ as:

$$\begin{aligned}
\mathcal{Z}_{k,i}(\bar{\theta}_{\mathsf{TX}}) &= \sum_{\tau \neq \theta_{\mathsf{TX}}} \exp\left(a_{kk} \log \boldsymbol{\psi}_{k,i}(\tau) + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \frac{1 - \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}{H-1}\right) \\
&\overset{\text{(a)}}{=} \sum_{\tau \neq \theta_{\mathsf{TX}}} \exp\left(\log \frac{\boldsymbol{\psi}_{k,i}(\tau)^{a_{kk}}}{(H-1)^{1-a_{kk}}} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})\right)
\end{aligned}$$

$$= \exp\left(\sum_{\substack{\ell=1\\ \ell\neq k}}^{K} a_{\ell k}\log \boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})\right)\frac{\sum_{\tau\neq\theta_{\mathsf{TX}}}\boldsymbol{\psi}_{k,i}(\tau)^{a_{kk}}}{(H-1)^{1-a_{kk}}}, \tag{4.68}$$

where in (a) we introduced $\boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}}) \triangleq 1-\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})$. Now, applying the sum-of-powers inequality[6] to the rightmost term in (4.68) results in:

$$\frac{\sum_{\tau\neq\theta_{\mathsf{TX}}}\boldsymbol{\psi}_{k,i}(\tau)^{a_{kk}}}{(H-1)^{1-a_{kk}}} \leq \left(\sum_{\tau\neq\theta_{\mathsf{TX}}}\boldsymbol{\psi}_{k,i}(\tau)\right)^{a_{kk}} = \exp\left(a_{kk}\log \boldsymbol{\psi}_{k,i}(\bar{\theta}_{\mathsf{TX}})\right). \tag{4.69}$$

Replacing (4.69) to (4.68) we get:

$$\mathfrak{Z}_{k,i}(\bar{\theta}_{\mathsf{TX}}) \leq \exp\left(\sum_{\ell=1}^{K} a_{\ell k}\log \boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})\right) \triangleq \mathfrak{Y}_{k,i}(\bar{\theta}_{\mathsf{TX}}). \tag{4.70}$$

In view of (4.70), we can lower bound the expression in (4.67):

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \geq \frac{\mathfrak{Y}_{k,i}(\theta_{\mathsf{TX}})}{\mathfrak{Y}_{k,i}(\theta_{\mathsf{TX}})+\mathfrak{Y}_{k,i}(\bar{\theta}_{\mathsf{TX}})} = \frac{1}{1+\frac{\mathfrak{Y}_{k,i}(\bar{\theta}_{\mathsf{TX}})}{\mathfrak{Y}_{k,i}(\theta_{\mathsf{TX}})}}. \tag{4.71}$$

Applying $\log(\cdot)$ to both sides of (4.71), and replacing back the definitions of $\mathfrak{Y}_{k,i}(\theta_{\mathsf{TX}})$ and $\mathfrak{Y}_{k,i}(\bar{\theta}_{\mathsf{TX}})$ from (4.65) and (4.70) respectively, we can write the following inequality:

$$\begin{aligned}\log \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) &\geq \log \frac{1}{1+\exp\left(\sum_{\ell=1}^{K} a_{\ell k}\log\frac{\boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}\right)}\\ &\triangleq f\left(\sum_{\ell=1}^{K} a_{\ell k}\log\frac{\boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}\right),\end{aligned} \tag{4.72}$$

[6] For $r, s \neq 0$ with $r < s$, and for positive values x_i, we have that [71] :

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^r\right)^{1/r} \leq \left(\frac{1}{n}\sum_{i=1}^{n} x_i^s\right)^{1/s}.$$

In particular, with the choice $s = 1$ we can write the following inequality:

$$\frac{1}{n^{1-r}}\sum_{i=1}^{n} x_i^r \leq \left(\sum_{i=1}^{n} x_i\right)^r.$$

where we defined the concave function[7]

$$f(x) \triangleq \log \frac{1}{1+e^x}. \tag{4.73}$$

Using Jensen's inequality, we have that

$$\begin{aligned}
\log \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) &\geq \sum_{\ell=1}^{K} a_{\ell k} \log \frac{1}{1+e^{\log \frac{\boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}}} = \sum_{\ell=1}^{K} a_{\ell k} \log \frac{1}{1+\frac{\boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}} \\
&= \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}}) + \boldsymbol{\psi}_{\ell,i}(\bar{\theta}_{\mathsf{TX}})} = \sum_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}}) \\
&\overset{\text{(a)}}{=} \sum_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}}) + \sum_{\ell=1}^{K} a_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}{\sum\limits_{\theta' \in \Theta} \boldsymbol{\mu}_{\ell,i-1}(\theta') L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta')},
\end{aligned} \tag{4.74}$$

where in (a), we replaced $\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})$ using the Bayesian update seen in (4.6).

Taking the expectation of both sides of (4.74) conditioned on $\mathcal{F}_{i-1}$, we have that:

$$\begin{aligned}
\mathbb{E}\left[\log \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \middle| \mathcal{F}_{i-1}\right] &\geq \sum_{\ell=1}^{K} a_{\ell k} \log \boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}}) \\
&\quad + \sum_{\ell=1}^{K} a_{\ell k} \mathbb{E}\left[\log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}{\sum\limits_{\theta' \in \Theta} \boldsymbol{\mu}_{\ell,i-1}(\theta') L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta')} \middle| \mathcal{F}_{i-1}\right].
\end{aligned} \tag{4.75}$$

Since the current signal profile $\boldsymbol{\xi}_i$ is independent from the past data vectors (and, hence, is independent from the past belief vector $\boldsymbol{\mu}_{\ell,i-1}$), we see that the second term on the RHS of (4.75) is the following KL divergence, as defined in (4.60) (we recall that we are considering the case $\theta_{\mathsf{TX}} = \theta_0$):

$$\mathbb{E}\left[\log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}{\sum\limits_{\theta' \in \Theta} \boldsymbol{\mu}_{\ell,i-1}(\theta') L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta')} \middle| \mathcal{F}_{i-1}\right] = \delta_\ell(\boldsymbol{\mu}_{\ell,i-1}). \tag{4.76}$$

Multiplying both sides of (4.75) by π_k, summing over k, and recalling that $\sum_{k=1}^{K} \pi_k a_{\ell k} = \pi_\ell$ because π is the Perron eigenvector, Eqs. (4.75) and (4.76) imply part 1) of the lemma.

Part 2) follows from part 1). In fact, $\boldsymbol{m}_i$ is nonpositive because $\boldsymbol{\mu}_{k,i} \leq 1$, and $\boldsymbol{m}_i$ is a submartingale because the KL divergence is nonnegative, and, hence, Eq. (4.62) implies:

$$\mathbb{E}\left[\boldsymbol{m}_i | \mathcal{F}_{i-1}\right] \geq \boldsymbol{m}_{i-1}. \tag{4.77}$$

[7]The concavity of the function $f(x)$ can be seen from its second derivative:

$$\frac{d^2 f(x)}{dx^2} = \frac{-e^x}{\left[1+e^x\right]^2} < 0,$$

for any $x \in \mathbb{R}$.

Part 3) follows from the martingale convergence theorem [18].

Finally, part 4) follows by taking the total expectation in (4.77), which yields:

$$0 \geq \mathbb{E}\left(\boldsymbol{m}_i\right) \geq \mathbb{E}\left(\boldsymbol{m}_{i-1}\right) \geq \ldots \geq m_0 = \sum_{k=1}^{K} \pi_k \log \mu_{k,0}(\theta_{\mathsf{TX}}), \tag{4.78}$$

which implies that the sequence of expectations is a (monotonically) convergent sequence. □

Using part 3) of Lemma 4.2, we can establish the following technical corollary which will be useful later in the analysis.

Corollary 4.1 (Expectation of log-beliefs $\boldsymbol{\psi}_{k,i}$). *Let $\theta_{\mathsf{TX}} = \theta_0$. For all $i \geq 1$ we have that:*

$$\mathbb{E}\left(\log \frac{1}{\boldsymbol{\psi}_{k,i}(\theta_{\mathsf{TX}})}\right) \leq \frac{1}{\pi_k} \sum_{\ell=1}^{K} \pi_\ell \log \frac{1}{\mu_{\ell,0}(\theta_{\mathsf{TX}})}. \tag{4.79}$$

Proof. Using the Bayesian update in (4.6) we can write:

$$\begin{aligned}
\mathbb{E}\left(\log \frac{1}{\boldsymbol{\psi}_{k,i}(\theta_{\mathsf{TX}})}\right) &= \mathbb{E}\left(\log \frac{1}{\boldsymbol{\mu}_{k,i-1}(\theta_{\mathsf{TX}})}\right) - \mathbb{E}\left(\log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta_{\mathsf{TX}})}{\sum\limits_{\theta\in\Theta} \boldsymbol{\mu}_{k,i-1}(\theta) L_k(\boldsymbol{\xi}_{k,i}|\theta)}\right) \\
&= \mathbb{E}\left(\log \frac{1}{\boldsymbol{\mu}_{k,i-1}(\theta_{\mathsf{TX}})}\right) - \mathbb{E}\left(\delta_k(\boldsymbol{\mu}_{k,i-1})\right) \\
&\leq \mathbb{E}\left(\log \frac{1}{\boldsymbol{\mu}_{k,i-1}(\theta_{\mathsf{TX}})}\right).
\end{aligned} \tag{4.80}$$

On the other hand, using (4.78) we can write:

$$\begin{aligned}
&\pi_k \log \boldsymbol{\mu}_{k,i-1}(\theta_{\mathsf{TX}}) \geq \sum_{\ell=1}^{K} \pi_\ell \log \mu_{\ell,i-1}(\theta_{\mathsf{TX}}) = \boldsymbol{m}_{i-1} \\
&\Rightarrow \mathbb{E}\left(\log \frac{1}{\boldsymbol{\mu}_{k,i-1}(\theta_{\mathsf{TX}})}\right) \leq \frac{1}{\pi_k} \sum_{\ell=1}^{K} \pi_\ell \log \frac{1}{\mu_{\ell,0}(\theta_{\mathsf{TX}})},
\end{aligned} \tag{4.81}$$

which combined with (4.80) yields the desired claim. □

Lemma 4.3 (The clear-sighted agent learns the truth). *Let $\theta_{\mathsf{TX}} = \theta_0$. Under Assumptions 2.4, 2.2 and 4.2 we have that:*

$$\boldsymbol{\mu}_{k^\star,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{p}} 1 \tag{4.82}$$

Proof. We start by considering an arbitrary agent k. Taking the total expectation in (4.62) we

get:

$$0 \geq \mathbb{E}\left(\boldsymbol{m}_i\right) \geq \mathbb{E}\left(\boldsymbol{m}_{i-1}\right) + \sum_{k=1}^{K} \pi_k \mathbb{E}\left(\delta_k(\boldsymbol{\mu}_{k,i-1})\right). \tag{4.83}$$

First of all, we remark that the last expectation in (4.83) is computed with respect to the only random quantity that appears within brackets, that is $\boldsymbol{\mu}_{k,i-1}$. Using (4.83) along with the fact that the KL divergence is nonnegative, we see that:

$$0 \leq \sum_{k=1}^{K} \pi_k \mathbb{E}\left(\delta_k(\boldsymbol{\mu}_{k,i-1})\right) \leq \mathbb{E}\left(\boldsymbol{m}_i\right) - \mathbb{E}\left(\boldsymbol{m}_{i-1}\right), \tag{4.84}$$

which, in view of part $4)$ of Lemma 4.2 implies that [72]:

$$\lim_{i\to\infty} \sum_{k=1}^{K} \pi_k \mathbb{E}\left(\delta_k(\boldsymbol{\mu}_{k,i-1})\right) = 0 \tag{4.85}$$

Recalling that $\pi_k > 0$, we conclude that $\delta_k(\boldsymbol{\mu}_{k,i-1})$ converges to zero in mean. This implies in particular that $\delta_k(\boldsymbol{\mu}_{k,i-1})$ converges to zero in probability, namely,

$$\delta_k(\boldsymbol{\mu}_{k,i-1}) \xrightarrow{\text{p}} 0. \tag{4.86}$$

Recalling that $\delta_k(\boldsymbol{\mu}_{k,i-1})$ is the KL divergence between $L_k(\theta_{\mathsf{TX}})$ and $\sum_{\theta\in\Theta} \boldsymbol{\mu}_{k,i-1}(\theta) L_k(\theta)$ as defined in (4.60), from Pinsker's inequality [16] we can write:

$$\delta_k(\boldsymbol{\mu}_{k,i-1}) \geq \frac{1}{2} \left\| L_k(\theta_{\mathsf{TX}}) - \sum_{\theta\in\Theta} \boldsymbol{\mu}_{k,i-1}(\theta) L_k(\theta) \right\|^2, \tag{4.87}$$

where $\|\cdot\|$ denotes the total variation norm.

Let us now specialize the analysis to the clear-sighted agent $k^\star$. From Assumption 4.2, the set of distinguishable hypotheses $\bar{\Theta}_{k^\star}$ is non-empty. Thus, we have that:

$$\begin{aligned}
&L_{k^\star}(\theta_{\mathsf{TX}}) - \sum_{\theta\in\Theta} \boldsymbol{\mu}_{k^\star,i-1}(\theta) L_{k^\star}(\theta) \\
&= \left(1 - \sum_{\theta\in\Theta_{k^\star}} \boldsymbol{\mu}_{k^\star,i-1}(\theta)\right) L_{k^\star}(\theta_{\mathsf{TX}}) - \sum_{\theta\in\bar{\Theta}_{k^\star}} \boldsymbol{\mu}_{k^\star,i-1}(\theta) L_{k^\star}(\theta) \\
&= \sum_{\theta\in\bar{\Theta}_{k^\star}} \boldsymbol{\mu}_{k^\star,i-1}(\theta) \left(L_{k^\star}(\theta_{\mathsf{TX}}) - \sum_{\tau\in\bar{\Theta}_{k^\star}} \alpha(\tau) L_{k^\star}(\tau) \right),
\end{aligned} \tag{4.88}$$

where we defined:

$$\alpha(\tau) = \frac{\boldsymbol{\mu}_{k^\star,i-1}(\tau)}{\sum\limits_{\theta\in\bar{\Theta}_{k^\star}} \boldsymbol{\mu}_{k^\star,i-1}(\theta)}. \tag{4.89}$$

Assumption 4.2 establishes a lower bound c on the KL divergence between the true likelihood and any convex combination of the distinguishable likelihoods, which implies that the true

likelihood is not in the convex hull of distinguishable likelihoods. This further implies that there exists some $c' > 0$, for which

$$\left\| L_{k^\star}(\theta_{\mathsf{TX}}) - \sum_{\tau \in \bar{\Theta}_{k^\star}} \alpha(\tau) L_{k^\star}(\tau) \right\| \geq c', \tag{4.90}$$

where $\|\cdot\|$ represents the total variation norm [18]. From (4.88) and (4.90) we can write:

$$\begin{aligned}
&\left\| L_{k^\star}(\theta_{\mathsf{TX}}) - \sum_{\theta \in \Theta} \boldsymbol{\mu}_{k^\star,i-1}(\theta) L_{k^\star}(\theta) \right\| \\
&= \left| \sum_{\theta \in \bar{\Theta}_{k^\star}} \boldsymbol{\mu}_{k^\star,i-1}(\theta) \right| \left\| L_{k^\star}(\theta_{\mathsf{TX}}) - \sum_{\tau \in \bar{\Theta}_{k^\star}} \alpha(\tau) L_{k^\star}(\tau) \right\| \\
&\geq c' \left| \sum_{\theta \in \bar{\Theta}_{k^\star}} \boldsymbol{\mu}_{k^\star,i-1}(\theta) \right|.
\end{aligned} \tag{4.91}$$

Joining the latter inequality with (4.87) we get:

$$\delta_{k^\star}(\boldsymbol{\mu}_{k^\star,i-1}) \geq \frac{c'^2}{2} \left| \sum_{\theta \in \bar{\Theta}_{k^\star}} \boldsymbol{\mu}_{k^\star,i-1}(\theta) \right|^2. \tag{4.92}$$

Since c' is strictly positive, we conclude from (4.86) that, for every $\theta \in \bar{\Theta}_{k^\star}$:

$$\boldsymbol{\mu}_{k^\star,i}(\theta) \xrightarrow{\text{p}} 0. \tag{4.93}$$

It remains to show that the same result holds for the indistinguishable non-transmitted hypotheses, i.e., for $\theta \in \Theta_{k^\star} \setminus \{\theta_{\mathsf{TX}}\}$. But this result comes directly from Lemma 4.6, under Assumptions 2.4 and 2.2. We have therefore shown that, for the clear-sighted agent $k^\star$, the beliefs for all $\theta \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$ vanish in probability, which finally yields the claim since the sum of the beliefs over Θ is equal to 1. □

Lemma 4.4 (Influence of a learning agent). *Let $\theta_{\mathsf{TX}} = \theta_0$. Under Assumptions 2.4, 2.2 and 4.2, if, for a certain agent h,*

$$\boldsymbol{\mu}_{h,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{p}} 1, \tag{4.94}$$

then the same result holds for all agents $k \neq h$.

Proof. Let h be an agent that fulfills (4.94). Consider that the combination weight a_{hk} is *strictly* positive. From (4.12) we can therefore write:

$$\begin{aligned}
&\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} \\
&= a_{kk} \log \boldsymbol{\psi}_{k,i}(\theta) + \sum_{\ell \neq k} a_{\ell k} \log \frac{1 - \boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}{H - 1} + \sum_{\ell=1}^{K} a_{\ell k} \log \frac{1}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}
\end{aligned}$$

$$\leq a_{hk} \log(1 - \boldsymbol{\psi}_{h,i}(\theta_{\mathsf{TX}})) + \sum_{\ell=1}^{K} a_{\ell k} \log \frac{1}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}. \tag{4.95}$$

By exponentiating (4.95) we can write:

$$\boldsymbol{\mu}_{k,i}(\theta) \leq \underbrace{\left(1 - \boldsymbol{\psi}_{h,i}(\theta_{\mathsf{TX}})\right)^{a_{hk}}}_{\triangleq \boldsymbol{x}_i} \underbrace{e^{\sum_{\ell=1}^{K} a_{\ell k} \log \frac{1}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}}}_{\triangleq \boldsymbol{y}_i}. \tag{4.96}$$

First, we prove that the term $\boldsymbol{x}_i$ in (4.96) goes to zero in probability. To this end, we observe that:

$$\begin{aligned} 1 - \boldsymbol{\psi}_{h,i}(\theta_{\mathsf{TX}}) &= \frac{\sum\limits_{\theta \neq \theta_{\mathsf{TX}}} \boldsymbol{\mu}_{h,i-1}(\theta) L_h(\boldsymbol{\xi}_{h,i}|\theta)}{\boldsymbol{\mu}_{h,i-1}(\theta_{\mathsf{TX}}) L_h(\boldsymbol{\xi}_{h,i}|\theta_{\mathsf{TX}}) + \sum\limits_{\theta \neq \theta_{\mathsf{TX}}} \boldsymbol{\mu}_{h,i-1}(\theta) L_h(\boldsymbol{\xi}_{h,i}|\theta)} \\ &\leq \sum_{\theta \neq \theta_{\mathsf{TX}}} \frac{\boldsymbol{\mu}_{h,i-1}(\theta)}{\boldsymbol{\mu}_{h,i-1}(\theta_{\mathsf{TX}})} \frac{L_h(\boldsymbol{\xi}_{h,i}|\theta)}{L_h(\boldsymbol{\xi}_{h,i}|\theta_{\mathsf{TX}})}. \end{aligned} \tag{4.97}$$

We now show that each individual term of the summation,

$$\underbrace{\frac{\boldsymbol{\mu}_{h,i-1}(\theta)}{\boldsymbol{\mu}_{h,i-1}(\theta_{\mathsf{TX}})}}_{\triangleq \boldsymbol{s}_i} \underbrace{\frac{L_h(\boldsymbol{\xi}_{h,i}|\theta)}{L_h(\boldsymbol{\xi}_{h,i}|\theta_{\mathsf{TX}})}}_{\triangleq \boldsymbol{t}_i}, \tag{4.98}$$

vanishes in probability as $i \to \infty$. Indeed, the term $\boldsymbol{s}_i$ in (4.98) vanishes in probability as $i \to \infty$ in view of Lemma 4.3. On the other hand, the random variables $\boldsymbol{t}_i$ are identically distributed.[8] By application of Slutsky's theorem [73], we conclude that the product $\boldsymbol{s}_i \boldsymbol{t}_i$ converges to 0 in distribution (and, hence, in probability).

Second, we show that $\boldsymbol{y}_i$ matches the conditions in (4.112) (see Lemma 4.7 in Appendix 4.D). By application of Markov's inequality we conclude that, for any $M > 0$:

$$\begin{aligned} \mathbb{P}\left(\boldsymbol{y}_i > M\right) &= \mathbb{P}\left(\sum_{\ell=1}^{K} a_{\ell k} \log \frac{1}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})} > \log M\right) \\ &\leq \frac{1}{\log M} \sum_{\ell=1}^{K} a_{\ell k} \mathbb{E}\left(\log \frac{1}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})}\right) \\ &\leq \frac{1}{\log M} \sum_{\ell=1}^{K} \frac{a_{\ell k}}{\pi_\ell} \sum_{m=1}^{K} \pi_m \log \frac{1}{\mu_{m,0}(\theta_{\mathsf{TX}})}, \end{aligned} \tag{4.99}$$

where the latter inequality follows by Corollary 4.1. Since the final upper bound in (4.99) does not depend on i, we see that $\boldsymbol{y}_i$ fulfills (4.112) with the choice $g(M) = C/\log M$ for some finite positive constant C.

[8] We remark that the random variables $\boldsymbol{t}_i$ are well-behaved since $\frac{L_h(\boldsymbol{\xi}_{h,i}|\theta)}{L_h(\boldsymbol{\xi}_{h,i}|\theta_{\mathsf{TX}})}$ is a (nonnegative) random variable with finite expectation equal to 1.

Therefore, we conclude from Lemma 4.7 that the product $\boldsymbol{x}_i \boldsymbol{y}_i$ appearing in the upper bound in (4.96) goes to zero in probability and, hence, that:

$$\boldsymbol{\mu}_{k,i}(\theta) \xrightarrow{\text{p}} 0, \tag{4.100}$$

for any agent k for which $a_{hk} > 0$. Since the network is strongly connected, given an agent h that fulfills (4.94), and an arbitrary agent k (not necessarily a neighbor of h), there will always be a path connecting h to k. Iterating the above reasoning along this path implies the desired result. □

We can now conclude the proof of Theorem 4.2. Under Assumption 4.2, there exists at least one clear-sighted agent $k^\star$. Lemma 4.3 guarantees that agent $k^\star$ learns the truth in probability, whereas Lemma 4.4 ensures that learning propagates across the network. It is therefore legitimate to write:

$$\sum_{k=1}^{K} \pi_k \log \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{p}} 0. \tag{4.101}$$

Using part 3) of Lemma 4.2 (and since almost-sure convergence implies convergence in probability), and applying jointly (4.63) and (4.101) we conclude that:

$$\sum_{k=1}^{K} \pi_k \log \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 0. \tag{4.102}$$

On the other hand, since $\pi_k > 0$ and $\log \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \leq 0$, the convergence in (4.102) implies that:

$$\log \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 0 \Rightarrow \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 1, \tag{4.103}$$

for all agents $k = 1, 2, \ldots, K$.

4.D Auxiliary Lemmas

Lemma 4.5 (Convergence for non-transmitted hypotheses). *Let $\theta, \theta' \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$, and define:*

$$\boldsymbol{q}_{k,i}(\theta, \theta') \triangleq \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')}. \tag{4.104}$$

For every $k = 1, 2, \ldots, K$, under Assumptions 2.4 and 2.2, there exists a random variable $\boldsymbol{q}_{k,\infty}(\theta, \theta')$ ensuring the following convergence in distribution:

$$\boldsymbol{q}_{k,i}(\theta, \theta') \xrightarrow{\text{d}} \boldsymbol{q}_{k,\infty}(\theta, \theta'). \tag{4.105}$$

Proof. Since θ and θ' are distinct from θ_{TX}, using (4.6), (4.7) and (4.11) we can write:

$$\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')} = a_{kk} \log \frac{\boldsymbol{\mu}_{k,i-1}(\theta)}{\boldsymbol{\mu}_{k,i-1}(\theta')} + a_{kk} \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta)}{L_k(\boldsymbol{\xi}_{k,i}|\theta')}. \tag{4.106}$$

The result in (4.105) follows from part 1) of auxiliary Lemma 4.9 by setting:

$$a = a_{kk}, \quad \boldsymbol{y}_i = \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')}, \quad \boldsymbol{x}_i = \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta)}{L_k(\boldsymbol{\xi}_{k,i}|\theta')}, \tag{4.107}$$

where Assumption 2.4 guarantees that $\boldsymbol{x}_i$ satisfies the conditions in Lemma 4.9, and Assumption 2.2 guarantees that y_0 assumes a finite value. □

From Lemma 4.5, we see that the log-ratio of belief components concerning non-transmitted hypotheses converges in distribution to a random variable $\boldsymbol{q}_{k,\infty}$. Investigating this limiting random variable in more detail, thanks to part 1) of Lemma 4.9, we are able to write it as

$$\boldsymbol{q}_{k,\infty}(\theta, \theta') = \sum_{i=1}^{\infty} a_{kk}^i \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta)}{L_k(\boldsymbol{\xi}_{k,i}|\theta')}. \tag{4.108}$$

For each realization of signal profiles $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots)$, the infinite summation in (4.108) will converge almost surely to a random value. The distribution with which these random values are generated will be the same distribution that governs the oscillatory behavior of $\boldsymbol{q}_{k,i}(\theta, \theta')$ as $i \to \infty$.

Although the characterization of the limiting random variable $\boldsymbol{q}_{k,\infty}(\theta, \theta')$, described in (4.108), does not appear intuitive, its mere existence will enable other (stronger) convergence results starting from the one presented in Lemma 4.6. We see now that if a certain agent k discards any non-transmitted hypothesis $\theta \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$, then the existence of the limiting random variable $\boldsymbol{q}_{k,\infty}(\theta, \theta')$ will allow it to discard all other non-transmitted hypotheses.

Lemma 4.6 (Rejection of non-transmitted hypotheses). *Assume, for a given agent k, and for one non-transmitted hypothesis $\theta' \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$:*

$$\boldsymbol{\mu}_{k,i}(\theta') \xrightarrow{\text{p}} 0, \tag{4.109}$$

and that Assumptions 2.4 and 2.2 hold. Then the same convergence holds for all hypotheses $\theta \in \Theta \setminus \{\theta', \theta_{\mathsf{TX}}\}$ for the same agent.

Proof. Let $\theta \neq \theta_{\mathsf{TX}}$ be a non-transmitted hypothesis that fulfills (4.109). In view of (4.104), for any $\theta' \in \Theta \setminus \{\theta, \theta_{\mathsf{TX}}\}$ we can write:

$$\boldsymbol{\mu}_{k,i}(\theta) = \boldsymbol{\mu}_{k,i}(\theta') e^{\boldsymbol{q}_{k,i}(\theta,\theta')}. \tag{4.110}$$

Now, under Assumptions 2.4 and 2.2, Lemma 4.5 reveals that $\boldsymbol{q}_{k,i}(\theta, \theta')$ converges in distribution to a certain random variable $\boldsymbol{q}_{k,\infty}(\theta, \theta')$. In view of the continuous mapping theorem [73], we conclude that:

$$e^{\boldsymbol{q}_{k,i}(\theta,\theta')} \xrightarrow{\text{d}} e^{\boldsymbol{q}_{k,\infty}(\theta,\theta')}. \tag{4.111}$$

Examining (4.110), we see that $\boldsymbol{\mu}_{k,i}(\theta)$ is given by the product of two random sequences: **i)** the first sequence, $\{\boldsymbol{\mu}_{k,i}(\theta')\}$, vanishes in probability as $i \to \infty$ in view of (4.109); **ii)** the second sequence, $\{e^{\boldsymbol{q}_{k,i}(\theta,\theta')}\}$, converges in distribution as $i \to \infty$ in view of (4.111). S Using

Slutsky's Theorem [73], we conclude that $\boldsymbol{\mu}_{k,i}(\theta)$ converges to zero in distribution, and, hence, in probability. □

In other words, whenever an agent discards a non-transmitted hypothesis, it will automatically discard all other non-transmitted hypotheses. This result will bind together the evolution of the non-transmitted hypotheses in the case when the respective beliefs components are converging in probability to zero, which we refer to as *parallel rejection* of non-transmitted hypotheses.

Finally we introduce a technical result, which is used in the proof of Lemma 4.4.

Lemma 4.7 (Useful convergence result). *Let $\boldsymbol{z}_i = \boldsymbol{x}_i \boldsymbol{y}_i$, where $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{y}_i\}$ are two sequences of nonnegative random variables such that $\boldsymbol{x}_i$ vanishes in probability, and:*

$$\mathbb{P}(\boldsymbol{y}_i > M) \leq g(M), \text{ with } \lim_{M \to \infty} g(M) = 0. \tag{4.112}$$

Then, we have that:

$$\boldsymbol{z}_i \xrightarrow{\text{p}} 0. \tag{4.113}$$

Proof. Let us consider the following implication of events, for any positive values M and γ:

$$\left\{\boldsymbol{x}_i \leq \frac{\gamma}{M}\right\} \bigcap \left\{\boldsymbol{y}_i \leq M\right\} \Rightarrow \left\{\boldsymbol{x}_i \boldsymbol{y}_i \leq \gamma\right\}, \tag{4.114}$$

which, using De Morgan's laws [18], is equivalent to:

$$\left\{\boldsymbol{x}_i \boldsymbol{y}_i > \gamma\right\} \Rightarrow \left\{\boldsymbol{x}_i > \frac{\gamma}{M}\right\} \bigcup \left\{\boldsymbol{y}_i > M\right\}. \tag{4.115}$$

Since, for any two events $\mathcal{A}, \mathcal{B}$, the condition $\mathcal{A} \Rightarrow \mathcal{B}$ implies that $\mathbb{P}(\mathcal{A}) \leq \mathbb{P}(\mathcal{B})$, from (4.115), and using the union bound, we conclude that:

$$\begin{aligned} \mathbb{P}(\boldsymbol{z}_i > \gamma) &\leq \mathbb{P}(\boldsymbol{x}_i > \gamma/M) + \mathbb{P}(\boldsymbol{y}_i > M) \\ &\leq \mathbb{P}(\boldsymbol{x}_i > \gamma/M) + g(M), \end{aligned} \tag{4.116}$$

where the latter inequality follows by the upper bound in (4.112). Now, let us fix a value $\varepsilon > 0$. For sufficiently large M, we have that $g(M) \leq \varepsilon/2$ in view of the limit appearing in (4.112). On the other hand, since by assumption $\boldsymbol{x}_i$ converges to zero in probability, for given values of M and γ there exists certainly a sufficiently large i_0 such that, for every $i \geq i_0$, also the quantity $\mathbb{P}(\boldsymbol{x}_i > \gamma/M)$ is upper bounded by $\varepsilon/2$, which implies, for $i \geq i_0$:

$$\mathbb{P}(\boldsymbol{z}_i > \gamma) \leq \varepsilon, \tag{4.117}$$

and the claim of the lemma is proved. □

4.E Proof of Theorem 4.3

From Lemma 4.1 (see Appendix 4.B), we see that the sign of the quantity on the RHS of (4.44) will dictate different convergence behaviors. Note that KL divergences are finite from Assumption 2.4. First, consider the case when

$$D(\bar{\theta}_{\mathsf{TX}}) > D(\theta_{\mathsf{TX}}), \tag{4.118}$$

which implies that the asymptotic rate of convergence seen in (4.44) is strictly negative. Since $\boldsymbol{\mu}_{k,i}(\theta)$ is bounded by 1 for any hypothesis θ, then

$$\begin{aligned}
&\frac{1}{i}\log\frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} \xrightarrow{\text{a.s.}} D(\theta_{\mathsf{TX}}) - D(\bar{\theta}_{\mathsf{TX}}) < 0 \\
&\Rightarrow \log\frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} \xrightarrow{\text{a.s.}} -\infty \\
&\Rightarrow \boldsymbol{\mu}_{k,i}(\theta) \xrightarrow{\text{a.s.}} 0,
\end{aligned} \tag{4.119}$$

which holds for all $\theta \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$. This, in turn, implies that

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 1. \tag{4.120}$$

Next, consider the case:

$$D(\bar{\theta}_{\mathsf{TX}}) < D(\theta_{\mathsf{TX}}), \tag{4.121}$$

implying that the asymptotic rate of convergence in (4.44) is strictly positive. In this case, since again $\boldsymbol{\mu}_{k,i}(\theta)$ is bounded, we have that

$$\begin{aligned}
&\frac{1}{i}\log\frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} \xrightarrow{\text{a.s.}} D(\theta_{\mathsf{TX}}) - D(\bar{\theta}_{\mathsf{TX}}) > 0 \\
&\Rightarrow \log\frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} \xrightarrow{\text{a.s.}} +\infty \\
&\Rightarrow \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 0,
\end{aligned} \tag{4.122}$$

which, in view of (3.74), implies that, for every $\theta \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$,

$$\boldsymbol{\mu}_{k,i}(\theta) \xrightarrow{\text{a.s.}} \frac{1}{H-1}. \tag{4.123}$$

4.F Proof of Theorem 4.4

We will start by addressing the first part of Theorem 4.4. Let us develop the recursion in (4.12) with $\theta = \theta_{\mathsf{TX}}$ and $\theta' = \theta_0$.

$$\log\frac{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})}{\boldsymbol{\mu}_{k,i}(\theta_0)} = a_{kk}\log\frac{\boldsymbol{\mu}_{k,i-1}(\theta_{\mathsf{TX}})}{\boldsymbol{\mu}_{k,i-1}(\theta_0)} + a_{kk}\log\frac{L_k(\boldsymbol{\xi}_{k,i}|\theta_{\mathsf{TX}})}{L_k(\boldsymbol{\xi}_{k,i}|\theta_0)}$$

$$
\begin{aligned}
&+\sum_{\substack{\ell=1\\ \ell\neq k}}^{K} a_{\ell k}\log\frac{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}{\sum_{\tau\neq\theta_{\mathsf{TX}}}\frac{1}{H-1}\boldsymbol{\mu}_{\ell,i-1}(\tau)L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)}\\
&\overset{\text{(a)}}{\leq}\sum_{\ell=1}^{K} a_{\ell k}\log\frac{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})}{\boldsymbol{\mu}_{\ell,i-1}(\theta_0)}+\sum_{\ell=1}^{K} a_{\ell k}\log\frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}\\
&+\sum_{\substack{\ell=1\\ \ell\neq k}}^{K} a_{\ell k}\log\Big(\boldsymbol{\mu}_{\ell,i-1}(\theta_0)L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)\Big)\\
&-\sum_{\substack{\ell=1\\ \ell\neq k}}^{K} a_{\ell k}\sum_{\tau\neq\theta_{\mathsf{TX}}}\frac{\log\Big(\boldsymbol{\mu}_{\ell,i-1}(\tau)L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)\Big)}{H-1}\\
&=\sum_{\ell=1}^{K} a_{\ell k}\log\frac{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})}{\boldsymbol{\mu}_{\ell,i-1}(\theta_0)}+\sum_{\ell=1}^{K} a_{\ell k}\log\frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}\\
&-\sum_{\substack{\ell=1\\ \ell\neq k}}^{K}\frac{a_{\ell k}}{H-1}\sum_{\tau\neq\theta_{\mathsf{TX}}}\left(\log\frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}+\log\frac{\boldsymbol{\mu}_{\ell,i-1}(\tau)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_0)}\right),
\end{aligned}
\tag{4.124}
$$

where (a) follows from Jensen's inequality applied as follows:

$$
\log\left(\sum_{\tau\neq\theta_{\mathsf{TX}}}\frac{1}{H-1}\boldsymbol{\mu}_{\ell,i-1}(\tau)L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)\right)\geq\sum_{\tau\neq\theta_{\mathsf{TX}}}\frac{1}{H-1}\log\Big(\boldsymbol{\mu}_{\ell,i-1}(\tau)L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)\Big). \tag{4.125}
$$

Setting $\boldsymbol{y}_{k,i}=\log\frac{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})}{\boldsymbol{\mu}_{k,i}(\theta_0)}$ and

$$
\begin{aligned}
\boldsymbol{x}_{k,i}=&\sum_{\ell=1}^{K} a_{\ell k}\log\frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}-\sum_{\substack{\ell=1\\ \ell\neq k}}^{K}\frac{a_{\ell k}}{H-1}\sum_{\tau\neq\theta_{\mathsf{TX}}}\log\frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}\\
&-\sum_{\substack{\ell=1\\ \ell\neq k}}^{K}\frac{a_{\ell k}}{H-1}\sum_{\tau\neq\theta_{\mathsf{TX}}}\log\frac{\boldsymbol{\mu}_{\ell,i-1}(\tau)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_0)},
\end{aligned}
\tag{4.126}
$$

we can write (4.124) in vector form as:

$$
\boldsymbol{y}_i\preccurlyeq A^{\mathsf{T}}\boldsymbol{y}_{i-1}+\boldsymbol{x}_i, \tag{4.127}
$$

where the symbol $\preccurlyeq$ denotes element-wise inequality. Therefore, the recursion in (4.127) matches the model in (4.143), but for the fact that we have an inequality in place of an equality. Since the matrix A has nonnegative entries, we can still develop the recursion preserving the inequality, allowing us to use the results from Lemma 4.8 (Appendix 4.G) in the form of an inequality.

We need now to show that $\boldsymbol{x}_i$, as defined in (4.126), satisfies the conditions (4.145)–(4.147) in Lemma 4.8. Regarding the log-likelihood ratio terms, i.e., the first two terms on the RHS of (4.126), since these terms satisfy Assumption 2.4 and since the observations $\boldsymbol{\xi}_{\ell,i}$ are i.i.d.

across time, the result of Lemma 4.8 can be applied to these terms, in view of Property 4.1 (Appendix 4.G). For these two terms, we have that

$$\bar{x}_k = \sum_{\ell=1}^{K} a_{\ell k} \mathbb{E}\left(\log \frac{L_\ell(\theta_{\mathsf{TX}})}{L_\ell(\theta_0)}\right) - \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} \frac{a_{\ell k}}{H-1} \sum_{\tau \neq \theta_{\mathsf{TX}}} \mathbb{E}\left(\log \frac{L_\ell(\tau)}{L_\ell(\theta_0)}\right). \tag{4.128}$$

For what concerns the log-belief ratio, i.e., the third term on the RHS of (4.126), we have that this term behaves like the recursion seen in (4.106), which reveals that the log-belief ratio for the non-transmitted hypotheses matches the model in (4.162). As a result, conditions (4.145)–(4.147) are automatically satisfied in view of Lemma 4.9.

From Lemma 4.8, we have that

$$\begin{aligned}\limsup_{i\to\infty} \frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})}{\boldsymbol{\mu}_{k,i}(\theta_0)} \overset{\text{a.s.}}{\leq} &- \sum_{\ell=1}^{K} \pi_\ell d_\ell(\theta_{\mathsf{TX}}) + \sum_{\ell=1}^{K} \pi_\ell \sum_{\substack{n=1 \\ n\neq\ell}}^{K} \frac{a_{n\ell}}{H-1} \sum_{\tau\neq\theta_{\mathsf{TX}}} d_n(\tau) \\ &- \sum_{\ell=1}^{K} \pi_\ell \sum_{\substack{n=1 \\ n\neq\ell}}^{K} \frac{a_{n\ell}}{H-1} \sum_{\tau\neq\theta_{\mathsf{TX}}} \lim_{i\to\infty} \frac{1}{i} \sum_{j=1}^{i} \log \frac{\boldsymbol{\mu}_{n,j-1}(\tau)}{\boldsymbol{\mu}_{n,j-1}(\theta_0)}. \end{aligned} \tag{4.129}$$

We recall that for $\tau, \theta_0 \neq \theta_{\mathsf{TX}}$ and $\tau \in \bar{\Theta}_n$, according to Lemma 4.9 (Appendix 4.G),

$$\begin{aligned}\frac{1}{i} \sum_{j=1}^{i} \log \frac{\boldsymbol{\mu}_{n,j}(\tau)}{\boldsymbol{\mu}_{n,j}(\theta_0)} &\xrightarrow{\text{a.s.}} \frac{a_{nn}}{1-a_{nn}} \mathbb{E}\left(\log \frac{L_n(\tau)}{L_n(\theta_0)}\right) \\ &= -\frac{a_{nn}}{1-a_{nn}} d_n(\tau). \end{aligned} \tag{4.130}$$

Thus replacing (4.130) into (4.129), yields

$$\begin{aligned}&\limsup_{i\to\infty} \frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})}{\boldsymbol{\mu}_{k,i}(\theta_0)} \\ &\overset{\text{a.s.}}{\leq} - \sum_{\ell=1}^{K} \pi_\ell d_\ell(\theta_{\mathsf{TX}}) + \sum_{\ell=1}^{K} \pi_\ell \sum_{\substack{n=1 \\ n\neq\ell}}^{K} a_{n\ell} \left(\frac{a_{nn}}{1-a_{nn}} + 1\right) \frac{1}{H-1} \sum_{\tau\neq\theta_{\mathsf{TX}}} d_n(\tau) \\ &= - \sum_{\ell=1}^{K} \pi_\ell d_\ell(\theta_{\mathsf{TX}}) + \frac{1}{H-1} \sum_{\tau\neq\theta_{\mathsf{TX}}} \sum_{\ell=1}^{K} \pi_\ell \sum_{\substack{n=1 \\ n\neq\ell}}^{K} a_{n\ell} \frac{1}{1-a_{nn}} d_n(\tau) \\ &= - \sum_{\ell=1}^{K} \pi_\ell d_\ell(\theta_{\mathsf{TX}}) + \frac{1}{H-1} \sum_{\tau\neq\theta_{\mathsf{TX}}} \sum_{\ell=1}^{K} \pi_\ell d_\ell(\tau), \end{aligned} \tag{4.131}$$

where (4.131) follows from algebraic manipulations, taking into account the left stochasticity of matrix A and the definition of the Perron eigenvector π. As long as the RHS of (4.131) assumes a negative value, this implies that

$$\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 0. \tag{4.132}$$

The proof for the first part of Theorem 4.4 is complete. We proceed now to examine the second part. Considering Assumption 4.3 and developing the recursion in (4.106) for $\theta = \tau$ and $\theta' = \tau'$, the boundedness of log-likelihood ratios is inherited by the ratio of the log-beliefs for any non-transmitted hypotheses $\tau, \tau' \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$. In fact, exploiting (4.106) and the upper bound in (4.28), and iterating over i, we can write:

$$\begin{aligned}\log \frac{\boldsymbol{\mu}_{k,i}(\tau)}{\boldsymbol{\mu}_{k,i}(\tau')} &\leq a_{kk}^i \log \frac{\boldsymbol{\mu}_{k,0}(\tau)}{\boldsymbol{\mu}_{k,0}(\tau')} + B \sum_{j=1}^{i} a_{kk}^{i-j+1} \\ &= a_{kk}^i \log \frac{\mu_{k,0}(\tau)}{\mu_{k,0}(\tau')} + a_{kk} \frac{1 - a_{kk}^i}{1 - a_{kk}} B. \end{aligned} \tag{4.133}$$

We know that a_{kk}^i converges to zero as $i \to \infty$. For an arbitrarily small $\varepsilon > 0$, there exists an instant i_0 such that for $i > i_0$ we have that:

$$\begin{aligned}&\log \frac{\boldsymbol{\mu}_{k,i}(\tau)}{\boldsymbol{\mu}_{k,i}(\tau')} \leq \frac{a_{kk}}{1 - a_{kk}} B + \varepsilon \log \frac{\mu_{k,0}(\tau)}{\mu_{k,0}(\tau')} \\ &\Rightarrow \boldsymbol{\mu}_{k,i}(\tau) \leq \boldsymbol{\mu}_{k,i}(\tau') e^{\frac{a_{kk}}{1-a_{kk}} B + \epsilon}. \end{aligned} \tag{4.134}$$

where we defined:

$$\epsilon \triangleq \varepsilon \log \frac{\mu_{k,0}(\tau)}{\mu_{k,0}(\tau')}. \tag{4.135}$$

Note that if ε is arbitrarily small, ϵ will also be arbitrarily close to zero due to Assumption 2.2.

Developing the recursion in (4.12) with $\theta \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$ and $\theta' = \theta_{\mathsf{TX}}$, we have that:

$$\begin{aligned}&\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} \\ &= a_{kk} \log \frac{\boldsymbol{\psi}_{k,i}(\theta)}{\boldsymbol{\psi}_{k,i}(\theta_{\mathsf{TX}})} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \frac{\sum_{\tau \neq \theta_{\mathsf{TX}}} L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau) \boldsymbol{\mu}_{\ell,i-1}(\tau)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}}) \boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})(H-1)} \\ &= \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\psi}_{\ell,i}(\theta)}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \frac{\sum_{\tau \neq \theta_{\mathsf{TX}}} L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau) \boldsymbol{\mu}_{\ell,i-1}(\tau)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta) \boldsymbol{\mu}_{\ell,i-1}(\theta)(H-1)} \\ &\overset{(a)}{\leq} \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\psi}_{\ell,i}(\theta)}{\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \frac{\sum_{\tau \neq \theta_{\mathsf{TX}}} L_\ell(\boldsymbol{\xi}_{\ell,i}|\tau)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)(H-1)} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \left(\frac{a_{\ell\ell}}{1 - a_{\ell\ell}} B + \epsilon \right) \\ &\overset{(b)}{=} \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\mu}_{\ell,i-1}(\theta)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})} + \sum_{\ell=1}^{K} a_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})} \\ &+ \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\bar{\theta}_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \left(\frac{a_{\ell\ell}}{1 - a_{\ell\ell}} B + \epsilon \right) \\ &= \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\mu}_{\ell,i-1}(\theta)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})} + \sum_{\ell=1}^{K} a_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})} + \sum_{\ell=1}^{K} a_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\bar{\theta}_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)} \end{aligned}$$

$$
\begin{aligned}
&- a_{kk} \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\bar{\theta}_{\mathsf{TX}})}{L_k(\boldsymbol{\xi}_{k,i}|\theta)} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \left(\frac{a_{\ell\ell}}{1-a_{\ell\ell}} B + \epsilon \right) \\
&= \sum_{\ell=1}^{K} a_{\ell k} \log \frac{\boldsymbol{\mu}_{\ell,i-1}(\theta)}{\boldsymbol{\mu}_{\ell,i-1}(\theta_{\mathsf{TX}})} + \sum_{\ell=1}^{K} a_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\bar{\theta}_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})} \\
&- a_{kk} \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\bar{\theta}_{\mathsf{TX}})}{L_k(\boldsymbol{\xi}_{k,i}|\theta)} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \left(\frac{a_{\ell\ell}}{1-a_{\ell\ell}} B + \epsilon \right),
\end{aligned} \tag{4.136}
$$

where in (a) we used the bound in (4.134) for $\tau' = \theta$, that is:

$$
\frac{\boldsymbol{\mu}_{\ell,i-1}(\tau)}{\boldsymbol{\mu}_{\ell,i-1}(\theta)} \leq e^{\frac{a_{\ell\ell}}{1-a_{\ell\ell}} B + \epsilon} \tag{4.137}
$$

and in (b) we used the definition of $L_\ell(\bar{\theta}_{\mathsf{TX}})$ seen in (4.10). Setting $\boldsymbol{y}_{k,i} = \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})}$ and:

$$
\boldsymbol{x}_{k,i} = \sum_{\ell=1}^{K} a_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\bar{\theta}_{\mathsf{TX}})}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_{\mathsf{TX}})} - a_{kk} \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\bar{\theta}_{\mathsf{TX}})}{L_k(\boldsymbol{\xi}_{k,i}|\theta)} + \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \left(\frac{a_{\ell\ell}}{1-a_{\ell\ell}} B + \epsilon \right), \tag{4.138}
$$

we can rewrite (4.136) in vector form as:

$$
\boldsymbol{y}_i \preccurlyeq A^{\mathsf{T}} \boldsymbol{y}_{i-1} + \boldsymbol{x}_i, \tag{4.139}
$$

for all $i > i_0$. Accordingly, the recursion in (4.139) satisfies the model in (4.143) with inequality, and with initial state $\boldsymbol{y}_{k,i_0}$. As we develop the recursion, the inequality in (4.139) is preserved. Regarding the conditions on $\boldsymbol{x}_i$ for applying Lemma 4.8, the first two terms on the RHS of (4.138) inherit the i.i.d. property of the observations $\boldsymbol{\xi}_i$ and have finite expectation. The third term on the RHS of (4.138) is deterministic and bounded. Applying Property 4.1, we see that $\boldsymbol{x}_i$ satisfies the conditions (4.145)–(4.147) in Lemma 4.8. From Lemma 4.8 and for each $\theta \neq \theta_{\mathsf{TX}}$, we have that

$$
\begin{aligned}
\limsup_{i\to\infty} \frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} \overset{\text{a.s.}}{\leq} &\sum_{k=1}^{K} \pi_k \sum_{\ell=1}^{K} a_{\ell k} \mathbb{E}\left(\log \frac{L_\ell(\bar{\theta}_{\mathsf{TX}})}{L_\ell(\theta_{\mathsf{TX}})} \right) - \sum_{k=1}^{K} \pi_k a_{kk} \mathbb{E}\left(\log \frac{L_k(\bar{\theta}_{\mathsf{TX}})}{L_k(\theta)} \right) \\
&+ \sum_{k=1}^{K} \pi_k \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \left(\frac{a_{\ell\ell}}{1-a_{\ell\ell}} B + \epsilon \right).
\end{aligned} \tag{4.140}
$$

Taking into account the arbitrariness of ϵ, we end up with the following result:

$$
\begin{aligned}
\limsup_{i\to\infty} \frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}})} \overset{\text{a.s.}}{\leq} &- \sum_{k=1}^{K} \pi_k d_k(\bar{\theta}_{\mathsf{TX}}) + \sum_{k=1}^{K} \pi_k d_k(\theta_{\mathsf{TX}}) + \sum_{k=1}^{K} \pi_k a_{kk} d_k(\bar{\theta}_{\mathsf{TX}}) \\
&- \sum_{k=1}^{K} \pi_k a_{kk} d_k(\theta) + B \sum_{k=1}^{K} \pi_k \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} a_{\ell k} \frac{a_{\ell\ell}}{1-a_{\ell\ell}}
\end{aligned}
$$

$$\begin{aligned}
&\overset{\text{(a)}}{\leq} -\sum_{k=1}^{K}\pi_k d_k(\bar{\theta}_{\mathsf{TX}}) + \sum_{k=1}^{K}\pi_k d_k(\theta_{\mathsf{TX}}) + \sum_{k=1}^{K}\pi_k a_{kk} d_k(\bar{\theta}_{\mathsf{TX}}) \\
&+ B\sum_{k=1}^{K}\pi_k \sum_{\substack{\ell=1\\ \ell\neq k}}^{K} a_{\ell k}\frac{a_{\ell\ell}}{1-a_{\ell\ell}} \\
&\overset{\text{(b)}}{=} \sum_{k=1}^{K}\pi_k d_k(\theta_{\mathsf{TX}}) - \sum_{k=1}^{K}\pi_k(1-a_{kk}) d_k(\bar{\theta}_{\mathsf{TX}}) + B\sum_{k=1}^{K}\pi_k a_{kk},
\end{aligned} \tag{4.141}$$

where in (a) we considered that $\sum_{k=1}^{K}\pi_k a_{kk} d_k(\theta) \geq 0$ from the nonnegativity of the KL divergences and of terms a_{kk} and the positivity of the Perron eigenvector. In (b), we considered the left stochasticity of matrix A and the definition of the Perron eigenvector. As long as the RHS of (4.141) assumes a negative value, it implies that for all $\theta \in \Theta \setminus \{\theta_{\mathsf{TX}}\}$:

$$\boldsymbol{\mu}_{k,i}(\theta) \xrightarrow{\text{a.s.}} 0 \Rightarrow \boldsymbol{\mu}_{k,i}(\theta_{\mathsf{TX}}) \xrightarrow{\text{a.s.}} 1. \tag{4.142}$$

4.G Auxiliary Results

Lemma 4.8 (Main vector recursion). *Let a sequence of random vectors $\boldsymbol{y}_i$ with dimension $K \times 1$ be defined through the following recursion, for $i = 1, 2, \ldots$*

$$\boldsymbol{y}_i = A^{\mathsf{T}}\boldsymbol{y}_{i-1} + \boldsymbol{x}_i. \tag{4.143}$$

where $\boldsymbol{y}_0$ is an initial (a.s. finite) random vector, and A is a primitive left-stochastic (deterministic) matrix satisfying:

$$\lim_{i\to\infty} A^i = \pi\mathbb{1}^{\mathsf{T}}, \tag{4.144}$$

for some vector π with positive entries such that $\mathbb{1}^{\mathsf{T}}\pi = 1$. Moreover $\boldsymbol{x}_i$ is a sequence of random vectors (with entries $\{\boldsymbol{x}_{\ell,i}\}$) possessing the following properties, for a certain deterministic vector $\bar{x}$:

$$\frac{1}{i}\sum_{j=1}^{i}\boldsymbol{x}_{\ell,j} \xrightarrow{\text{a.s.}} \bar{x}_\ell, \tag{4.145}$$

$$\limsup_{i\to\infty}\frac{1}{i}\sum_{j=1}^{i}|\boldsymbol{x}_{\ell,j}| \overset{\text{a.s}}{=} \boldsymbol{M}, \tag{4.146}$$

$$\frac{\boldsymbol{x}_{\ell,i}}{i} \xrightarrow{\text{a.s.}} 0, \tag{4.147}$$

where $\boldsymbol{M}$ is a nonnegative (a.s. finite) random variable. Then we have that,

$$\frac{1}{i}\boldsymbol{y}_i \xrightarrow{\text{a.s.}} \mathbb{1}\pi^{\mathsf{T}}\bar{x}. \tag{4.148}$$

Proof. Iterating the recursion in (4.143), we get:

$$\boldsymbol{y}_i = (A^i)^{\mathsf{T}}\boldsymbol{y}_0 + \sum_{j=0}^{i-1}(A^j)^{\mathsf{T}}\boldsymbol{x}_{i-j}. \tag{4.149}$$

Once scaled by i, the first term on the RHS vanishes almost surely when i tends to infinity in view of the properties of A. We focus on the second term. It is useful to rewrite the summation in (4.149) as follows:

$$\frac{1}{i}\sum_{j=0}^{i-1}(A^j)^{\mathsf{T}}\boldsymbol{x}_{i-j} = \frac{1}{i}\sum_{j=0}^{i-1}(A^j - \pi\mathbb{1}^{\mathsf{T}})^{\mathsf{T}}\boldsymbol{x}_{i-j} + \frac{1}{i}\sum_{j=0}^{i-1}\mathbb{1}\pi^{\mathsf{T}}\boldsymbol{x}_{i-j}. \tag{4.150}$$

Regarding the last term on the RHS of (4.150), in view of (4.145), we have that:

$$\frac{1}{i}\sum_{j=0}^{i-1}\mathbb{1}\pi^{\mathsf{T}}\boldsymbol{x}_{i-j} = \mathbb{1}\pi^{\mathsf{T}}\frac{1}{i}\sum_{j=1}^{i}\boldsymbol{x}_j \xrightarrow{\text{a.s.}} \mathbb{1}\pi^{\mathsf{T}}\bar{x}. \tag{4.151}$$

Accordingly, the claim of the lemma will be proved if we show that the first term on the RHS of (4.150) vanishes with probability one. From (4.144), for some $\varepsilon > 0$, there exists an index i_0 such that, for all $j > i_0$:

$$\Big|[A^j]_{\ell k} - \pi_\ell\Big| < \varepsilon. \tag{4.152}$$

Let us therefore split the term of interest as:

$$\frac{1}{i}\sum_{j=0}^{i-1}(A^j - \pi\mathbb{1}^{\mathsf{T}})^{\mathsf{T}}\boldsymbol{x}_{i-j} = \frac{1}{i}\sum_{j=0}^{i_0}(A^j - \pi\mathbb{1}^{\mathsf{T}})^{\mathsf{T}}\boldsymbol{x}_{i-j} + \frac{1}{i}\sum_{j=i_0+1}^{i-1}(A^j - \pi\mathbb{1}^{\mathsf{T}})^{\mathsf{T}}\boldsymbol{x}_{i-j}. \tag{4.153}$$

Regarding the first term on the RHS of (4.153), we can write the absolute value of its $k-$th component as:

$$\begin{aligned}\frac{1}{i}\left|\sum_{j=0}^{i_0}\sum_{\ell=1}^{K}\Big([A^j]_{\ell k} - \pi_k\Big)\boldsymbol{x}_{\ell,i-j}\right| &\leq \frac{1}{i}\sum_{j=0}^{i_0}\sum_{\ell=1}^{K}\Big|[A^j]_{\ell k} - \pi_k\Big||\boldsymbol{x}_{\ell,i-j}| \\ &\overset{\text{(a)}}{\leq} \sum_{j=0}^{i_0}\sum_{\ell=1}^{K}\frac{|\boldsymbol{x}_{\ell,i-j}|}{i} \xrightarrow{\text{a.s.}} 0,\end{aligned} \tag{4.154}$$

where the inequality in (a) follows because A is left-stochastic and π is the Perron eigenvector and the almost sure convergence to 0 is due to (4.147).

Let us address the second term on the RHS of (4.153). Considering its k-th component, we can write its absolute value as:

$$\begin{aligned}\frac{1}{i}\left|\sum_{j=i_0+1}^{i-1}\sum_{\ell=1}^{K}\Big([A^j]_{\ell k} - \pi_\ell\Big)\boldsymbol{x}_{\ell,i-j}\right| &\overset{\text{(a)}}{\leq} \varepsilon\sum_{\ell=1}^{K}\frac{1}{i}\sum_{j=i_0+1}^{i-1}|\boldsymbol{x}_{\ell,i-j}| \\ &= \varepsilon\sum_{\ell=1}^{K}\frac{1}{i}\sum_{j=1}^{i-i_0-1}|\boldsymbol{x}_{\ell,j}|,\end{aligned} \tag{4.155}$$

where in (a) we used the bound in (4.152). From (4.155), in view of (4.146), it follows that

$$\limsup_{i\to\infty} \frac{1}{i} \left| \sum_{j=i_0+1}^{i-1} \sum_{\ell=1}^{K} \left([A^j]_{\ell k} - \pi_\ell \right) \boldsymbol{x}_{\ell,i-j} \right| \overset{\text{a.s.}}{\leq} \varepsilon \boldsymbol{M}. \tag{4.156}$$

Finally, in view of (4.154) and (4.156) we can write the absolute value of the $k-$th component of (4.153) as:

$$\limsup_{i\to\infty} \frac{1}{i} \left| \sum_{j=0}^{i} \sum_{\ell=1}^{K} \left([A^j]_{\ell k} - \pi_k \right) \boldsymbol{x}_{\ell,i-j} \right| \overset{\text{a.s.}}{\leq} \varepsilon \boldsymbol{M}. \tag{4.157}$$

From (4.157), due to the arbitrariness of ε, the term on the LHS of (4.153) vanishes and the proof is complete. □

The following property shows that conditions (4.145)–(4.147) in Lemma 4.8 are satisfied for the particular case in which the random vectors $\{\boldsymbol{x}_i\}$ are i.i.d. and have finite expectation.

Property 4.1 (Properties of random variables with finite expectation). Consider the sequence of i.i.d. integrable random vectors $\{\boldsymbol{x}_i\}$ with $\mathbb{E}(\boldsymbol{x}_i) = \bar{x}$. Then, the following properties are satisfied:

$$\frac{1}{i} \sum_{j=1}^{i} \boldsymbol{x}_j \xrightarrow{\text{a.s.}} \bar{x}, \tag{4.158}$$

$$\frac{1}{i} \sum_{j=1}^{i} |\boldsymbol{x}_j| \xrightarrow{\text{a.s.}} \mathbb{E}(|\boldsymbol{x}_j|) < \infty \tag{4.159}$$

$$\frac{\boldsymbol{x}_i}{i} \xrightarrow{\text{a.s.}} 0, \tag{4.160}$$

where the almost sure convergence holds element-wise for the random vector summations. Eqs. (4.158) and (4.159) follow from the Strong Law of Large Numbers (SLLN) [18] and (4.160) follows from integrability[9].

Lemma 4.9 (Scalar recursion). *Let $\boldsymbol{y}_i$ be a (scalar) random variable satisfying, for $0 < a < 1$ and $i = 1, 2, \ldots$:*

$$\boldsymbol{y}_i = a\boldsymbol{y}_{i-1} + a\boldsymbol{x}_i, \tag{4.162}$$

where $\{\boldsymbol{x}_i\}$ are i.i.d. integrable *random variables whose expectation is given by $\mathbb{E}(\boldsymbol{x}_i) = \mathsf{m}_{\mathsf{x}}$, and $\boldsymbol{y}_0$ is an initial (a.s. finite) random variable. We have that:*

[9] For any integrable random variable $\boldsymbol{z}$, and any $\varepsilon > 0$, we have that [74][Theorem 3.2.1]:

$$\varepsilon \sum_{i=1}^{\infty} \mathbb{P}(|\boldsymbol{z}| > \varepsilon i) \leq \mathbb{E}(|\boldsymbol{z}|) < \infty. \tag{4.161}$$

Since $\boldsymbol{x}_i$ are integrable and identically distributed, from (4.161), we have $\sum_{i=1}^{\infty} \mathbb{P}(|\boldsymbol{x}_i| > \varepsilon i) < \infty$. Therefore, condition (4.160) follows from the Borel-Cantelli lemma[18].

1. *$\boldsymbol{y}_i$ converges in distribution, as $i \to \infty$, to a random variable $\boldsymbol{y}_\infty$ that can be defined as:*

$$\boldsymbol{y}_i \xrightarrow{\text{d}} \boldsymbol{y}_\infty \triangleq \sum_{j=1}^{\infty} a^j \boldsymbol{x}_j. \tag{4.163}$$

2. *The following conditions are satisfied:*

$$\frac{\boldsymbol{y}_i}{i} \xrightarrow{\text{a.s.}} 0, \tag{4.164}$$

$$\frac{1}{i}\sum_{j=1}^{i} \boldsymbol{y}_j \xrightarrow{\text{a.s.}} \frac{a}{1-a}\mathsf{m}_{\mathsf{x}}, \tag{4.165}$$

$$\limsup_{i\to\infty} \frac{1}{i}\sum_{j=1}^{i} |\boldsymbol{y}_j| \overset{\text{a.s.}}{\leq} \frac{a}{1-a}\mathbb{E}\left(|\boldsymbol{x}_j|\right). \tag{4.166}$$

Proof. For item 1), we develop the recursion in (4.162):

$$\boldsymbol{y}_i = a^i \boldsymbol{y}_0 + \sum_{j=1}^{i} a^j \boldsymbol{x}_{i-j+1}. \tag{4.167}$$

As i goes to infinity, the first term on the RHS of (4.167) vanishes almost surely. Regarding the second term on the RHS of (4.167), since $\boldsymbol{x}_i$ are i.i.d. across i, we can write the following equality in distribution for $i = 1, 2, \ldots$:

$$\sum_{j=1}^{i} a^j \boldsymbol{x}_{i-j+1} \overset{\text{d}}{=} \sum_{j=1}^{i} a^j \boldsymbol{x}_j. \tag{4.168}$$

The random series on the RHS of (4.168) is the sum of independent random variables, with

$$\sum_{j=1}^{\infty} \mathbb{E}\left(|a^j \boldsymbol{x}_j|\right) = \mathbb{E}\left(|\boldsymbol{x}|\right) \sum_{j=1}^{\infty} a^j = \mathbb{E}\left(|\boldsymbol{x}|\right) \frac{a}{1-a} < \infty, \tag{4.169}$$

where index j was suppressed due to identical distribution across time. This condition is sufficient to conclude that the random series is almost-surely (and absolutely) convergent [75, Lemma 3.6′]. Denoting the value of the series by $\boldsymbol{y}_\infty$, we conclude that:

$$\boldsymbol{y}_i \xrightarrow{\text{d}} \boldsymbol{y}_\infty. \tag{4.170}$$

For item 2), we will first show the result in (4.164). To do that, consider again the recursion in (4.167):

$$\begin{aligned} \boldsymbol{y}_i &= a^i \boldsymbol{y}_0 + \sum_{j=1}^{i} a^j \boldsymbol{x}_{i-j+1} \\ \Rightarrow \frac{1}{i}\boldsymbol{y}_i &= \frac{1}{i} a^i \boldsymbol{y}_0 + \frac{1}{i}\sum_{j=1}^{i} a^j \boldsymbol{x}_{i-j+1}. \end{aligned} \tag{4.171}$$

The first term on the RHS of (4.171) converges to zero almost surely as i goes to infinity. Since $0 < a < 1$, we know that $\{a^j\}$ forms a converging sequence, which implies that for some $\varepsilon > 0$, there exists an index i_0 such that, for all $j > i_0$:

$$|a^j| < \varepsilon. \tag{4.172}$$

We can therefore rewrite the second term on the RHS of (4.171) as

$$\frac{1}{i}\sum_{j=1}^{i} a^j \boldsymbol{x}_{i-j+1} = \frac{1}{i}\sum_{j=1}^{i_0} a^j \boldsymbol{x}_{i-j+1} + \frac{1}{i}\sum_{j=i_0+1}^{i} a^j \boldsymbol{x}_{i-j+1}. \tag{4.173}$$

Let us address the first term on the RHS of (4.173), but considering its absolute value:

$$\frac{1}{i}\left|\sum_{j=1}^{i_0} a^j \boldsymbol{x}_{i-j+1}\right| \leq \frac{1}{i}\sum_{j=1}^{i_0} |\boldsymbol{x}_{i-j+1}| = \sum_{j=i-i_0+1}^{i} \frac{|\boldsymbol{x}_j|}{i} \xrightarrow{\text{a.s.}} 0, \tag{4.174}$$

which vanishes almost surely in view of Property 4.1 and similar arguments as the ones used in (4.161). Now consider the second term on the RHS of (4.173). In view of (4.172), the term can be bounded as:

$$\frac{1}{i}\left|\sum_{j=i_0+1}^{i} a^j \boldsymbol{x}_{i-j+1}\right| \leq \frac{1}{i}\,\varepsilon \sum_{j=i_0+1}^{i} |\boldsymbol{x}_{i-j+1}| = \frac{1}{i}\,\varepsilon \sum_{j=1}^{i-i_0} |\boldsymbol{x}_j| \tag{4.175}$$

$$\Rightarrow \limsup_{i\to\infty} \frac{1}{i}\left|\sum_{j=i_0+1}^{i} a^j \boldsymbol{x}_{i-j+1}\right| \overset{\text{a.s.}}{\leq} \varepsilon\mathbb{E}\,(|\boldsymbol{x}_j|), \tag{4.176}$$

where the RHS of (4.175) converges to the RHS of (4.176) in view of the SLLN (since $\boldsymbol{x}_j$ is integrable). Taking the absolute value of the LHS of (4.173) and using (4.174) and (4.176), we can write:

$$\limsup_{i\to\infty} \frac{1}{i}\left|\sum_{j=1}^{i} a^j \boldsymbol{x}_{i-j+1}\right| \overset{\text{a.s.}}{\leq} \varepsilon\mathbb{E}\,(|\boldsymbol{x}_j|). \tag{4.177}$$

Due to the arbitrariness of ε in (4.177), we conclude that the limit superior in (4.176) vanishes, and therefore (4.164) holds.

Let us now show the result in (4.165), but considering the original recursion in (4.162):

$$\begin{aligned} &\boldsymbol{y}_i = a\boldsymbol{y}_{i-1} + a\boldsymbol{x}_i \\ &\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i = \frac{a}{n}\sum_{i=1}^{n} \boldsymbol{y}_{i-1} + \frac{a}{n}\sum_{i=1}^{n} \boldsymbol{x}_i. \end{aligned} \tag{4.178}$$

Note that the first term on the RHS of (4.178) can be written as

$$\frac{1}{n}a\sum_{i=1}^{n} \boldsymbol{y}_{i-1} = \frac{a}{n}\sum_{k=0}^{n-1} \boldsymbol{y}_k = \frac{a}{n}\sum_{k=1}^{n} \boldsymbol{y}_k - a\frac{\boldsymbol{y}_n}{n} + a\frac{\boldsymbol{y}_0}{n}, \tag{4.179}$$

and therefore (4.178) can be rewritten as

$$(1-a)\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i = -a\frac{\boldsymbol{y}_n}{n} + a\frac{\boldsymbol{y}_0}{n} + a\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \xrightarrow{\text{a.s.}} a\mathsf{m}_{\mathsf{x}}, \tag{4.180}$$

where the first term on the RHS vanishes almost surely in view of (4.164) and so does the second term, whereas the third term converges almost surely to $a\mathsf{m}_{\mathsf{x}}$ from the SLLN. It remains to verify condition (4.166). To this aim, it is useful to introduce the recursion:

$$\boldsymbol{s}_i = a\boldsymbol{s}_{i-1} + |\boldsymbol{x}_i|, \text{ with initial condition } s_0 = |\boldsymbol{y}_0|. \tag{4.181}$$

From (4.181) we can write:

$$\boldsymbol{s}_i = a^i|\boldsymbol{y}_0| + \sum_{j=1}^{i} a^j|\boldsymbol{x}_{i-j+1}|. \tag{4.182}$$

Comparing (4.182) against (4.167), by application of the triangle inequality we conclude that $|\boldsymbol{y}_i| \leq \boldsymbol{s}_i$. On the other hand, $\boldsymbol{s}_i$ matches the model in (4.162) and, hence, in view of (4.165) we can write ($\mathbb{E}\,(|\boldsymbol{x}|)$ is the common mean of the random variables $|\boldsymbol{x}_j|$):

$$\lim_{i\to\infty}\frac{1}{i}\sum_{j=1}^{i}\boldsymbol{s}_j = \frac{a}{1-a}\mathbb{E}\,(|\boldsymbol{x}|). \tag{4.183}$$

Moreover, since $|\boldsymbol{y}_i| \leq \boldsymbol{s}_i$ we have that:

$$\limsup_{i\to\infty}\frac{1}{i}\sum_{j=1}^{i}|\boldsymbol{y}_j| \overset{\text{a.s.}}{\leq} \lim_{i\to\infty}\frac{1}{i}\sum_{j=1}^{i}\boldsymbol{s}_j \overset{\text{a.s.}}{=} \frac{a}{1-a}\mathbb{E}\,(|\boldsymbol{x}|) < \infty, \tag{4.184}$$

which reveals that condition (4.166) holds. □

Non-Stationary World Part II

5 Adaptive Social Networks

5.1 Introduction[1]

In previous chapters, we have considered social learning in the context of a stationary world. The real world is however constantly changing over time, and engineering systems should account for this nonstationarity. For example, a real-time weather forecast system should be able to detect changes in air pressure and humidity and *adapt* its prediction based on these measurements.

Although existing social learning strategies incorporate streaming observations from the world, they are designed under the assumption that the operating conditions (e.g., the underlying state of nature, the network topology, the quality of data, the statistical models,...) are fixed over time. In this stationary setting, agents manage to concentrate their beliefs around the true state of the world, often at an exponentially fast rate of convergence. Such superior convergence properties have however the collateral effect of hindering adaptation in face of *nonstationary world conditions*, which we illustrate next by means of an example.

Consider a network of agents aiming to solve a weather forecast problem using a social learning algorithm. At each instant, these agents collect data arising from one among three possible hypotheses: "*sunny*", "*cloudy*", "*rainy*". At first, data is consistent with the hypothesis "*sunny*", but, in view of a weather change at instant $i = 200$, data then indicates that the correct forecast is "*rainy*". As we see in Figure 5.1 (the curves illustrate the behavior of Agent 1), the social learning algorithm reacts with a considerable inertia to the hypothesis drift.

In fact, Figure 5.1 shows clearly that the agent learns well until instant $i = 200$, whereas from $i = 200$ onward, the situation changes greatly. The classic social learning algorithm has a delayed reaction. First, agents perceive a change only at $i \approx 350$, but start opting for the wrong hypothesis "*cloudy*". Then, after a prohibitive number of iterations, at $i \approx 550$, agents manage to overcome their stubbornness and opt for the correct hypothesis "*rainy*". To tackle this problem, this chapter proposes an Adaptive Social Learning (ASL) strategy, whose performance is shown in the second column of Figure 5.1 for the same example. We see that the ASL algorithm manages to track the target change at instant $i \approx 200$, exhibiting a higher

[1] This chapter is adapted from [76], [77].

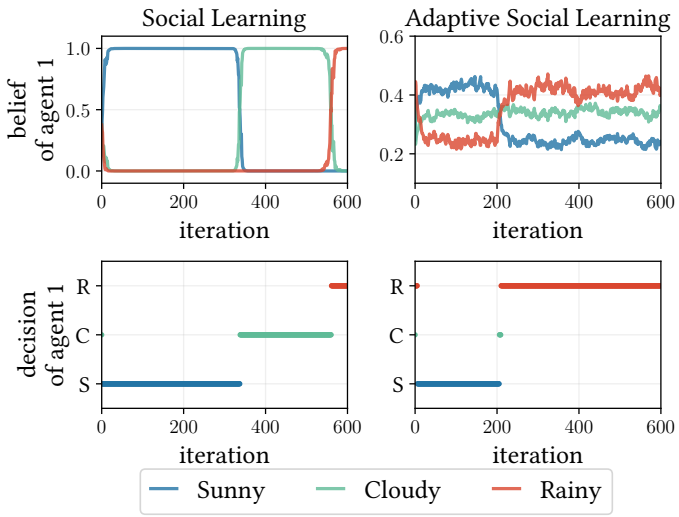


Figure 5.1: *Classic* social learning vs. *adaptive* social learning. *Top panels*: Belief evolution of agent 1, with θ_0 changing at time $i = 200$. *Bottom panels*: The instantaneous decision taken by agent 1 by choosing the hypothesis that maximizes the current belief.

adaptation capacity than that of the classic social learning algorithm.

In this chapter, we first introduce a novel social learning strategy that enables adaptation. Then, we provide a detailed analysis of this strategy. In particular, by exploiting recent advances in the field of distributed detection over adaptive networks—see [78] for an overview—we present a characterization of the social learning performance at each individual agent, in terms of **i)** convergence of the system at the steady-state (Theorem 5.1); **ii)** achievability of consistent learning (Theorem 5.2); **iii)** a Gaussian approximation for the learning performance (Theorem 5.3); **iv)** the error exponents for the learning error probabilities (Theorem 5.4); and **v)** the transient evolution for the instantaneous error probabilities. As the analysis will show, the ASL model allows the user to design the adaptation time, at the expense of losing some learning accuracy, i.e., agents no longer achieve full confidence around the true hypothesis. Instead, agents maintain some skepticism regarding the true hypothesis, as illustrated in the belief curves of Figure 5.1.

5.2 Problem Setting

We consider a strongly connected network, where agents observe the world under objective evidence—see Section 2.1.3. In other words, the observations measured by agent k at every instant are generated from one of the likelihood models $L_k(\xi|\theta_0)$ with $\theta_0 \in \Theta$.

We would like to account for nonstationary world conditions, e.g., for changing θ_0 over time. To that purpose, we devise an *adaptive* social learning strategy, which should be able to react promptly in view of environment changes and deliver proper inference performance in a reasonable reaction time. It is thus necessary to first identify formally the concepts of adaptation and learning, and the technical framework that will be used to characterize these concepts.

- **Learning**: In the context of social learning, “learning" means “guessing the right hypothesis". In order to quantify the learning performance, we specialize the standard prescriptions of adaptation theory to the social learning context. Given that the data are steadily generated according to a certain true likelihood model, what is the probability that an agent guesses the true state of nature? In the theory of adaptation, this analysis is commonly referred to as *steady-state* analysis [37].
- **Adaptation**: Assume that the system has been in operation for an arbitrary time. During this time, several phenomena can have occurred, i.e., variations of the true hypothesis, variations in the statistical conditions (i.e., malfunctioning of the system giving rise to distributions different from the nominal ones), missing observations, and so on. Due to the recursive nature of the social learning algorithms, at a given time i_0 all these variations are simply summarized in a certain initial belief vector μ_{i_0}. From $i_0 + 1$ onward, assume that the system becomes stable and the data are steadily generated according to a given likelihood model. Accordingly, the adaptation ability will be quantified by measuring how long it takes (*adaptation time*), given an arbitrary initial belief μ_{i_0}, for an agent to enter the steady-state regime and reach a prescribed probability of guessing the true hypothesis. In the theory of adaptation, this analysis is commonly referred to as *transient* analysis [37].

5.3 ASL Strategy

Examining the Bayesian update in (2.2), we see that it incorporates the new information into the past belief by giving *equal* weight to both $\boldsymbol{\mu}_{k,i-1}$ and the new information contained in $L_k(\boldsymbol{\xi}_{k,i}|\theta)$. In order to promote adaptation, it is necessary to increase the relative credit given to the new data with respect to the belief accumulated over time by learning from past data. To this end, we turn the Bayesian update step into the following *adaptive* form:

$$\boldsymbol{\psi}_{k,i}(\theta) = \frac{\boldsymbol{\mu}_{k,i-1}^{1-\delta}(\theta) L_k^{\delta}(\boldsymbol{\xi}_{k,i}|\theta)}{\sum_{\theta' \in \Theta} \boldsymbol{\mu}_{k,i-1}^{1-\delta}(\theta') L_k^{\delta}(\boldsymbol{\xi}_{k,i}|\theta')}, \tag{5.1}$$

where $0 < \delta < 1$ is a design parameter employed by each agent to modulate the relative weights assigned to the past and new information. In particular, relatively large values for δ give more importance to the new data, whereas small values for δ give more importance to the past beliefs. In this way, as we will show later in Section 5.7, the step-size parameter δ infuses the social learning algorithm with an adaptation mechanism.

The intermediate belief resulting from (5.1) is propagated across neighboring agents, and locally aggregated using the combination rule in (2.3) that we report here for ease of reference:

$$\boldsymbol{\mu}_{k,i}(\theta) = \frac{\prod_{\ell \in \mathcal{N}_k} \boldsymbol{\psi}_{\ell,i}(\theta)^{a_{\ell k}}}{\sum_{\theta' \in \Theta} \prod_{\ell \in \mathcal{N}_k} \boldsymbol{\psi}_{\ell,i}(\theta)^{a_{\ell k}}} \tag{5.2}$$

Examining (5.1), we see that the ASL strategy implements a convex combination of probability functions at the exponent, by discounting both the past belief and the new likelihood through the weights $1 - \delta$ and δ, respectively. However, the update (5.1) cannot be considered a *Bayesian*

update because the likelihood exponentiated to δ does not integrate to one (w.r.t. ξ).

We can however modify (5.1) to get an *adaptive Bayesian* update:

$$\boldsymbol{\psi}_{k,i}(\theta) = \frac{\boldsymbol{\mu}_{k,i-1}^{1-\delta}(\theta) L_k(\boldsymbol{\xi}_{k,i}|\theta)}{\sum_{\theta' \in \Theta} \boldsymbol{\mu}_{k,i-1}^{1-\delta}(\theta') L_k(\boldsymbol{\xi}_{k,i}|\theta')}. \tag{5.3}$$

Observe from (5.3) that the limiting choice $\delta = 0$ (i.e., no adaptation) gives back the classic Bayesian update in (2.2). In contrast, the update in (5.1) cannot be reduced to (2.2) for any selection of $\delta \in (0, 1)$. This notwithstanding, we now argue that the ASL strategies (5.1) and (5.3) are in fact equivalent. To this aim, we can develop the recursion obtained by combining (5.2) and (5.3) and get:

$$\begin{aligned}
\log \frac{\boldsymbol{\mu}_{k,i}(\theta)}{\boldsymbol{\mu}_{k,i}(\theta')} &= (1-\delta) \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \log \frac{\boldsymbol{\mu}_{\ell,i-1}(\theta)}{\boldsymbol{\mu}_{\ell,i-1}(\theta')} + \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta')} \\
&= (1-\delta)^i \sum_{\ell=1}^{K} [A^i]_{\ell k} \log \frac{\boldsymbol{\mu}_{\ell,0}(\theta)}{\boldsymbol{\mu}_{\ell,0}(\theta')} + \sum_{m=0}^{i-1} \sum_{\ell=1}^{K} (1-\delta)^m [A^{m+1}]_{\ell k} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i-m}|\theta)}{L_\ell(\boldsymbol{\xi}_{\ell,i-m}|\theta')}.
\end{aligned} \tag{5.4}$$

Doing the same for (5.1) and (5.2) in an expression similar to (5.4), except that an additional term δ multiplies the second term on the RHS of (5.4). In both cases, the first term on the RHS dies out exponentially fast with time, whereas the relevant term that determines the algorithms' evolution over time is given by the second term on the RHS, which differ only on the scaling factor δ.

As a result, we conclude that the time-evolution of the log-belief ratios for the two ASL strategies is equivalent. For example, the opinion that maximizes the belief function would be the same under both strategies, implying the same error probability. In fact, proportionality of the log-belief ratios implies that the belief function of one strategy is simply an exponentiated (and normalized) version of the belief function of the other strategy. This does not mean that the beliefs of the two strategies would take on the same values. In particular, our results will show that, as $\delta \to 0$, the steady-state log-belief ratios are stable under (5.1), which immediately implies that they diverge (i.e., achieving a belief close to 1 at the true hypothesis) under (5.3). While immaterial from a technical perspective, these differences might matter from a *behavioral* perspective [43], namely, to understand which update strategy reflects better the way of reasoning that an individual agent uses in social learning environments. For the sake of clarity, in the presentation of our technical results we opt for sticking to the update rule in (5.1), since the log-belief ratio is stable.

5.4 Statistical Descriptors of Performance

Assume that the algorithm has been running until a certain time i_0, with the evolution of the system up to i_0 being summarized in the "initial" belief vectors $\boldsymbol{\mu}_{k,i_0}$. Starting from i_0, the ASL algorithm behavior will exhibit two important phases: a *transient* phase where, given the (possibly wrong) initial belief, each agent must suddenly adapt in order to depart from $\boldsymbol{\mu}_{k,i_0}$ and start learning the correct hypothesis; and a *steady-state* phase where, given sufficient time

to learn ($i \to \infty$), each agent must achieve high confidence in learning the correct hypothesis. According to the theory of adaptive inference, the performance of an adaptive learning strategy is characterized under the *steady-state* regime.

By examining the algorithm recursions (5.1) and (5.2), in light of Assumptions 2.4 and 2.2, the belief remains always nonzero at any θ during the algorithm evolution—a similar argument is presented in Chapter 2. Now, assume that the algorithm has been running up to time i_0, and that from $i_0 + 1$ onward the system remains stationary for sufficiently long time, with the data being generated according to hypothesis θ_0. In order to perform a steady-state analysis from $i_0 + 1$ onward, we need to consider $\boldsymbol{\mu}_{k,i_0}$ as initial state. Since we have observed that the beliefs are always nonzero, we can see that the initial belief vector $\boldsymbol{\mu}_{k,i_0}$ fulfills Assumption 2.2.

In summary, for the purpose of the steady-state analysis and without loss of generality, we will assume that the steady-state analysis starts at time $i_0 = 0$ and consider an initial belief vector $\boldsymbol{\mu}_{k,0}$ that fulfills Assumption 2.2. The true hypothesis θ_0 is kept constant over time, yielding:

$$\boldsymbol{\xi}_{k,i} \sim L_k(\xi|\theta_0), \;\; k = 1, 2, \ldots, K, \;\; i = 1, 2, \ldots \tag{5.5}$$

Therefore, for the purpose of the steady-state analysis, we will always imply that expectations and probabilities are evaluated under the distributions $L_k(\xi|\theta_0)$. Note also that the observations $\{\boldsymbol{\xi}_{k,i}\}$ are independent and identically distributed (i.i.d.) over time, i.e., over the index i. We will assume that they can have different distributions across the agents, i.e., across the index k. Statistical independence across the agents will be only used to prove some of the forthcoming results (Theorems 5.3 and 5.4 further ahead).

Log-Belief Ratios: In order to characterize the learning performance, it is convenient to introduce the logarithm of the ratio between the belief evaluated at θ_0 and the belief evaluated at a generic hypothesis $\theta \neq \theta_0$:

$$\boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta) \triangleq \log \frac{\boldsymbol{\mu}_{k,i}(\theta_0)}{\boldsymbol{\mu}_{k,i}(\theta)}, \tag{5.6}$$

which is well-defined since, as already remarked, the belief remains nonzero at any θ during the algorithm evolution. With the symbol $\boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta)$ we denote a random function of: the agent index $k = 1, 2, \ldots, K$, the time index $i = 0, 1, \ldots$, the hypothesis $\theta \in \Theta \setminus \theta_0$, and the adaptation parameter δ. When we omit the argument θ and write $\boldsymbol{\lambda}_{k,i}^{(\delta)}$, we will be referring to the $(H-1)\times 1$ vector of log-belief ratios, namely,

$$\boldsymbol{\lambda}_{k,i}^{(\delta)} = \left[\boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta_1), \boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta_2), \ldots, \boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta_{H-1})\right]^{\mathsf{T}}, \tag{5.7}$$

where the elements in the set of wrong-hypotheses have been indexed as:

$$\Theta \setminus \theta_0 = \{\theta_1, \theta_2, \ldots, \theta_{H-1}\}. \tag{5.8}$$

Error Probability: One natural way for the agents to choose a hypothesis is to select the hypothesis that maximizes the belief. Therefore, the error probability at each time i can be

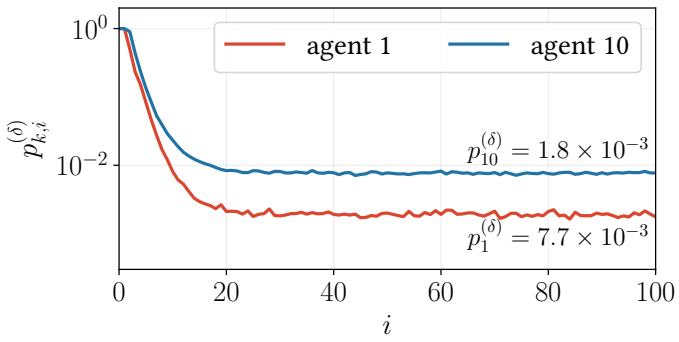


Figure 5.2: Evolution of the error probability of two agents in a network running the ASL algorithm.

expressed as

$$p_{k,i}^{(\delta)} = \mathbb{P}\left(\arg\max_{\theta\in\Theta} \boldsymbol{\mu}_{k,i}(\theta) \neq \theta_0\right). \tag{5.9}$$

It is useful to rewrite the error probability as a function of the log-belief ratios. To this end, observe that the event within brackets in (5.9) corresponds to saying that the belief is not maximized at θ_0, which in turn corresponds to saying that the log-belief ratios in (5.6) are less than or equal to zero for at least one $\theta \neq \theta_0$. Therefore, the *instantaneous* error probability can be equivalently rewritten as:

$$p_{k,i}^{(\delta)} = \mathbb{P}\left(\exists\theta \neq \theta_0 : \boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta) \leq 0\right). \tag{5.10}$$

Finally, we introduce the *steady-state* error probability:

$$p_k^{(\delta)} \triangleq \lim_{i\to\infty} p_{k,i}^{(\delta)}. \tag{5.11}$$

Theorem 5.1 will show that the steady-state error probability *exists* and can be characterized by the steady-state behavior of the log-belief ratios. To *evaluate* this probability, we will perform an asymptotic analysis in the regime of small δ, which will allow us to obtain reliable predictions of the steady-state performance.

In Figure 5.2 we show an example of evolution for the error probability of two agents in a network implementing the ASL strategy.[2] All the probabilities are estimated empirically by Monte Carlo simulation. We see how the instantaneous error probability $p_{k,i}^{(\delta)}$ converges to a steady-state *nonzero* value $p_k^{(\delta)}$ as i increases. It is useful to remark that this behavior is different from that of classic social learning, where, *under stationary conditions*, the error probability of each agent vanishes as time elapses. This is one instance of the adaptation/learning trade-off: Non-adaptive strategies can increase their accuracy indefinitely under stationary conditions. However, astronomically low values of the error probabilities lead to a detrimental inertia in responding to possible changes.

Log-Likelihood Ratios: For $k = 1, 2, \ldots, K$, $i = 0, 1, \ldots$, and $\theta \neq \theta_0$, we introduce the

[2]The details of the network topology as well as of the statistical learning problem are immaterial at this stage of the presentation.

log-likelihood ratio:

$$\boldsymbol{x}_{k,i}(\theta) \triangleq \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta_0)}{L_k(\boldsymbol{\xi}_{k,i}|\theta)}, \tag{5.12}$$

and its expectation:

$$d_k(\theta) \triangleq \mathbb{E}\,(\boldsymbol{x}_{k,i}(\theta)) < \infty, \tag{5.13}$$

namely, the KL divergence between $L_k(\xi|\theta_0)$ and $L_k(\xi|\theta)$, which is finite in view of Assumption 2.4, implying that the log-likelihood ratios cannot diverge (but for an ensemble of realizations with zero probability). We recall that the expectation in (5.13) is computed assuming that the random variable $\boldsymbol{\xi}_{k,i}$ is distributed according to model $L_k(\xi|\theta_0)$. Since we focus on the steady state, this distribution is constant over time, which explains why $d_k(\theta)$ does not depend on i. Furthermore, since the true hypothesis θ_0 is held fixed during the steady-state analysis, in order to avoid a heavier notation we are not emphasizing the dependence of the KL divergence $d_k(\theta)$ on θ_0.

We continue by introducing an average variable that will play a role in the forthcoming results, namely, the *network average* of log-likelihood ratios, for all $\theta \neq \theta_0$:

$$\boldsymbol{x}_{\mathrm{ave},i}(\theta) = \sum_{\ell=1}^{K} \pi_\ell \boldsymbol{x}_{\ell,i}(\theta). \tag{5.14}$$

The random variable $\boldsymbol{x}_{\mathrm{ave},i}(\theta)$ appearing in (5.14) is obtained by combining linearly the local log-likelihood ratios $\boldsymbol{x}_{\ell,i}(\theta)$. The combination weight assigned to the log-likelihood ratio of the ℓ-th agent is given by the *limiting* combination weight, i.e., by the ℓ-th entry, π_ℓ, of the Perron eigenvector. We will see in the following that the asymptotic properties of the ASL strategy as $\delta \to 0$ are directly related to the statistical properties of the vector of average variables, $\boldsymbol{x}_{\mathrm{ave},i}$.

5.5 Steady-State Analysis

Different from the classic social learning setting, in the *adaptive* setting the belief will not converge as $i \to \infty$. In contrast, the belief of each agent will preserve a *random* behavior—this can be seen in the example shown in Figure 5.1. This everlasting randomness is critical to ensure that the algorithm will adapt quickly to a change in the environment. On the other hand, it makes the steady-state analysis more difficult, since the beliefs preserve a random character even when $i \to \infty$. The first step in the steady-state analysis is to establish whether such random fluctuations lead to *stable* random variables as $i \to \infty$, which is shown in Theorem 5.1.

Before stating the theorem, let us examine the evolution of the log-belief ratios. Exploiting (5.2) and (5.1), we end up with the following recursion, for every $\theta \neq \theta_0$:

$$\boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta) = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \left\{ (1-\delta)\boldsymbol{\lambda}_{\ell,i-1}^{(\delta)}(\theta) + \delta \boldsymbol{x}_{\ell,i}(\theta) \right\}, \tag{5.15}$$

which can be rewritten as the following two-step recursion:

$$\boldsymbol{\nu}_{\ell,i}^{(\delta)}(\theta) = (1-\delta)\boldsymbol{\lambda}_{\ell,i-1}^{(\delta)}(\theta) + \delta\, \boldsymbol{x}_{\ell,i}(\theta), \tag{5.16}$$

$$\boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta) = \sum_{\ell \in \mathcal{N}_k} a_{\ell k}\, \boldsymbol{\nu}_{\ell,i}^{(\delta)}(\theta). \tag{5.17}$$

The time-evolution of the log-belief ratios in (5.16) and (5.17) is in the form of a *diffusion* algorithm with constant *step-size* δ—see, e.g., [37]. This is why we refer to δ as the step-size.

Developing the recursion in (5.15) and recalling that $A = [a_{\ell k}]$ is the combination matrix we can write, for all $\theta \neq \theta_0$:

$$\boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta) = \underbrace{(1-\delta)^i \sum_{\ell=1}^{K} [A^i]_{\ell k} \boldsymbol{\lambda}_{\ell,0}(\theta)}_{\text{transient term}} + \delta \sum_{m=0}^{i-1} \sum_{\ell=1}^{K} (1-\delta)^m [A^{m+1}]_{\ell k}\, \boldsymbol{x}_{\ell,i-m}(\theta). \tag{5.18}$$

Since the transient term dies out as $i \to \infty$, in order to evaluate the steady-state behavior of $\boldsymbol{\lambda}_{k,i}(\theta)$, we can ignore it and focus on the second term:

$$\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) = \delta \sum_{\ell=1}^{K} \sum_{m=0}^{i-1} (1-\delta)^m [A^{m+1}]_{\ell k}\, \boldsymbol{x}_{\ell,i-m}(\theta). \tag{5.19}$$

5.5.1 Steady-State Log-Belief Ratios

The goal of the steady-state analysis is to evaluate the performance (i.e., the error probability) for large i. For this evaluation to be meaningful, we must ascertain that the error probability in (5.10) converges as $i \to \infty$. To this end, we will now establish that there exists a certain limiting random vector, $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$, such that the *probability distribution* of the vector of log-belief ratios, $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}$, converges, as $i \to \infty$, to the probability distribution of $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$. This notion of convergence can be formally defined as follows.

We say that the sequence (over the index i) of random vectors $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}$ converges *in distribution* or *weakly* as $i \to \infty$ if we can define a random vector $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$ such that [73]:

$$\lim_{i \to \infty} \mathbb{P}\left(\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)} \in \mathcal{B}\right) = \mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)} \in \mathcal{B}\right) \tag{5.20}$$

for all measurable sets $\mathcal{B}$ whose boundary $\partial\mathcal{B}$ has zero probability under the limiting distribution, namely, for all measurable sets $\mathcal{B}$ fulfilling the condition:

$$\mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)} \in \partial\mathcal{B}\right) = 0. \tag{5.21}$$

In the following, weak convergence will be compactly denoted as:

$$\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)} \xrightarrow{\text{d}} \widetilde{\boldsymbol{\lambda}}_k^{(\delta)}, \tag{5.22}$$

and the vector $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$ will be referred to as the *steady-state* log-belief vector, since it provides the statistical characterization of the log-belief vector $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}$ as $i \to \infty$.

We are now ready to present the theorem that establishes the existence of steady-state log-belief ratios.

Theorem 5.1 (Steady-state log-belief ratios). *Let Assumptions 2.4 and 2.2 hold, and let*

$$\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) \triangleq \delta \sum_{\ell=1}^{K} \sum_{m=0}^{i-1} (1-\delta)^m [A^{m+1}]_{\ell k}\, \boldsymbol{x}_{\ell,m+1}(\theta) \tag{5.23}$$

be the random sum obtained from (5.19) *by taking the summands in reversed order.*

First, we have that all the K inner sums in (5.23) *are almost-surely absolutely convergent as $i \to \infty$, implying that $\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta)$ converges almost surely to the random series:*

$$\widetilde{\boldsymbol{\lambda}}_{k}^{(\delta)}(\theta) \triangleq \delta \sum_{\ell=1}^{K} \sum_{m=0}^{\infty} (1-\delta)^m [A^{m+1}]_{\ell k}\, \boldsymbol{x}_{\ell,m+1}(\theta). \tag{5.24}$$

Second, we have that the vector of log-belief ratios $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}$ (with the original, i.e., non-reversed ordering of summation) converges in distribution to the vector $\widetilde{\boldsymbol{\lambda}}_{k}^{(\delta)}$, namely,

$$\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)} \xrightarrow{\mathrm{d}} \widetilde{\boldsymbol{\lambda}}_{k}^{(\delta)}. \tag{5.25}$$

Proof. See Appendix 5.B. □

It is useful to make some comments on Theorem 5.1. First, finiteness of the expectation of $\boldsymbol{x}_{k,i}$ is sufficient (through Assumption 2.4) to guarantee the existence of a steady-state random variable. No assumption is made on higher-order moments.

Second, it is important to notice that (5.24) does not correspond to letting $i \to \infty$ in the summation in (5.19). To explain why, let us compare the random sums:

$$\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) = \delta \sum_{m=0}^{i-1} \sum_{\ell=1}^{K} (1-\delta)^m [A^{m+1}]_{\ell k}\, \boldsymbol{x}_{\ell,i-m}(\theta), \tag{5.26}$$

and

$$\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) = \delta \sum_{m=0}^{i-1} \sum_{\ell=1}^{K} (1-\delta)^m [A^{m+1}]_{\ell k}\, \boldsymbol{x}_{\ell,m+1}(\theta). \tag{5.27}$$

In Figure 5.3 we examine a sample path for these sums, and we can see that they exhibit different behavior. The random sum in (5.26), displayed with solid line in Figure 5.3, exhibits steadily random fluctuations as time elapses. In contrast, the random sum in (5.27), displayed with dashed line, converges as time elapses, specifically to the random value $\widetilde{\boldsymbol{\lambda}}_{k}^{(\delta)}(\theta)$ defined in (5.24). Both behaviors are consistent with what we have already shown in Theorem 5.1. These profoundly different behaviors depend on the different ordering of the summands in (5.26) and (5.27). In particular, in (5.27) the most recent term, $\boldsymbol{x}_{\ell,i}(\theta)$, takes the smallest weight $(1-\delta)^{i-1}$, which lets the remainder of the series vanish (almost surely). In contrast, in (5.26) the most

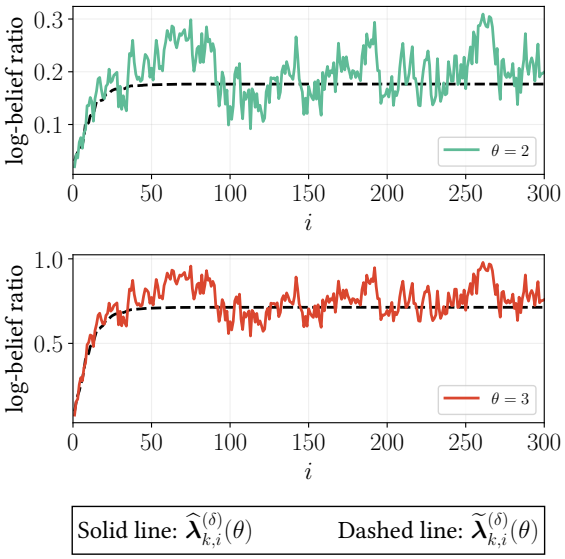


Figure 5.3: Comparison of the random sequences $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta)$ and $\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta)$ for $\delta = 0.1$, for the Gaussian setting described in Section 5.8 further ahead.

recent term, $\boldsymbol{x}_{\ell,i}(\theta)$ takes the highest weight $(1-\delta)^0 = 1$, thus keeping fluctuations (hence, adaptation) alive.

Even though the sums in (5.26) and (5.27) exhibit a markedly different behavior in terms of their time-evolution (i.e., on the sample paths), one notable conclusion from Theorem 5.1 is that their *probability distributions* converge to the same distribution, that is the distribution of the limiting variable $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$. This equivalence can be explained as follows. With reference to the top panel in Figure 5.3, consider a sufficiently large i (say, $i = 300$) and take the corresponding values of the dashed curve and the solid curve, namely, $\widehat{\boldsymbol{\lambda}}_{k,300}(2)$ and $\widetilde{\boldsymbol{\lambda}}_{k,300}(2)$. These values are different. However, if we now repeat the experiment in Figure 5.3 several times, the realizations of $\widehat{\boldsymbol{\lambda}}_{k,300}(2)$ across different experiments will be distributed in the same way as the realizations of $\widetilde{\boldsymbol{\lambda}}_{k,300}(2)$.

The existence of a limiting distribution for the log-belief vector $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}$ makes the definition of a *steady-state* error probability meaningful, since from Eqs. (5.10) and (5.11) we see that the steady-state error probability can be computed as:[3]

$$p_k^{(\delta)} = \mathbb{P}\left(\exists \theta \neq \theta_0 : \widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) \leq 0\right). \tag{5.28}$$

However, it should be noticed that Theorem 5.1 constitutes only a first, albeit fundamental step towards the characterization of the ASL performance, since it establishes only the existence of a steady-state error probability without providing any explicit characterization thereof. Such

[3]According to the definition of convergence in distribution, the result in (5.28) holds provided that the limiting random variable $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$ has no point mass at 0. However, we rule out such pathological case that is in practice the exception rather than the rule.

characterization is in general not available. In the next sections we tackle this challenging problem by focusing on an asymptotic characterization of $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$ in the regime of small δ.

5.6 Small-δ Analysis

We have ascertained that there exist *steady-state* random variables characterizing the log-belief ratios. Then, the steady-state learning performance can be determined by examining the probability that these random variables fulfill certain conditions. For example, the steady-state probability that an agent learns the truth is the probability that the steady-state log-belief ratio of that agent is positive only at the true value θ_0. In general, the exact characterization of these steady-state variables is a formidable task. For this reason, we resort to an asymptotic analysis in the regime of small δ. We will provide three types of asymptotic results.

- **Section 5.6.1: Weak law of small step-sizes (Theorem 5.2)**. We show that, for small δ, the steady-state vector $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$ concentrates around the weighted average of the agents' KL divergences defined in (5.29). This concentration property guarantees that, with high probability as $\delta \to 0$, the true hypothesis is chosen by each agent. This result requires only finiteness of the first moments of the log-likelihood ratios, i.e., finiteness of the KL divergences.
- **Section 5.6.2: Asymptotic normality (Theorem 5.3)**. We obtain a Central Limit Theorem (CLT) that provides a normal approximation, holding for small δ, for the error probabilities of each individual agent. This result is proven assuming independence across agents and requires finiteness of the variance of the log-likelihood ratios. We remark that previous results of asymptotic normality for adaptive distributed detection assumed finiteness of higher-order moments [79]. To the best of our knowledge, the result in Theorem 5.3 (which is based on part 5 of Lemma 5.1) is the first result that assumes the minimal requirement of finiteness of second moments.
- **Section 5.6.3: Large deviations analysis (Theorem 5.4)**. We characterize the exponential rate of decay of the error probabilities as $\delta \to 0$. This result is proven assuming independence across agents and requires the existence of the moment generating function of the log-likelihood ratios.

Notably, the above three steps reflect perfectly a classic path in asymptotic statistics. However, in order to avoid misunderstandings, it is necessary to clarify one fundamental difference between the small-δ analysis and classic results. In order to illustrate this difference let us refer, for example, to the CLT result. In the classic setting of asymptotic statistics, one examines the asymptotic behavior of sums of random variables when the number of terms of the sum goes to infinity. In contrast, the CLT proved in this work does *not* affirm that the sums involved in (5.19) converge to a Gaussian as $i \to \infty$. As a matter of fact, we have shown in Theorem 5.1 that the sums in (5.19) converge to certain random variables, but these variables are *not Gaussian*, in general. The CLT that we prove deals instead with the behavior, as δ goes to zero, of the steady-state random vector $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$. The same distinction applies to the other two types of asymptotic results, namely, the weak law and the large deviations analysis. For this reason, as explained

in [78], the correct way to deal with the asymptotic regime of small step-sizes in the adaptation context is made of two steps:

- First, introduce a proper steady-state vector $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$, which already embodies the effect of combining an infinite number of summands. This steady-state vector will be non-degenerate (i.e., no weak law as $i \to \infty$), will be non-Gaussian (i.e., no CLT as $i \to \infty$), and will be non-vanishing (i.e., no large deviations as $i \to \infty$).
- Then, characterize the asymptotic behavior of the steady-state random vector $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$ as δ goes to zero.

It is worth noticing that, in the adaptation literature, the critical role of the first step is usually not emphasized. This is because the adaptation literature mostly focuses on *estimation* problems, where one usually quantifies the performance by evaluating convergence of the *moments* [37]. In contrast, when dealing with *decision* problems (as in our case), the performance is quantified through *probabilities*, namely, the probabilities of making a wrong (or correct) decision. In order to evaluate probabilities at the steady state, it is critical to obtain first a representation of the steady-state random variables [78].

5.6.1 Consistent Social Learning

Before stating the consistency result, we introduce the expectation of the average log-likelihood ratio in (5.14):

$$\mathsf{m}_{\text{ave}}(\theta) \triangleq \mathbb{E}\left(\boldsymbol{x}_{\text{ave},i}(\theta)\right) = \sum_{\ell=1}^{K} \pi_\ell d_\ell(\theta), \tag{5.29}$$

which does not depend on i owing to the identical distribution over time implied by the steady-state analysis. We will rely on the assumptions presented in Chapter 2 for the classical social learning setting.

Theorem 5.2 (Consistency of ASL). *Under Assumptions 2.4 and 2.2, we have the following convergence:*

$$\widetilde{\boldsymbol{\lambda}}_k^{(\delta)} \xrightarrow[\delta \to 0]{\text{p}} \mathsf{m}_{\text{ave}} \tag{5.30}$$

Since under Assumption 2.5 all entries of m_{ave} *are strictly positive, Eq.* (5.30) *implies that each agent learns correctly the true hypothesis as* $\delta \to 0$*, namely, for all* $\theta \neq \theta_0$ *we have that the steady-state error probability of all agents* $k = 1, 2, \ldots, K$ *converges to zero as* δ *approaches zero:*

$$\lim_{\delta \to 0} p_k^{(\delta)} = 0. \tag{5.31}$$

Proof. See Appendix 5.C. □

The result of Theorem 5.2 relies on the weak law of small step-sizes proved in Lemma 5.1, part 3. Technically, this law requires finiteness of only the first moments $d_\ell(\theta)$, which is guaranteed by

Assumption 2.4. Moreover, the result of Theorem 5.2 requires that $\mathsf{m}_{\rm ave}(\theta) > 0$ for all $\theta \neq \theta_0$. Since the entries of the Perron eigenvector are all strictly positive, we see that $\mathsf{m}_{\rm ave}(\theta)$ is strictly greater than zero for every θ if, for every θ, there exists at least one agent ℓ for which the KL divergence $d_\ell(\theta)$ is strictly positive. In other words, in order to achieve consistent learning, it is sufficient that at least one of the first moments (i.e., the KL divergence) is nonzero, which is guaranteed by Assumption 2.5. As already mentioned in previous chapters, although individual agents might not be able to learn properly on their own, under a *global* identifiability condition, agents are encouraged to collaborate since the network possesses sufficient information to learn the true hypothesis.

Theorem 5.2 establishes that the error probability vanishes as $\delta \to 0$. On the other hand, it does not establish *how* it vanishes. We will see that the ASL strategy is characterized by an exponential law, since the error probability of each individual agent decays exponentially fast as a function of the inverse step-size $1/\delta$.

5.6.2 Normal Approximation for Small δ

We will now prove a central limit theorem for the steady-state random vector $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$. To this end, we will assume finiteness of second-order moments for the log-likelihoods. We furthermore assume statistical independence across the agents.

In order to state the CLT, it is convenient to define some useful quantities. First, we introduce the covariance between the log-likelihood ratios at θ and θ', that is:

$$\rho_\ell(\theta,\theta') = \mathbb{E}\left[\Big(\boldsymbol{x}_{\ell,i}(\theta) - d_\ell(\theta)\Big)\Big(\boldsymbol{x}_{\ell,i}(\theta') - d_\ell(\theta')\Big)\right]. \tag{5.32}$$

Then we introduce the covariance between the average variables $\boldsymbol{x}_{{\rm ave},i}(\theta)$ and $\boldsymbol{x}_{{\rm ave},i}(\theta')$ which, exploiting independence across agents, can be evaluated as:

$$\mathsf{c}_{\rm ave}(\theta,\theta') \triangleq \sum_{\ell=1}^{K} \pi_\ell^2 \rho_\ell(\theta,\theta'). \tag{5.33}$$

Next, it is necessary to examine the behavior of the first two moments of the log-belief ratios. In view of Lemma 5.1, part 2, it is possible to conclude that the expectation of the steady-state random vector $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$ can be expressed as:

$$\mathsf{m}_k^{(\delta)}(\theta) \triangleq \mathbb{E}\left(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta)\right) = \mathsf{m}_{\rm ave}(\theta) + O(\delta), \tag{5.34}$$

where $O(\delta)$ is a quantity such that the ratio $O(\delta)/\delta$ remains bounded as $\delta \to 0$. Likewise, using part 4 of Lemma 5.1, we conclude that the covariance of the steady-state random vector $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$ is:

$$\begin{aligned} c_k^{(\delta)}(\theta,\theta') &\triangleq \mathbb{E}\left[\Big(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) - \mathsf{m}_k^{(\delta)}(\theta)\Big)\Big(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta') - \mathsf{m}_k^{(\delta)}(\theta')\Big)\right] \\ &= \frac{\mathsf{c}_{\rm ave}(\theta,\theta')}{2}\,\delta + O(\delta^2). \end{aligned} \tag{5.35}$$

Equations (5.34) and (5.35) can be rewritten in vector and matrix form, respectively as:

$$\mathsf{m}_k^{(\delta)} = \mathsf{m}_{\text{ave}} + O(\delta), \quad \mathsf{C}_k^{(\delta)} = \frac{\mathsf{C}_{\text{ave}}}{2}\,\delta + O(\delta^2), \tag{5.36}$$

where $\mathsf{C}_k^{(\delta)} = [c_k^{(\delta)}(\theta, \theta')]$ and $\mathsf{C}_{\text{ave}} = [\mathsf{c}_{\text{ave}}(\theta, \theta')]$ are the matrices that collect the individual covariances. We see from (5.36) that, as $\delta \to 0$, there is a leading term that does not depend on the agent index k (whose impact is implicitly included in the higher order corrections, i.e., the $O(\cdot)$ terms).

The first relation in (5.36) reveals that the expectation vector of the steady-state log-belief ratios, $\mathsf{m}_k^{(\delta)}$, approximates, for small δ, the expectation vector of the *average* log-likelihood ratios, m_{ave}. In comparison, the second relation in (5.36) reveals that the covariance matrix of the steady-state log-belief ratios, $\mathsf{C}_k^{(\delta)}$, goes to zero as $\mathsf{C}_{\text{ave}}\,\delta/2$, where C_{ave} is the covariance matrix of the *average* log-likelihood ratios, namely,

$$\lim_{\delta \to 0} \frac{2\mathsf{C}_k^{(\delta)}}{\delta} = \mathsf{C}_{\text{ave}}. \tag{5.37}$$

We are now ready to state our central limit theorem.

Theorem 5.3 (Asymptotic normality). *Assume that the data $\{\boldsymbol{\xi}_{k,i}\}$ are independent across the agents (recall that they are always assumed i.i.d. over time), and that the log-likelihood ratios have finite variance. Then, under Assumptions 2.4, 2.2 and 2.5, the following convergence holds:*

$$\frac{\widetilde{\boldsymbol{\lambda}}_k^{(\delta)} - \mathsf{m}_{\text{ave}}}{\sqrt{\delta}} \xrightarrow[\delta \to 0]{\text{d}} \mathscr{G}\left(0, \frac{\mathsf{C}_{\text{ave}}}{2}\right), \tag{5.38}$$

where $\mathscr{G}(0, C)$ is a zero-mean multivariate Gaussian with covariance matrix equal to C.

Proof. See Appendix 5.D. □

Theorem 5.3 entails the following approximation, holding for $\delta \approx 0$:

$$\widetilde{\boldsymbol{\lambda}}_k^{(\delta)} \approx \mathscr{G}\left(\mathsf{m}_{\text{ave}}, \frac{\mathsf{C}_{\text{ave}}}{2}\,\delta\right). \tag{5.39}$$

We see that such approximation does *not* depend on the agent index k. As shown in [78], in order to capture differences in performance across the agents, it is possible to replace the limiting expectation vector m_{ave} and the limiting covariance matrix $\mathsf{C}_{\text{ave}}\,\delta/2$ with their exact counterparts, i.e., with the series appearing in (5.34) and (5.35), yielding the refined approximation:

$$\widetilde{\boldsymbol{\lambda}}_k^{(\delta)} \approx \mathscr{G}\left(\mathsf{m}_k^{(\delta)}, \mathsf{C}_k^{(\delta)}\right). \tag{5.40}$$

The approximations in (5.39) and (5.40) will be tested in the section devoted to numerical experiments.

5.6.3 Large Deviations for Small δ

In this section we focus on another relevant type of asymptotic analysis, namely, a *large deviations* analysis [80], [81]. The application of large deviations to adaptive networks was used in [78], [79], [82].

The basic aim of the LD analysis is to estimate the exponential decay rate of the probabilities associated to certain *rare* events. In our setting, the rare event is the probability that an agent opts for the wrong hypothesis. We will show that, at the steady state, this type of event becomes in fact rare as δ approaches zero.

More formally, the LD analysis will furnish the following type of representation for the steady-state error probability [80], [81]:

$$p_k^{(\delta)} \doteq e^{-\Phi/\delta}, \tag{5.41}$$

where the notation $\doteq$ means equality to the leading exponential order (as $\delta \to 0$) or, more explicitly:

$$\lim_{\delta \to 0} \delta \log p_k^{(\delta)} = -\Phi, \tag{5.42}$$

for a certain value Φ that is called the *error exponent*. Notably, in the exponent Φ we did not put any dependence on the agent index k. This is because, as shown in Theorem 5.4 further ahead, *all agents will exhibit the same error exponent.*

On the other hand, it should be remarked that the equality at the leading exponential order in (5.41) does *not* imply that we can approximate the probability of error as $e^{-\Phi/\delta}$, namely,

$$p_k^{(\delta)} \not\approx e^{-\Phi/\delta}. \tag{5.43}$$

This is because any LD analysis neglects sub-exponential corrections. For example, it is immediate to check that the probabilities $e^{-\Phi/\delta}$ and $100\, e^{-\Phi/\delta}$ have the same LD exponent (equal to Φ), but the second probability is two orders of magnitude larger. These sub-exponential corrections embody higher-order differences in the error probabilities (see, e.g., Figure 5.2) that can arise across the agents due to different factors, for example, due to differences between very central agents, with a high number of neighbors, as opposed to peripheral agents, with few neighbors. To compensate for sub-exponential corrections, a refined LD framework exists, referred to as "exact asymptotics", which has been applied to binary adaptive detection [78], [82].

In summary, the aim of a large deviations analysis is to evaluate the asymptotic decay rate of the error probabilities, which is a meaningful and significant index of the inferential performance. Before stating the main result about the LD analysis, it is necessary to introduce the Logarithmic Moment Generating Function (LMGF), a.k.a. cumulant generating function, of the log-likelihood ratios:

$$\Lambda_k(t;\theta) = \log \mathbb{E}\left(e^{t\, \boldsymbol{x}_{k,i}(\theta)}\right). \tag{5.44}$$

We recall that, in the steady-state regime, the expectation is computed under the true model $L_k(\xi|\theta_0)$, which does not change over time, and this explains why $\Lambda_k(t;\theta)$ does not depend on i. It is also useful to introduce the LMGF of the average variable $\boldsymbol{x}_{\mathrm{ave},i}(\theta)$ which, under the

assumption that the data are independent across the agents, is:

$$\Lambda_{\rm ave}(t;\theta) = \log \mathbb{E}\left(e^{t\,\boldsymbol{x}_{{\rm ave},i}(\theta)}\right) = \sum_{\ell=1}^{K} \Lambda_\ell(\pi_\ell t;\theta). \tag{5.45}$$

Theorem 5.4 (Error exponents). *Assume that the data $\{\boldsymbol{\xi}_{k,i}\}$ are independent across the agents (recall that they are always assumed i.i.d. over time), and that the logarithmic moment generating function of $\boldsymbol{x}_{k,i}(\theta)$ exists everywhere, namely, for all $k = 1, 2, \ldots, K$ and $\theta \neq \theta_0$:*

$$\Lambda_k(t;\theta) < +\infty \;\; \forall t \in \mathbb{R}. \tag{5.46}$$

Let

$$\phi(t;\theta) = \int_0^t \frac{\Lambda_{\rm ave}(\tau;\theta)}{\tau} d\tau. \tag{5.47}$$

Then, under Assumptions 2.4, 2.2 and 2.5 we have the following two results holding for every agent $k = 1, 2, \ldots, K$. First, we have that:

$$\mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) \leq 0\right) \dot{=} e^{-\Phi(\theta)/\delta}, \quad \Phi(\theta) = -\inf_{t\in\mathbb{R}} \phi(t;\theta). \tag{5.48}$$

Second, the error probability is dominated by the worst-case (i.e., smaller) exponent:

$$p_k^{(\delta)} \dot{=} e^{-\Phi/\delta}, \quad \Phi = \min_{\theta\neq\theta_0} \Phi(\theta). \tag{5.49}$$

Proof. See Appendix 5.E □

The main message conveyed by Theorem 5.4 is that the steady-state error probability of each individual agent converges to zero as $\delta \to 0$, exponentially fast as a function of $1/\delta$. This exponential law provides a *universal* law for adaptive social learning, which reflects the universal scaling law of distributed adaptive detection—see [78]. The exponent Φ governing such an exponential decay is computed from the logarithmic moment generating function of the average log-likelihood, where the weights of this average are the limiting weights, i.e., the entries of the Perron eigenvector.

The need for cooperation has been already motivated in relation to social learning problems that are locally non-identifiable. Theorem 5.4 implies another potential benefit of cooperation, namely, that *cooperation improves the learning accuracy*. We will illustrate this aspect through one example. Assume the most favorable case where all agents could learn the true hypothesis individually. Consider then a doubly-stochastic combination matrix, yielding a Perron eigenvector with uniform entries $\pi_\ell = 1/K$ for all $\ell = 1, 2, \ldots, K$. Exploiting (5.49), we can easily see that in this particular case the error exponent of the network is given by:

$$\Phi = K\Phi_{\rm ind}, \tag{5.50}$$

where $\Phi_{\rm ind}$ is the error exponent of an individual agent. According to (5.50), we see that the network error exponent is K times larger than the individual error exponent, which in turn

implies an K-fold exponential improvement in the learning accuracy. Intuitively, a network of K agents observes K times as much data as a single agent at each time instant. The strong-connectivity of the network allows for the data to fully propagate across agents and yields the aforementioned learning performance improvement.

5.7 Transient Analysis

5.7.1 Qualitative Description of the Transient Phase

Preliminarily, we deem it is useful to provide a qualitative overview of the transient behavior of adaptive social learning in comparison to classic social learning. To this end, we consider initially a simple example consisting of a single-agent (indices k and ℓ dropped) binary ($\Theta = \{1,\, 2\}$) problem, with symmetric KL divergences:

$$\mathbb{E}_1 \left(\log \frac{L(\boldsymbol{\xi}_i|1)}{L(\boldsymbol{\xi}_i|2)} \right) = -\mathbb{E}_2 \left(\log \frac{L(\boldsymbol{\xi}_i|1)}{L(\boldsymbol{\xi}_i|2)} \right) \triangleq x > 0, \tag{5.51}$$

where $\mathbb{E}_\theta$ denotes expectation under the distribution $L(\xi|\theta)$. We assume that at time $i = 1$, the true underlying hypothesis is $\theta_0 = 1$, and the situation remains stationary until a certain time T_1, after which data start being generated according to $\theta_0 = 2$, and that is why a transient analysis is necessary to see how the learning algorithm is able to track this drift.

To examine how the learning process progresses over time, it is sufficient to consider the time-evolution of the log-belief ratio:

$$\boldsymbol{r}_i \triangleq \log \frac{\boldsymbol{\mu}_i(1)}{\boldsymbol{\mu}_i(2)}, \tag{5.52}$$

whose positive (resp., negative) values will let the agent opt for $\theta = 1$ (resp., $\theta = 2$). Specializing (2.2) and (2.3) to the single-agent binary setting, classic social learning evolves according to the recursion (we add a superscript to distinguish classic from adaptive social learning):

$$\boldsymbol{r}_i^{\mathsf{SL}} = \boldsymbol{r}_{i-1}^{\mathsf{SL}} + \log \frac{L(\boldsymbol{\xi}_i|1)}{L(\boldsymbol{\xi}_i|2)}, \qquad \boldsymbol{r}_0^{\mathsf{SL}} = 0. \tag{5.53}$$

Likewise, replacing (2.2) with (5.1), the adaptive social learning strategy in this single-agent binary case evolves according to the recursion:

$$\boldsymbol{r}_i^{\mathsf{ASL}} = (1-\delta)\boldsymbol{r}_{i-1}^{\mathsf{ASL}} + \delta \log \frac{L(\boldsymbol{\xi}_i|1)}{L(\boldsymbol{\xi}_i|2)}, \qquad \boldsymbol{r}_0^{\mathsf{ASL}} = 0. \tag{5.54}$$

In both (5.53) and (5.54), we assume flat initial priors (i.e., $\boldsymbol{r}_0^{\mathsf{SL}} = \boldsymbol{r}_0^{\mathsf{ASL}} = 0$). To get a flavor of the main trade-offs involved in the transient behavior, let us focus on the time-evolution of the *expected values*. Taking expectations in (5.53), at time T_1 we have:

$$\mathbb{E}\left(\boldsymbol{r}_{\mathsf{T}_1}^{\mathsf{SL}}\right) = \mathsf{T}_1 x, \tag{5.55}$$

where x is the symmetric KL divergence introduced in (5.51). Equation (5.55) shows that the

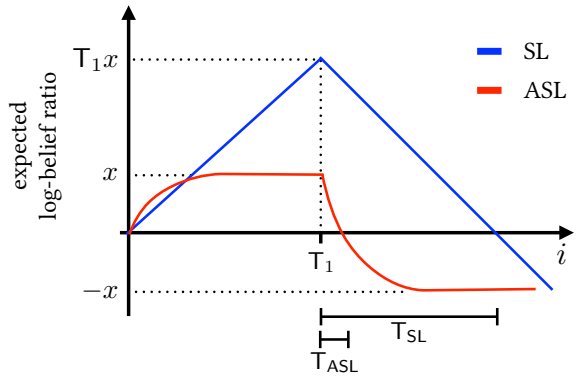


Figure 5.4: Diagram of the time evolution of the log-belief ratio in expectation for the classic social learning strategy (in blue) and for the ASL strategy (in red) within the single-agent case.

expected value of the log-belief ratio grows linearly with the stationarity interval T_1. This linear growth is a reflection of the increasing knowledge acquired by the agent as it aggregates new information represented by the log-likelihood ratio $\log \frac{L(\boldsymbol{\xi}_i|1)}{L(\boldsymbol{\xi}_i|2)}$. In a virtual asymptotic regime, this knowledge becomes a certainty, i.e., as $\mathsf{T}_1 \longrightarrow +\infty$, $\boldsymbol{r}_{\mathsf{T}_1}^{\mathsf{SL}} \to +\infty$, which implies that if hypothesis 1 remains in force indefinitely, the belief of the agent regarding this hypothesis achieves full confidence. Unfortunately, this increasing confidence comes at the cost of a slow adaptation regime. Indeed, since from time $\mathsf{T}_1 + 1$ the true hypothesis is $\theta_0 = 2$, from (5.53) and (5.55) we have that:

$$\mathbb{E}\left(\boldsymbol{r}_i^{\mathsf{SL}}\right) = \mathbb{E}\left(\boldsymbol{r}_{\mathsf{T}_1}^{\mathsf{SL}}\right) - ix = (\mathsf{T}_1 - i)x. \tag{5.56}$$

Now, the adaptation time can be roughly identified by considering the time necessary to overcome the initial bias towards hypothesis 1 once the true hypothesis switches from 1 to 2. In terms of our qualitative mean-value analysis, this is the time necessary for the expected log-belief ratio to change from positive to negative, which, in view of (5.56) implies that the adaptation time for the classic social learning strategy is on the order of:

$$\mathsf{T}_{\mathsf{SL}} = \mathsf{T}_1. \tag{5.57}$$

This behavior is clearly not admissible for an adaptive algorithm, since it implies that the time necessary to recover from a wrong opinion is proportional to the stationarity interval where this opinion was actually true.

Let us switch to the adaptive strategy. Developing the recursion until time T_1, from (5.54) we get, respectively:

$$\mathbb{E}\left(\boldsymbol{r}_{\mathsf{T}_1}^{\mathsf{ASL}}\right) = \delta \sum_{m=0}^{i-1} (1-\delta)^m x = \left(1 - (1-\delta)^{\mathsf{T}_1}\right) x \approx x, \tag{5.58}$$

where the approximation is motivated from assuming a sufficiently large T_1. Considering then that from time $\mathsf{T}_1 + 1$ onward the true hypothesis is $\theta_0 = 2$, Eqs. (5.54) and (5.58) yield, for any

$i > \mathsf{T}_1$:

$$\begin{aligned}\mathbb{E}\left(\boldsymbol{r}_i^{\mathsf{ASL}}\right) &= (1-\delta)^i \mathbb{E}\left(\boldsymbol{r}_{\mathsf{T}_1}^{\mathsf{ASL}}\right) - \delta \sum_{m=0}^{i-1} (1-\delta)^m x \\ &\approx -\left(1 - 2(1-\delta)^i\right) x. \end{aligned} \tag{5.59}$$

Now, equating (5.59) to zero to evaluate the adaptation time, we obtain:

$$\mathsf{T}_{\mathsf{ASL}} = \frac{\log 2}{\log(1-\delta)^{-1}} \approx \frac{\log 2}{\delta}. \tag{5.60}$$

A visual comparison of the enhanced adaptation provided by the ASL strategy is exemplified in Figure 5.4.

Comparing (5.60) against (5.57), we see that, in contrast to the undesirable behavior exhibited by classic social learning, the adaptive formulation exhibits a controlled initial bias. This is because, after a relatively long stationarity interval T_1, the expected log-belief is concentrated around a fixed value x, and the adaptation time will then increase roughly as $1/\delta$. In a nutshell, while the reaction capacity of classic social learning is not controlled by design and is severely affected by the duration of previous stationarity intervals, in adaptive social learning the adaptation time is not affected by previous stationarity intervals and is controlled through the step-size. This enhanced adaptation comes at the expense of learning accuracy. In fact, as we have established in the previous sections, the steady-state error probability does not converge to zero as time elapses, but converges to some stable value. However, this value vanishes exponentially fast as a function of $1/\delta$, highlighting the fundamental trade-off of adaptive social learning: The smaller the step-size δ, the smaller the error probability and the slower the adaptation.

In the theory of adaptation and learning, the transient analysis is typically performed by characterizing the evolution of suitable higher-order moments, such as second or fourth order moments of the pertinent statistics [37]. However, this analysis is more appropriate for estimation/regression problems where the focus of the transient analysis is to ascertain how long it takes for the pertinent system state to attain a prescribed neighborhood of the expected value. In our social learning setting, it is more appropriate to identify an adaptation time in terms of *error probabilities*. As established in Theorem 5.4, the behavior of these probabilities is governed by the logarithmic moment generating function of the observations which, as the name itself suggests, incorporates dependence upon *all moments*. Accordingly, a meaningful way to perform the transient analysis is to examine the time-evolution of logarithmic moment generating functions, rather than individual moments. This characterization constitutes the core of Theorem 5.5, which is introduced in the next section.

5.7.2 Quantitative Description of the Transient Phase

In this section, we provide a rigorous analysis to support the qualitative description of the transient behavior, seen in Section 5.7.1. We assume that the ASL strategy has been in operation for a certain arbitrary time i_0. All the knowledge accumulated by the agents until this time is summarized in the belief vector $\boldsymbol{\mu}_{i_0}$. We remark that the evolution of the statistical models

from $i = 0$ to $i = i_0$ is left completely arbitrary, that is, the system could have experienced several drifts in the statistical conditions, including change of the underlying hypotheses, data generated according to models that do not match the assumed likelihoods, and so on. From the ASL algorithm viewpoint, all these effects are summarized in the belief vector $\boldsymbol{\mu}_{i_0}$ that acts as initial state at time i_0. In order to perform the transient analysis, we assume that from $i_0 + 1$ onward, the true hypothesis is steadily equal to θ_0, and will establish how much time is necessary to stay sufficiently close to the steady-state learning performance starting from a given (arbitrary) realization μ_{i_0}. As done before, to simplify the notation we set $i_0 = 0$ and the initial state becomes μ_0.

In a social learning problem the adaptation time should be properly related to the time-evolution of the error probability, and particularly to the time necessary for the *instantaneous* error probability to approach the *steady-state* error probability. Accordingly, in the next theorem we start by providing an upper bound on the *instantaneous* error probability introduced in (5.10).

Theorem 5.5 (Bounds on the instantaneous error probability). *The claim of the theorem holds under the same assumptions of Theorem 5.4. Let κ and β be the constants defined in Property 2.1, and let $t_\theta^\star < 0$ be the unique solution to the equation:*

$$\frac{\Lambda_{\text{ave}}(t_\theta^\star;\theta)}{t_\theta^\star} = 0. \tag{5.61}$$

Let

$$\lambda_{\text{ave},0}(\theta) = \sum_{\ell=1}^{K} \pi_\ell \lambda_{\ell,0}(\theta) \tag{5.62}$$

be the network average of the initial log-belief ratios $\lambda_{\ell,0}(\theta)$, and let, for all $\theta \neq \theta_0$:

$$\mathsf{K}_1(\theta) \triangleq |t_\theta^\star| \left(\mathsf{m}_{\text{ave}}(\theta) - \lambda_{\text{ave},0}(\theta) \right), \tag{5.63}$$

$$\mathsf{K}_2(\theta) \triangleq \kappa |t_\theta^\star| \sum_{\ell=1}^{K} |\lambda_{\ell,0}(\theta)|. \tag{5.64}$$

Then, the instantaneous error probability $p_{k,i}^{(\delta)}$ is upper bounded as:

$$p_{k,i}^{(\delta)} \leq \sum_{\theta \neq \theta_0} e^{\frac{1}{\delta}\left(-\Phi(\theta) + \mathsf{K}_1(\theta)(1-\delta)^i + \mathsf{K}_2(\theta)(1-\delta)^i \beta^i + O(\delta)\right)}, \tag{5.65}$$

where the notation $O(\delta)$ signifies that the ratio $O(\delta)/\delta$ stays bounded as $\delta \to 0$.

Proof. See Appendix 5.F. □

Theorem 5.5 reveals the main behavior of the transient error probability. Examining the error exponent of the upper bound in (5.65) we see, up to higher-order small-δ corrections embodied in the term $O(\delta)$, the emergence of three terms: the *steady-state* error exponent $\Phi(\theta)$ already identified in Theorem 5.4, and two other terms that characterize the *transient* behavior. The

first transient term decays as $(1-\delta)^i$, and is thus influenced solely by the step-size. The second transient term, $(1-\delta)^i\beta^i$, decays faster and is influenced also by the parameter β. This parameter, according to Property 2.1, is determined by the second largest-magnitude eigenvalue of A, and accordingly determines the mixing properties of A (i.e., the convergence rate of $[A^i]_{\ell k}$ to the Perron eigenvector entry π_ℓ). Therefore, the second transient term, with rate $(1-\delta)^i\beta^i$, determines a transient phenomenon that is related to the convergence of the matrix-powers to a "centralized" solution with combination weights π_ℓ. In comparison, the first term, with rate $(1-\delta)^i$, determines a transient phenomenon ruled by the step-size only.

In summary, Theorem 5.5 provides an upper bound on the instantaneous error probability that converges, as $i \to \infty$, to a sum of exponential terms with steady-state error exponent $\Phi = \min_{\theta\neq\theta_0} \Phi(\theta)$. Accordingly, we identify as a meaningful definition for the *adaptation time* the critical time instant after which the error probability decays with an error exponent $(1-\epsilon)\Phi$, for some small ϵ. This is made precise in the following corollary.

Corollary 5.1 (Adaptation time). *Under the same notation and assumptions of Theorem 5.5, let*

$$\begin{aligned}
\mathsf{K}_1 &\triangleq \max_{\theta\neq\theta_0} \mathsf{K}_1(\theta) = \max_{\theta\neq\theta_0} \left\{ |t_\theta^\star| \left[\mathsf{m}_{\rm ave}(\theta) - \lambda_{{\rm ave},0}(\theta) \right] \right\}, \\
\mathsf{K}_2 &\triangleq \max_{\theta\neq\theta_0} \mathsf{K}_2(\theta) = \kappa \max_{\theta\neq\theta_0} \left\{ |t_\theta^\star| \sum_{\ell=1}^{K} |\lambda_{\ell,0}(\theta)| \right\}.
\end{aligned} \tag{5.66}$$

Then, the upper bound:

$$p_{k,i}^{(\delta)} \leq e^{-\frac{1}{\delta}[(1-\epsilon)\Phi + O(\delta)]} \tag{5.67}$$

holds for all $i > \mathsf{T}_{\mathsf{ASL}}$, where $\mathsf{T}_{\mathsf{ASL}}$ is given by the following rules:

i) (Favorable case, all initial states are good).

If $\lambda_{{\rm ave},0}(\theta) \geq \mathsf{m}_{\rm ave}(\theta)$ for all $\theta \neq \theta_0$:

$$\mathsf{T}_{\mathsf{ASL}} = \frac{1}{\log \beta^{-1}} \log \frac{\mathsf{K}_2}{\epsilon\,\Phi}, \qquad \epsilon < \frac{\mathsf{K}_2}{\Phi}. \tag{5.68}$$

ii) (Unfavorable case, at least one initial state is bad).

If $\lambda_{{\rm ave},0}(\theta) < \mathsf{m}_{\rm ave}(\theta)$ for at least one $\theta \neq \theta_0$:

$$\mathsf{T}_{\mathsf{ASL}} = \frac{1}{\log(1-\delta)^{-1}} \log \frac{\mathsf{K}_1}{\epsilon\,\Phi}, \qquad \epsilon < \frac{\mathsf{K}_1}{\Phi}. \tag{5.69}$$

Proof. See Appendix 5.G. □

Let us now examine the main parameters and phenomena affecting the adaptation time $\mathsf{T}_{\mathsf{ASL}}$.

- **Memory**: The memory coming from the past algorithm evolution is summarized in the starting belief vector μ_0, which in turn determines the average log-belief $\lambda_{\mathrm{ave},0}(\theta)$.

  First of all, we notice that an average initial state $\lambda_{\mathrm{ave},0}(\theta)$ greater than $\mathsf{m}_{\mathrm{ave}}(\theta)$ creates already a (favorable) bias toward the true hypothesis. Accordingly, when $\lambda_{\mathrm{ave},0}(\theta) \geq \mathsf{m}_{\mathrm{ave}}(\theta)$ the transient term $\mathsf{K}_1(\theta)(1-\delta)^i$ reduces the error probability since $\mathsf{K}_1(\theta) < 0$. In this case, the dominant transient term is $(1-\delta)^i\beta^i$, and the corresponding adaptation time in (5.68) is essentially determined by the mixing parameter β, i.e., by how fast the combination weights converge to the Perron eigenvector. Under this regime, the adaptation time *does not depend critically on the step-size.*

  In comparison, the case where $\lambda_{\mathrm{ave},0}(\theta) < \mathsf{m}_{\mathrm{ave}}(\theta)$ is the unfavorable case where we are, as $\lambda_{\mathrm{ave},0}(\theta)$ decreases, progressively far from the steady-state. Under this regime, for small δ the dominant transient term is $\mathsf{K}_1(\theta)(1-\delta)^i$, and the adaptation time *scales with the step-size as* $1/\log(1-\delta)^{-1} \approx 1/\delta$.

  One particularly interesting case is when the average initial state is negative. This happens, for example, when the initial state comes from a previous learning cycle where the agent converged to a certain hypothesis that has then changed at the beginning of the subsequent learning cycle. In line with intuition, the adaptation time (5.69) increases with increasing size of the wrong starting conditions. Moreover, this dependence upon the past states is only logarithmic, which reveals that the past algorithm evolution has not a dramatic impact on the adaptation time.

- **KL divergences and error exponent**: By ignoring the initial state, Eq. (5.69) becomes:

$$\mathsf{T}_{\mathsf{ASL}} = \frac{1}{\log(1-\delta)^{-1}} \log \frac{\max_{\theta\neq\theta_0} \left[|t^\star_\theta| \mathsf{m}_{\mathrm{ave}}(\theta)\right]}{\epsilon\, \Phi}. \tag{5.70}$$

  From Property P2) in Lemma 5.2 (see Appendix 5.F), we know that:

$$\Phi(\theta) \leq |t^\star_\theta| \mathsf{m}_{\mathrm{ave}}(\theta), \tag{5.71}$$

  which shows that the ratio $\max_{\theta\neq\theta_0} \left[|t^\star_\theta| \mathsf{m}_{\mathrm{ave}}(\theta)\right] / \Phi$ appearing in (5.70) is greater than 1. Even if declaring a general behavior for this ratio for all statistical models is not obvious, we see that the numerator and the denominator are not independent. For example, having an "easier" detection problem where the KL divergences (numerator) increase typically corresponds to an increase of the error exponent (denominator) as well. However, in all cases the dependence on these parameters is not critical, since it is logarithmic.

- **Parameter** $t^\star_\theta$: First of all, to evaluate and interpret the bound on the adaptation time it is useful to remark that the term $|t^\star_\theta|$ is comprised between $1/\pi_{\max}$ and $1/\pi_{\min}$—see property P3) in Lemma 5.2. Apparently, these bounds introduce a dependence on the network parameters (i.e., on the Perron eigenvector). However, we should be careful here, and recall that the network error exponent Φ depends on the whole network as well. In order to get insights on this dependence, let us ignore the initial state and consider the case where all likelihoods are equal across agents and the combination matrix is doubly stochastic (yielding a uniform Perron eigenvector). Under these assumptions,

from property P3) in Lemma 5.2 we get $t_\theta^\star = -K$, and using (5.50) we obtain:

$$\mathsf{T}_{\mathsf{ASL}} = \frac{1}{\log(1-\delta)^{-1}} \log \frac{\max_{\theta \neq \theta_0} \left[\mathsf{m}_{\mathrm{ave}}(\theta) - \lambda_{\mathrm{ave},0}(\theta)\right]}{\epsilon\, \Phi_{\mathrm{ind}}}, \tag{5.72}$$

which shows how the network size appearing in the parameter $t_\theta^\star = -K$ is perfectly compensated by the network size embodied in the network exponent $\Phi = K\Phi_{\mathrm{ind}}$. Accordingly, we expect that the network parameters have a reduced impact on the transient time in (5.69), while, as observed before, the effect of the network is embodied in the parameter β controlling the higher-order transient term $(1-\delta)^i \beta^i$ in (5.65), which is neglected in the small-δ regime.

- **Parameter** ϵ: The smaller ϵ is, the closer the error exponent to the steady-state exponent Φ will be. Remarkably, the dependence is logarithmic in $1/\epsilon$, which means that this parameter is not critical.

- **Step-size**: Finally, in the (more interesting) case where the initial state is not good, see (5.69), the adaptation time scales as $1/\delta$. We remark that this behavior matches well the qualitative analysis of Section 5.7.1.

  The bottom line of Corollary 5.1 is that the adaptive capabilities of the ASL strategy are enhanced by a larger value of δ, by yielding a reduced adaptation time. A larger δ however is not always desirable, since it can reduce the accuracy in the decision-making process (as seen in Theorem 5.4, the steady-state probability of error is increased for larger δ). Both phenomena represent the trade-off adaptation vs. learning present in the ASL strategy and should be taken into account when designing δ. Such trade-off can be better summarized by combining Theorem 5.4 and Corollary 5.1, which shows that the error probability decays exponentially fast with the adaptation time, roughly as:

$$p_k^{(\delta)} \approx \exp\left\{ -\frac{\Phi}{\log[\mathsf{K}_1 \times (\epsilon\, \Phi)^{-1}]}\, \mathsf{T}_{\mathsf{ASL}} \right\}. \tag{5.73}$$

Stability over successive learning cycles: The characterization of the transient stage provided by Theorem 5.5 and the related corollary is valid under an arbitrary choice of the starting state $\lambda_{k,0}$. However, as we have commented in the previous section, if we start from a wrong state the level of this state affects adversely the adaptation time. Therefore, some fundamental questions arise. Assume that the time axis is divided into successive intervals (learning cycles) wherein the system evolves under stationary conditions. Then, the belief accumulated at the end of a learning cycle can be wrong in relation to the subsequent learning cycle. How "wrong" are the initial beliefs at the beginning of a learning cycle as the algorithm progresses? Do these initial states compromise the learning capability of the algorithm over successive cycles? These fundamental questions can be answered by combined steady-state and transient analyses. In fact, from the steady-state analysis carried out in the previous sections, we learned that the steady-state log-belief ratios fluctuate in a small neighborhood (of size $\sim \sqrt{\delta}$) of the expected values of the pertinent KL divergences. This means that at the end of each cycle the ASL strategy converges to some *stable* state, i.e., a state that does not diverge as the step-size δ becomes small. As a result, the initial states of each learning cycle would evolve in a stable manner and, hence, do not compromise the learning performance of the algorithm, provided that the

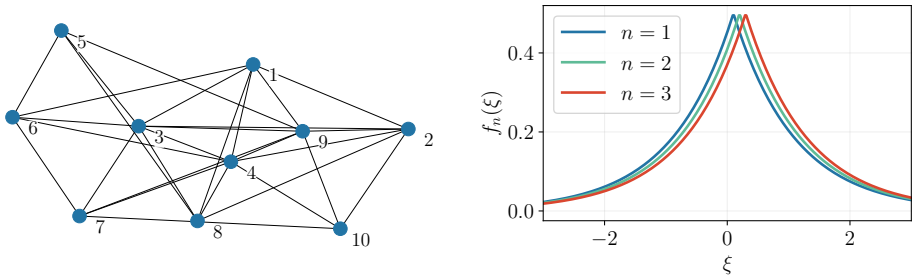


Figure 5.5: *(Left)* Strongly connected network topology with $K = 10$ agents. *(Right)* Family of Laplace likelihood functions.

adaptation time is smaller than the duration of the learning cycles. These aspects will be more quantitatively illustrated in Section 5.9, with reference to specific illustrative examples.

5.8 Illustrative Examples

We consider the strongly connected network of $K = 10$ agents displayed in the left panel of Figure 5.5, where all agents have a self-loop (not displayed in the figure). Besides, the combination matrix is designed using an averaging rule, resulting in a left-stochastic matrix [37].

The network is faced with the following statistical learning problem. We consider a family of Laplace likelihood functions with scale parameter 1, seen in the right panel of Figure 5.5. Formally, we are given three Laplace densities:

$$f_n(\xi) = \frac{1}{2} \exp\left\{-|\xi - 0.1n|\right\}, \tag{5.74}$$

for $n \in \{1, 2, 3\}$. The likelihoods of the data collected by the agents are chosen from among these Laplace densities.

To make things more interesting, we assume that the inference problem is *not locally identifiable*. The setup for each agent's family of likelihood functions can be seen in Table 5.1.

Table 5.1: Identifiability setup for the network in the left panel of Figure 5.5.

Agent k	**Likelihood function**: $L_k(\xi\|\theta)$		
	$\theta = 1$	$\theta = 2$	$\theta = 3$
$1 - 3$	$f_1(\xi)$	$f_1(\xi)$	$f_3(\xi)$
$4 - 6$	$f_1(\xi)$	$f_3(\xi)$	$f_3(\xi)$
$7 - 10$	$f_1(\xi)$	$f_2(\xi)$	$f_1(\xi)$

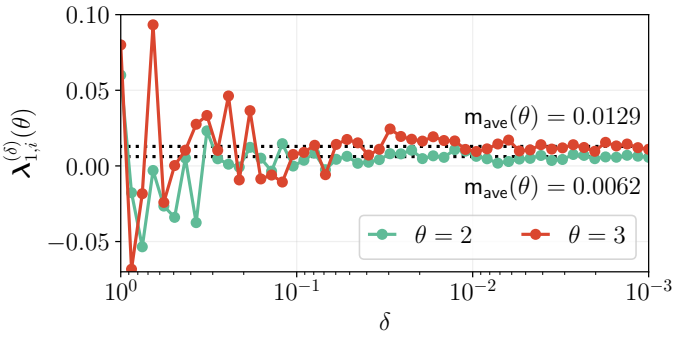


Figure 5.6: Consistency of the ASL strategy (Theorem 5.2). According to the weak-law of small step-sizes, the steady-state log-belief ratios for agent 1 concentrate around the predicted expectation values in $\mathsf{m}_{\mathrm{ave}}$ as δ approaches zero.

In summary, the data $\{\boldsymbol{\xi}_{k,i}\}$ are i.i.d. (across time and agents) Laplace random variables, with expectations that depend both on the agent k and the hypothesis θ. Accordingly, we will use the notation $e_k(\theta)$ to denote the expectation of $\boldsymbol{\xi}_{k,i}$, computed under likelihood $L_k(\xi|\theta)$. For example, using Table 5.1, we see that:

$$e_1(1) = 0.1, \quad e_4(3) = 0.3, \quad e_7(2) = 0.2. \tag{5.75}$$

We are now ready to delve into a detailed illustration of the numerical experiments. In particular, in this section we will test how the empirical performance matches the steady-state performance as characterized in Theorems 5.1–5.4. In order to examine the steady-state behavior *empirically*, we need that the ASL algorithm run for a sufficiently long period of time. In line with the prescriptions from Section 5.7, the duration of this this period is chosen as at least one order of magnitude larger than the inverse of the step-size, $1/\delta$.

5.8.1 Consistency

We consider that all agents are running the ASL algorithm for a fixed $\theta_0 = 1$ over 8000 time samples (after which we consider that they achieved the steady state). From Theorem 5.2, we saw that as δ approaches zero, all agents k are able to consistently learn—see (5.30). In order to show this effect, for each value of δ (50 sample points in the interval $\delta \in [0.001, 1]$ are taken), we consider a different realization of the observations. In Figure 5.6, for agent 1 and $\theta = 2, 3$, we show how the log-belief ratios $\boldsymbol{\lambda}_1^{(\delta)}(\theta)$ behave for decreasing values of δ. We see the weak-law of small step-sizes arising, since the limiting log-belief ratios tend to concentrate around $\mathsf{m}_{\mathsf{ave}}$.

5.8.2 Asymptotic Normality

We consider 10000 time samples, where again all agents are collecting data under a true hypothesis $\theta_0 = 1$. From Theorem 5.3, we saw that in steady state we can approximate the log-belief ratios distribution by a multivariate Gaussian pdf, see Eqs. (5.39) and (5.40). In Figure 5.7, we assume that the ASL algorithm has reached the steady state at $i = 10000$, and display the log-likelihood ratios corresponding to instant $i = 10000$. The experiment is repeated over

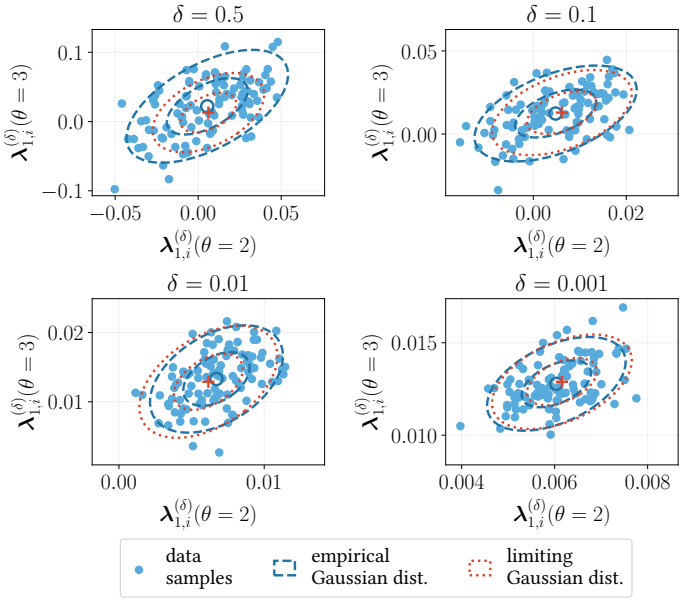


Figure 5.7: Distribution of data samples at steady state compared with the limiting and empirical Gaussian distributions.

100 Monte Carlo runs, such that we obtain 100 realizations of the steady-state variable $\boldsymbol{\lambda}_k^{(\delta)}$. Moreover, we consider 4 values of δ.

In dashed blue lines we see the ellipses representing the confidence intervals relative to one and two standard deviations computed for the empirical Gaussian approximation seen in (5.40): the smaller ellipse encompasses approximately 68% of the samples whereas the larger ellipse encompasses 95%. In red dotted lines, we see the corresponding ellipses for the limiting theoretical Gaussian approximation seen in (5.39), with the red cross indicating the limiting theoretical expectation $\mathsf{m}_{\mathsf{ave}}$. Note how as δ decreases, the ellipses tend to be smaller, which is in accordance with the scaling of the covariance matrices by δ in (5.39) and (5.40), and the distributions tend to overlap, which is in accordance with the behavior predicted by Theorem 5.3.

5.8.3 Error Exponents

We start by evaluating the theoretical exponents for the Laplace example at hand. To this aim, we need to compute first the logarithmic moment generating function of the log-likelihood ratios $\boldsymbol{x}_{k,i}(\theta)$ in (5.12). Since the data follow a Laplace distribution, the log-likelihood ratio is:

$$\boldsymbol{x}_{k,i}(\theta) = |\boldsymbol{\xi}_{k,i} - e_k(\theta)| - |\boldsymbol{\xi}_{k,i} - e_k(\theta_0)|. \tag{5.76}$$

Before we proceed to characterize the random variable $\boldsymbol{x}_{k,i}(\theta)$, let us define the auxiliary quantity:

$$\Delta_{k,\theta} \triangleq e_k(\theta) - e_k(\theta_0). \tag{5.77}$$

We also introduce the centered variable $\widetilde{\boldsymbol{\xi}}_{k,i} = \boldsymbol{\xi}_{k,i} - e_k(\theta_0)$, and therefore we can write:

$$\boldsymbol{x}_{k,i}(\theta) = |\widetilde{\boldsymbol{\xi}}_{k,i} - \Delta_{k,\theta}| - |\widetilde{\boldsymbol{\xi}}_{k,i}|. \tag{5.78}$$

For the case in which $\Delta_{k,\theta} > 0$, the random variable $\boldsymbol{x}_{k,i}(\theta)$ depends on the random variable $\widetilde{\boldsymbol{\xi}}_{k,i}$ in the following manner:

$$\boldsymbol{x}_{k,i}(\theta) = \begin{cases} -\Delta_{k,\theta}, & \text{if } \widetilde{\boldsymbol{\xi}}_{k,i} > \Delta_{k,\theta}, \\ \Delta_{k,\theta} - 2\widetilde{\boldsymbol{\xi}}_{k,i}, & \text{if } \widetilde{\boldsymbol{\xi}}_{k,i} \in [0, \Delta_{k,\theta}], \\ \Delta_{k,\theta}, & \text{if } \widetilde{\boldsymbol{\xi}}_{k,i} < 0. \end{cases} \tag{5.79}$$

We can then express the cumulative distribution function of $\boldsymbol{x}_{k,i}(\theta)$ as

$$\mathbb{P}[\boldsymbol{x}_{k,i}(\theta) \leq x] = \begin{cases} 0, & \text{if } x < -\Delta_{k,\theta}, \\ \mathbb{P}\left(\widetilde{\boldsymbol{\xi}}_{k,i} \geq \frac{\Delta_{k,\theta} - x}{2}\right), & \text{if } x \in [-\Delta_{k,\theta}, \Delta_{k,\theta}], \\ 1, & \text{if } x > \Delta_{k,\theta}, \end{cases} \tag{5.80}$$

where $\mathbb{P}[\mathcal{A}]$ is the probability of event $\mathcal{A}$, computed from the distribution of $\widetilde{\boldsymbol{\xi}}_{k,i}$. Note that its probability density function is given by $L_k(\xi + e_k(\theta_0)|\theta_0)$, which is a Laplace distribution with zero mean and scale parameter 1.

From the cumulative distribution function in (5.80), we can derive the density function of $\boldsymbol{x}_{k,i}(\theta)$ as:

$$\begin{aligned} p(x) &= \mathbb{P}\left(\widetilde{\boldsymbol{\xi}}_{k,i} > \Delta_{k,\theta}\right) \delta(x + \Delta_{k,\theta}) + \mathbb{P}\left(\widetilde{\boldsymbol{\xi}}_{k,i} < 0\right) \delta(x - \Delta_{k,\theta}) \\ &\quad + \frac{1}{2} L_k\left(\frac{\Delta_{k,\theta} - x}{2} + e_k(\theta_0) \middle| \theta_0\right) \operatorname{rect}\left(\frac{x}{2\Delta_{k,\theta}}\right), \\ &= \frac{1}{2} \exp\left[-\Delta_{k,\theta}\right] \delta(x + \Delta_{k,\theta}) + \frac{1}{2} \delta(x - \Delta_{k,\theta}) \\ &\quad + \frac{1}{4} \exp\left[-\frac{(\Delta_{k,\theta} - x)}{2}\right] \operatorname{rect}\left(\frac{x}{2\Delta_{k,\theta}}\right), \end{aligned} \tag{5.81}$$

where $\operatorname{rect}(\cdot)$ is the rectangle function, i.e., it is equal to 1 in the interval $]-\frac{1}{2}, \frac{1}{2}[$ and 0 elsewhere. Also we should distinguish the notation $\delta(x)$, which represents the Dirac delta-function, from the notation δ, which refers to the step-size parameter.

The LMGF of variable $\boldsymbol{x}_{k,i}(\theta)$, whose expression was seen in (5.44), can be explicitly computed using (5.81):

$$\begin{aligned} \Lambda_k(t; \theta) &= \log\left(\int_{\mathbb{R}} e^{tx} p(x) dx\right) \\ &= \log\left[\frac{1}{2} \exp(-\Delta_{k,\theta}(t+1)) + \frac{1}{2} \exp(\Delta_{k,\theta} t)\right. \\ &\quad \left. + \frac{1}{2} \exp\left(-\frac{\Delta_{k,\theta}}{2}\right) \frac{\sinh(\Delta_{k,\theta}(t + 1/2))}{t + 1/2}\right]. \end{aligned} \tag{5.82}$$

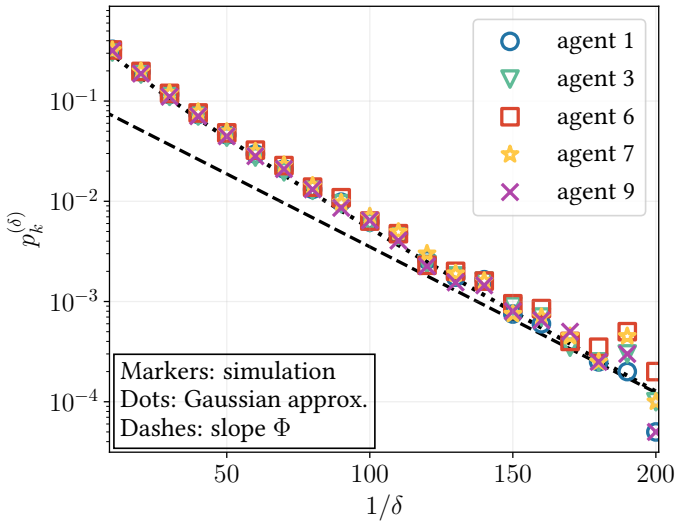


Figure 5.8: Steady-state error probability. Markers refer to the empirical probability curves estimated via Monte Carlo simulation. The dotted line refers to the theoretical error probability in (5.10) computed using the Gaussian approximation in (5.39). The slope of the probability curves is compared against the slope Φ (i.e., the error exponent) predicted by Theorem 5.4, and shown with dashed line.

If similar steps are followed for the case $\Delta_{k,\theta} < 0$, we would find the following expression for the LMGF:

$$\Lambda_k(t;\theta) = \log\left[\frac{1}{2}\exp(\Delta_{k,\theta}(t+1)) + \frac{1}{2}\exp(-\Delta_{k,\theta}t) - \frac{1}{2}\exp\left(\frac{\Delta_{k,\theta}}{2}\right)\frac{\sinh(\Delta_{k,\theta}(t+1/2))}{t+1/2}\right]. \tag{5.83}$$

Assuming that the true state is $\theta_0 = 1$, we can then evaluate numerically $\Phi(\theta)$ by employing the expressions in Theorem 5.4, for $\theta = 2$ and $\theta = 3$, from which we obtain $\Phi(2) = 0.03348$ and $\Phi(3) = 0.05051$. Finally, the error probability dominant exponent is given by:

$$\Phi = \min_{\theta\in\{2,3\}} \Phi(\theta) = 0.03348 \tag{5.84}$$

Now we illustrate the details of the numerical experiments. We consider that the true state of nature is set as $\theta_0 = 1$, and we let all agents execute the ASL algorithm for 3000 iterations and for 20 values of δ in the interval $[1/150, 1/10]$. We run 20000 Monte Carlo experiments and we compute the steady-state empirical probability of error for each agent and each value of δ. In Figure 5.8, the empirical probability curves of agents $1, 3, 6, 7, 9$ are compared against the theoretical error probability in (5.10) computed using the Gaussian approximation in (5.39). The slope of these curves is compared against the slope Φ (i.e., the error exponent) predicted by Theorem 5.4.

5.9 Evolution over Successive Learning Cycles

In this section we focus on a specific nonstationary setting to illustrate in more detail the role of adaptation. We consider the time axis can be divided into successive *random* intervals (*learning cycles*) wherein the system conditions remain stationary. We do not focus here on situations where the system parameters can vary smoothly at each time instant following some "trajectory", as happens, e.g., in tracking applications. While from the analysis of similar algorithms we can expect that the ASL strategy possesses some inherent tracking ability, the study of this scenario is left for future work.

We examine an environment where there are three different sources of nonstationarity, which will be modeled as (mutually independent) homogeneous Markov chains, as now specified:

- The true hypothesis can change over time. For $i = 1, 2, \ldots$, the true state of nature at time i, denoted by $\boldsymbol{\theta}_0(i)$, follows a Markov process with possible states in $\Theta = \{1, 2, 3\}$ and with transition probabilities described by the finite-state diagram in Figure 5.9 (where only transition probabilities are displayed, with the complementary probabilities of remaining in a state being omitted).

- The combination policy can change over time. We assume that the agents employ two possible combination matrices, one doubly-stochastic (DS), the other left-stochastic (LS). For $i = 1, 2, \ldots$, the combination matrix in force at time i, denoted by $\boldsymbol{A}(i)$, follows a Markov process with transition matrix represented by the corresponding finite-state diagram in Figure 5.9.

- The system can be in one of three possible functioning states, namely, nominal (N), perturbed (P), and bad (B). For $i = 1, 2, \ldots$, the operating state at time i is denoted by $\boldsymbol{f}(i)$. Under state $\boldsymbol{f}(i) =$ nominal, the data are generated according to the true likelihood corresponding to hypothesis $\theta_0(i)$. Under state $\boldsymbol{f}(i) =$ perturbed, some noise is added to perturb the true data model (while the agents still rely on the nominal likelihood to run their ASL strategy). State $\boldsymbol{f}(i) =$ bad corresponds to a failure of the system, where a large amount of noise is added to the data so as to impair the learning process. The transition matrix of the functioning process is encoded in the pertinent finite-state diagram in Figure 5.9.

Let us evaluate the average duration of a learning cycle. In order to be conservative, we focus on the worst case, i.e., on the shorter average duration, which is obtained when the system is in the most unstable case (i.e., the state where transitions are more frequent). Examining Figure 5.9, the most unstable state is obtained when: **i)** the hypothesis in force is $\boldsymbol{\theta}_0(i) = 2$, since from such intermediate state the Markov chain can move leftward or rightward, while from the other states it cannot; **ii)** the combination policy is either left stochastic or doubly stochastic; and **iii)** the system works under a perturbed state of functioning, for the same reasons as in point **i)**. Now, given that the overall system is in the joint state $\{\boldsymbol{\theta}_0(i) = 2, \boldsymbol{A}(i) =$ left stochastic, $\boldsymbol{f}(i) =$ perturbed$\}$, the probability that the system remains stable for a single step is equal to:

$$q^\star = (1 - 2\,q_{\rm hyp})(1 - q_{\rm mat})(1 - 2\,q_{\rm fun}). \tag{5.85}$$

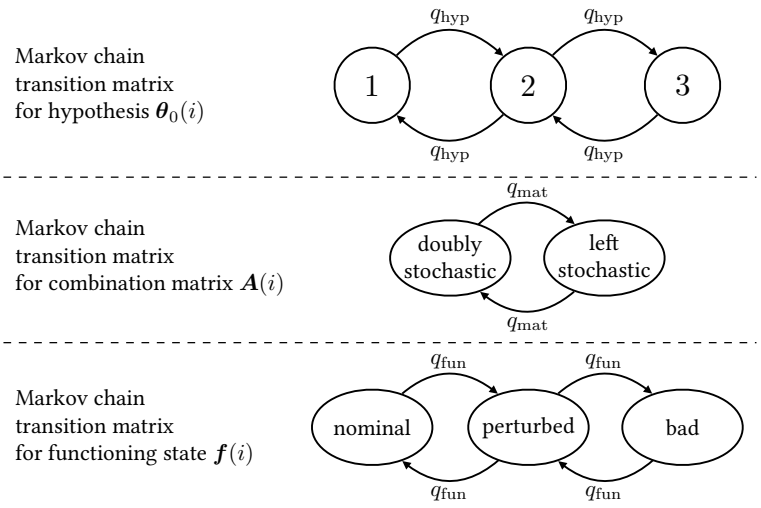


Figure 5.9: Transition matrices of the Markov chains corresponding to the sources of nonstationarity illustrated in Section 5.9.

Likewise, the probability that the system remains stable for a certain number of steps is ruled by a geometric distribution of parameter $q^\star$, yielding the following average duration for the worst-case learning cycle:

$$\mathsf{T}_{\mathsf{LC}} = \frac{q^\star}{1 - q^\star}. \tag{5.86}$$

In order to model a nonstationary environment where the system parameters remain stable during the learning cycles, we take inspiration from the Gilbert-Elliott model typically employed to model random bursts of errors over communication channels [83], [84]. According to the Gilbert-Elliott model, the transition probabilities between states of the chain are kept small so as to ensure that the chain remains in the same state for some contiguous time samples (i.e., we have "bursts" where the same state is repeatedly observed).

For what concerns the nominal likelihood models, we use the following family of Laplace likelihood functions, for $n \in \{1, 2, 3\}$:

$$f_n(\xi) = \frac{1}{2} \exp\left\{-|\xi - n|\right\}, \tag{5.87}$$

under the same identifiability setup as in Table 5.1. The network topology is the same as in the left panel of Figure 5.5, on top of which we build two possible combination matrices: a left-stochastic matrix obtained through a uniform-averaging combination policy, and a doubly-stochastic matrix obtained through a Laplacian combination policy [37]. Under this setting, we evaluate the adaptation time exploiting (5.69). Regarding the initial states appearing in (5.69), we assume that in a given learning cycle the system comes from a previous learning cycle where the agents converged to a hypothesis different from that in force during the current learning cycle. Then we consider the worst-case initial state, and further the worst-case over all possible θ and θ_0. With these conservative choices, the time necessary to stay at 3 dB from the exponent

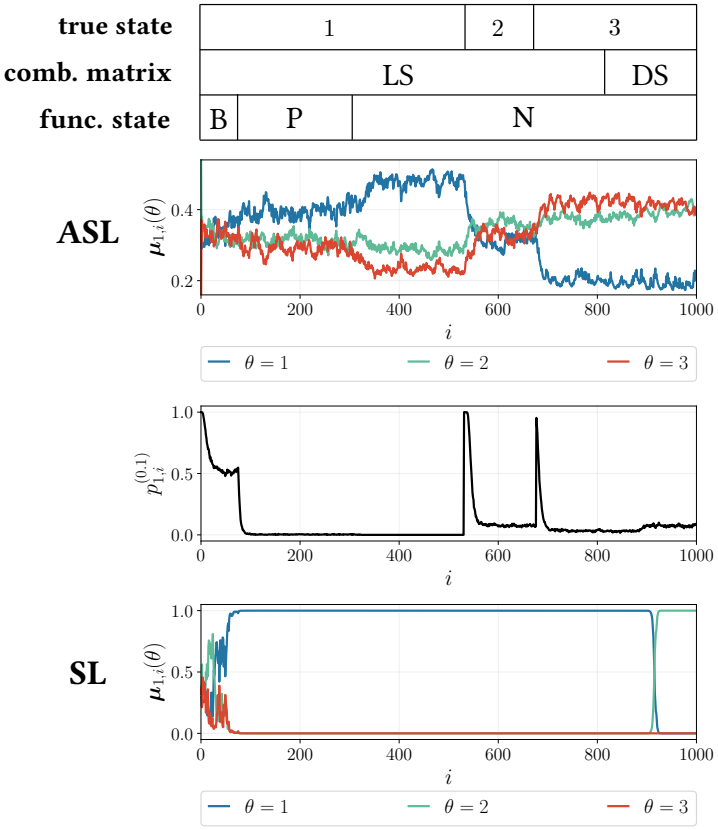


Figure 5.10: Evolution of the learning strategies over successive learning cycles, with step-size $\delta = 0.1$ and average learning-cycle duration $\mathsf{T}_{\mathsf{LC}} \approx 100$. *First (top) row.* Observed transitions for the three sources of nonstationarity illustrated in the main text, namely, state of functioning, combination matrix, and hypothesis. *Second row.* Time-evolution of the belief at agent 1 for the *adaptive* social learning strategy. *Third row.* Time-evolution of the error probability at agent 1 for the *adaptive* social learning strategy. *Fourth row.* Time-evolution of the belief at agent 1 for the *classic* social learning strategy.

Φ is equal to:

$$\mathsf{T}_{\mathsf{ASL}} \approx \frac{2.7286}{\delta}. \tag{5.88}$$

We now examine two settings that correspond to (relatively) short and long learning cycles, respectively.

– *"Short" Learning Cycles.* First of all, we consider that malfunctioning events and variations of the combination matrix are rare as compared to changes in the hypothesis. In particular, we set:

$$q_{\text{hyp}} = 5 \times 10^{-3}, \;\; q_{\text{mat}} = 10^{-3}, \;\; q_{\text{fun}} = 10^{-3}. \tag{5.89}$$

Exploiting (5.86), the average duration of a learning cycle can be approximated as $\mathsf{T}_{\mathsf{LC}} \approx 76$. If we equate the value found for T_{LC} to the adaptation time in (5.88), we get $\delta = 0.035$. For proper learning, we need that the adaptation time is smaller than the average duration of a learning cycle to ensure convergence to the correct hypothesis. In the experiments shown in

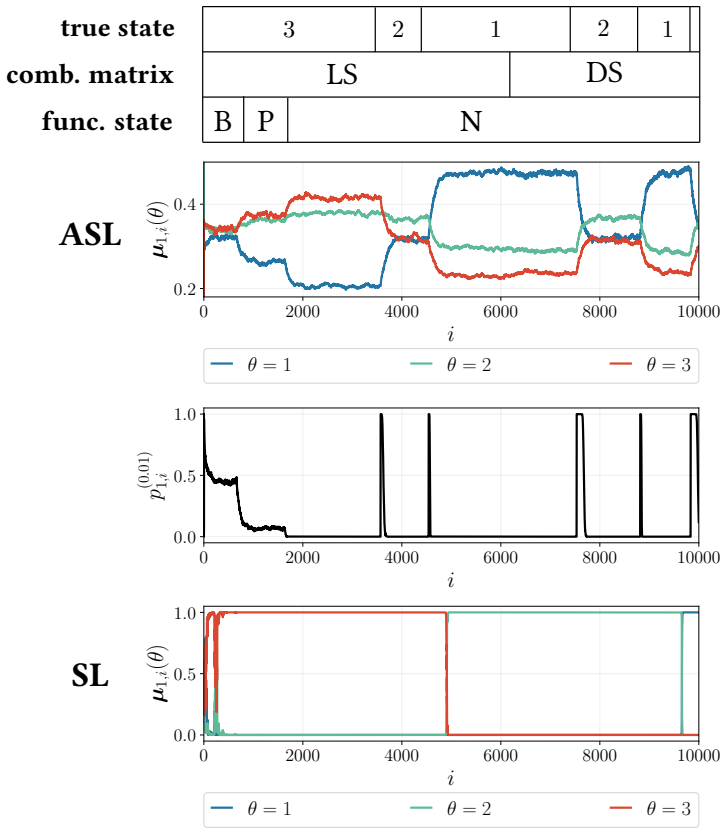


Figure 5.11: Evolution of the learning strategies over successive learning cycles, with step-size $\delta = 0.01$ and average learning-cycle duration $\mathsf{T}_{\mathsf{LC}} \approx 1000$. *First (top) row.* Observed transitions for the three sources of nonstationarity illustrated in the main text, namely, state of functioning, combination matrix, and hypothesis. *Second row.* Time-evolution of the belief at agent 1 for the *adaptive* social learning strategy. *Third row.* Time-evolution of the error probability at agent 1 for the *adaptive* social learning strategy. *Fourth row.* Time-evolution of the belief at agent 1 for the *classic* social learning strategy.

Figure 5.10 we made the choice:

$$\delta = 0.1, \tag{5.90}$$

which corresponds to an adaptation time not larger than one third of the average worst-case learning cycle. In Figure 5.10, we display: in the second row, the time-evolution of the beliefs at agent 1 corresponding to one realization of the process; in the third row, the corresponding error probability; and in the fourth row, the time-evolution of the beliefs at agent 1 for a classic social learning strategy (same realization considered for the ASL strategy). During the considered time interval, several variations occurred, according to the nonstationary model described before.

First, we see that, except for the learning cycle corresponding to a bad state of functioning, the ASL strategy is able to learn well in all learning cycles, after a relatively short transient at the beginning of each cycle. The ability of learning well is showed by the time-evolution of the beliefs (second row), which shows how the maximum belief corresponds to the true hypothesis, after relatively short adaptation intervals necessary to react in face of nonstationarities. More quantitatively, the ability of learning is showed by the time-evolution of the error probabilities

(third row), where we see some peaks (error probability close to 1) that clearly correspond to the changes, and that have a short duration dictated by the adaptation times. In sharp contrast, the classic social learning strategy looses irremediably its learning ability yet after the first learning cycle.

Zooming in on Figure 5.10, we see that nonstationarities in the hypotheses induce a perceivable change in the learning performance, whereas nonstationarities in the combination policy or in the state of functioning deserve a separate analysis.

For what concerns the combination policies, we see that the learning ability is preserved in face of a change, i.e., the system does *not* undergo an interval of poor performance. This behavior makes perfect sense, since from the theoretical analysis we know that the ASL strategy must consistently learn both with a left-stochastic or a doubly-stochastic matrix. What can be different are the "steady-state" beliefs, which depends on the Perron eigenvector. In this particular example, we have verified that, as opposed to the uniform Perron eigenvector corresponding to the doubly-stochastic matrix, the eigenvector of the left-stochastic matrix features higher weights corresponding to more informative agents (i.e., agents with higher KL divergences), which provides an explanation of the slightly distinct belief levels observed in Figure 5.10.

Regarding the state of functioning, we see that during the "bad" functioning state the data does not provide useful information, and the system undergoes an interval of failure (error probability ≈ 0.5). The adaptivity of the ASL strategy allows the agents to recover from this failure state in the successive learning cycles. In particular, the agents are able to recover and learn well already during the "perturbed" state of functioning. Actually, this regime of operation where the data does not follow any of the nominal likelihoods is not covered by our steady-state analysis. Our results could be in principle extended by allowing arbitrary distributions for the true data—this is actually carried out partly in [85]. In this case, it is expected that, for reasonable amounts of perturbation, the agents are still able to learn, as happens in the considered example. Moreover, we expect that passing from a perturbed to a nominal state, the performance improves. Visually, this effect can be more clearly appreciated in the subsequent example shown in Figure 5.11.

In summary, we see that the starting values at the beginning of each learning cycle are stable, since they arise as steady-state limiting values from at the end of the previous learning cycle. As such, these starting values do not diverge as time elapses, guaranteeing proper learning over successive learning cycles. This is a critical property, since it reveals that the number of variations of the underlying statistical conditions occurring during the entire algorithm evolution does not impair learning with the ASL strategy. What really matters is that the duration of the learning cycle is sufficiently large to allow a (small) value of δ to enable accurate learning.

– *"Long" Learning Cycles.* In Figure 5.11, we consider the more favorable situation where the average duration of the learning cycle is increased by one order of magnitude, using the following transition probabilities for the pertinent Markov chains:

$$q_{\mathrm{hyp}} = 5 \times 10^{-4}, \;\; q_{\mathrm{mat}} = 10^{-4}, \;\; q_{\mathrm{fun}} = 10^{-4}. \tag{5.91}$$

Accordingly, we expect that the adaptation properties of the system will be preserved if we

reduce the step-size by one order of magnitude, yielding:

$$\delta = 0.01. \tag{5.92}$$

Comparing Figure 5.11 against Figure 5.10, we see that the general behavior is perfectly confirmed, and two notable effects emerge. First, the adaptation properties are preserved, i.e., the system is able to adapt to the changes sufficiently fast to guarantee a stable evolution over successive learning cycles. Second, the fluctuations around the limiting steady-state are reduced w.r.t. Figure 5.10, yielding a smaller error probability, as it must be according to the theoretical analysis carried out in the previous sections since we are now using a smaller step-size $\delta = 0.01$.

5.10 Concluding Remarks

Existing social learning implementations do not operate well in *nonstationary* environments. For example, even if the agents learned correctly the true state, when this state changes, agents in classic social learning tend to be stubborn and keep on believing the old state. In this chapter we proposed an adaptive social learning strategy, which overcomes this issue, and examined its performance and provided convergence guarantees in great detail. The key insight is the introduction of an *adaptive update* depending on a step-size parameter δ that allows to tune the degree of adaptation. The introduction of the step-size δ allows the user to explore the trade-off between accuracy in decision making and adaptation time.

In the steady-state phase, with focus on the small step-size regime, we have ascertained that the ASL strategy is able to learn consistently, and we have provided reliable performance characterization of the learning performance at each individual agent. In the transient phase, we have shown how the learning performance evolves over time and how the choice of the step-size affects the adaptation time.

The strategy proposed is able to infuse the network with adaptation capabilities, without any assumptions on the nature of the nonstationarity. Assuming additional knowledge, we could derive more specialized strategies, tailored to a given nature of nonstationarity. An extension in this direction can be found in [86], where the true state of the world is assumed to evolve according to a Markov chain. Inspired by hidden-Markov-model filtering, the authors propose a modification to the social learning algorithm, resulting in superior tracking performance.

The work described in this chapter has motivated interesting scientific ramifications. In [87], [88], the problems of topology learning and graph explainability are investigated in the context of adaptive social learning. Meanwhile, in [85], [89], the authors find that doubly-stochastic combination policy is optimal in the sense that it minimizes the steady-state probability of error. This implies that the optimal way of combining the beliefs in adaptive social learning is to attribute uniform centrality scores to all agents.

5.A Main Lemma

In the following, the symbols $\mathcal{S}^o$ and $\overline{\mathcal{S}}$ denote the interior and the closure of set $\mathcal{S}$, respectively.

Lemma 5.1 (Asymptotic properties of random series useful for adaptation). *For $m = 0, 1, \ldots$, let $\{\boldsymbol{z}_m\}$ be a sequence of i.i.d. integrable random variables with:*

$$\mathsf{m}_z \triangleq \mathbb{E}(\boldsymbol{z}_m), \qquad \mathsf{m}_z^{\text{abs}} \triangleq \mathbb{E}(|\boldsymbol{z}_m|) < \infty. \tag{5.93}$$

Let also $0 < \delta < 1$, and consider the following partial sums:

$$\boldsymbol{s}_i(\delta) = \delta \sum_{m=0}^{i} (1-\delta)^m \alpha_m \boldsymbol{z}_m, \tag{5.94}$$

where $0 < \alpha_m \leq 1$, with α_m converging to some value $\alpha > 0$ and obeying the following upper bound for all m:

$$|\alpha_m - \alpha| \leq \kappa \beta^m, \tag{5.95}$$

for some constant $\kappa > 0$ and for some $0 < \beta < 1$. Then, we have the following asymptotic properties.

1. Steady-state stability. *The partial sums in* (5.94) *are almost-surely absolutely convergent, namely, we can define the (almost-surely) convergent series:*

$$\boldsymbol{s}^{\text{abs}}(\delta) \triangleq \delta \sum_{m=0}^{\infty} (1-\delta)^m \alpha_m |\boldsymbol{z}_m|, \tag{5.96}$$

$$\boldsymbol{s}(\delta) \triangleq \delta \sum_{m=0}^{\infty} (1-\delta)^m \alpha_m \boldsymbol{z}_m. \tag{5.97}$$

2. First moment. *The expectation of $\boldsymbol{s}(\delta)$ is:*

$$\mathbb{E}(\boldsymbol{s}(\delta)) = \mathsf{m}_z \delta \sum_{m=0}^{\infty} (1-\delta)^m \alpha_m = \alpha\, \mathsf{m}_z + O(\delta), \tag{5.98}$$

where $O(\delta)$ is a quantity such that the ratio $O(\delta)/\delta$ remains bounded as $\delta \to 0$.

3. Weak law of small step-sizes. *The series $\boldsymbol{s}(\delta)$ converges to $\alpha\, \mathsf{m}_z$ in probability as $\delta \to 0$, namely, for all $\epsilon > 0$ we have that:*

$$\lim_{\delta \to 0} \mathbb{P}\left[|\boldsymbol{s}(\delta) - \alpha\, \mathsf{m}_z| > \epsilon\right] = 0. \tag{5.99}$$

4. Second moment. *If:*

$$\sigma_z^2 \triangleq \mathsf{VAR}[\boldsymbol{z}_m] < \infty, \tag{5.100}$$

then:

$$\begin{aligned} \mathsf{VAR}(\boldsymbol{s}(\delta)) &= \sigma_z^2 \delta^2 \sum_{m=0}^{\infty} (1-\delta)^{2m} \alpha_m^2 \\ &= \frac{\alpha^2 \sigma_z^2}{2}\, \delta + O(\delta^2). \end{aligned} \tag{5.101}$$

5. Asymptotic normality. *If $\boldsymbol{z}_m$ has finite variance σ_z^2, then the following convergence in distribution holds:*

$$\frac{\boldsymbol{s}(\delta) - \mathsf{m}_z}{\sqrt{\delta}} \xrightarrow[\delta \to 0]{\mathrm{d}} \mathscr{G}\left(0, \alpha^2 \sigma_z^2/2\right), \tag{5.102}$$

and, hence, $\boldsymbol{s}(\delta)$ is asymptotically normal as $\delta \to 0$.

6. Large deviations. *Assume that $\boldsymbol{z}_m$ is non-deterministic and has LMGF finite everywhere:*

$$\Lambda_z(t) = \log \mathbb{E}\left(e^{\boldsymbol{z}_m t}\right) < +\infty, \ \ \forall t \in \mathbb{R}. \tag{5.103}$$

Let $\Lambda_{\alpha z}(t) = \Lambda_z(\alpha t)$ be the LMGF of the scaled variable $\alpha \boldsymbol{z}_m$, where α is defined in (5.95)*. Denoting by $\Lambda_\delta(t)$ the LMGF of $\boldsymbol{s}(\delta)$, we have that:*

$$\lim_{\delta \to 0} \delta \Lambda_\delta(t/\delta) = \phi(t) = \int_0^t \frac{\Lambda_{\alpha z}(\tau)}{\tau} d\tau. \tag{5.104}$$

Then the following Large Deviations Principle (LDP) holds for any measurable set $\mathcal{S}$ (the infimum over an empty set is taken as $+\infty$):

$$\liminf_{\delta \to 0} \ \delta \log \mathbb{P}\left(\boldsymbol{s}(\delta) \in \mathcal{S}\right) \geq -\inf_{\gamma \in \mathcal{S}^o} \phi^\star(\gamma), \tag{5.105}$$

$$\limsup_{\delta \to 0} \delta \log \mathbb{P}\left(\boldsymbol{s}(\delta) \in \mathcal{S}\right) \leq -\inf_{\gamma \in \overline{\mathcal{S}}} \phi^\star(\gamma), \tag{5.106}$$

where

$$\phi^\star(\gamma) = \sup_{t \in \mathbb{R}} [\gamma t - \phi(t)] \tag{5.107}$$

is the Fenchel-Legendre transform of $\phi(t)$ [80], [81]. The function $\phi^\star(\gamma)$ (which is allowed to be an extended real number) is usually called rate function [80], [81] and has the following properties.

- *Let z_- and z_+ be the boundaries of the support of $\boldsymbol{z}_m$, and let $\mathcal{D} = \{\gamma \in \mathbb{R} : \phi^\star(\gamma) < +\infty\}$. Then $\mathcal{D}$ is given by the following open interval:*

$$\mathcal{D} = (\alpha z_-, \alpha z_+). \tag{5.108}$$

- *The function $\phi^\star(\gamma)$ is smooth and strictly convex on $\mathcal{D}$, and diverges to $+\infty$ at the boundaries of $\mathcal{D}$. In particular, if a boundary is finite, the rate function is equal to $+\infty$ at that boundary.*
- *$\phi^\star(\gamma) \geq 0$, with equality if, and only if, $\gamma = \alpha \mathsf{m}_z$.*

A typical shape of the rate function is illustrated in Figure 5.12. Exploiting the aforementioned regularity properties of $\phi^\star(\gamma)$, from (5.105)–(5.106) *we have in particular that:*

$$\lim_{\delta \to 0} \delta \log \mathbb{P}\left(\boldsymbol{s}(\delta) \geq \gamma\right) = -\phi^\star(\gamma), \quad \forall \gamma \geq \alpha \mathsf{m}_z, \tag{5.109}$$

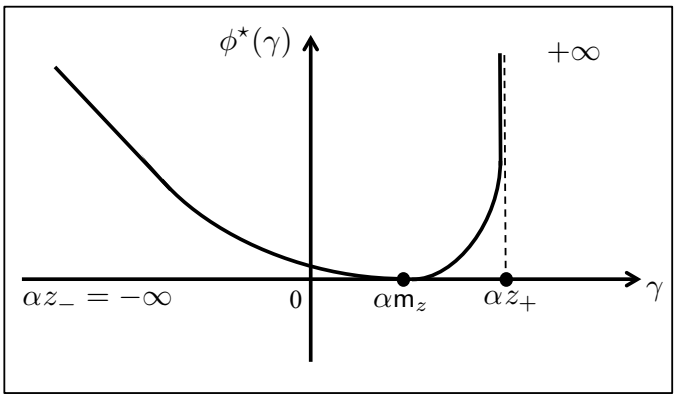


Figure 5.12: Typical shape of the rate function.

$$\lim_{\delta\to 0} \delta \log \mathbb{P}\left(\boldsymbol{s}(\delta) \leq \gamma\right) = -\phi^\star(\gamma), \quad \forall \gamma \leq \alpha \mathsf{m}_z. \tag{5.110}$$

Proof. We prove sequentially the six parts of the lemma.

Part 1. In view of (5.93), the following series of (absolute) expectations is convergent:

$$\begin{aligned}\delta \sum_{m=0}^{\infty} (1-\delta)^m \alpha_m \mathbb{E}\left(|\boldsymbol{z}_m|\right) &= \mathsf{m}_z^{\mathrm{abs}} \delta \sum_{m=0}^{\infty} (1-\delta)^m \alpha_m \\ &\leq \mathsf{m}_z^{\mathrm{abs}} \delta \sum_{m=0}^{\infty} (1-\delta)^m \\ &= \mathsf{m}_z^{\mathrm{abs}} < +\infty. \end{aligned} \tag{5.111}$$

In view of [75][Lemma 3.6′], convergence of the series of absolute first moments implies that the random series $\boldsymbol{s}^{\mathrm{abs}}(\delta)$ is almost-surely finite, which in turn implies that so is $\boldsymbol{s}(\delta)$, and part 1 is proved.

Part 2. Since the series of (absolute) expectations is convergent, so is the series of expectations:

$$\sum_{m=0}^{\infty} (1-\delta)^m \alpha_m \mathbb{E}\left(\boldsymbol{z}_m\right) = \mathsf{m}_z \sum_{m=0}^{\infty} (1-\delta)^m \alpha_m. \tag{5.112}$$

On the other hand, by triangle inequality we have the following upper bound:

$$|\boldsymbol{s}_i(\delta)| \leq \delta \sum_{m=0}^{i} (1-\delta)^m \alpha_m |\boldsymbol{z}_m| \leq \boldsymbol{s}^{\mathrm{abs}}(\delta). \tag{5.113}$$

Now we observe that $\boldsymbol{s}^{\mathrm{abs}}(\delta)$ is a proper random variable in view of part 1. Furthermore, it is an integrable random variable from Beppo Levi's monotone convergence theorem [90][Th. 1.5.7, p. 27], thanks to the convergence of absolute expectations in (5.112).

We conclude that the random sequence $\boldsymbol{s}_i(\delta)$ is upper bounded by an integrable random variable.

Therefore, the dominated convergence theorem [90][Th. 1.5.8, p. 27] implies that the expectation of the almost-sure limit $\boldsymbol{s}(\delta)$ is equal to the convergent series of expectations, and the first equality in (5.98) follows. Moreover, we can write:

$$\delta \sum_{m=0}^{\infty} (1-\delta)^m \alpha_m = \delta \sum_{m=0}^{\infty} (1-\delta)^m (\alpha_m - \alpha) + \alpha \underbrace{\delta \sum_{m=0}^{\infty} (1-\delta)^m}_{=1}. \tag{5.114}$$

In view of (5.95), the absolute value of the first summation on the RHS in (5.114) is dominated by:

$$\kappa \delta \sum_{m=0}^{\infty} \Big(\beta(1-\delta)\Big)^m = \frac{\kappa \delta}{1-\beta(1-\delta)} = O(\delta). \tag{5.115}$$

We conclude from (5.112), (5.114) and (5.115) that the second equality in (5.98) holds.

Part 3. Let

$$\zeta_m \triangleq \delta(1-\delta)^m \alpha_m, \tag{5.116}$$

and consider the following centered variables:

$$\widetilde{\boldsymbol{s}}(\delta) = \boldsymbol{s}(\delta) - \mathbb{E}\,(\boldsymbol{s}(\delta)), \qquad \widetilde{\boldsymbol{z}}_m = \boldsymbol{z}_m - \mathbb{E}\,(\boldsymbol{z}_m). \tag{5.117}$$

In view of parts 1 and 2, the centered partial sums:

$$\boldsymbol{s}_i(\delta) - \mathbb{E}\,(\boldsymbol{s}_i(\delta)) = \sum_{m=0}^{i} \zeta_m \widetilde{\boldsymbol{z}}_m \tag{5.118}$$

converge in distribution to $\widetilde{\boldsymbol{s}}(\delta)$ as $i \to \infty$. By Lévy's continuity theorem, the corresponding characteristic functions must converge [91][Th. 2, p. 431]. Since the $\boldsymbol{z}_m$'s are i.i.d. we can write:

$$\varphi_{\tilde{s}}(t) \triangleq \mathbb{E}\left(e^{j\widetilde{\boldsymbol{s}}(\delta)t}\right) = \prod_{m=0}^{\infty} \varphi_{\tilde{z}}(\zeta_m t), \tag{5.119}$$

where $j = \sqrt{-1}$. We want to show that $\widetilde{\boldsymbol{s}}(\delta)$ converges in probability to 0 as $\delta \to 0$. In view of Lévy's continuity Theorem this is tantamount to showing that $\varphi_{\tilde{s}}(t)$ converges to 1 as $\delta \to 0$. Using (5.119) we can write:[4]

$$|\varphi_{\tilde{s}}(t) - 1| \leq \sum_{m=0}^{\infty} |\varphi_{\tilde{z}}(\zeta_m t) - 1|. \tag{5.121}$$

[4]The following inequality is known for complex numbers x_m, y_m, with $|x_m| \leq 1$ and $|y_m| \leq 1$ [91]:

$$\left| \prod_{m=0}^{i} x_m - \prod_{m=0}^{i} y_m \right| \leq \sum_{m=0}^{i} |x_m - y_m|, \tag{5.120}$$

Consider, without loss of generality, a positive t. Since the random variables $\widetilde{\boldsymbol{z}}_m$ have finite expectation, the first derivative of the characteristic function, $\varphi'_{\widetilde{z}}(t)$, is a continuous function, and by the mean-value theorem we can write (since in particular $\mathbb{E}\,(\widetilde{\boldsymbol{z}}_m) = 0$):

$$\varphi_{\widetilde{z}}(\zeta_m t) = 1 + \zeta_m t\, \varphi'_{\widetilde{z}}(t_m), \text{ for some } t_m \in (0, \zeta_m t). \tag{5.122}$$

Accordingly we can write:

$$|\varphi_{\widetilde{z}}(\zeta_m t) - 1| \leq \zeta_m |t| \max_{\tau \in [0, \delta t]} |\varphi'_{\widetilde{z}}(\tau)|, \tag{5.123}$$

where the latter inequality follows from the fact that $\zeta_m \leq \delta$, see (5.116). Applying (5.123) to (5.121) we get:

$$|\varphi_{\tilde{s}}(t) - 1| \leq |t| \max_{\tau \in [0, \delta t]} |\varphi'_{\widetilde{z}}(\tau)| \underbrace{\sum_{m=0}^{\infty} \zeta_m}_{\leq 1}. \tag{5.124}$$

On the other hand, since $\varphi'_{\widetilde{z}}(0) = \mathbb{E}\,(\widetilde{\boldsymbol{z}}_m) = 0$, from the continuity of $\varphi'_{\widetilde{z}}(t)$ it follows that:

$$\lim_{\delta \to 0} \max_{\tau \in [0, \delta t]} |\varphi'_{\widetilde{z}}(\tau)| = 0, \tag{5.125}$$

which proves that $\boldsymbol{s}(\delta)$ converges to $\mathbb{E}\,(\boldsymbol{s}(\delta))$ in probability as $\delta \to 0$. The claim in (5.99) then follows from (5.98).

Part 4. Since the variables $\boldsymbol{z}_m$ have common finite variance σ_z^2 and are independent, it is immediate to see that:

$$\lim_{i \to \infty} \mathsf{VAR}(\boldsymbol{s}_i(\delta)) = \sigma_z^2\, \delta^2 \sum_{m=0}^{\infty} (1-\delta)^{2m} \alpha_m^2 < \infty. \tag{5.126}$$

Consider now the squared and centered variables:

$$\Big(\boldsymbol{s}_i(\delta) - \mathbb{E}\,(\boldsymbol{s}_i(\delta))\Big)^2 = \delta^2 \left(\sum_{m=0}^{i} (1-\delta)^m \alpha_m (\boldsymbol{z}_m - \mathsf{m}_z)\right)^2. \tag{5.127}$$

In view of parts 1 and 2 the quantity on the LHS converges almost surely, as $i \to \infty$, to:

$$\Big(\boldsymbol{s}(\delta) - \mathbb{E}\,(\boldsymbol{s}(\delta))\Big)^2. \tag{5.128}$$

Given the convergence of the variance of the partial sums in (5.126), by Fatou's lemma we conclude that [90][Th. 1.5.5, p. 26]:

$$\mathsf{VAR}(\boldsymbol{s}(\delta)) \leq \lim_{i \to \infty} \mathsf{VAR}(\boldsymbol{s}_i(\delta)), \tag{5.129}$$

i.e., the limiting variable $\boldsymbol{s}(\delta)$ has finite variance. But since the limiting variable $\boldsymbol{s}(\delta)$ can be

written as:

$$\boldsymbol{s}(\delta) = \boldsymbol{s}_i(\delta) + \delta \sum_{m=i+1}^{\infty} (1-\delta)^m \alpha_m \boldsymbol{z}_m, \tag{5.130}$$

with the two quantities on the RHS being statistically independent, the variance of $\boldsymbol{s}(\delta)$ cannot be smaller than the variance of $\boldsymbol{s}_i(\delta)$ for all i, implying that:

$$\mathsf{VAR}(\boldsymbol{s}(\delta)) \geq \lim_{i\to\infty} \mathsf{VAR}(\boldsymbol{s}_i(\delta)). \tag{5.131}$$

Combining (5.129) with (5.131) we see that the variance of the almost-sure limit $\boldsymbol{s}(\delta)$ is equal to the convergent series of variances, which is the first equality in (5.101).

In order to prove the second equality in (5.101) we write:

$$\begin{aligned}\mathsf{VAR}\left(\delta \sum_{m=0}^{\infty} (1-\delta)^m \alpha_m \boldsymbol{z}_m\right) &= \sigma_z^2 \delta^2 \sum_{m=0}^{\infty} (1-\delta)^{2m} \alpha_m^2 \\ &= \sigma_z^2 \delta^2 \sum_{m=0}^{\infty} (1-\delta)^{2m} \left(\alpha_m^2 - \alpha^2\right) \\ &+ \alpha^2 \sigma_z^2 \delta^2 \sum_{m=0}^{\infty} (1-\delta)^{2m}. \end{aligned} \tag{5.132}$$

Reasoning as done to prove part 2, we can easily show that the first summation on the RHS in (5.132) is $O(\delta^2)$. The second summation is instead equal to:

$$\frac{\alpha^2 \sigma_z^2 \delta^2}{1-(1-\delta)^2} = \frac{\alpha^2 \sigma_z^2 \delta}{2-\delta}, \tag{5.133}$$

and the second equality in (5.101) follows.

Part 5. Let

$$\sigma_{\lim}^2 \triangleq \frac{\alpha^2 \sigma_z^2}{2}. \tag{5.134}$$

The claim in (5.102) is equivalent to prove that the random variable $\frac{\boldsymbol{s}(\delta) - \mathsf{m}_z}{\sqrt{\delta}\sigma_{\lim}}$ converges in distribution to a standard Gaussian. On the other hand, we have that:

$$\frac{\boldsymbol{s}(\delta) - \mathsf{m}_z}{\sqrt{\delta}\sigma_{\lim}} = \frac{\boldsymbol{s}(\delta) - \mathbb{E}\left(\boldsymbol{s}(\delta)\right)}{\sqrt{\delta}\sigma_{\lim}} + \frac{\mathbb{E}\left(\boldsymbol{s}(\delta)\right) - \mathsf{m}_z}{\sqrt{\delta}\sigma_{\lim}}. \tag{5.135}$$

Since the second term in (5.135) converges to zero in view of (5.98), from Slutsky's theorem [73][Th. 1.11, p. 60] it suffices to show that the random variable $\frac{\boldsymbol{s}(\delta) - \mathbb{E}\left(\boldsymbol{s}(\delta)\right)}{\sqrt{\delta}\sigma_{\lim}}$ converges in distribution to a standard Gaussian. To this end, we start by introducing, with slight abuse of notation w.r.t. (5.116) and (5.117), the quantities:

$$\zeta_m \triangleq \frac{\sqrt{2\delta}(1-\delta)^m \alpha_m}{\alpha}, \tag{5.136}$$

and:

$$\widetilde{\boldsymbol{s}}(\delta) = \frac{\boldsymbol{s}(\delta) - \mathbb{E}\left(\boldsymbol{s}(\delta)\right)}{\sqrt{\delta}\sigma_{\lim}}, \qquad \widetilde{\boldsymbol{z}}_m = \frac{\boldsymbol{z}_m - \mathbb{E}\left(\boldsymbol{z}_m\right)}{\sigma_z}. \tag{5.137}$$

We notice that $\widetilde{\boldsymbol{z}}_m$ has zero mean and unit variance.

We will now show that $\widetilde{\boldsymbol{s}}(\delta)$ converges in distribution to a standard Gaussian. In view of Lévy's continuity theorem, this claim is equivalent to the convergence, as $\delta \to 0$, of the characteristic function of $\widetilde{\boldsymbol{s}}(\delta)$ to the characteristic function $e^{-\frac{t^2}{2}}$. From (5.97), (5.134), (5.136) and (5.137) we see that:

$$\widetilde{\boldsymbol{s}}(\delta) = \sum_{m=0}^{\infty} \zeta_m \widetilde{\boldsymbol{z}}_m. \tag{5.138}$$

Reasoning as done to compute (5.119), the characteristic function of $\widetilde{\boldsymbol{s}}(\delta)$ in (5.137) can be written as:

$$\varphi_{\tilde{s}}(t) = \prod_{m=0}^{\infty} \varphi_{\tilde{z}}(\zeta_m t). \tag{5.139}$$

Using the triangle inequality for complex numbers we can write:

$$\left|\varphi_{\tilde{s}}(t) - e^{-\frac{t^2}{2}}\right| \leq \left|\varphi_{\tilde{s}}(t) - e^{-\frac{\sum_{m=0}^{\infty}\zeta_m^2 t^2}{2}}\right| + \left|e^{-\frac{\sum_{m=0}^{\infty}\zeta_m^2 t^2}{2}} - e^{-\frac{t^2}{2}}\right|. \tag{5.140}$$

Now, that the second term on the RHS of (5.140) converges to zero follows from part 4), since from (5.101) and the definition of ζ_m in (5.136) we conclude that:

$$\lim_{\delta\to 0} \sum_{m=0}^{\infty} \zeta_m^2 = 1. \tag{5.141}$$

Let us now focus on the first term on the RHS of (5.140). Since the characteristic functions have magnitude not greater than 1, in view of (5.120) and (5.139) we can write:

$$\begin{aligned}\left|\varphi_{\tilde{s}}(t) - e^{-\frac{\sum_{m=0}^{\infty}\zeta_m^2 t^2}{2}}\right| &\leq \sum_{m=0}^{\infty}\left|\varphi_{\tilde{z}}(\zeta_m t) - e^{-\frac{\zeta_m^2 t^2}{2}}\right| \\ &\leq \sum_{m=0}^{\infty}\left|\varphi_{\tilde{z}}(\zeta_m t) - 1 + \frac{\zeta_m^2 t^2}{2}\right| \\ &+ \sum_{m=0}^{\infty}\left|e^{-\frac{\zeta_m^2 t^2}{2}} - 1 + \frac{\zeta_m^2 t^2}{2}\right|,\end{aligned} \tag{5.142}$$

where in the latter step we applied the triangle inequality. Now, the last term in (5.142) converges to zero since for any positive s we have $|e^{-s} - 1 + s| \leq s^2/2$, and since it is immediate to show that (see the proof in [79]):

$$\lim_{\delta\to 0} \sum_{m=0}^{\infty} \zeta_m^4 = 0. \tag{5.143}$$

On the other hand, using [90][Lemma 3.3.19, p. 134] we can write, for an arbitrarily small $\epsilon > 0$:

$$\left|e^{j\widetilde{\boldsymbol{z}}_m \zeta_m t} - 1 - j\widetilde{\boldsymbol{z}}_m \zeta_m t + \frac{1}{2}\widetilde{\boldsymbol{z}}_m^2 \zeta_m^2 t^2\right|$$

$$\leq \mathbb{I}\left[|\widetilde{\boldsymbol{z}}_m|\zeta_m \leq \epsilon\right]\frac{|\widetilde{\boldsymbol{z}}_m\zeta_m t|^3}{6} + \mathbb{I}\left[|\widetilde{\boldsymbol{z}}_m|\zeta_m > \epsilon\right](\widetilde{\boldsymbol{z}}_m\zeta_m t)^2$$
$$\leq \epsilon\widetilde{\boldsymbol{z}}_m^2\zeta_m^2\,\frac{|t|^3}{6} + \widetilde{\boldsymbol{z}}_m^2\mathbb{I}\left[|\widetilde{\boldsymbol{z}}_m|\zeta_m > \epsilon\right]\zeta_m^2 t^2$$
$$\leq \epsilon\widetilde{\boldsymbol{z}}_m^2\zeta_m^2\,\frac{|t|^3}{6} + \widetilde{\boldsymbol{z}}_m^2\mathbb{I}\left[|\widetilde{\boldsymbol{z}}_m| > \epsilon\alpha/\sqrt{2\delta}\right]\zeta_m^2 t^2, \tag{5.144}$$

where $\mathbb{I}[\mathcal{E}]$ is the indicator of event $\mathcal{E}$, and the last inequality follows because $\zeta_m \leq \sqrt{2\delta}/\alpha$—see (5.136). Let now:

$$g(\delta) = \mathbb{E}\left(\widetilde{\boldsymbol{z}}_m^2\mathbb{I}\left[\widetilde{\boldsymbol{z}}_m^2 > \epsilon\alpha/\sqrt{2\delta}\right]\right). \tag{5.145}$$

Owing to identical distribution of $\widetilde{\boldsymbol{z}}_m$ across index m, the function $g(\delta)$ does not depend on m. Since $\widetilde{\boldsymbol{z}}_m$ has finite variance, we have that $g(\delta) \to 0$ as $\delta \to 0$. In view of (5.144), recalling that the magnitude of the expectation is upper bounded by the expectation of the magnitude, and that $\widetilde{\boldsymbol{z}}_m$ has zero mean and unit variance, we have that:

$$\left|\varphi_{\tilde{z}}(\zeta_m t) - 1 + \frac{\zeta_m^2 t^2}{2}\right| \leq \sum_{m=0}^{\infty}\zeta_m^2\left(\epsilon\frac{|t|^3}{6} + t^2 g(\delta)\right), \tag{5.146}$$

and, hence,

$$\limsup_{\delta\to 0}\left|\varphi_{\tilde{z}}(\zeta_m t) - 1 + \frac{\zeta_m^2 t^2}{2}\right| \leq \epsilon\,\frac{|t|^3}{6}, \tag{5.147}$$

finally implying, due to the arbitrariness of ϵ, that $\varphi_{\tilde{\boldsymbol{s}}}(t)$ converges to $e^{-t^2/2}$ as $\delta \to 0$. We have therefore shown that $\widetilde{\boldsymbol{s}}(\delta)$ in (5.137) converges to a standard Gaussian as $\delta \to 0$, and this completes the proof of part 5.

Part 6. The convergence in (5.104) can be proved as done in [79, Appendix C]. Then the convergence in (5.104) implies the LDP in (5.105)–(5.106) in view of the Gärtner-Ellis theorem [80][Th. 2.3.6, p. 44], [81][Th. V.6, p. 54].

Next we focus on the regularity properties of the Fenchel-Legendre transform $\phi^\star(\gamma)$. Following the development used in [79, Appendix C], we can prove that $\mathcal{D}^o$ is an interval, that $\phi^\star(\gamma)$ is smooth and strictly convex for $\gamma \in \mathcal{D}^o$, and that $\phi^\star(\gamma) \geq 0$ with equality if, and only if, $\gamma = \alpha\mathsf{m}_z$.

Thus, it remains to characterize the boundaries of $\mathcal{D}^o$ and the behavior of the rate function at these boundaries. To this end, it is sufficient to prove the claim with $\alpha = 1$ and for the right boundary, since the proof for other values of α and for the left boundary is simply obtained using the scaling and reflection properties of the LMGF [80], [81].

Now, since it has been shown in [79, Appendix C] that the right boundary of $\mathcal{D}^o$ is equal to $\lim_{t\to\infty}\Lambda_z(t)/t$, we must now prove that this limit equals z_+ (recall that we are working with $\alpha = 1$). We start by noticing that, letting $z_- < \overline{z} < z_+$, the LMGF $\Lambda_z(t)$ can be written as:

$$\Lambda_z(t) = \log\left(\mathbb{E}\left(\mathbb{I}[\boldsymbol{z}_m \leq \overline{z}]e^{\boldsymbol{z}_m t}\right) + \mathbb{E}\left(\mathbb{I}[\boldsymbol{z}_m > \overline{z}]e^{\boldsymbol{z}_m t}\right)\right). \tag{5.148}$$

From (5.148) we get, for all $t > 0$:

$$\frac{\Lambda_z(t)}{t} \geq \frac{\log\left(e^{\overline{z}t}\,\mathbb{E}\;\left(\mathbb{I}\left[\boldsymbol{z}_m > \overline{z}\right]\right)\right)}{t} = \overline{z} + \frac{\log q}{t}, \tag{5.149}$$

where we set $q = \mathbb{P}\left(\boldsymbol{z}_m > \overline{z}\right)$. We remark that $0 < q < 1$ since $\overline{z}$ is internal to the support of $\boldsymbol{z}_m$. From (5.149) we get:

$$\liminf_{t\to\infty} \frac{\Lambda_z(t)}{t} \geq \overline{z}. \tag{5.150}$$

If $z_+ = +\infty$ the result is proved due to arbitrariness of $\overline{z}$. If $z_+ < +\infty$, we can choose $\overline{z} = z_+ - \epsilon$, and conclude that the limit inferior in (5.150) is equal to z_+. The fact that the corresponding limit superior is equal to z_+ follows by observing that, in view of (5.148), for all $t > 0$ the quantity $\Lambda_z(t)/t$ is upper bounded by z_+.

Finally, we characterize the behavior of the rate function at the boundaries of $\mathcal{D}^o$. We focus again on the right boundary z_+. When $z_+ = +\infty$, it suffices to notice that the rate function $\phi^\star(\gamma)$ is strictly convex in $\mathcal{D}^o$ and is strictly increasing for $\gamma > \mathsf{m}_z$ (see Figure 5.12) to conclude that the rate function diverges to $+\infty$ as $\gamma \to z_+$.

We move on to examine the case $z_+ < +\infty$. Exploiting (5.148) we can write, for all $t > 0$:

$$\begin{aligned}\Lambda_z(t) &\leq \log\left((1-q)e^{\overline{z}t} + qe^{z_+ t}\right)\\ &= z_+ t + \log\left((1-q)e^{-(z_+ - \overline{z})t} + q\right).\end{aligned} \tag{5.151}$$

Since $z_+ > \overline{z}$, for any $\epsilon > 0$ there exists $t_\epsilon > 0$ such that:

$$(1-q)e^{-(z_+-\overline{z})t} \leq \epsilon q, \;\; \text{for all } t \geq t_\epsilon, \tag{5.152}$$

implying, in view of (5.151):

$$\Lambda_z(t) \leq z_+ t + \log((1+\epsilon)q), \;\; \text{for all } t \geq t_\epsilon. \tag{5.153}$$

Using (5.153) in (5.104) we can thus write:

$$\begin{aligned}\phi(t) = \int_0^t \frac{\Lambda_z(\tau)}{\tau} d\tau &= \int_0^{t_\epsilon} \frac{\Lambda_z(\tau)}{\tau} d\tau + \int_{t_\epsilon}^t \frac{\Lambda_z(\tau)}{\tau} d\tau\\ &\leq \phi(t_\epsilon) + z_+(t - t_\epsilon) + \int_{t_\epsilon}^t \frac{\log\left((1+\epsilon)q\right)}{\tau} d\tau\\ &= \phi(t_\epsilon) + z_+(t - t_\epsilon) + \log\left((1+\epsilon)q\right)\,\log\frac{t}{t_\epsilon}.\end{aligned} \tag{5.154}$$

Plugging the latter inequality in (5.107) we get:

$$\begin{aligned}\phi^\star(z_+) &\geq \sup_{t\geq t_\epsilon}[z_+ t - \phi(t)] \geq -\phi(t_\epsilon) + z_+ t_\epsilon\\ &\quad + \log\frac{1}{(1+\epsilon)q}\,\sup_{t\geq t_\epsilon}\log\frac{t}{t_\epsilon} = +\infty,\end{aligned} \tag{5.155}$$

where we have chosen ϵ so small to ensure that $(1+\epsilon)q < 1$. Finally, in view of (5.107) we can write, for a generic $t \in \mathbb{R}$:

$$\lim_{\gamma \to z_+} \phi^\star(\gamma) \geq \lim_{\gamma \to z_+} [\gamma t - \phi(t)] = [z_+ t - \phi(t)], \tag{5.156}$$

and from (5.155) we conclude that $\phi^\star(\gamma) \to +\infty$ as $\gamma \to z_+$. □

5.B Proof of Theorem 5.1

We are interested in characterizing, for each agent k, the *joint* behavior of the random variables $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta)$ for all values of $\theta \neq \theta_0$. To this end, it is useful to consider the $(H-1) \times 1$ vector $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}$ similarly defined as the vector in (5.7). We also introduce, for a fixed time epoch i, the $K \times (H-1)$ data matrix $\boldsymbol{X}_i$, whose entries, for $\ell = 1, 2, \ldots, K$ and $\theta \neq \theta_0$, are:

$$[\boldsymbol{X}_i]_{\ell\theta} = \boldsymbol{x}_{\ell,i}(\theta). \tag{5.157}$$

In light of (5.19) we can write:

$$\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)} = f_{k,i}^{(\delta)}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_i), \tag{5.158}$$

to highlight that the random vector $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}$ is a certain function $f_{k,i}^{(\delta)}$ of the data matrices $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_i$. Since the data are i.i.d. over time, reversing the order of the data matrices in (5.158) does not change the distribution of the resulting random vector, i.e.:

$$\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)} = f_{k,i}^{(\delta)}(\boldsymbol{X}_i, \boldsymbol{X}_{i-1}, \ldots, \boldsymbol{X}_1) \overset{\mathrm{d}}{=} \widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}, \tag{5.159}$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution. Considering this reversed order of the data matrices in (5.19) and exchanging the order of summation we obtain:

$$\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) = \sum_{\ell=1}^{K} \delta \sum_{m=0}^{i-1} (1-\delta)^m [A^{m+1}]_{\ell k}\, \boldsymbol{x}_{\ell,m+1}(\theta). \tag{5.160}$$

From part 1) of Lemma 5.1 in the Appendix, each of the K inner partial sums (scaled by δ) converges *almost surely*. In fact, the random variables $\boldsymbol{x}_{\ell,m+1}(\theta)$ have finite first moment in view of Assumption 2.4, and the weights $[A^{m+1}]_{\ell k}$ fulfill condition (5.95) in view of Property 2.1. It makes thus sense to define a proper random variable as the (almost-surely convergent) value of the random series in (5.160), which corresponds to (5.24). This in turn implies the following almost-sure convergence, as $i \to \infty$, of the *vector* with reversed ordering, $\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}$, to the limiting random vector $\widetilde{\boldsymbol{\lambda}}_{k}^{(\delta)}$. In view of (5.159), this almost-sure convergence implies the convergence *in distribution* of the original (i.e., with correct ordering of the data matrices $\boldsymbol{X}_i$) vector $\widehat{\boldsymbol{\lambda}}_{k,i}$, finally yielding the claim of the theorem.

5.C Proof of Theorem 5.2

We start by proving (5.30). Examining (5.24) we see that each one of the K inner series matches the conditions in Lemma 5.1, part 3, implying that the ℓ-th inner series converges in probability, as $\delta \to 0$, to the expected value $\pi_\ell \mathbb{E}\left(\boldsymbol{x}_{\ell,m+1}(\theta)\right) = \pi_\ell d_\ell(\theta)$. As a result, $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta)$ converges in probability to $\mathsf{m}_{\text{ave}}(\theta)$, which implies, for any $\epsilon > 0$:

$$\lim_{\delta \to 0} \mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) < \mathsf{m}_{\text{ave}}(\theta) - \epsilon\right) = 0. \tag{5.161}$$

Since under Assumption 2.5 the quantity $\mathsf{m}_{\text{ave}}(\theta)$ is strictly positive, we conclude that:

$$\lim_{\delta \to 0} \mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) \leq 0\right) = 0, \tag{5.162}$$

which, by application of the union bound, in light of (5.10) gives:

$$\begin{aligned} p_k^{(\delta)} &= \mathbb{P}\left(\exists \theta \neq \theta_0 : \widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) \leq 0\right) \\ &\leq \sum_{\theta \neq \theta_0} \mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) \leq 0\right) \overset{\delta \to 0}{\longrightarrow} 0, \end{aligned} \tag{5.163}$$

and the claim of the theorem is proved.

5.D Proof of Theorem 5.3

In the following we will refer to the elements $\theta_1, \theta_2, \ldots, \theta_{H-1}$ in the set $\Theta \setminus \{\theta_0\}$—see (5.8). Consider a zero-mean Gaussian random vector:

$$\boldsymbol{g} = [\boldsymbol{g}(\theta_1), \boldsymbol{g}(\theta_2), \ldots, \boldsymbol{g}(\theta_{H-1})]^{\mathsf{T}}, \tag{5.164}$$

with covariance matrix equal to $\mathsf{C}_{\text{ave}}/2$. We recall that the (θ, θ')-th entry of C_{ave} is the covariance $\mathsf{c}_{\text{ave}}(\theta, \theta')$ defined in (5.33). What we want to show is that the random vector:

$$\frac{\widetilde{\boldsymbol{\lambda}}_k^{(\delta)} - \mathsf{m}_{\text{ave}}}{\sqrt{\delta}} \tag{5.165}$$

converges in distribution to $\boldsymbol{g}$.

When dealing with convergence in distribution of random vectors, the standard path is to reduce the vector problem to a scalar problem through the following argument. In view of Lévy's continuity theorem for random vectors, convergence in distribution takes place if, and only if, convergence of the pertinent (multivariate) characteristic functions takes place [73]. This implies that[5] our claim will be proved if we show that, for any sequence of real numbers

[5]This corollary of Lévy's continuity theorem is also known as Cramér-Wold device or theorem [73][Th. 1.9, p. 56].

$t(\theta_1), t(\theta_2), \ldots, t(\theta_{H-1})$:

$$\sum_{\theta\neq\theta_0} t(\theta)\,\frac{\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta)-\mathsf{m}_{\rm ave}(\theta)}{\sqrt{\delta}} \xrightarrow[\delta\to 0]{\rm d} \sum_{\theta\neq\theta_0} t(\theta)\boldsymbol{g}(\theta). \tag{5.166}$$

Obviously, the linear combination on the RHS in (5.166) is a Gaussian random variable with zero mean and with variance:

$$\mathsf{VAR}\left(\sum_{\theta\neq\theta_0} t(\theta)\boldsymbol{g}(\theta)\right) = \sum_{\theta\neq\theta_0}\sum_{\theta'\neq\theta_0} t(\theta)t(\theta')\frac{\mathsf{c}_{\rm ave}(\theta,\theta')}{2}. \tag{5.167}$$

Let us now examine the LHS in (5.166). Using (5.160) we get:

$$\sum_{\theta\neq\theta_0} t(\theta)\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) = \sum_{\ell=1}^{K}\delta\sum_{m=0}^{\infty}(1-\delta)^m[A^{m+1}]_{\ell k}\sum_{\theta\neq\theta_0} t(\theta)\boldsymbol{x}_{\ell,m+1}(\theta), \tag{5.168}$$

whereas using (5.13) we have:

$$\sum_{\theta\neq\theta_0} t(\theta)\mathsf{m}_{\rm ave}(\theta) = \sum_{\ell=1}^{K}\pi_\ell\sum_{\theta\neq\theta_0} t(\theta)d_\ell(\theta). \tag{5.169}$$

Let us now set:

$$\boldsymbol{z}_m^{(\ell)} \triangleq \sum_{\theta\neq\theta_0} t(\theta)\boldsymbol{x}_{\ell,m+1}(\theta), \tag{5.170}$$

$$\alpha_m^{(\ell)} \triangleq [A^{m+1}]_{\ell k}, \tag{5.171}$$

$$\boldsymbol{s}^{(\ell)}(\delta) \triangleq \delta\sum_{m=0}^{\infty}(1-\delta)^m\alpha_m^{(\ell)}\boldsymbol{z}_m^{(\ell)}. \tag{5.172}$$

We observe that:

$$\mathbb{E}\left(\boldsymbol{z}_m^{(\ell)}\right) = \sum_{\theta\neq\theta_0} t(\theta)d_\ell(\theta), \tag{5.173}$$

$$\mathsf{VAR}\left(\boldsymbol{z}_m^{(\ell)}\right) = \sum_{\theta\neq\theta_0}\sum_{\theta'\neq\theta_0} t(\theta)t(\theta')\rho_\ell(\theta,\theta'). \tag{5.174}$$

Exploiting Eqs. (5.170)–(5.173), the LHS in (5.166) can be cast in the form:

$$\sum_{\ell=1}^{K}\frac{\boldsymbol{s}^{(\ell)}(\delta)-\mathbb{E}\left(\boldsymbol{z}_m^{(\ell)}\right)}{\sqrt{\delta}}. \tag{5.175}$$

We see from Eqs. (5.170)–(5.172) that the random variables $\boldsymbol{s}^{(\ell)}(\delta)$ match the structure of the random series used in Lemma 5.1. We now verify that $\boldsymbol{s}^{(\ell)}(\delta)$ fulfills the conditions of part 5 in Lemma 5.1, for every $\ell = 1, 2, \ldots, K$. First we note that $\boldsymbol{z}_m^{(\ell)}$ has finite variance since it is a linear combination of random variables that have finite variance. Second we see that condition (5.95) is verified in view of Property 2.1. We conclude then from part 5 of Lemma 5.1 that the

following convergence in distribution holds:

$$\frac{\boldsymbol{s}^{(\ell)}(\delta) - \mathbb{E}\left(\boldsymbol{z}_m^{(\ell)}\right)}{\sqrt{\delta}} \xrightarrow[\delta\to 0]{\mathrm{d}} \mathscr{G}\left(0, \frac{\pi_\ell^2}{2}\mathsf{VAR}\left(\boldsymbol{z}_m^{(\ell)}\right)\right). \tag{5.176}$$

Since the data are independent across agents, we have that the random variables $\boldsymbol{s}^{(\ell)}(\delta)$ are independent across index ℓ. For this reason, and in view of (5.176), we conclude that the LHS in (5.166) is asymptotically normal, with zero mean and with variance given by:

$$\begin{aligned}\frac{\pi_\ell^2}{2}\sum_{\ell=1}^{K}\mathsf{VAR}\left(\boldsymbol{z}_m^{(\ell)}\right) &= \sum_{\theta\neq\theta_0}\sum_{\theta'\neq\theta_0} t(\theta)t(\theta')\sum_{\ell=1}^{K}\frac{\pi_\ell^2}{2}\rho_\ell(\theta,\theta') \\ &= \sum_{\theta\neq\theta_0}\sum_{\theta'\neq\theta_0} t(\theta)t(\theta')\frac{\mathsf{c}_{\mathrm{ave}}(\theta,\theta')}{2},\end{aligned} \tag{5.177}$$

where we have used (5.174). Since the RHS in (5.177) coincides with the variance in (5.167), the proof is complete.

5.E Proof of Theorem 5.4

In light of (5.10), the error probability of *not* choosing θ_0 can be bounded as follows (with the lower bound holding for every $\theta \neq \theta_0$):

$$\mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) \leq 0\right) \leq p_{k,i}^{(\delta)} \leq \sum_{\theta\neq\theta_0}\mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) \leq 0\right), \tag{5.178}$$

where the upper bound is the union bound. At the steady state, Eq. (5.178) implies:

$$\mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_{k}^{(\delta)}(\theta) \leq 0\right) \leq p_{k}^{(\delta)} \leq \sum_{\theta\neq\theta_0}\mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_{k}^{(\delta)}(\theta) \leq 0\right). \tag{5.179}$$

One key point to prove the claim of the theorem is the exponential characterization of the probability $\mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_{k}^{(\delta)}(\theta) \leq 0\right)$. Preliminarily, let us set:

$$\boldsymbol{z}_m^{(\ell)} \triangleq \boldsymbol{x}_{\ell,m+1}(\theta), \tag{5.180}$$

$$\alpha_m^{(\ell)} \triangleq [A^{m+1}]_{\ell k}, \tag{5.181}$$

$$\boldsymbol{s}^{(\ell)}(\delta) \triangleq \delta\sum_{m=0}^{\infty}(1-\delta)^m\alpha_m^{(\ell)}\boldsymbol{z}_m^{(\ell)}, \tag{5.182}$$

which yields:

$$\widetilde{\boldsymbol{\lambda}}_{k}^{(\delta)}(\theta) = \sum_{\ell=1}^{K}\boldsymbol{s}^{(\ell)}(\delta). \tag{5.183}$$

Recall that the log-likelihood ratios $\boldsymbol{x}_{\ell,m}(\theta)$ are assumed to be independent across agents (i.e., across ℓ). Thus $\boldsymbol{s}^{(\ell)}(\delta)$ are also independent random variables. Now, part 6 of Lemma 5.1

would provide the required exponential characterization for the individual variable $\boldsymbol{s}^{(\ell)}(\delta)$. We need instead the characterization for $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta)$, which is the sum of the (independent) variables $\boldsymbol{s}^{(\ell)}(\delta)$. Let us elaborate on this aspect. The starting point to prove part 6 in Lemma 5.1 is the convergence in (5.104). Exploiting additivity of the LMGF for independent variables, we conclude that the LMGF of $\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}$, scaled by δ and evaluated at t/δ, converges to the sum:

$$\sum_{\ell=1}^{K} \int_0^t \frac{\Lambda_\ell(\pi_\ell \tau;\theta)}{\tau} d\tau = \int_0^t \frac{\Lambda_{\text{ave}}(\tau;\theta)}{\tau} d\tau \triangleq \phi(t;\theta), \tag{5.184}$$

where: **i)** we used the fact that the LMGF of $\boldsymbol{z}_m^{(\ell)}$ is $\Lambda_\ell(t;\theta)$; **ii)** the intermediate equality comes from (5.45) (having exchanged the integral with the sum); and **iii)** the last equality comes from (5.47). Moreover, the properties of the rate function in part 6 of Lemma 5.1 depend only on the fact that $\Lambda_{\alpha z}(t)$ is a logarithmic moment generating function that is finite for all $t \in \mathbb{R}$. Since $\Lambda_{\text{ave}}(\tau;\theta)$ is the LMGF of the average variable $\boldsymbol{x}_{\text{ave},i}(\theta)$ (and is finite for all $t \in \mathbb{R}$ by assumption), all the remaining results in part 6 of Lemma 5.1 hold true, provided that the properties pertaining to $\alpha \boldsymbol{z}_m$ are now referred to $\boldsymbol{x}_{\text{ave},i}(\theta)$.

We conclude that it is legitimate to use the exponential characterization provided in Lemma 5.1. In particular, since we have $\gamma = 0 < \mathsf{m}_{\text{ave}}(\theta)$, the pertinent relation is given by (5.110) with the choice $\gamma = 0$, yielding:

$$\lim_{\delta \to 0} \delta \log \mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) \leq 0 \right) = -\Phi(\theta), \tag{5.185}$$

where the exponent $\Phi(\theta)$ is accordingly computed as the value of the rate function at $\gamma = 0$, namely,

$$\Phi(\theta) = \sup_{t \in \mathbb{R}} [-\phi(t;\theta)] = -\inf_{t \in \mathbb{R}} \phi(t;\theta). \tag{5.186}$$

Using the lower bound in (5.179), we can readily conclude from (5.185) and from the definitions appearing in (5.49) and (5.186) that:

$$\liminf_{\delta \to 0} \delta \log p_k^{(\delta)} \geq \max_{\theta \neq \theta_0} \Big(-\Phi(\theta) \Big) = -\min_{\theta \neq \theta_0} \Phi(\theta) = -\Phi. \tag{5.187}$$

Let us now focus on the upper bound in (5.179). By definition, for all $\theta \neq \theta_0$ we have that $\Phi \leq \Phi(\theta)$. Accordingly, the convergence in (5.110) implies that, given an arbitrary $\epsilon > 0$, for sufficiently small δ we can write:

$$\mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_k^{(\delta)}(\theta) \leq 0 \right) \leq e^{-(1/\delta)(\Phi - \epsilon)}. \tag{5.188}$$

Exploiting (5.188), the upper bound in (5.179) yields:

$$\delta \log p_k^{(\delta)} \leq \delta \log(H-1) - \Phi + \epsilon, \tag{5.189}$$

where we recall that H is the number of hypotheses or admissible models. Due to the arbitrariness of ϵ, we have:

$$\limsup_{\delta \to 0} \delta \log p_k^{(\delta)} \leq -\Phi. \tag{5.190}$$

Bridging (5.187) and (5.190) implies the desired claim.

5.F Proof of Theorem 5.5

We start by proving an auxiliary lemma.

Lemma 5.2 (Useful properties of the LMGF $\Lambda_\ell(t;\theta)$). *The lemma is proved under the same assumptions used in Theorem 5.4. Let*

$$\Lambda_\ell(t;\theta) = \log \mathbb{E}\left(e^{t\,\boldsymbol{x}_{\ell,i}(\theta)}\right) = \log \mathbb{E}\left(e^{t \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)}}\right) \tag{5.191}$$

be the LMGF of the log-likelihood at the ℓ-th agent, let

$$\Lambda_{\text{ave}}(t;\theta) = \log \mathbb{E}\left(e^{t\,\boldsymbol{x}_{\text{ave},i}(\theta)}\right) = \sum_{\ell=1}^{K} \Lambda_\ell(\pi_\ell t;\theta) \tag{5.192}$$

be the LMGF of the network average of log-likelihoods, $\boldsymbol{x}_{\text{ave},i}(\theta) = \sum_{\ell=1}^{K} \pi_\ell \boldsymbol{x}_{\ell,i}(\theta)$, and let:

$$\phi(t;\theta) = \int_0^t \frac{\Lambda_{\text{ave}}(\tau;\theta)}{\tau} d\tau. \tag{5.193}$$

Then, we have the following properties:

P1) The error exponent $\Phi(\theta)$ is given by:

$$\Phi(\theta) = -\inf_{t\in\mathbb{R}} \phi(t;\theta) = -\phi(t^\star_\theta;\theta), \tag{5.194}$$

where $t^\star_\theta < 0$ is the unique solution to:

$$\frac{\Lambda_{\text{ave}}(t^\star_\theta;\theta)}{t^\star_\theta} = 0. \tag{5.195}$$

P2) For all $t \in \mathbb{R}$ we have:

$$\Lambda_\ell(t;\theta) \geq d_\ell(\theta) t, \tag{5.196}$$

implying in particular that:

$$\Phi(\theta) \leq |t^\star_\theta| \mathsf{m}_{\text{ave}}(\theta). \tag{5.197}$$

P3) Let $\pi_{\min}$ and $\pi_{\max}$ be the minimum and maximum entry of the Perron eigenvector, respectively. Then we have:

$$\frac{1}{\pi_{\max}} \leq |t^\star_\theta| \leq \frac{1}{\pi_{\min}}. \tag{5.198}$$

Proof. From the convexity properties of $\phi(t;\theta)$ (see [79], [82] for a detailed summary) we know

that the infimum of $\phi(t;\theta)$ in (5.194) is in fact a unique minimum located at the solution $t_\theta^\star$ to the stationary equation:

$$\phi'(t_\theta^\star;\theta) = 0, \tag{5.199}$$

where $'$ denotes derivative w.r.t. t. Therefore, Eq. (5.195) follows from (5.193). On the other hand, in view of the convexity properties of $\phi(t;\theta)$, the function $\phi'(t;\theta)$ is strictly increasing in t, and since $\phi'(0;\theta) = \Lambda'_{\rm ave}(0;\theta) = \mathsf{m}_{\rm ave}(\theta) > 0$ (we use the fact that the first derivative of the LMGF evaluated in 0 is equal to the mean of the relative random variable), from (5.199) we conclude that the value $t_\theta^\star$ that minimizes the function $\phi(t)$ in (5.194) cannot but be negative, and the proof of property P1) is complete [79], [82].

Regarding property P2), from the convexity of the local LMGF $\Lambda_\ell(t;\theta)$ we can write, for all $t \in \mathbb{R}$:

$$\Lambda_\ell(t;\theta) \geq t\Lambda'_\ell(0;\theta) = t d_\ell(\theta). \tag{5.200}$$

Exploiting (5.192), (5.193), (5.194), and (5.200), we obtain:

$$\begin{aligned}\Phi(\theta) = -\phi(t_\theta^\star;\theta) &= -\int_0^{t_\theta^\star} \frac{\Lambda_{\rm ave}(\tau;\theta)}{\tau} d\tau \\ &= \sum_{\ell=1}^K \int_{t_\theta^\star}^0 \frac{\Lambda_\ell(\pi_\ell\tau;\theta)}{\tau} d\tau \\ &\leq |t_\theta^\star| \sum_{\ell=1}^K \pi_\ell d_\ell(\theta) = |t_\theta^\star| \mathsf{m}_{\rm ave}(\theta),\end{aligned} \tag{5.201}$$

and property P2) is proved.

Finally we prove property P3). Making explicit the definition of $\Lambda_{\rm ave}(t;\theta)$, Eq. (5.195) can be written as:

$$\frac{\sum_{\ell=1}^K \Lambda_\ell(\pi_\ell t_\theta^\star;\theta)}{t_\theta^\star} = 0. \tag{5.202}$$

In view of (5.191), with expectation computed under the model $L_\ell(\xi|\theta_0)$, we have that $\Lambda_\ell(-1;\theta) = 0$. Accordingly, when $\pi_\ell = 1/K$ for all ℓ, Eq. (5.198) is obvious. Let us focus on the case where the Perron eigenvector is not uniform. From the strict convexity of $\Lambda_\ell(t;\theta)$, we know that:

$$\begin{aligned}&\Lambda_\ell(\pi_\ell t;\theta) > 0 \ \text{ for } t < -\frac{1}{\pi_\ell}, \\ &\Lambda_\ell(\pi_\ell t;\theta) < 0 \ \text{ for } -\frac{1}{\pi_\ell} < t < 0,\end{aligned} \tag{5.203}$$

see Figure 5.13. Since the equality in (5.202) requires that $\Lambda_\ell(\pi_\ell t_\theta^\star;\theta)$ takes on at least one positive and one negative value, Eq. (5.203) implies property P3). □

Proof of Theorem 5.5. From (5.159) we know that $\widehat{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta)$ and $\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta)$ share the same distribution.

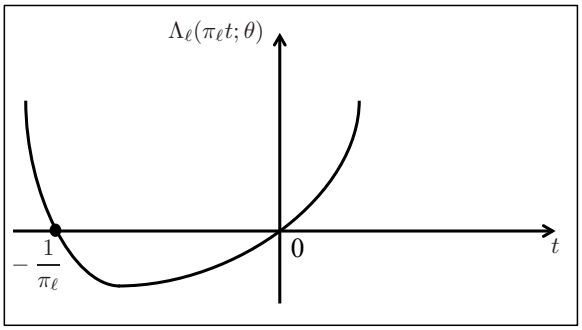


Figure 5.13: Typical shape of the LMGF of the ℓ-th likelihood.

Thus, from (5.18) and (5.19) we have that (we recall that $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution):

$$\begin{aligned}\boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta) &\stackrel{\mathrm{d}}{=} \widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) + (1-\delta)^i \sum_{\ell=1}^{K} [A^i]_{\ell k} \lambda_{\ell,0}(\theta) \\ &\geq \widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) + (1-\delta)^i \sum_{\ell=1}^{K} \pi_\ell \lambda_{\ell,0}(\theta) - \kappa(1-\delta)^i \beta^i \sum_{\ell=1}^{K} |\lambda_{\ell,0}(\theta)| \\ &= \widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) + (1-\delta)^i \sum_{\ell=1}^{K} \pi_\ell \lambda_{\ell,0}(\theta) - \frac{\mathsf{K}_2(\theta)}{|t_\theta^\star|}(1-\delta)^i \beta^i, \end{aligned} \tag{5.204}$$

where the inequality follows from Property 2.1, and in the last equality we used (5.64). In view of (5.204), and since $t_\theta^\star < 0$, we can write:

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{\lambda}_{k,i}^{(\delta)}(\theta) \leq 0\right) &\leq \mathbb{P}\left(\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) \leq -(1-\delta)^i \lambda_{\mathrm{ave},0}(\theta) + \frac{\mathsf{K}_2(\theta)}{|t_\theta^\star|}(1-\delta)^i \beta^i\right) \\ &\stackrel{\text{(a)}}{=} \mathbb{P}\left(\frac{t_\theta^\star}{\delta}\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta) \geq \frac{|t_\theta^\star|}{\delta}(1-\delta)^i \lambda_{\mathrm{ave},0}(\theta) - \frac{\mathsf{K}_2(\theta)}{\delta}(1-\delta)^i \beta^i\right) \\ &\stackrel{\text{(b)}}{\leq} \frac{\mathbb{E}\left(\exp\left\{\frac{t_\theta^\star}{\delta}\,\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta)\right\}\right)}{\exp\left\{\frac{|t_\theta^\star|}{\delta}\,(1-\delta)^i \lambda_{\mathrm{ave},0}(\theta) - \frac{\mathsf{K}_2(\theta)}{\delta}(1-\delta)^i \beta^i\right\}} \\ &\stackrel{\text{(c)}}{=} e^{\frac{1}{\delta}\left[\delta\Lambda_{k,i}^{(\delta)}\left(\frac{t_\theta^\star}{\delta};\theta\right) - (1-\delta)^i |t_\theta^\star| \lambda_{\mathrm{ave},0} + \mathsf{K}_2(\theta)(1-\delta)^i \beta^i\right]}, \end{aligned} \tag{5.205}$$

where (a) follows from multiplying by $t_\theta^\star/\delta$ both sides of the inequality in the probability brackets and taking into account the fact that $t_\theta^\star < 0$; (b) follows from applying Chernoff's bound; and in (c) we applied Property P3) and introduced the LMGF of $\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta)$, which can be explicitely defined as:

$$\begin{aligned}\Lambda_{k,i}^{(\delta)}(t;\theta) &= \log \mathbb{E}\left(e^{t\widetilde{\boldsymbol{\lambda}}_{k,i}^{(\delta)}(\theta)}\right) \\ &= \sum_{\ell=1}^{K} \sum_{m=0}^{i-1} \Lambda_\ell\left(\delta(1-\delta)^m [A^{m+1}]_{\ell k} t\,;\,\theta\right), \end{aligned} \tag{5.206}$$

with $\Lambda_\ell(t;\theta)$ being the LMGF of the log-likelihood ratio $\boldsymbol{x}_{\ell,m+1}(\theta)$. Now, letting

$$c_i \triangleq (1-\delta)^{i-1}, \tag{5.207}$$

and applying Eqs. (85) and (86) from [82] to the inner summation in (5.206), we have the following representation:

$$\begin{aligned}
\Lambda_{k,i}^{(\delta)}\left(\frac{t_\theta^\star}{\delta};\theta\right) &= \frac{1}{\delta}\left[\sum_{\ell=1}^{K}\int_{c_i\pi_\ell t_\theta^\star}^{\pi_\ell t_\theta^\star}\frac{\Lambda_\ell(\tau;\theta)}{\tau}d\tau + O(\delta)\right] \\
&\overset{\text{(a)}}{=} \frac{1}{\delta}\left[\phi(t_\theta^\star;\theta) - \sum_{\ell=1}^{K}\int_{0}^{c_i\pi_\ell t_\theta^\star}\frac{\Lambda_\ell(\tau;\theta)}{\tau}d\tau + O(\delta)\right] \\
&\overset{\text{(b)}}{=} \frac{1}{\delta}\left[-\Phi(\theta) + \int_{-c_i\pi_\ell|t_\theta^\star|}^{0}\frac{\Lambda_\ell(\tau;\theta)}{\tau}d\tau + O(\delta)\right] \\
&\overset{\text{(c)}}{\leq} \frac{1}{\delta}\left[-\Phi(\theta) + c_i|t_\theta^\star|\sum_{\ell=1}^{K}\pi_\ell d_\ell(\theta) + O(\delta)\right],
\end{aligned} \tag{5.208}$$

where (a) follows from (5.193), while (b) and (c) from properties P1) and P2) in Lemma 5.2, respectively. Using now (5.208) in (5.205) and using the definition of $\mathsf{K}_1(\theta)$ in (5.63) we get the upper bound in (5.65). □

5.G Proof of Corollary 5.1

We now determine the adaptation time as the critical instant after which we stay close to the exponent Φ, in the precise sense specified by (5.67). Let us consider first the case where $\lambda_{\text{ave}}(\theta) \geq \mathsf{m}_{\text{ave}}(\theta)$ for all $\theta \neq \theta_0$. In this case, we have $\mathsf{K}_1(\theta) \leq 0$ for all θ and, hence, in view of (5.65), condition (5.67) will be met if we ensure that:

$$i > \frac{1}{\log\beta^{-1}}\log\frac{\mathsf{K}_2}{\epsilon\,\Phi} \Rightarrow \mathsf{K}_2\,\beta^i < \epsilon\Phi, \tag{5.209}$$

which shows that the choice for $\mathsf{T}_{\mathsf{ASL}}$ in (5.68) guarantees (5.67) for all $i > \mathsf{T}_{\mathsf{ASL}}$.

We continue by examining the unfavorable case where $\lambda_{\text{ave}}(\theta) < \mathsf{m}_{\text{ave}}(\theta)$ for at least one value $\theta \neq \theta_0$. In this case we have $\mathsf{K}_1 = \max_{\theta\neq\theta_0}\mathsf{K}_1(\theta) > 0$, and we can write:

$$i > \frac{1}{\log(1-\delta)^{-1}}\log\frac{\mathsf{K}_1}{\epsilon\,\Phi} \Rightarrow (1-\delta)^i\mathsf{K}_1 < \epsilon\Phi. \tag{5.210}$$

Then, if we set the adaptation time $\mathsf{T}_{\mathsf{ASL}}$ according to the law in (5.210), the quantity β^i appearing in (5.65) would decay to zero as $\approx \beta^{1/\delta}$, and, hence, would be incorporated into the higher-order term $O(\delta)$, and the claim of the corollary is proved.

6 Learning with Imperfect Models

6.1 Introduction[1]

The social learning problem, introduced in Chapter 1, can be cast into the problem of decentralized classification of streaming observations. In this framework, hypotheses are replaced by *classes* and observations by *features*. The network of agents aims to find the class that best explains the growing number of observed features. For example, a network of cameras is recording a particular road intersection, and the network is trying to detect whether at any time an accident takes place. In this case, the possible classes are {accident, normal traffic} and the features are RGB frames captured by the cameras.

Social learning solutions, discussed in the previous chapters, require however prior knowledge of the true probability distributions characterizing the received features, referred to as *likelihoods*, which are in general not available in real-world applications. In practice, these models are only approximate, oftentimes the result of a previous *training* stage, where, from limited data, a parameterized model is learned.

In this chapter, we propose the Social Machine Learning (SML) strategy, which is a decentralized algorithm for combining the outputs of a *heterogeneous* network of classifiers over space and time, based on the social learning algorithms proposed in [41], [42], [76], [77]. The network is heterogeneous in two main aspects: First, agents may be observing different (possibly non-overlapping) sets of attributes of the same observed scene; Second, their statistical models need not be the same, e.g., agents may be observing the same attribute from different perspectives, which allows for a *distribution diversity* across agents. The strategy consists of two phases: A *training phase*, in which the classifiers are independently trained given a finite set of labeled data samples, and a *prediction phase*, in which the trained classifiers are deployed in a collaborative structure while observing streaming unlabeled samples.

The SML strategy proposed in this chapter inherits the following qualities from social learning: **i)** It is able to combine *heterogeneous* classifiers, i.e., using features with different dimensions and statistical models; **ii)** It can *adapt* in view of non-stationary conditions, i.e., under changing real-time measurements; **iii)** It has asymptotic performance guarantees, achieving *consistent*

[1] This chapter is adapted from [92], [93].

learning with high probability despite the imperfectly trained models; **iv)** It allows for continuous *accuracy improvement* as the number of prediction samples grows.

6.1.1 Related Work

The issue of considering imperfect likelihood models in social learning is recognized in the works [94], [95], where the authors propose a framework for incorporating uncertainty into non-Bayesian social learning. While [94] focuses only on sets of Gaussian distributions, their proposed strategy in [95] broadens the approach, but requires nonetheless prior knowledge about the structure of the likelihoods models, i.e., the exact parameterization of the distributions. While relevant for numerically generated data, in practical applications there is generally little *a priori* evidence regarding the structure of likelihoods, e.g., in the distributed classification of images or videos. Our proposed SML strategy, on the other hand, has the advantage of allowing the use of a fairly general class of distributions, which is relevant in practical machine learning tasks.

In the strategy we propose, agents (or classifiers) cooperate with neighbors to overcome *local* spatial limitations. They also aggregate their *instantaneous* opinions from streaming observations, strengthening their decision-making capabilities over time. These two aspects, i.e., information aggregation over space and time, are common topics of research in the fields of ensemble [96] and multi-view learning [97].

Popular examples of ensemble approaches are bagging [98] and boosting [99], in which classifiers combine weighted decisions across *space*. However, such combination takes place in a *centralized* manner, namely, it is assumed that all agents communicate their decisions to a fusion center. This mechanism is fundamentally different from the *fully decentralized* setting addressed here, where only local cooperation between neighboring agents is permitted, and each individual agent is eventually able to learn the correct class. Moreover, both bagging and boosting methods do not address the streaming data case, i.e., they do not leverage the temporal quality of the online observations.

In multi-view learning, multiple views of the same data are available, which are jointly used to improve generalization performance. Multi-view co-training approaches [100] are notably suitable for semi-supervised learning, where a substantial number of unlabeled samples are available. In these approaches, distinct classifiers are trained on different views, and one classifier's predictions on new unlabeled examples are used to enlarge the labeled training set of the other. The procedure is repeated over the unlabeled samples, improving their accuracy over successive iterations. However, multi-view learning does not address the *decentralized* and *streaming-data* aspects. Regarding the former aspect, in multi-view learning the classifiers are not spatially distributed or, if they are, it is simply assumed that they can share their beliefs without any constraints (i.e., as if they were co-located). Regarding the latter aspect, multi-view learning does not assume that streaming data are available for prediction.

6.2 Problem Setting

In this section, we revisit some concepts introduced in Chapter 2 and present the social learning problem from a classification perspective.

6.2.1 Inference Problem

We consider a network of K agents or classifiers, indexed by $k \in \{1, 2, \ldots, K\}$, trying to identify the true state of nature γ_0 out of a binary set of hypotheses or classes $\Gamma = \{-1, +1\}$. The true state characterizes the *scene* all agents are observing. To make a decision on the true state, each agent relies on the observation of streaming private data, which are features reflecting on the observed scene. Data are qualified as private due to the implicit assumption that raw observations cannot be shared among agents in order to, for example, minimize communication costs or preserve secrecy.

More specifically, each agent k observes at each instant i the feature vector $\boldsymbol{h}_{k,i} \in \mathcal{H}_k$. The feature vectors are assumed to be independent and identically distributed (i.i.d.) over time. Moreover, the features $\boldsymbol{h}_{k,i}$ at agent k given the state γ_0 form a sequence of i.i.d. random vectors distributed according to some conditional distribution (or likelihood):

$$\boldsymbol{h}_{k,i} \sim L_k(h|\gamma_0), \quad h \in \mathcal{H}_k, \gamma_0 \in \Gamma. \tag{6.1}$$

Notably, the model allows the features to be dependent across agents. The feature set $\mathcal{H}_k$ is particular to agent k, allowing agents to observe different attributes from the same scene; in particular, the dimension of $\mathcal{H}_k$ can be generally different across the agents. For example, an agent might be observing RGB video frames while another might be receiving infrared imagery taken both from the same street scene. Another source of heterogeneity is the likelihood model $L_k(h|\gamma)$, which differs across agents and reflects their individual perceptions. Within the previous street scene example, agents might observe frames captured under different, possibly non-overlapping, fields of view.

We can treat the true state of nature as a random variable $\boldsymbol{\gamma}_0$ and furthermore establish that the pair $(\boldsymbol{h}_{k,i}, \boldsymbol{\gamma}_0)$ is distributed according to the following joint distribution:

$$(\boldsymbol{h}_{k,i}, \boldsymbol{\gamma}_0) \sim p_k(h, \gamma) = L_k(h|\gamma)p_k(\gamma), \tag{6.2}$$

with $h \in \mathcal{H}_k,\ \gamma \in \Gamma$, for every $i = 1, 2, \ldots$ due to the i.i.d. assumption over time. Here, the notation $p_k(\gamma)$ corresponds to the prior distribution at agent k for $\boldsymbol{\gamma}_0$ over the discrete set of hypotheses Γ.

If the likelihood and prior distributions are perfectly known to agent k, different strategies can be deployed to enable truth learning. In a noncooperative framework, where each agent has enough information to solve the problem on their own, we can resort to the *Bayes classifier*. Alternatively, to leverage the data spread across different agents of the network, a cooperative strategy can be used, such as one of the existing *social learning* methods. We discuss each of the two strategies in more detail in the next paragraphs.

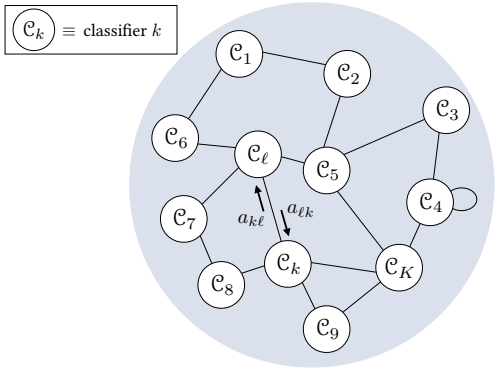


Figure 6.1: Diagram of the network of classifiers.

6.2.2 Bayes Classifier

When the KL divergence [16] between likelihoods $L_k(h|+1)$ and $L_k(h|-1)$ is strictly positive, we say that agent k possesses *informative* likelihoods and can therefore distinguish classes $+1$ and -1. Therefore, if the likelihood and prior distributions are known to agent k and its likelihoods are informative, the agent can employ the Bayes classifier to solve the following maximum-a-posteriori (MAP) problem given an observed sequence of features $\{\boldsymbol{h}_{k,j}\}$ with $j = 1, 2, \ldots, i$:

$$\boldsymbol{\gamma}_{k,i}^{\text{Bayes}} = \arg\max_{\gamma \in \Gamma} p_k(\gamma | \boldsymbol{h}_{k,1}, \boldsymbol{h}_{k,2}, \ldots, \boldsymbol{h}_{k,i}), \tag{6.3}$$

where $p_k(\gamma|\boldsymbol{h}_{k,1}, \boldsymbol{h}_{k,2}, \ldots, \boldsymbol{h}_{k,i})$ indicates the posterior probability of the event $\{\boldsymbol{\gamma} = \gamma\}$ given the sequence $\{\boldsymbol{h}_{k,j}\}$ with $j = 1, 2, \ldots, i$.

In Section 1.1.1, we have seen that the Bayes classifier, or MAP estimator, can be obtained from the recursive Bayesian update, and it learns the true underlying class with probability one asymptotically. While this result ensures consistent learning, it requires each individual agent to have informative likelihoods and thus to be able to distinguish both hypotheses. This restriction motivates the pursuit of *collaborative* social learning schemes, where agents exchange information to resolve ambiguities arising from incomplete information.

6.2.3 Social Learning

In a multi-agent setup, a network of agents is modeled as a strongly connected graph (Figure 6.1)—see Chapter 2 for a more detailed discussion on strongly connected networks. Since the $K-$agent network is collectively observing streaming features

$$\boldsymbol{h}_{1,i}, \boldsymbol{h}_{2,i}, \ldots, \boldsymbol{h}_{K,i}, \tag{6.4}$$

it is beneficial for agents to cooperate to solve the inference problem. Cooperation allows agents to aggregate K times more data at every instant and it furthermore enables successful learning even when the decision problem is not identifiable for the *individual* agents, but is *globally* identifiable at the network level.

In *non-adaptive social learning* [40]–[43], at every instant i, each agent k updates its *belief* $\varphi_{k,i}(\gamma)$, i.e., a probability mass function over the set of classes Γ, according to a two-step protocol:

$$\boldsymbol{\psi}_{k,i}(\gamma) = \frac{\varphi_{k,i-1}(\gamma)L_k(\boldsymbol{h}_{k,i}|\gamma)}{\sum_{\gamma'\in\Gamma}\varphi_{k,i-1}(\gamma')L_k(\boldsymbol{h}_{k,i}|\gamma')}, \tag{6.5}$$

$$\boldsymbol{\varphi}_{k,i}(\gamma) = \frac{\exp\left\{\sum_{\ell=1}^{K} a_{\ell k}\log\boldsymbol{\psi}_{\ell,i}(\gamma)\right\}}{\sum_{\gamma'\in\Gamma}\exp\left\{\sum_{\ell=1}^{K} a_{\ell k}\log\boldsymbol{\psi}_{\ell,i}(\gamma')\right\}}, \tag{6.6}$$

where in the first step (Eq. (6.5)) agent k updates its *intermediate belief* $\boldsymbol{\psi}_{k,i}$ using the observed feature vector $\boldsymbol{h}_{k,i}$. Then in the second step (Eq. (6.6)), agents share their intermediate beliefs with neighboring agents and update their beliefs using a geometric averaging rule.

An equivalent linear way of representing (6.5) and (6.6) is in the form of the *diffusion strategy* [36], [37]:

$$\boldsymbol{\eta}_{k,i} = \boldsymbol{\lambda}_{k,i-1} + c_k(\boldsymbol{h}_{k,i}), \tag{6.7}$$

$$\boldsymbol{\lambda}_{k,i} = \sum_{\ell=1}^{K} a_{\ell k}\boldsymbol{\eta}_{\ell,i}, \tag{6.8}$$

in terms of the following scalar quantities:

$$\boldsymbol{\lambda}_{k,i} \triangleq \log\frac{\boldsymbol{\varphi}_{k,i}(+1)}{\boldsymbol{\varphi}_{k,i}(-1)}, \quad \boldsymbol{\eta}_{k,i} \triangleq \log\frac{\boldsymbol{\psi}_{k,i}(+1)}{\boldsymbol{\psi}_{k,i}(-1)}, \tag{6.9}$$

$$c_k(\boldsymbol{h}_{k,i}) \triangleq \log\frac{L_k(\boldsymbol{h}_{k,i}|+1)}{L_k(\boldsymbol{h}_{k,i}|-1)}. \tag{6.10}$$

Equations (6.9) and (6.10) can be joined into a single equation as:

$$\boldsymbol{\lambda}_{k,i} = \sum_{\ell=1}^{K} a_{\ell k}\Big(\boldsymbol{\lambda}_{\ell,i-1} + c_\ell(\boldsymbol{h}_{\ell,i})\Big). \tag{6.11}$$

Following the discussion in Chapter 2, by developing the recursion in (6.11) it is possible to show that, as $i \to \infty$, the belief function is maximized at the true hypothesis, provided that a weighted combination (through the Perron eigenvector weights) of the detection statistics $c_\ell(\boldsymbol{h}_{\ell,i})$ has positive expectation under hypothesis $+1$ and negative expectation under -1,

namely,[2]

$$\boxed{\sum_{\ell=1}^{K} \pi_\ell \mathbb{E}_{\,L_\ell(+1)} c_\ell(\boldsymbol{h}_{\ell,i}) > 0, \;\; \sum_{\ell=1}^{K} \pi_\ell \mathbb{E}_{\,L_\ell(-1)} c_\ell(\boldsymbol{h}_{\ell,i}) < 0} \tag{6.13}$$

where $\mathbb{E}_{\,L_\ell(\gamma)}$ indicates that the expectation is computed with respect to the distribution $L_\ell(h|\gamma)$. We remark that the condition for consistency in (6.13) would apply to general detection statistics $c_\ell(\cdot)$, and not only to log-likelihood ratios as in (6.10). For example, detection statistics different from (6.10) may arise because the agents compute *mismatched* log-likelihood ratios due to imperfect knowledge. This observation is particularly relevant in our work since, when we will examine the social machine learning setting (where the likelihoods are unknown) we will need to work with general detection statistics learned from a training set. On the other hand, for the specific case where the likelihoods are known and (6.10) is employed, the conditions in (6.13) are satisfied whenever the network satisfies the *global identifiability* assumption(Assumption 2.5), i.e., at least one agent in the network is able to distinguish the hypotheses. In this case, for at least one agent k, it follows that

$$\mathbb{E}_{\,L_k(+1)} c_k(\boldsymbol{h}_{k,i}) = D(L_k(+1)||L_k(-1)) > 0 \tag{6.14}$$

$$\mathbb{E}_{\,L_k(-1)} c_k(\boldsymbol{h}_{k,i}) = -D(L_k(-1)||L_k(+1)) < 0. \tag{6.15}$$

From the positivity of the Perron eigenvector π, (6.14) and (6.15) imply that (6.13) is satisfied. Therefore, the strategy in (6.5) and (6.6) allows agents to learn the truth asymptotically, as i tends to infinity, with probability one [40], [41]. As already discussed in Chapter 5, the implementation above is suitable only for a *stationary* world.

In a real-time application, we expect the environment conditions to change with time, and the learning strategy should be able to track the drifting conditions within a reasonable response time. In Chapter 5, an *adaptive social learning* strategy was proposed to overcome the lack of adaptation in traditional social learning under *non-stationary* conditions [76], [77].

In one of the formulations seen in Chapter 5, the first step of the update rule in (6.16) is replaced by the *adaptive* update seen in (5.3):

$$\boldsymbol{\psi}_{k,i}(\gamma) = \frac{\varphi_{k,i-1}^{1-\delta}(\gamma) L_k(\boldsymbol{h}_{k,i}|\gamma)}{\sum_{\gamma' \in \Gamma} \varphi_{k,i-1}^{1-\delta}(\gamma') L_k(\boldsymbol{h}_{k,i}|\gamma')}, \tag{6.16}$$

where $0 < \delta \ll 1$ is a small step-size (or learning) parameter. The introduction of a step-size to the local update in (6.16) infuses the algorithm with the ability to adapt in face of non-stationary conditions with an *adaptation time* that scales as $\mathcal{O}(1/\delta)$ [77]. In the limit case, when $\delta \to 0$, we recover the Bayesian update in (6.5).

[2]The sufficient condition for consistent learning in (6.13) can be reached by following similar arguments as in Appendix **??**. Developing the recursion in (6.11) and dividing by i, we can conclude that

$$\frac{1}{i} \boldsymbol{\lambda}_{k,i} \xrightarrow{\text{a.s.}} \sum_{\ell=1}^{K} \pi_\ell \mathbb{E}_{\,L_\ell(\gamma_0)} c_\ell(\boldsymbol{h}_{\ell,i}). \tag{6.12}$$

If $\sum_{\ell=1}^{K} \pi_\ell \mathbb{E}_{\,L_\ell(\gamma_0)} c_\ell(\boldsymbol{h}_{\ell,i}) > 0$, then $\boldsymbol{\lambda}_{k,i}$ goes to $+\infty$, and thus agents decide for class $+1$. Otherwise if $\sum_{\ell=1}^{K} \pi_\ell \mathbb{E}_{\,L_\ell(\gamma_0)} c_\ell(\boldsymbol{h}_{\ell,i}) < 0$, then $\boldsymbol{\lambda}_{k,i}$ goes to $-\infty$, and thus agents decide for class -1 .

Similarly to (6.8), we can represent (6.6) and (6.16) in the form of an *adaptive diffusion strategy* [78]:

$$\boldsymbol{\eta}_{k,i} = (1-\delta)\boldsymbol{\lambda}_{k,i-1} + c_k(\boldsymbol{h}_{k,i}), \tag{6.17}$$

$$\boldsymbol{\lambda}_{k,i} = \sum_{\ell=1}^{K} a_{\ell k}\boldsymbol{\eta}_{\ell,i}, \tag{6.18}$$

which yields:

$$\boldsymbol{\lambda}_{k,i} = \sum_{\ell=1}^{K} a_{\ell k}\Big((1-\delta)\boldsymbol{\lambda}_{\ell,i-1} + c_\ell(\boldsymbol{h}_{\ell,i})\Big). \tag{6.19}$$

From (6.19), we see clearly that the step-size δ attenuates the influence of past data, embodied by $\boldsymbol{\lambda}_{\ell,i-1}$. As long as δ is strictly greater than zero and smaller than one, the recursion in (6.19) can be shown to be stable, i.e., $\boldsymbol{\lambda}_{k,i}$ does not degenerate to $\pm\infty$ as $i \to \infty$. This non-degenerate behavior is the reason why the adaptive social learning algorithm can quickly recover from a previous state when faced with changes in the environment.

The price for this improved adaptation is reflected on the learning accuracy. In contrast with the almost sure convergence found in traditional social learning, consistent learning now occurs asymptotically (as $i \to \infty$) with high probability in the regime of small step-sizes (as $\delta \to 0$) [76], [77]. The same sufficient condition for attaining consistent truth learning enunciated in (6.13), applies for the adaptive social learning algorithm as well, even for general detection statistics $c_\ell(\cdot)$ [78].

In both social learning strategies detailed above, one important statistic diffused across agents and over time is the log-ratio of likelihoods, $c_\ell(\cdot)$ as in (6.10), which is classically employed as the basic building block to design other types of distributed detection strategies [101], [102]. In real-world applications, these likelihood models are generally unavailable. Instead, they are obtained as the result of a prior training step in which (parameterized) models are trained using a finite set of data examples.

In view of this practical limitation of social learning, we propose in this work a two-phase learning strategy, which we refer to as Social Machine Learning (SML). In this strategy, the likelihood models are assumed to be *unknown*.

6.3 Social Machine Learning

The SML strategy is designed as a two-step approach. In the *training phase*, the classifiers are trained individually given private finite datasets. In the *prediction phase*, classifiers are deployed in a cooperative social learning structure. In Figure 6.2, we show a diagram depicting the SML approach. These distinct *learning phases* are detailed in the following sections.

To avoid confusion, random variables related to the training phase are topped with a symbol $\sim$. Furthermore, data samples pertaining to the training datasets are indexed by n, whereas, in the prediction phase, they are indexed by i. For example, $\widetilde{\boldsymbol{h}}_{k,n} \in \mathcal{H}_k$ represents the $n-$th feature vector available in the training dataset of agent k, whereas $\boldsymbol{h}_{k,i} \in \mathcal{H}_k$ represents the

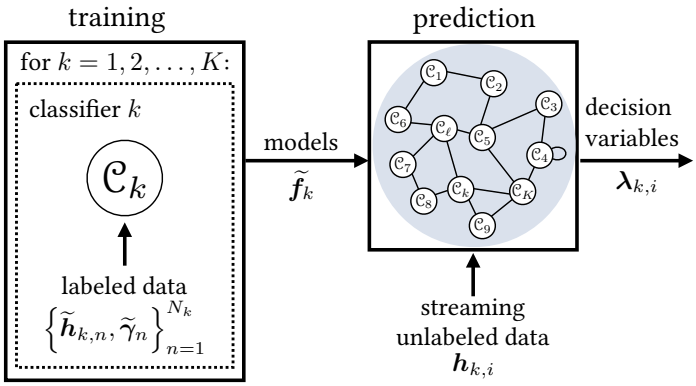


Figure 6.2: Social Machine Learning (SML) diagram.

feature vector observed by agent k at instant i during the prediction phase. We assume that the random variables are independent between different learning phases.

6.3.1 Training Phase

During training, each agent k has access to N_k examples consisting of pairs $\{\widetilde{\boldsymbol{h}}_{k,n}, \widetilde{\boldsymbol{\gamma}}_n\}_{n=1}^{N_k}$. We assume that the training set is balanced so that both classes are sufficiently explored, namely, we assume that, *during training*, labels $\widetilde{\boldsymbol{\gamma}}_n$ are uniformly distributed over $\Gamma = \{-1, +1\}$. This is a standard technical assumption that will be useful to obtain readable bounds for the social machine learning strategy. However, we remark that the assumption of a balanced dataset during the training phase does not impose any constraint on the behavior of the true hypothesis during the *prediction* phase. In particular, the data observed during prediction are all coming from a certain true state of nature γ_0, and we will establish that the SML strategy achieves vanishing error regardless of the particular hypothesis being in force.

The pair $(\widetilde{\boldsymbol{h}}_{k,n}, \widetilde{\boldsymbol{\gamma}}_n)$ is distributed according to the joint distribution:

$$(\widetilde{\boldsymbol{h}}_{k,n}, \widetilde{\boldsymbol{\gamma}}_n) \sim \widetilde{p}_k(h, \gamma) = L_k(h|\gamma)\widetilde{p}_k(\gamma), \tag{6.20}$$

where $h \in \mathcal{H}_k$, $\gamma \in \Gamma$ and $\widetilde{p}_k(\gamma) = 1/2$ for $\gamma \in \Gamma$. Note that, given a label γ, the corresponding feature vector $\widetilde{\boldsymbol{h}}_{k,n}$ is distributed according to

$$\widetilde{\boldsymbol{h}}_{k,n} \sim L_k(h|\gamma), \quad h \in \mathcal{H}_k, \gamma \in \Gamma. \tag{6.21}$$

Using these training samples, we wish to deploy a fully data-driven solution inspired by the social learning algorithms presented in Section 6.2.3. To accomplish this, we first need to approximate the key unknown function c_k used in (6.11) and (6.19) by a quantity resulting from some statistical learning method. The choice of method and the characteristics of the problem at hand, e.g., training samples and class of models considered, will heavily determine the quality of the resulting approximation. These decisive factors and their impact will be examined later in this work.

Let us delve into the details of our training setup. First, note that, under the assumption of uniform priors during training, and using Bayes' rule, we can write:

$$c_k(h) = \log \frac{L_k(h|+1)}{L_k(h|-1)} = \log \frac{\widetilde{p}_k(+1|h)}{\widetilde{p}_k(-1|h)}, \tag{6.22}$$

where $\widetilde{p}_k(\gamma|h)$ represents the posterior probability of $\{\widetilde{\boldsymbol{\gamma}}_n = \gamma\}$ given $\{\widetilde{\boldsymbol{h}}_{k,n} = h\}$ using the joint model seen in (6.20). The significance of the log-ratio of posterior probabilities on the RHS of (6.22) can be interpreted in an intuitive manner: the log-ratio is positive whenever class $+1$ is more likely to be the true state of nature given the observation of h and negative when class -1 is more likely to explain the same data evidence. It is therefore reasonable that we seek an approximation for the log-ratio of posteriors in (6.22) during the training phase.

One relevant machine learning paradigm to approximate the posterior distribution is the *discriminative* paradigm (which includes, e.g., logistic regression and neural networks), where the output of the classifier is in the form of *approximate* posterior probabilities for each class, namely, $\widehat{p}_k(+1|h)$ and $\widehat{p}_k(-1|h)$. In order to illustrate this paradigm, it is convenient to introduce the *logit* statistic:

$$\log \frac{\widehat{p}_k(+1|h)}{\widehat{p}_k(-1|h)} = \log \frac{\widehat{p}_k(+1|h)}{1 - \widehat{p}_k(+1|h)} \triangleq f_k(h), \tag{6.23}$$

where the function f_k can be chosen from an admissible class $\mathcal{F}_k$, namely,

$$f_k \in \mathcal{F}_k : \mathcal{H}_k \mapsto \mathbb{R}. \tag{6.24}$$

The choice of the class $\mathcal{F}_k$ depends on the choice of classifier. For example, in linear logistic regression with $h \in \mathbb{R}^M$, $\mathcal{F}_k$ is parameterized by a vector $w \in \mathbb{R}^M$, and we have the linear logit function [103]:

$$f_k(h; w) = w^{\mathsf{T}} h. \tag{6.25}$$

Another example is to consider MultiLayer Perceptrons (MLPs) with L hidden layers and a softmax output layer, whose weight matrices are given by $\{W_\ell\}$ over layers $\ell = 1, 2, \ldots, L$. In the binary classification case, the network outputs two approximate posterior quantities [104], namely $\widehat{p}_k(+1|h; W)$ and $\widehat{p}_k(-1|h; W)$, where W represents the parameterization of the classifier w.r.t. matrices $\{W_\ell\}$. In this case, the logit function is given by the expression:

$$f_k(h; W) = \log \frac{\widehat{p}_k(+1|h; W)}{\widehat{p}_k(-1|h; W)}, \tag{6.26}$$

where the class of functions $\mathcal{F}_k$ is parameterized by matrices $\{W_\ell\}$ in a nonlinear manner. Note that the forthcoming analysis does not assume a specific model for the logit function, and applies instead to general classes $\mathcal{F}_k$.

The logit functions f_k are trained by each classifier $k = 1, 2, \ldots, K$ by finding the function f_k within $\mathcal{F}_k$ that minimizes a suitable risk function $R_k(f_k)$. For example, in the already mentioned logistic regression and MLP cases, the training process results in optimal parameters w and W_ℓ for $\ell = 1, 2, \ldots, L$, respectively.

One common risk function adopted in binary classification is the logistic risk:

$$R_k(f_k) = \mathbb{E}_{\tilde{h}_k,\tilde{\gamma}} \log\left(1 + e^{-\widetilde{\gamma}_n f_k(\widetilde{\boldsymbol{h}}_{k,n})}\right), \tag{6.27}$$

where $\mathbb{E}_{\tilde{h}_k,\tilde{\gamma}}$ corresponds to the expectation computed under the (unknown) joint distribution $\widetilde{p}_k(h,\gamma)$ seen in (6.20). We remark that the logistic risk can be used either in association with the linear model in (6.25) or with more complex structures such as neural networks with softmax output layers. The logistic risk can be shown to be equivalent in the binary case to the cross-entropy risk function [105].

We define the *target risk* at every agent and the weighted network average according to:

$$\mathsf{R}_k^o \triangleq \inf_{f_k \in \mathcal{F}_k} R_k(f_k), \qquad \mathsf{R}^o \triangleq \sum_{k=1}^K \pi_k \mathsf{R}_k^o. \tag{6.28}$$

Unfortunately, in practice the expectation in (6.27) cannot be computed since the underlying feature/label distribution is unknown. The agents rely instead on a finite set of training samples to minimize an *empirical* risk:

$$\widetilde{\boldsymbol{f}}_k \triangleq \arg\min_{f_k \in \mathcal{F}_k} \widetilde{\boldsymbol{R}}_k(f_k), \tag{6.29}$$

given by

$$\widetilde{\boldsymbol{R}}_k(f_k) = \frac{1}{N_k} \sum_{n=1}^{N_k} \log\left(1 + e^{-\widetilde{\gamma}_n f_k(\widetilde{\boldsymbol{h}}_{k,n})}\right), \tag{6.30}$$

which is computed over the training set. The resulting function $\widetilde{\boldsymbol{f}}_k$ can then be used by the agents to approximate the logit statistic in (6.23). For future use, we also define the *network average* for the expected risk and the empirical risk expressions:

$$R(f) \triangleq \sum_{k=1}^K \pi_k R_k(f_k), \qquad \widetilde{\boldsymbol{R}}(f) \triangleq \sum_{k=1}^K \pi_k \widetilde{\boldsymbol{R}}_k(f_k), \tag{6.31}$$

where the argument f represents the dependence of the risk expressions on the collection of functions $\{f_k\}$, i.e., $R(f) = R(f_1, f_2, \ldots, f_K)$. This concise notation will be used whenever we are dealing with network-averaged quantities.

We will detail in the next section how the trained models can be deployed in the prediction phase, when agents are faced with streaming unlabeled feature vectors.

6.3.2 Prediction Phase

In the prediction phase, agents find themselves in the setup described in Section 6.2.1. They aim at solving the inference problem of determining the true state $\gamma_0 \in \Gamma$, given streaming *unlabeled* private features $\boldsymbol{h}_{k,i}$, $i = 1, 2, \ldots$. The difference now is that they are equipped with the trained models $\{\widetilde{\boldsymbol{f}}_k\}$, which are constructed so as to provide a reasonable approximation for the log-ratio of posterior probabilities (see Figure 6.2 for an illustrative diagram of this process).

During prediction, agents deploy one of the SL algorithms enunciated in Section 6.2.3 using the following approximation for the function c_k:

$$\widetilde{\boldsymbol{c}}_k(h) = \widetilde{\boldsymbol{f}}_k(h) - \widetilde{\boldsymbol{\mu}}_k(\widetilde{\boldsymbol{f}}_k), \tag{6.32}$$

where the second term on the RHS of (6.32) is called the *empirical training mean* and is defined for any function $f_k \in \mathcal{F}_k$ as:

$$\widetilde{\boldsymbol{\mu}}_k(f_k) = \frac{1}{N_k}\sum_{n=1}^{N_k} f_k(\widetilde{\boldsymbol{h}}_{k,n}), \tag{6.33}$$

i.e., it is defined as the average of function f_k over the training samples. Discounting the empirical training mean in (6.32) prevents the logit statistic from being biased towards one class or another. This is relevant considering that the decision of each agent is taken according to the rule

$$\boldsymbol{\gamma}_{k,i}^{\mathsf{SML}} \triangleq \text{sign}(\boldsymbol{\lambda}_{k,i}), \tag{6.34}$$

where $\text{sign}(x) = +1$, if $x \geq 0$ and $\text{sign}(x) = -1$ otherwise, i.e., the decision threshold is zero. Note that $\widetilde{\boldsymbol{c}}_k$ is a random function, whose randomness stems from the training phase.

Next, we illustrate how the debiasing operation used in (6.32) helps preventing biased decisions, but first let us define for any function $f_k \in \mathcal{F}_k$ the following conditional means:

$$\mu_k^+(f_k) \triangleq \mathbb{E}_{L_k(+1)} f_k(\boldsymbol{h}_{k,i}), \quad \mu_k^-(f_k) \triangleq \mathbb{E}_{L_k(-1)} f_k(\boldsymbol{h}_{k,i}). \tag{6.35}$$

Assume that f_k is fixed, i.e., $\widetilde{\boldsymbol{c}}_k(h) = f_k(h) - \widetilde{\boldsymbol{\mu}}_k(f_k)$, and that N_k is sufficiently large. Then, the empirical mean $\widetilde{\boldsymbol{\mu}}_k(f_k)$ approximates the expected value of $f_k(\widetilde{\boldsymbol{h}}_{k,n})$ in the training phase, namely, $[\mu_k^+(f_k) + \mu_k^-(f_k)]/2$ (see Eq. (6.68) in Appendix 6.A). In this case, function $\widetilde{\boldsymbol{c}}_k$ is deterministic, and we can write:

$$\widetilde{c}_k(h) = f_k(h) - \frac{\mu_k^+(f_k) + \mu_k^-(f_k)}{2}. \tag{6.36}$$

Taking the conditional expectation of $\widetilde{c}_k(\boldsymbol{h}_{k,i})$, computed w.r.t. the prediction samples $\boldsymbol{h}_{k,i}$ given classes $+1$ and -1, yields:

$$\mathbb{E}_{L_k(+1)} \widetilde{c}_k(\boldsymbol{h}_{k,i}) = \frac{\mu_k^+(f_k) - \mu_k^-(f_k)}{2}, \tag{6.37}$$

$$\mathbb{E}_{L_k(-1)} \widetilde{c}_k(\boldsymbol{h}_{k,i}) = -\frac{\mu_k^+(f_k) - \mu_k^-(f_k)}{2}. \tag{6.38}$$

The approximation $\widetilde{c}_k$ satisfies the conditions for consistent learning in (6.13) if (6.37) is strictly positive and if (6.38) is strictly negative. Note that the debiasing operation introduces a symmetry to (6.37) and (6.38). Therefore both consistent learning conditions in (6.13) are satisfied by ensuring that the weaker condition $\mu_k^+(f_k) > \mu_k^-(f_k)$ holds, regardless of the sign of the individual terms $\mu_k^+(f_k)$ and $\mu_k^-(f_k)$. Thus, even in the biased case, in which $\mu_k^+(f_k) > \mu_k^-(f_k) > 0$, consistent learning using $\widetilde{c}_k$ can be achieved.

We proceed now to formally translate the consistent learning conditions for the SL algorithms

seen in (6.13), considering the approximation $\widetilde{\boldsymbol{c}}_k(\boldsymbol{h}_{k,i})$ in (6.32). First, we define the network average of the conditional means in (6.35):

$$\mu^+(f) \triangleq \sum_{k=1}^{K} \pi_k \mu_k^+(f_k), \quad \mu^-(f) \triangleq \sum_{k=1}^{K} \pi_k \mu_k^-(f_k), \tag{6.39}$$

and the network average of the *empirical* training mean:

$$\widetilde{\boldsymbol{\mu}}(f) = \sum_{k=1}^{K} \pi_k \widetilde{\boldsymbol{\mu}}_k(f_k). \tag{6.40}$$

The training phase will generate the set of models $\{\widetilde{\boldsymbol{f}}_k\}$, which are random with respect to the training datasets. Given a particular training setup, we can "freeze" the randomness of the training set and work conditionally on a particular realization of learned models $\{\widetilde{f}_k\}$.

We are now interested in ascertaining whether or not these particular learned models allow for consistent learning *during the prediction phase*. To this end, we can apply the condition for consistent learning seen in (6.13) to the functions $\{\widetilde{c}_k\}$ in (6.32), for a frozen set of trained models $\{\widetilde{f}_k\}$, resulting in the following two conditions:

$$\sum_{k=1}^{K} \pi_k \mathbb{E}_{L_k(+1)} \widetilde{f}_k(\boldsymbol{h}_{k,i}) > \sum_{k=1}^{K} \pi_k \widetilde{\mu}_k(\widetilde{f}_k), \tag{6.41}$$

$$\sum_{k=1}^{K} \pi_k \mathbb{E}_{L_k(-1)} \widetilde{f}_k(\boldsymbol{h}_{k,i}) < \sum_{k=1}^{K} \pi_k \widetilde{\mu}_k(\widetilde{f}_k). \tag{6.42}$$

where we recall that $\mathbb{E}_{L_k(\gamma)}$ is the expectation computed with respect to the *prediction* samples $\boldsymbol{h}_{k,i}$ under the distribution $L_k(h|\gamma)$, and the prediction samples are independent of any random variable generated in the training phase. Finally, substituting the definitions in (6.35) and (6.39) respectively into (6.41) and (6.42), yields the following necessary conditions for consistent learning within the SML paradigm, conditionally on a given set of trained models $\{\widetilde{f}_k\}$.

$$\boxed{\mu^+(\widetilde{f}) > \widetilde{\mu}(\widetilde{f}) \quad \text{and} \quad \mu^-(\widetilde{f}) < \widetilde{\mu}(\widetilde{f})} \tag{6.43}$$

Since, the above description is given conditioned on a set of trained models $\{\widetilde{f}_k\}$, the conditions in (6.43) depend on the randomness stemming from the training phase. Therefore, characterizing the consistency of learning requires characterizing probabilistically the occurrence of both events described in (6.43). More precisely, we can define the *probability of consistent learning*, namely,

$$P_c \triangleq \mathbb{P}\left(\mu^+(\widetilde{\boldsymbol{f}}) > \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\,,\, \mu^-(\widetilde{\boldsymbol{f}}) < \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right), \tag{6.44}$$

where boldface fonts now highlight the randomness in the training set. In the next section, we provide the characterization of (6.44) for classifiers belonging to general classes of bounded real-valued functions $\mathcal{F}_k$. In this case, we assume that there exists some real value $\beta > 0$ such that:

$$|f_k(h)| \leq \beta, \quad f_k \in \mathcal{F}_k,\, h \in \mathcal{H}_k. \tag{6.45}$$

For example, consider the linear logistic regression case seen in (6.25). In practical applications, features belong to a bounded set $\mathcal{H}_k$, and thus condition (6.45) would be satisfied if the vector of weights w is constrained according to $\|w\|_2 \leq b$, where b is some positive real value. Similarly, in the multilayer perceptron example seen in (6.26), the condition in (6.45) is satisfied for norm-constrained neural networks [106], i.e., where the weight matrices are bounded in norm by a certain positive real value b.

6.4 Consistency of Social Machine Learning

We will need to call upon well-established statistical learning paradigms (e.g., the Vapnik-Chervonenkis theory) and adapt them to the *distributed network* setting considered in this work [107], [108]. More specifically, we will move along the path summarized below.

- We will assume that the individual agents minimize an *empirical risk*, producing a collection of K learned models, namely, the functions $\{\widetilde{\boldsymbol{f}}_k\}$. As usual, these functions are random due to the randomness of the training samples.
- We will examine the prediction (i.e., classification) performance obtained with the learned models $\{\widetilde{\boldsymbol{f}}_k\}$. In particular, we will establish technical conditions for the social learning algorithm to predict reliably the correct label as the number of streaming data gathered during the *prediction phase* increases.
- Since the learned models inherit the randomness of the training set, the consistency guarantees must be formulated in a probabilistic manner—see (6.44). Specifically, we guarantee a high probability that the samples in the training set lead to models $\{\widetilde{\boldsymbol{f}}_k\}$ that enable correct classification.
- As it happens in classical statistical learning frameworks, the interplay between empirical and optimal risk will be critical to ascertain the learning and prediction ability of the classifiers. However, differently from what is obtained in classical statistical learning frameworks, our results will depend significantly on the *graph* properties. In particular, a major role will be played by *weighted combinations of the individual risk functions*. The combination weights are the entries of the Perron eigenvector reflecting the combination matrix that governs the social learning interactions among the agents. This property leads to novel and interesting phenomena, for example, consistent classification can be achieved even if some of the agents learn bad models, but the plurality of the agents is able to reach a satisfying *aggregate* risk value.

Under the framework described above, the nontrivial interplay between the training and prediction phases might lead to some confusion. Therefore, it is useful to clarify the main path followed in the forthcoming analysis. We will focus on the probability of consistent learning P_c in (6.44), namely, the probability that the training set produces, at the end of the *training* phase, a consistent classifier. By "consistent", we mean that the classifier is able to mark the unlabeled data observed during the *prediction* phase correctly as $i \to \infty$. The probability P_c will be shown to be close to 1 if the training set size is large enough, namely, we will show that consistent learning is achievable provided that *sufficient training* is allowed. As Theorem 6.1

will show, in order to quantify the qualification "sufficient", it is critical to introduce a formal way to characterize the classifier structure.

The complexity of the classifier structure is related to the complexity of the class of functions $\mathcal{F}_k$. The latter is quantified by using the concept of *Rademacher complexity* (initially introduced as Rademacher penalty in [109]). We follow the definition in [110] and [106] and consider a class of functions $\mathcal{F}$ and a set x with N training samples, namely, $x \triangleq \{x_1, x_2 \ldots, x_N\}$, where $x_n \in \mathcal{X}$ for all $n = 1, 2, \ldots, N$. We also introduce the set of vectors $\mathcal{F}(x)$ defined as:

$$\mathcal{F}(x) \triangleq \left\{ [f(x_1), f(x_2), \ldots, f(x_N)] \,\middle|\, x_n \in \mathcal{X}\,, f \in \mathcal{F} \right\}. \tag{6.46}$$

Then, the (empirical) Rademacher complexity associated with $\mathcal{F}(x)$ is:

$$\mathcal{R}(\mathcal{F}(x)) \triangleq \mathbb{E}_r \sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{r}_i f(x_n) \right|, \tag{6.47}$$

where $\boldsymbol{r}_n$ are independent and identically distributed Rademacher random variables, i.e., with $\mathbb{P}(\boldsymbol{r}_n = +1) = \mathbb{P}(\boldsymbol{r}_n = -1) = 1/2$. This quantity can be seen as a measure of *overfitting* during training over the class of functions $\mathcal{F}$ [111]. In general, to avoid overfitting during training, and to ensure an improved generalization performance, we choose models with small classifier complexity.

Applying the above definition to our multi-agent case, we define the individual empirical Rademacher complexity of agent k for samples $h^{(k)} \triangleq \{h_{k,1}, \ldots, h_{k,N_k}\}$ as

$$\mathcal{R}(\mathcal{F}_k(h^{(k)})) = \mathbb{E}_r \sup_{f_k \in \mathcal{F}_k} \left| \frac{1}{N_k} \sum_{n=1}^{N_k} \boldsymbol{r}_n f_k(h_{k,n}) \right|, \tag{6.48}$$

and its expected Rademacher complexity, for features $\boldsymbol{h}_{k,1}, \boldsymbol{h}_{k,2}, \ldots, \boldsymbol{h}_{k,N}$ as

$$\rho_k \triangleq \mathbb{E}_{h_k} \mathcal{R}(\mathcal{F}_k(\boldsymbol{h}^{(k)})), \tag{6.49}$$

which represents the Rademacher complexity of the k-th classifier structure, *averaged* over the feature distribution. We also define the (expected) *network Rademacher complexity* according to:

$$\rho \triangleq \sum_{k=1}^{K} \pi_k \rho_k, \tag{6.50}$$

which represents an average complexity across all agents in the network, weighted by their centrality scores (given by the elements of the Perron eigenvector π).

6.4.1 Learning Consistency

In Theorem 6.1, we show that the SML strategy consistently learns the truth during the prediction phase, with high probability as the number of training samples grows and for a moderately complex classifier structure. Before introducing the theorem, we define the following two

quantities (we assume $N_k > 0$ for all k):

$$\alpha_k \triangleq \frac{N_{\max}}{N_k}, \qquad \alpha \triangleq \sum_{k=1}^{K} \pi_k \alpha_k, \tag{6.51}$$

with $N_{\max} \triangleq \max_k N_k$. The *individual imbalance penalty* α_k quantifies how distinct the number of training samples of agent k is compared with $N_{\max}$. The *network imbalance penalty* α is the average of α_k over the network, and it quantifies how unequal the training samples are across different agents. For example, if all agents possess the same number of training samples, i.e., $N_k = N_{\max}$, for all $k = 1, 2, \ldots, K$, then α assumes minimal value with $\alpha = 1$. The value of α tends to grow when agents have very different number of training samples, e.g., when, for some k, $N_k \ll N_{\max}$.

Moreover, we assume that the target risk R^o is strictly smaller than $\log 2$. To understand the meaning of such assumption, we first consider a single agent k, for which $\mathsf{R}_k^o < \log 2$. This assumption eliminates the case where the classifier makes *uninformed* decisions of the form:

$$\widehat{p}_k(\gamma|h) = \frac{1}{2}, \text{ for any } h \in \mathcal{H}_k \text{ and } \gamma \in \Gamma, \tag{6.52}$$

i.e., where the classification decision is independent of the input feature vector. In this case, from (6.23), $f_k(h) = 0$ for any $h \in \mathcal{H}_k$, which in view of (6.27) implies $R_k(f_k) = \log 2$. This situation arises, for example, when the classifier structure is not complex enough to address the classification task at hand. Requiring $\mathsf{R}_k^o < \log 2$ guarantees that the classifier k performs better than a classifier that randomly assigns labels $+1$ and -1 with equal probability. Requiring that the *network* target risk satisfies $\mathsf{R}^o < \log 2$ is an even weaker assumption, since it establishes this bound to the risk values averaged over the graph. For example, suppose that, in a $K-$agent network, $K-1$ classifiers yield uninformed decisions like in (6.52), for which $\mathsf{R}_k^o = \log 2$. To satisfy $\mathsf{R}^o < \log 2$ on a network level, it suffices that one classifier performs better than the uninformed ones.

The next theorem characterizes the consistency of the SML strategy during the prediction phase in terms of an exponential lower bound on the probability of consistent learning in (6.44).

Theorem 6.1 (SML consistency). *For the logistic risk, assume that* $\mathsf{R}^o < \log 2$ *and that* $f_k(h) \leq \beta$ *for every* $h \in \mathcal{H}_k$, $f_k \in \mathcal{F}_k$ *and* $k = 1, 2, \ldots, K$, *with* $\beta > 0$. *Assume* $\rho < \mathscr{E}(\mathsf{R}^o)$, *where* $\mathscr{E}(\mathsf{R}^o)$ *is exactly computed in* (6.93) *and can be approximated as (see Figure 6.8 in Appendix 6.A):*

$$\mathscr{E}(\mathsf{R}^o) \approx 0.2812 \left(1 - \frac{\mathsf{R}^o}{\log 2}\right). \tag{6.53}$$

Then, we have the following bound for the probability of consistent learning, defined in (6.44)*:*

$$P_c \geq 1 - 2\exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\left(\mathscr{E}(\mathsf{R}^o) - \rho\right)^2\right\}. \tag{6.54}$$

Proof. See Appendix 6.A. □

Theorem 6.1 has at least two important implications. First, if the network-average Rademacher complexity ρ is smaller than the function $\mathscr{E}(\mathsf{R}^o)$, then the probability of consistent learning is bounded in an exponential way. Now, the function $\mathscr{E}(\mathsf{R}^o)$ is an error exponent that determines how fast the probability of consistent learning approaches 1. It is a function of the optimal risk R^o—see the definition in (6.93). An excellent approximation for $\mathscr{E}(\mathsf{R}^o)$ is (6.53), showing that such exponent quantifies how close the target risk is to the $\log 2$ risk boundary. As already discussed, the $\log 2$ risk boundary corresponds to the risk associated with a binary classifier that randomly classifies samples with labels $+1$ and -1. The closer the target model is to the $\log 2$ risk, the smaller the value of $\mathscr{E}(\mathsf{R}^o)$. In other words, smaller values of $\mathscr{E}(\mathsf{R}^o)$ are symptomatic of more difficult classification problems. Therefore, Eq. (6.54) reveals a remarkable interplay between the inherent difficulty of the classification problem (quantified inversely by $\mathscr{E}(\mathsf{R}^o)$) and the complexity of the classifier structure (quantified by ρ). Ideally, we would like to have simple classification problems (i.e., higher values of $\mathscr{E}(\mathsf{R}^o)$) and low Rademacher complexity ρ. Notably, both indices are *network* indices that embody the network structure inside them.

Second, the exponent characterizing the bound in (6.54) depends on the size of the training sets at the individual agents (through the network imbalance penalty α and the maximum training-set size), and the bounding constant β. In particular, we see from (6.54) that the exponent (and, hence, the probability of consistent learning) increases if we have larger training sets (i.e., larger N_{max} and/or smaller α) and more constrained class of functions (i.e., smaller β).

In summary, the bound in (6.54) can be used to establish conditions under which the probability of consistent learning approaches 1 exponentially fast as the training-set sizes increase. To this end, we must observe that the quantity ρ itself depends on the training-set sizes. Accordingly, it is necessary to obtain an estimate (or a bound) for the network-average Rademacher complexity. Once this is done, we will be in the position of evaluating the sample complexity of the SML strategy, namely, of evaluating how many samples are necessary to achieve a target probability of consistency. This analysis will be pursued in the next section.

6.4.2 Sample Complexity

Under typical classifier structures, the Rademacher complexity scales as $C_k/\sqrt{N_k}$, where N_k is the number of training samples pertaining to agent k, and C_k is a constant quantifying the inherent complexity of the $k-$th classifier structure [106]. As an example, we will show in the next section how the Rademacher complexity behaves for the particular structure of multilayer perceptrons, and provide an upper bound for a given design of number of hidden layers and hidden units.

Now, assuming that the Rademacher complexity of each classifier k is bounded as $C_k/\sqrt{N_k}$, the *network* Rademacher complexity will be bounded as:

$$\rho \leq \sum_{\ell=1}^{K} \pi_k \frac{C_k}{\sqrt{N_k}} = \frac{1}{\sqrt{N_{\mathsf{max}}}} \underbrace{\sum_{k=1}^{K} \pi_k C_k \sqrt{\alpha_k}}_{\triangleq \mathsf{C}}, \tag{6.55}$$

where C is an average constant that mixes the individual complexity constants C_k, accounting

for the Perron eigenvector entries π_k and the individual imbalance penalties α_k. In the case where (6.55) is satisfied, exploiting (6.54) we obtain the bound:

$$P_c \geq 1 - 2\exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\left(\mathscr{E}(\mathsf{R}^o) - \frac{\mathsf{C}}{\sqrt{N_{\max}}}\right)^2\right\}. \tag{6.56}$$

Equation (6.56) shows that when $N_{\max}$ scales to infinity (with the relative proportions between $N_{\max}$ and N_k kept fixed, i.e., α kept constant), the probability of consistent learning approaches 1 exponentially fast. Moreover, Eq. (6.56) can be used to carry out a sample-complexity analysis of the SML strategy, as stated in the forthcoming theorem.

Theorem 6.2 (SML sample complexity). *Assume $\rho_k \leq C_k/\sqrt{N_k}$ for some constant $C_k > 0$ for all $k = 1, 2, \ldots, K$, and let*

$$\mathsf{C} \triangleq \sum_{k=1}^{K} \pi_k C_k \sqrt{\alpha_k}. \tag{6.57}$$

Then, for the logistic risk, consistent learning takes place with probability at least $1 - \varepsilon$, if the maximum number of training samples across the network satisfies:

$$N_{\max} > \left(\frac{\mathsf{C}}{\mathscr{E}(\mathsf{R}^o)}\right)^2 \left(1 + \frac{\alpha\beta}{2\mathsf{C}}\sqrt{\frac{1}{2}\log\left(\frac{2}{\varepsilon}\right)}\right)^2. \tag{6.58}$$

Proof. See Appendix 6.C. □

We now examine how the relevant system parameters appearing in (6.58) influence the sample complexity.

- **Target performance**: The desired probability of consistent learning, $1 - \varepsilon$, influences the bound in (6.58) only logarithmically, and, hence, has a mild effect on the necessary number of training samples.
- **Term α**: Term α quantifies how unequal the number of training samples is across agents. Larger values of α imply that agents have a more uneven number of samples, and thus require that $N_{\max}$ be increased to compensate for the lack of data at some agents in the network.
- **Term β**: Term β corresponds to the bound of the output of the logit function f_k and, hence, increasing β corresponds to increasing the possible logit functions to choose from. Accordingly, from (6.58) we see that the larger the value of β, the larger the number of training samples necessary to result in highly probable consistent learning.
- **Term C**: The constant C quantifies the complexity of the chosen classifier structure. The necessary number of training samples grows quadratically with an increase in the classifiers' complexity.

- **Term** $\mathscr{E}(\mathsf{R}^o)$: As explained before, the term $\mathscr{E}(\mathsf{R}^o)$ quantifies (inversely) the difficulty of the classification problem. Smaller values of $\mathscr{E}(\mathsf{R}^o)$ are representative of more difficult classification problems, and accordingly correspond to the necessity of acquiring more training samples.

- **Role of the network**: Given the networked nature of our inference problem, described in the early Section 6.2.1, and the fact that the conditions for consistent learning are given with respect to network average values as seen in (6.43), it is expected that the network structure plays a significant role in the results of Theorems 6.1 and 6.2. The network influence, as well as the graph topology, are captured by the presence of the Perron eigenvector π in the probability expression for consistent learning, through the network terms α, ρ and R^o, namely, the network imbalance penalty, the network Rademacher complexity and the network target risk.

  The Perron eigenvector represents the centrality or influence of each agent in determining the values of the pertinent network terms, e.g., a more influential agent k has more power to steer the value of the network target risk R^o towards its own private target risk R_k^o. For doubly-stochastic combination matrices, the vector π is a vector with elements $1/K$ [37], thus influence is uniform across agents. While the dependence on the structure connecting the classifiers is not found in existing statistical bounds in the literature for ensembles of classifiers [110], [112], similar network average dependences are key quantities in distributed estimation and social learning [37], [41], [77]. For example, in social learning, convergence occurs around the hypothesis $\gamma \in \Gamma$ that minimizes the network average KL divergence, i.e., $\sum_{k=1}^{K} \pi_k D(L_k(\gamma_0)|L_k(\gamma))$ [41], [42], [77].

In summary, Eq. (6.58) quantifies how the main system parameters act on the SML sample complexity. Specifically, we see that: **i)** owing to the exponential bound, the dependence on the target error probability ε is mild; **ii)** the number of samples to achieve a prescribed performance increases with the "size" of the class of functions (higher β), the heterogeneity among classifiers (higher α), the complexity of classifiers (higher C), and the difficulty of the learning problem (lower $\mathscr{E}(\mathsf{R}^o)$); and $iii)$ the network role is encoded in the Perron eigenvector that appears in the *network-averaged* values α, C, and R^o.

In the next section, we discuss in greater detail the expression of the classifier complexity ρ for feedforward neural networks as a function of the classifier structure, i.e., number of hidden layers (depth of the neural network) and weight of hidden units (width of the neural network), and the size of the training dataset.

6.4.3 Neural Network Complexity

In this section, we complement the result from Theorem 6.1 by showing that the term ρ in (6.50), which depends on the Rademacher complexity of the classifier, vanishes with an increasing number of training samples in the case of the MultiLayer Perceptron (MLP). Assume that one classifier has the structure of a MLP with L layers (excluding the input layer) and activation function σ. We drop index k as we are referring to a single MLP. Each layer ℓ consists of n_ℓ nodes, equivalently the size of layer ℓ is given by n_ℓ.

At each node $m = 1, 2, \ldots, n_\ell$ of layers $\ell = 2, 3, \ldots, L$, the following function $g_m^{(\ell)}$ is implemented:

$$g_m^{(\ell)}(h) = \sum_{j=1}^{n_{\ell-1}} w_{mj}^{(\ell)} \sigma\left(g_j^{(\ell-1)}(h)\right). \tag{6.59}$$

The parameters $w_{mj}^{(\ell)}$ correspond to the elements of the weight matrix W_ℓ of dimension $n_\ell \times n_{\ell-1}$. For the first layer, the function implemented at node m is of the form:

$$g_m^{(1)}(h) = \sum_{j=1}^{n_0} w_{mj}^{(1)} h_j, \tag{6.60}$$

where the input vector h has dimension n_0. A bias parameter can be incorporated in (6.60) by considering an additional input element $h_{n_0+1} = 1$.

For a MLP whose purpose is to solve a binary classification problem, we denote the output at layer L by $z \in \mathbb{R}^2$, where $z_m = g_m^{(L)}(h)$ for $m = 1, 2$. The final output is given by applying the softmax function to z, that is,

$$\widehat{p}(+1|h) = \frac{e^{z_1}}{e^{z_1} + e^{z_2}}, \quad \widehat{p}(-1|h) = \frac{e^{z_2}}{e^{z_1} + e^{z_2}}. \tag{6.61}$$

In this case, the logit function is given by:

$$f^{\mathsf{NN}}(h) = \log \frac{\widehat{p}(+1|h)}{\widehat{p}(-1|h)} = z_1 - z_2, \tag{6.62}$$

where we say that f^{NN} belongs to a class of functions $\mathcal{F}^{\mathsf{NN}}$, which is parameterized by matrices W_ℓ, for $\ell = 1, 2, \ldots, L$, according to (6.59), (6.60) and (6.62).

The general evolution for the Rademacher complexity of class $\mathcal{F}^{\mathsf{NN}}$ described above is well known in the literature as scaling with $C/\sqrt{N}$ [106]. We would like nonetheless to obtain an expression for this complexity, which depends explicitly on the design choices for the MLP, i.e., depending on the depth and weights of the network. The objective is to provide the user with a general guideline on how to choose these parameters for a desired complexity value. With this purpose, a formal upper bound for this complexity is enunciated in Proposition 6.1 inspired by results from [106], [113].

Proposition 6.1 (Rademacher complexity of norm-constrained MLPs). *Consider an L-layered multilayer perceptron, satisfying[3] $\|W_\ell\|_1 \leq b$, for every layer $\ell = 1, 2, \ldots, L$. Assume that the input vector $x \in \mathbb{R}^{n_0}$ satisfies $\max_i |x_i| \leq c$, and that the activation function $\sigma(x)$ is Lipschitz with constant L_σ with $\sigma(0) = 0$. Then the Rademacher complexity for the set of vectors $\mathcal{F}^{\mathsf{NN}}(x)$ is bounded as:*

$$\mathcal{R}(\mathcal{F}^{\mathsf{NN}}(x)) \leq \frac{4}{\sqrt{N}} \left[(2bL_\sigma)^{L-1} bc\sqrt{\log(2n_0)}\right]. \tag{6.63}$$

Proof. See Appendix 6.D. □

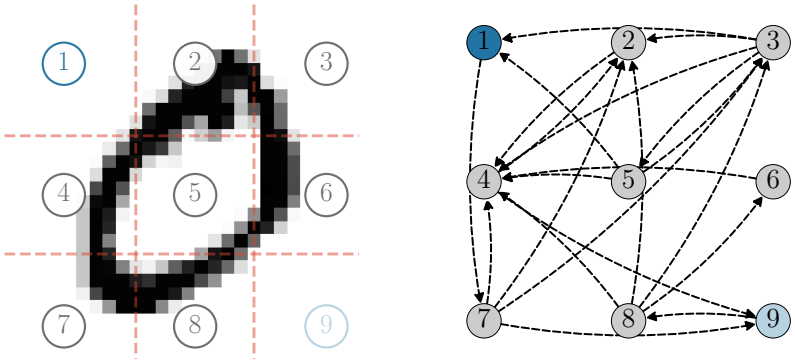


Figure 6.3: Each fraction of the image is observed by a different agent. *(Left)* Agents 1 and 9, highlighted in blue, correspond to the least informed agents. *(Right)* Topology of the network of agents.

Assume we have a network of K classifiers, each with a MLP structure. Given Proposition 6.1, we can explicitly characterize the constant C_k found in (6.55) as:

$$C_k = 4\left[(2b^{(k)}L_\sigma^{(k)})^{(L^{(k)}-1)}b^{(k)}c^{(k)}\sqrt{\log(2n_0^{(k)})}\right], \tag{6.64}$$

where we introduce superscript (k) to indicate that the classifier structural parameters can change across different agents. This characterization in association with Theorem 6.2 can be used to design the MLP architecture, according to the available training samples, or yet to select the number of samples needed for a given set of previously fixed architectures.

6.5 Simulation Results

6.5.1 MNIST Dataset

In the simulations, we consider the MNIST dataset [114], building a binary classification problem aimed at distinguishing digits 0 and 1. We employ a network of 9 spatially distributed agents, where each agent observes only a part of the image (see left panel of Figure 6.3). These agents wish to collaborate and discover which digit corresponds to the image they are collectively observing.

As we can see in the left panel of Figure 6.3, different agents will observe data with different levels of informativeness, e.g., agents 1 and 9 will dispose of little or no information within their attributed image patch. To overcome this lack of local information, agents are connected through a strongly connected network, whose combination matrix was generated using an averaging rule [37]. In the right panel of Figure 6.3, we show the network topology.

In the training phase, each agent is provided with a balanced set of 212 labeled images. Using this set of examples, classifiers are independently trained using a MLP with 2 hidden layers, each with 64 hidden units and $\tanh$ activation function, over 30 training epochs. The updates

[3]Note that $\|W\|_1$ corresponds to the maximum column sum matrix norm of matrix W.

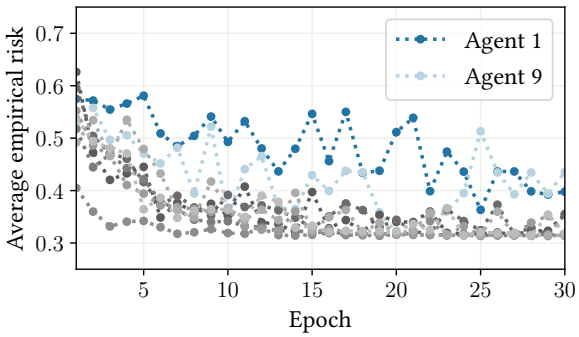


Figure 6.4: Empirical training risk averaged over 5 repetitions. The risk corresponding to agents 1 and 9 are highlighted in blue.

are performed using a batch size of 10 with learning rate 0.0001. The training is repeated 5 times for each agent. The empirical training risk for each classifier over the training epochs can be seen in Figure 6.4, where the risk was averaged over the 5 training repetitions.

As expected, in Figure 6.4 we see that classifiers 1 and 9 result in the least reliable training performances, i.e., their empirical risks exhibit the most variance across training. This could be problematic if these agents were to solve the classification problem on their own, but we will see that their individual poor classification performance is mitigated when collaborating within the network.

In the prediction phase, agents observe unlabeled images over time. The nature of images switches at every *prediction cycle*: In the cycle corresponding to interval $i \in [0, 1000)$ agents start observing digits 0. In the following cycle, i.e., $i \in [1000, 2000)$, the nature of images changes to depict digits 1. Then, from instant $i = 3000$ it switches back to digits 0, and so on. We implement the SML strategy based on the adaptive social learning algorithm described in Section 6.2.3. In Figure 6.5, we see the evolution of the decision variable $\boldsymbol{\lambda}_{1,i}$ for agent 1 with $\delta = 0.01$.

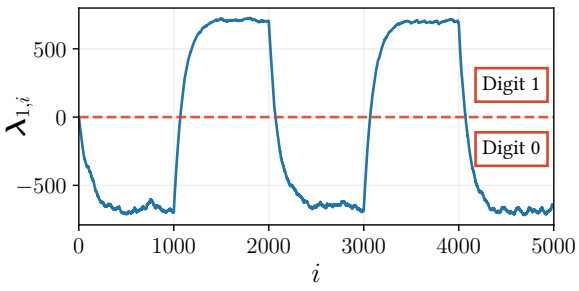


Figure 6.5: Evolution of the decision variable for agent 1 over the test phase. The observed digit is 0 within interval [0,1000), then it switches every 1000 time instants.

In Figure 6.5, we see how, despite the limited information available during training, agent 1 is able to clearly distinguish digits 0 and 1. The instantaneous decision of agent 1 is given by the sign of the decision variable $\boldsymbol{\lambda}_{1,i}$ at any given time, i.e., whether the decision variable lies

above or below the decision threshold (the orange dashed line in Figure 6.5). Moreover, within the same prediction cycle, we can see in Figure 6.5 that the decision variable moves away from the decision threshold in the correct sense, i.e., it becomes more positive under digit 1 and more negative under digit 0.

6.5.2 Comparison with AdaBoost

We compare the performance of the Social Machine Learning strategy with the classical AdaBoost strategy, as presented in [115]. In Boosting strategies, agents are trained sequentially, yielding a logit statistic $\widetilde{\boldsymbol{f}}_k^{\mathsf{Boost}}$ for each classifier k. The agents in this case are neural network classifiers, with the same architecture as described in the previous example. Once each agent is trained, its performance on the training dataset is evaluated and results in a boosting weight a_k (see [115] for further details on the implementation of AdaBoost). Larger values of a_k indicate that agent k has a better accuracy in the training dataset and makes less mistakes.

During the prediction phase, as agents observe unlabeled data $\boldsymbol{h}_{k,i}$, the decision of an individual agent is given by:

$$\boldsymbol{\gamma}_{k,i}^{\mathsf{Boost}} = \operatorname{sign}\left(\widetilde{\boldsymbol{f}}_k^{\mathsf{Boost}}(\boldsymbol{h}_{k,i})\right), \tag{6.65}$$

and the collective decision is performed using the boosting weights determined during training, according to:

$$\boldsymbol{\gamma}_{i}^{\mathsf{Boost}} = \operatorname{sign}\left(\sum_{\ell=1}^{K} a_\ell \boldsymbol{\gamma}_{\ell,i}^{\mathsf{Boost}}\right). \tag{6.66}$$

Note that computing $\boldsymbol{\gamma}_{i}^{\mathsf{Boost}}$ requires centralized information, i.e., knowledge of the instantaneous decisions of all agents. We compare this centralized boosting decision with the individual instantaneous decision of agent 1 from the SML strategy, whose decision variable was seen in Figure 6.5. The comparison can be seen in Figure 6.6, for a similar prediction setup as previously described. As a result from training AdaBoost, the lowest boosting weights were obtained for agents $1, 3, 4, 9$. This result is expected since these agents are observing less relevant information (see Figure 6.3) and can be regarded as the weakest agents.

In Figure 6.6, we see how the SML strategy results in virtually no misclassified samples when detecting the true class, whereas the AdaBoost solution makes mistakes throughout the prediction phase. We highlight that SML achieves such superior performance in a fully decentralized environment, where agents only communicate with their neighbors, whereas AdaBoost requires sequential training of each agent and centralized processing to establish the combined classification decisions.

In the second simulation setup, we can observe how the SML strategy improves its learning performance (i.e., the error probability decreases) over time during the prediction phase. To emphasize this behavior, we reduce the number of hidden units in the neural network structure to 10, and the number of training samples available at each agent to 40. The SML strategy and AdaBoost are trained for the new setup, and deployed in two prediction scenarios: a *stationary* scenario, in which the true underlying class corresponds to digit 0 throughout the simulation period; and a *nonstationary* scenario, when the true underlying class starts at digit 0 and at

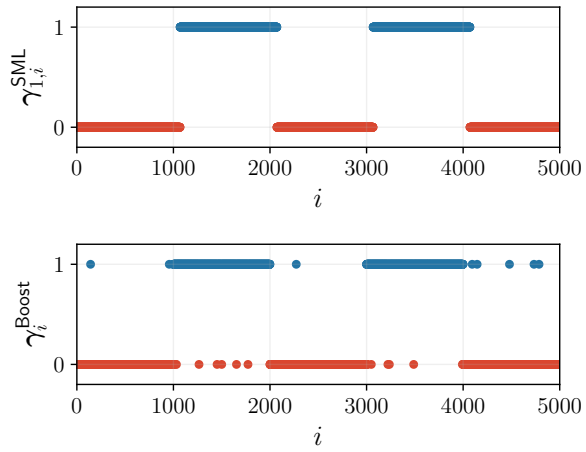


Figure 6.6: Comparison of the individual decision of Agent 1 within the SML framework and the collective AdaBoost decision. The observed digit is 0 within interval [0,1000), then it switches every 1000 time instants.

instant $i = 20$ switches to digit 1.

The SML strategy is implemented considering distinct social learning approaches. In the stationary scenario, we use traditional social learning (SL), implemented with the Bayesian update in (6.5) and combination rule in (6.6). In the nonstationary scenario, we use adaptive social learning (ASL), implemented with the adaptive Bayesian update in (6.16) and combination rule in (6.6). In Figure 6.7, we depict the probability of error of the centralized Boosting algorithm and the SML strategy at agent 1 (now with the choice of parameter $\delta = 0.1$) for the two scenarios. The probability is empirically estimated from 1000 Monte Carlo runs.

In the top panel of Figure 6.7, we note that the SML strategy, associated with the SL protocol, quickly surpasses AdaBoost's performance and attains a significantly improved accuracy over time. Notably this improvement in accuracy exhibits a linear behavior as times progresses. The traditional social learning strategy, although powerful, is not suitable to operate under nonstationary conditions, as discussed in [77]. This is why, in the bottom panel of Figure 6.7, we consider instead the SML strategy with the ASL protocol. In this scenario, we can clearly distinguish two prediction cycles, corresponding to the period under different underlying classes of digits. Compared with AdaBoost, SML yields the best performance as time passes. It is able to adapt its predictive behavior in view of the change in the observed digit, and it eventually surpasses the performance of AdaBoost with an adaptation time that scales with the chosen step-size $1/\delta$.

This improved performance can be explained by noticing the following aspect. The SML strategy leverages not only information distributed across agents, but also knowledge accumulated over time. We see that by considering, for example, the SML strategy associated with the SL protocol in (6.11), the evolution of the decision variable of agent k over the prediction phase is governed

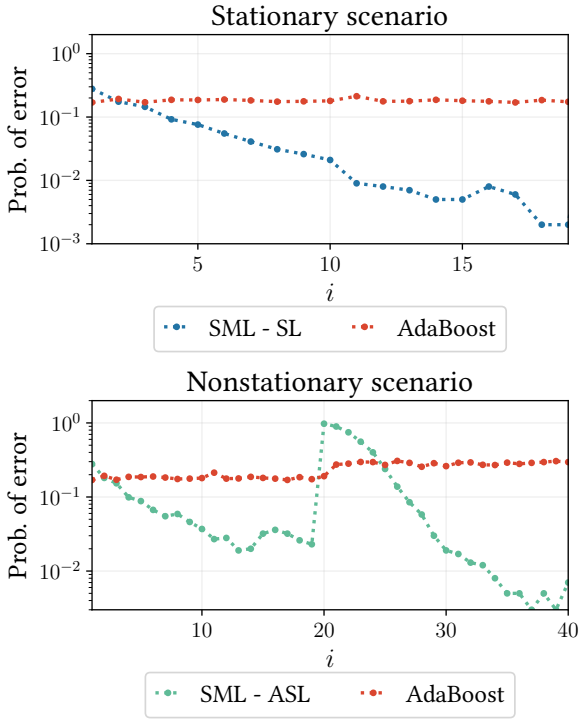


Figure 6.7: Evolution of the probability of error for the SML strategy and AdaBoost (centralized decision) estimated from 1000 Monte Carlo runs. (*Top*) SML is run with the traditional social learning rule (SL). The true state corresponds to digit 0. (*Bottom*) SML is run with the adaptive social learning rule (ASL). Until instant $i = 20$, the true state corresponds to digit 0, after which the true state is digit 1.

by the recursion:

$$\boldsymbol{\lambda}_{k,i} = \sum_{\ell=1}^{K} a_{\ell k} \Big(\boldsymbol{\lambda}_{\ell,i-1} + \widetilde{\boldsymbol{c}}_{\ell}(\boldsymbol{h}_{\ell,i}) \Big). \tag{6.67}$$

As we can see in (6.67), at every instant i, $\boldsymbol{\lambda}_{k,i}$ aggregates information from the past through the term $\boldsymbol{\lambda}_{\ell,i-1}$. The aggregation of past decision variables leverages the fact that the underlying class changes slowly during the prediction phase, thus allowing the classifiers to grow in confidence over time.

We should also note that, in face of a single observation, i.e., at instant $i = 1$ in Figure 6.7, AdaBoost outperforms SML in its classification accuracy. This can be explained by the fact that SML is a decentralized algorithm, i.e., at each iteration, agent k communicates only with its one-hop neighbors. If the agent's neighbors happen to be poorly informed classifiers, then their $1-$iteration decision will also be unreliable. As the time passes, that is, as i grows, this challenge is overcome due to the strong-connectivity of the graph topology, which enables the diffusion of information across all agents. In the example above, this is accomplished around instant $i = 4$, when SML surpasses AdaBoost in performance.

6.6 Concluding Remarks

In this chapter, we focused on the following classification problem. A network of spatially distributed agents observes an event and all agents wish to determine the underlying class exploiting a growing number of streaming observations collected over time. Such problem has been thoroughly studied within the *social learning* literature, where agents possess a set of possible models to explain their observations. By cooperating with neighbors, these agents are able to overcome local limitations and achieve collective consistent learning of the true underlying class.

These methods, however powerful, depend on the prior knowledge of the set of possible models, or *likelihoods*, which characterize the distribution of observations given different underlying classes. In this chapter, we provided a fully data-driven solution to the aforementioned problem. We introduced Social Machine Learning, which is a two-step framework that allows aggregating the information perceived by heterogeneous classifiers to improve their decision performance over time, as the classifiers observe streaming data. The classifiers are heterogeneous in the sense that their private observations originate from different distributions. In our approach, we introduce a training phase that, with a finite training dataset, results in approximate models for the unknown data logit statistics. These models are deployed in a prediction phase, where one of the available social learning algorithms can be used.

We show that consistent learning in the prediction phase can be achieved with high probability, and we describe how the number of training samples should scale to yield the desired consistency. Furthermore, the decentralized collaboration among agents results in an increased robustness in face of poorly informed agents, as seen in the simulation results. Simulations also show that our solution continually improves performance over time, leveraging past acquired knowledge to make better informed decisions in the present.

6.A Proof of Theorem 6.1

Before detailing the proof of Theorem 6.1, we enunciate Lemma 6.1, which provides a lower bound on the probability of consistent learning. We denote the total expected value of $f_k(\widetilde{\boldsymbol{h}}_{k,n})$ by:

$$\mu_k(f_k) \triangleq \mathbb{E}_{\tilde{h}_k} f_k(\widetilde{\boldsymbol{h}}_{k,n}) = \frac{\mu_k^+(f_k) + \mu_k^-(f_k)}{2}, \tag{6.68}$$

where we considered equal priors over the two classes $+1$ and -1. We also denote its average across the network by:

$$\mu(f) \triangleq \sum_{\ell=1}^{K} \pi_k \mu_k(f_k) \overset{\text{(a)}}{=} \frac{\mu^+(f) + \mu^-(f)}{2}, \tag{6.69}$$

where (a) follows from using (6.68) and the definition of $\mu^+(f)$ and $\mu^-(f)$ found in (6.39).

Lemma 6.1 (Probability bound for consistent learning). *For any $d > 0$, we have that:*

$$P_c \geq 1 - \mathbb{P}\left(\left|\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}})\right| \geq d\right) - \mathbb{P}\left(R(\widetilde{\boldsymbol{f}}) \geq \Delta\right), \tag{6.70}$$

where $\Delta \triangleq \log(1 + e^{-d})$ and P_c is the probability of consistent learning defined in (6.44).

Proof. Define the following events, which will be used in the proof:

$$\mathcal{A} \triangleq \left\{\left|\mu(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right| \geq \frac{\mu^+(\widetilde{\boldsymbol{f}}) - \mu^-(\widetilde{\boldsymbol{f}})}{2}\right\}, \tag{6.71}$$

$$\mathcal{B} \triangleq \left\{\frac{\mu^+(\widetilde{\boldsymbol{f}}) - \mu^-(\widetilde{\boldsymbol{f}})}{2} > d\right\}. \tag{6.72}$$

First, in view of (6.44) and using de Morgan's law [18], we can write:

$$\begin{aligned}
1 - P_c &= \mathbb{P}\left(\left\{\mu^+(\widetilde{\boldsymbol{f}}) \leq \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right\} \bigcup \left\{\mu^-(\widetilde{\boldsymbol{f}}) \geq \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right\}\right) \\
&\overset{\text{(a)}}{=} \mathbb{P}\left(\left\{\mu^+(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}}) \leq \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}})\right\}\right. \\
&\qquad \left. \bigcup \left\{\mu^-(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}}) \geq \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}})\right\}\right) \\
&\overset{\text{(b)}}{=} \mathbb{P}\left(\left\{\mu^+(\widetilde{\boldsymbol{f}}) - \frac{\mu^+(\widetilde{\boldsymbol{f}}) + \mu^-(\widetilde{\boldsymbol{f}})}{2} \leq \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}})\right\}\right. \\
&\qquad \left. \bigcup \left\{\mu^-(\widetilde{\boldsymbol{f}}) - \frac{\mu^+(\widetilde{\boldsymbol{f}}) + \mu^-(\widetilde{\boldsymbol{f}})}{2} \geq \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}})\right\}\right) \\
&= \mathbb{P}\left(\left\{\frac{\mu^+(\widetilde{\boldsymbol{f}}) - \mu^-(\widetilde{\boldsymbol{f}})}{2} \leq -\left(\mu(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right)\right\}\right. \\
&\qquad \left. \bigcup \left\{\frac{\mu^+(\widetilde{\boldsymbol{f}}) - \mu^-(\widetilde{\boldsymbol{f}})}{2} \leq \mu(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right\}\right) \\
&\overset{\text{(c)}}{=} \mathbb{P}(\mathcal{A}) \\
&\overset{\text{(d)}}{=} \mathbb{P}\left(\mathcal{A}\,,\,\mathcal{B}\right) + \mathbb{P}\left(\mathcal{A}\,,\,\overline{\mathcal{B}}\right) \\
&\overset{\text{(e)}}{\leq} \mathbb{P}\left(\left|\mu(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right| \geq d\right) + \mathbb{P}\left(\frac{\mu^+(\widetilde{\boldsymbol{f}}) - \mu^-(\widetilde{\boldsymbol{f}})}{2} \leq d\right),
\end{aligned} \tag{6.73}$$

where in (a) we subtract $\mu(\widetilde{\boldsymbol{f}})$ from both terms within the probability operator, in (b) we replace

$\mu(\widetilde{\boldsymbol{f}})$ with (6.69), and (c) follows from the following relation:

$$\begin{aligned}
&\left\{\frac{\mu^+(\widetilde{\boldsymbol{f}})-\mu^-(\widetilde{\boldsymbol{f}})}{2} \le -\Big(\mu(\widetilde{\boldsymbol{f}})-\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\Big)\right\} \\
&\bigcup \left\{\frac{\mu^+(\widetilde{\boldsymbol{f}})-\mu^-(\widetilde{\boldsymbol{f}})}{2} \le \mu(\widetilde{\boldsymbol{f}})-\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right\} \\
&\Leftrightarrow \left\{\left|\mu(\widetilde{\boldsymbol{f}})-\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right| \ge \frac{\mu^+(\widetilde{\boldsymbol{f}})-\mu^-(\widetilde{\boldsymbol{f}})}{2}\right\} \triangleq \mathcal{A}. \qquad (6.74)
\end{aligned}$$

In (d), we used the law of total probability, and (e) follows from:

$$\mathcal{A} \bigcap \mathcal{B} \Rightarrow \left\{\left|\mu(\widetilde{\boldsymbol{f}})-\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right| \ge d\right\}, \qquad (6.75)$$

where $\mathcal{B}$ is defined in (6.72), and also from the fact that the probability of intersection of two events is upper bounded by the probability of one of the events.

Let us address the second probability term on the RHS of (6.73). Consider the average network risk evaluated on the training samples $(\widetilde{\boldsymbol{h}}_{k,n}, \widetilde{\boldsymbol{\gamma}}_n)$, computed for a given function $f_k \in \mathcal{F}_k$:

$$\begin{aligned}
\sum_{k=1}^{K} \pi_k R_k(f_k) &= \sum_{k=1}^{K} \pi_k \mathbb{E}_{\tilde{h}_k,\tilde{\gamma}} \log\left(1+\exp\Big(-\widetilde{\gamma}_n f_k(\widetilde{\boldsymbol{h}}_{k,n})\Big)\right) \\
&\overset{(a)}{\ge} \sum_{k=1}^{K} \pi_k \log\left(1+\exp\Big(-\mathbb{E}_{\tilde{h}_k,\tilde{\gamma}}\widetilde{\gamma}_n f_k(\widetilde{\boldsymbol{h}}_{k,n})\Big)\right) \\
&\overset{(b)}{\ge} \log\left(1+\exp\Big(-\sum_{k=1}^{K}\pi_k \mathbb{E}_{\tilde{h}_k,\tilde{\gamma}}\widetilde{\gamma}_n f_k(\widetilde{\boldsymbol{h}}_{k,n})\Big)\right) \\
&\overset{(c)}{\ge} \log\left(1+\exp\left(\frac{1}{2}\sum_{k=1}^{K}\pi_k \mathbb{E}_{L_k(-1)} f_k(\widetilde{\boldsymbol{h}}_{k,n}) - \frac{1}{2}\sum_{k=1}^{K}\pi_k \mathbb{E}_{L_k(+1)} f_k(\widetilde{\boldsymbol{h}}_{k,n})\right)\right) \\
&\overset{(d)}{=} \log\left(1+\exp\Big(-\frac{\mu^+(f)-\mu^-(f)}{2}\Big)\right), \qquad (6.76)
\end{aligned}$$

where in (a) and (b) we used Jensen's inequality with the convexity of function $\log(1+e^x)$. In (c), we used the assumption of uniform priors during training, and in (d) we used the definition of the conditional means averaged over the network found in (6.35) and (6.39). From (6.76), we have the following implication for a given $f_k \in \mathcal{F}_k$ for $k = 1, 2, \ldots, K$:

$$\frac{\mu^+(f)-\mu^-(f)}{2} \le d \Rightarrow \sum_{k=1}^{N} \pi_k R_k(f_k) \ge \log\left(1+e^{-d}\right). \qquad (6.77)$$

Replace f_k in (6.77) by $\widetilde{\boldsymbol{f}}_k$ (i.e., the models obtained in the training phase). Then (6.77) results

in:

$$\mathbb{P}\left(\sum_{k=1}^{N}\pi_k R_k(\widetilde{\boldsymbol{f}}_k) \geq \log\left(1+e^{-d}\right)\right) \geq \mathbb{P}\left(\frac{\mu^+(\widetilde{\boldsymbol{f}})-\mu^-(\widetilde{\boldsymbol{f}})}{2} \leq d\right) = \mathbb{P}\left(\overline{\mathcal{B}}\right). \tag{6.78}$$

Using (6.78) in (6.73) and defining $\Delta \triangleq \log(1+e^{-d})$ yields the bound in (6.70). □

Proof of Theorem 6.1. From Lemma 6.1, we obtain the lower bound in (6.70) for the probability of consistent learning. Next, we need to examine each of the terms on the RHS of (6.70).

Regarding the first term, we can write:

$$\left|\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}})\right| \leq \sup_{f\in\mathcal{F}} |\widetilde{\boldsymbol{\mu}}(f)-\mu(f)|\,, \tag{6.79}$$

which implies that

$$\mathbb{P}\left(\left|\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}})\right| \geq d\right) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} |\widetilde{\boldsymbol{\mu}}(f)-\mu(f)| \geq d\right), \tag{6.80}$$

providing us with a *uniform* bound for the first term on the RHS of (6.70). We will now call upon Theorem 6.3 (Appendix 6.B) to obtain an exponential upper bound on the RHS of (6.80). Using (6.103) with the choice $x = d$ in (6.80) yields:

$$\mathbb{P}\left(\left|\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}})\right| \geq d\right) \leq \exp\left\{\frac{-(d-4\rho)^2 N_{\mathsf{max}}}{2\alpha^2\beta^2}\right\}, \tag{6.81}$$

for any positive d such that $d > 4\rho$.

Next, we examine the second term on the RHS of (6.70). Using Lemma 6.2 (Appendix 4.G), with the choice $x = \Delta - \mathsf{R}^o$, we can derive the following uniform upper bound:

$$\mathbb{P}\left(R(\widetilde{\boldsymbol{f}}) \geq \Delta\right) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\widetilde{\boldsymbol{R}}(f) - R(f)\right| \geq \frac{\Delta-\mathsf{R}^o}{2}\right), \tag{6.82}$$

for any positive d such that $\Delta = \log(1+e^{-d}) > \mathsf{R}^o$. Such d exists since, by assumption, $\mathsf{R}^o < \log 2$.

Next, consider Eq. (6.102) of Theorem 6.3 (Appendix 6.B), with the choice $x = (\Delta - \mathsf{R}^o)/2$ and function $\phi(x) = \log(1+e^x)$, which is a function with Lipschitz constant $L_\phi = 1$. Replacing (6.102) with these choices into (6.82) results in the bound:

$$\mathbb{P}\left(R(\widetilde{\boldsymbol{f}}) \geq \Delta\right) \leq \exp\left\{\frac{-(\frac{\Delta-\mathsf{R}^o}{2} - 4\rho)^2 N_{\mathsf{max}}}{2\alpha^2\beta^2}\right\}, \tag{6.83}$$

for any d such that $(\Delta - \mathsf{R}^o)/2 > 4\rho$. Using (6.81) and (6.83) in (6.70) results in the following

bound on the probability of consistent learning

$$P_c \geq 1 - \exp\left\{\frac{-8(\frac{d}{4}-\rho)^2 N_{\mathsf{max}}}{\alpha^2\beta^2}\right\} - \exp\left\{\frac{-8(\frac{\Delta-\mathsf{R}^o}{8}-\rho)^2 N_{\mathsf{max}}}{\alpha^2\beta^2}\right\}, \tag{6.84}$$

for any d satisfying

$$\frac{d}{4} - \rho > 0 \quad \text{and} \quad \frac{\Delta - \mathsf{R}^o}{8} - \rho > 0, \tag{6.85}$$

i.e., for any d contained in the following interval:

$$d \in (4\rho, -\log(e^{8\rho+\mathsf{R}^o} - 1)). \tag{6.86}$$

For simplicity, we can rewrite (6.84) in the following manner:

$$P_c \geq 1 - \exp\left\{\frac{-8E_1^2(x) N_{\mathsf{max}}}{\alpha^2\beta^2}\right\} - \exp\left\{\frac{-8E_2^2(x) N_{\mathsf{max}}}{\alpha^2\beta^2}\right\}, \tag{6.87}$$

where we introduced the auxiliary functions:

$$E_1(x) \triangleq x - \rho, \tag{6.88}$$

$$E_2(x) \triangleq \frac{\log(1+e^{-4x}) - \mathsf{R}^o}{8} - \rho, \tag{6.89}$$

and the free variable d was replaced by $x \triangleq \frac{d}{4}$. We can now maximize the minimum exponent, i.e., the slowest decay rate, over the free parameter x. To this end, let us consider the value $x^\star$ that solves the equation:

$$E_1(x^\star) = E_2(x^\star), \tag{6.90}$$

which corresponds to:

$$x^\star = \frac{\log(1+e^{-4x^\star}) - \mathsf{R}^o}{8} \Leftrightarrow e^{\mathsf{R}^o}\, e^{12x^\star} - e^{4x^\star} - 1 = 0. \tag{6.91}$$

Setting $e^{4x^\star} = y$, we have to solve the third-order equation:

$$e^{\mathsf{R}^o}\, y^3 - y - 1 = 0, \tag{6.92}$$

whose unique real-valued solution $y^\star$ is available in closed form. Within the range $\mathsf{R}^o \in [0, \log 2]$, $y^\star$ is strictly greater than 1, yielding:

$$\begin{aligned} x^\star &= \frac{1}{4}\log(y^\star) \\ &= \frac{1}{4}\log\left(\frac{2\times 3^{\frac{1}{3}} + 2^{\frac{1}{3}} e^{-\mathsf{R}^o}[Z(\mathsf{R}^o)]^{\frac{2}{3}}}{6^{\frac{2}{3}}[Z(\mathsf{R}^o)]^{\frac{1}{3}}}\right) \triangleq \mathscr{E}(\mathsf{R}^o), \end{aligned} \tag{6.93}$$

where

$$Z(\mathsf{R}^o) = 9e^{2\mathsf{R}^o} + \sqrt{3e^{3\mathsf{R}^o}(-4+27e^{\mathsf{R}^o})}. \tag{6.94}$$

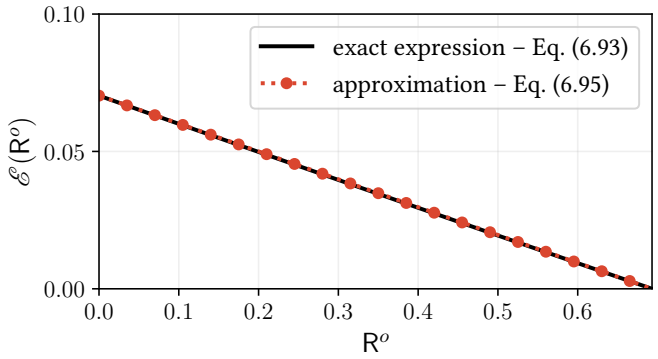


Figure 6.8: Comparison between the exact expression in (6.93) and the approximation in (6.95).

A good approximation for the function $\mathscr{E}(\mathsf{R}^o)$ is the linear fit—see Figure 6.8:

$$\mathscr{E}(\mathsf{R}^o) \approx 4\mathscr{E}(0)\left(1 - \frac{\mathsf{R}^o}{\log 2}\right), \tag{6.95}$$

where the maximum allowed complexity corresponding to a zero risk is:

$$4\mathscr{E}(0) = 0.2812, \tag{6.96}$$

which is related to the solution of the third-order equation:

$$y^3 - y - 1 = 0. \tag{6.97}$$

Figure 6.8 shows how accurate the linear approximation in (6.95) is with respect to the exact expression for $\mathscr{E}(\mathsf{R}^o)$ in (6.93) within the interval $\mathsf{R}^o \in [0, \log 2]$.

Now, since $E_1(x)$ is an increasing function of x, while $E_2(x)$ is a decreasing function of x, we conclude that if we choose a value $x \neq x^\star$ the minimum exponent necessarily decreases. Accordingly, the minimum exponent is maximized at the value $x^\star = \mathscr{E}(\mathsf{R}^o)$.

Finally, letting

$$\rho < \mathscr{E}(\mathsf{R}^o), \tag{6.98}$$

we end up with the following bound:

$$P_c \geq 1 - 2\exp\left\{-\frac{8N_{\mathsf{max}}}{\alpha^2\beta^2}\Big(\mathscr{E}(\mathsf{R}^o) - \rho\Big)^2\right\}, \tag{6.99}$$

and the proof is complete. □

6.B Auxiliary Theorem

To develop the forthcoming result, we consider a L_ϕ-Lipschitz loss function $\phi : \mathbb{R} \mapsto \mathbb{R}_+$. The individual expected and empirical risks are written accordingly as:

$$R_k(f_k) = \mathbb{E}_{h_k,\gamma_k} \phi(-\boldsymbol{\gamma}_{k,n} f_k(\boldsymbol{h}_{k,n})), \tag{6.100}$$

$$\widetilde{\boldsymbol{R}}_k(f_k) = \frac{1}{N_k} \sum_{n=1}^{N_k} \phi(-\boldsymbol{\gamma}_{k,n} f_k(\boldsymbol{h}_{k,n})), \tag{6.101}$$

where we removed the symbol $\sim$ from the top of random variables $\boldsymbol{\gamma}_{k,n}$ and $\boldsymbol{h}_{k,n}$ for simplicity of notation. Their network averages $R(f)$ and $\widetilde{\boldsymbol{R}}(f)$ are defined as shown in (6.31).

Theorem 6.3 (Uniform law of large numbers). *Assume that the loss function $\phi : \mathbb{R} \mapsto \mathbb{R}_+$ is $L_\phi-$Lipschitz and that there exists $\beta > 0$ such that $f_k(h) \leq \beta$ for every $h \in \mathcal{H}_k$, and $f_k \in \mathcal{F}_k$ and $k = 1, 2, \ldots, K$. Then we have the following two results. First,*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \widetilde{\boldsymbol{R}}(f) - R(f) \right| \geq x \right) \leq \exp\left\{ \frac{-N_{\mathsf{max}} \left(x - 4L_\phi \rho\right)^2}{2\alpha^2 L_\phi^2 \beta^2} \right\}, \tag{6.102}$$

for any $x > 4L_\phi \rho$. Second,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\widetilde{\boldsymbol{\mu}}(f) - \mu(f)| \geq x \right) \leq \exp\left\{ \frac{-N_{\mathsf{max}} \left(x - 4\rho\right)^2}{2\alpha^2 \beta^2} \right\}, \tag{6.103}$$

for any $x > 4\rho$, where $N_{\mathsf{max}} \triangleq \max_k N_k$, ρ is the network Rademacher complexity defined in (6.50)*, and α is defined as* (6.51)*.*

Proof. In the proof, we use the known *independent bounded differences inequality*, which is also known as *McDiarmid's inequality* [116]. The inequality is reproduced here without proof to facilitate its reference in the forthcoming results.

McDiarmid's Inequality. Let $\boldsymbol{x}$ represent a sequence of independent random variables $\boldsymbol{x}_n$, with $n = 1, 2, \ldots, N$ and $\boldsymbol{x}_n \in \mathcal{X}_n$ for all n. Suppose that the function $g : \prod_{n=1}^N \mathcal{X}_n \mapsto \mathbb{R}$ satisfies for every $j = 1, 2, \ldots, N$:

$$|g(x) - g(\check{x})| \leq c_j \tag{6.104}$$

whenever the sequences x and $\check{x}$ differ only in the $j-$th component. Then we have for $t > 0$:

$$\mathbb{P}\Big(g(\boldsymbol{x}) - \mathbb{E}\, g(\boldsymbol{x}) \geq t \Big) \leq e^{-2t^2 / \sum_{j=1}^N c_j^2}, \tag{6.105}$$

$$\mathbb{P}\Big(g(\boldsymbol{x}) - \mathbb{E}\, g(\boldsymbol{x}) \leq -t \Big) \leq e^{-2t^2 / \sum_{j=1}^N c_j^2}. \tag{6.106}$$

□

We now develop the proof of (6.102) and (6.103) in Theorem 6.3 separately as follows.

Proof of (6.102): Consider that the sequence of samples $\boldsymbol{x}_n$ is replaced by a sequence of random pairs $(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)$, with $n = 1, 2, \ldots, N_{\mathsf{max}}$, where $N_{\mathsf{max}} \triangleq \max_k N_k$. The quantity $\boldsymbol{h}_n$ is a sequence collecting random variables (or vectors) $\boldsymbol{h}_{k,n}$ for $k = 1, 2, \ldots, K$:

$$\boldsymbol{h}_n \triangleq \{\boldsymbol{h}_{1,n}, \boldsymbol{h}_{2,n}, \ldots, \boldsymbol{h}_{K,n}\}, \tag{6.107}$$

and $\boldsymbol{\gamma}_n$ is a sequence of random variables $\boldsymbol{\gamma}_{k,n}$ for $k = 1, 2, \ldots, K$:

$$\boldsymbol{\gamma}_n \triangleq \{\boldsymbol{\gamma}_{1,n}, \boldsymbol{\gamma}_{2,n}, \ldots, \boldsymbol{\gamma}_{K,n}\}. \tag{6.108}$$

The pairs $(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)$ are independent and identically distributed over time, i.e., for all n.

Define the following auxiliary quantity:

$$\chi_k(f_k) \triangleq \mathbb{E}_{h_k, \gamma_k} \phi(-\boldsymbol{\gamma}_{k,n} f_k(\boldsymbol{h}_{k,n})), \tag{6.109}$$

where we recall that $\mathbb{E}_{h_k,\gamma_k}$ is the expectation computed according to the joint distribution of $\boldsymbol{h}_{k,n}$ and $\boldsymbol{\gamma}_{k,n}$. Our function of interest is the following:

$$g(h, \gamma) = \sup_{f \in \mathcal{F}} \left| \sum_{k=1}^{K} \pi_k \left[\chi_k(f_k) - \frac{1}{N_k} \sum_{n=1}^{N_k} \phi(-\gamma_{k,n} f_k(h_{k,n})) \right] \right|, \tag{6.110}$$

where, to keep a concise notation, the arguments h, γ indicate that the function $g(\cdot)$ depends on the collection of sequences h_n (defined in (6.107)) and γ_n (defined in (6.108)) for $n = 1, 2, \ldots, N_k$. The argument f represents the ensemble of functions $\{f_k\}$, where $f_k \in \mathcal{F}_k$, and we define the global space of functions:

$$\mathcal{F} \triangleq \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_K. \tag{6.111}$$

From the collections h and γ, we can construct collections $\check{h}$ and $\check{\gamma}$, by replacing $h_{k,j}$ and $\gamma_{k,j}$ respectively with the distinct samples $\check{h}_{k,j}$ and $\check{\gamma}_{k,j}$ for all $k = 1, 2, \ldots, K$. If $j > N_k$, the inner summand in (6.110) is not altered, then, using the indicator function, we can write

$$\begin{aligned}
g(\check{h}, \check{\gamma}) = \sup_{f \in \mathcal{F}} \Bigg| \sum_{k=1}^{K} \pi_k \Bigg[& \chi_k(f_k) - \frac{1}{N_k} \sum_{\substack{n=1 \\ n \neq j}}^{N_k} \phi(-\gamma_{k,n} f_k(h_{k,n})) \\
& - \frac{\mathbb{I}\left[j \leq N_k\right]}{N_k} \phi(-\check{\gamma}_{k,j} f_k(\check{h}_{k,j})) - \frac{\mathbb{I}\left[j > N_k\right]}{N_k} \phi(-\gamma_{k,j} f_k(h_{k,j})) \Bigg] \Bigg| \\
= \sup_{f \in \mathcal{F}} \Bigg| \sum_{k=1}^{K} \pi_k \Bigg[& \chi_k(f_k) - \frac{1}{N_k} \sum_{n=1}^{N_k} \phi(-\gamma_{k,n} f_k(h_{k,n})) \\
& + \frac{\mathbb{I}\left[j \leq N_k\right]}{N_k} \Big(\phi(-\gamma_{k,j} f_k(h_{k,j})) - \phi(-\check{\gamma}_{k,j} f_k(\check{h}_{k,j})) \Big) \Bigg] \Bigg|,
\end{aligned} \tag{6.112}$$

where $\mathbb{I}[\mathcal{E}]$ is the indicator function defined as: $\mathbb{I}[\mathcal{E}] = 1$, if event $\mathcal{E}$ takes place, $\mathbb{I}[\mathcal{E}] = 0$ otherwise. It is convenient to introduce the following quantities:

$$u_k(f_k) \triangleq \chi_k(f_k) - \frac{1}{N_k}\sum_{n=1}^{N_k}\phi(-\gamma_{k,n}f_k(h_{k,n})), \tag{6.113}$$

$$v_k(f_k) \triangleq \frac{\mathbb{I}[j \le N_k]}{N_k}\Big[\phi(-\gamma_{k,j}f_k(h_{k,j})) - \phi(-\check{\gamma}_{k,j}f_k(\check{h}_{k,j}))\Big], \tag{6.114}$$

where the dependence of $u_k(\cdot)$ upon (h,γ) and of $v_k(\cdot)$ upon $(\check{h},\check{\gamma})$ has been skipped for ease of notation. In view of the definitions in (6.113) and (6.114), we can rewrite (6.110) and (6.112) as:

$$g(h,\gamma) = \sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{K}\pi_k u_k(f_k)\right| \tag{6.115}$$

$$g(\check{h},\check{\gamma}) = \sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{K}\pi_k u_k(f_k) + \sum_{k=1}^{K}\pi_k v_k(f_k)\right|. \tag{6.116}$$

Applying Lemma 6.4 (Appendix 6.E) with the choices $s_1 = g(h,\gamma)$, $s_2 = g(\check{h},\check{\gamma})$, and

$$S(f) = \sum_{k=1}^{K}\pi_k u_k(f_k), \quad T(f) = \sum_{k=1}^{K}\pi_k v_k(f_k), \tag{6.117}$$

we obtain:

$$|g(h,\gamma) - g(\check{h},\check{\gamma})| \le \sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{K}\pi_k v_k(f_k)\right| \overset{\text{(a)}}{\le} \sum_{k=1}^{K}\pi_k \sup_{f_k\in\mathcal{F}_k}\Big|v_k(f_k)\Big|, \tag{6.118}$$

where (a) follows from the triangle inequality and the subadditive property of the supremum operator. Replacing (6.114) into (6.118) yields

$$\begin{aligned}
|g(h,\gamma) - g(\check{h},\check{\gamma})| &\le \sum_{k=1}^{K}\pi_k \sup_{f_k\in\mathcal{F}_k}\left|\frac{\mathbb{I}[j\le N_k]}{N_k}\Big[\phi(-\gamma_{k,j}f_k(h_{k,j})) - \phi(-\check{\gamma}_{k,j}f_k(\check{h}_{k,j}))\Big]\right| \\
&\le \sum_{k=1}^{K}\pi_k \sup_{f_k\in\mathcal{F}_k}\left|\frac{1}{N_k}\Big[\phi(-\gamma_{k,j}f_k(h_{k,j})) - \phi(-\check{\gamma}_{k,j}f_k(\check{h}_{k,j}))\Big]\right| \\
&\overset{\text{(a)}}{\le} L_\phi\sum_{k=1}^{K}\frac{\pi_k}{N_k}\sup_{f_k\in\mathcal{F}_k}\left|\gamma_{k,j}f_k(h_{k,j}) - \check{\gamma}_{k,j}f_k(\check{h}_{k,j})\right| \\
&\overset{\text{(b)}}{\le} L_\phi\sum_{k=1}^{K}\frac{\pi_k}{N_k}\sup_{f_k\in\mathcal{F}_k}\left\{\Big|\gamma_{k,j}\Big|\Big|f_k(h_{k,j})\Big| + \Big|\check{\gamma}_{k,j}\Big|\Big|f_k(\check{h}_{k,j})\Big|\right\} \\
&\overset{\text{(c)}}{\le} 2L_\phi\beta\sum_{k=1}^{K}\frac{\pi_k}{N_k} \overset{\text{(d)}}{=} \frac{2\alpha L_\phi\beta}{N_{\mathsf{max}}}.
\end{aligned} \tag{6.119}$$

where (a) follows from the Lipschitz property of ϕ, (b) follows from the triangle inequality, (c) follows from the boundedness assumption $f_k(h) \leq \beta$ and the fact that $|\gamma_{k,n}| = 1$ for all k and i. Finally, in (d) we used the definition in (6.51), namely,

$$\alpha \triangleq \sum_{k=1}^{K} \pi_k \frac{N_{\mathsf{max}}}{N_k}. \tag{6.120}$$

Applying McDiarmid's Inequality in (6.105) with $c_j = 2\alpha L_\phi \beta / N_{\mathsf{max}}$, we obtain the following deviation bound:

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |R(f) - \widetilde{\boldsymbol{R}}(f)| - \mathbb{E} \sup_{f\in\mathcal{F}} |R(f) - \widetilde{\boldsymbol{R}}(f)| \geq t \right) \leq e^{-t^2 N_{\mathsf{max}}/(2\alpha^2 L_\phi^2 \beta^2)}, \tag{6.121}$$

holding for all $t > 0$. To conclude the proof, we seek to upper bound the second term inside the probability operator in (6.121). The result from Lemma 6.3 (Appendix 6.E) can be directly employed to conclude that:

$$\mathbb{E} \sup_{f\in\mathcal{F}} |R(f) - \widetilde{\boldsymbol{R}}(f)| \leq 4 L_\phi \rho. \tag{6.122}$$

In view of (6.122), we have that

$$\begin{aligned} &\sup_{f\in\mathcal{F}} |R(f) - \widetilde{\boldsymbol{R}}(f)| \geq t + 4L_\phi\rho \\ &\Rightarrow \sup_{f\in\mathcal{F}} |R(f) - \widetilde{\boldsymbol{R}}(f)| - \mathbb{E} \sup_{f\in\mathcal{F}} |R(f) - \widetilde{\boldsymbol{R}}(f)|) \geq t. \end{aligned} \tag{6.123}$$

From (6.123) and (6.121), we can conclude that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |R(f) - \widetilde{\boldsymbol{R}}(f)| \geq t + 4L_\phi\rho \right) \leq e^{-t^2 N_{\mathsf{max}}/(2\alpha^2 L_\phi^2 \beta^2)}. \tag{6.124}$$

Defining $x = t + 4L_\phi\rho$, and noting that $x > 4L_\phi\rho$ since $t > 0$, completes the proof of (6.102).

Proof of (6.103): The proof for the uniform bound in (6.103) follows similar arguments and will be thus presented in a concise manner. We start by using McDiarmid's Inequality with the following choice of function g:

$$g(h) = \sup_{f\in\mathcal{F}} \left| \sum_{k=1}^{K} \pi_k \left[\nu_k(f_k) - \frac{1}{N_k} \sum_{n=1}^{N_k} f_k(h_{k,n}) \right] \right|, \tag{6.125}$$

where we define the auxiliary quantity:

$$\nu_k(f_k) \triangleq \mathbb{E}_{h_k} f_k(\boldsymbol{h}_{k,n}). \tag{6.126}$$

We follow similar steps as the ones used to prove (6.102), which results in the following bound:

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\mu(f) - \widetilde{\boldsymbol{\mu}}(f)| - \mathbb{E} \sup_{f\in\mathcal{F}} |\mu(f) - \widetilde{\boldsymbol{\mu}}(f)| \geq t \right) \leq e^{-t^2 N_{\mathsf{max}}/(2\alpha^2 \beta^2)}. \tag{6.127}$$

We use again Lemma 6.3 (Appendix 6.E) to bound the second term inside the probability operation in (6.127). For this we take $L_\phi = 1$ and we take $\boldsymbol{\gamma}_n = 1$ as a deterministic variable, which allows us to derive the result:

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\mu(f) - \widetilde{\boldsymbol{\mu}}(f)| \leq 4\rho. \tag{6.128}$$

Replacing this bound in (6.127), defining $x = t + 4\rho$, with $x > 4\rho$, yields the final result. □

6.C Proof of Theorem 6.2

Assuming $\rho_k \leq C_k/\sqrt{N_k}$, it follows that (6.55) holds, i.e.,

$$\rho \leq \frac{\mathsf{C}}{\sqrt{N_{\mathsf{max}}}}. \tag{6.129}$$

For the bound in Theorem 6.1 to hold, the Rademacher complexity must satisfy

$$\rho \leq \mathscr{E}(\mathsf{R}^o). \tag{6.130}$$

In view of (6.129), (6.130) is met if we choose:

$$\frac{\mathsf{C}}{\sqrt{N_{\mathsf{max}}}} < \mathscr{E}(\mathsf{R}^o) \Leftarrow N_{\mathsf{max}} > \left(\frac{\mathsf{C}}{\mathscr{E}(\mathsf{R}^o)}\right)^2. \tag{6.131}$$

Next, for a desired minimum probability of consistent learning we should consider the bound found in (6.56). We have that:

$$\begin{aligned} & P_c \geq 1 - \varepsilon \\ & \Leftrightarrow 2\exp\left\{-\frac{8N_{\mathsf{max}}}{\alpha^2\beta^2}\left(\mathscr{E}(\mathsf{R}^o) - \frac{\mathsf{C}}{\sqrt{N_{\mathsf{max}}}}\right)^2\right\} \leq \varepsilon \\ & \Leftrightarrow N_{\mathsf{max}}\left(\mathscr{E}(\mathsf{R}^o) - \frac{\mathsf{C}}{\sqrt{N_{\mathsf{max}}}}\right)^2 \geq \frac{\alpha^2\beta^2}{8}\log\left(\frac{2}{\varepsilon}\right). \end{aligned} \tag{6.132}$$

We can develop the quadratic term in the LHS of (6.132) as

$$\begin{aligned} & N_{\mathsf{max}}\left(\mathscr{E}(\mathsf{R}^o) - \frac{\mathsf{C}}{\sqrt{N_{\mathsf{max}}}}\right)^2 \\ & = N_{\mathsf{max}}\left[\mathscr{E}(\mathsf{R}^o)\right]^2 - 2\sqrt{N_{\mathsf{max}}}\,\mathsf{C}\,\mathscr{E}(\mathsf{R}^o) + \mathsf{C}^2. \end{aligned} \tag{6.133}$$

Let

$$z = \sqrt{N_{\mathsf{max}}}\,\mathscr{E}(\mathsf{R}^o), \quad b = \mathsf{C}^2 - \frac{\alpha^2\beta^2}{8}\log\left(\frac{2}{\varepsilon}\right). \tag{6.134}$$

To solve the inequality in (6.132), we must study the following quadratic equality:

$$z^2 - 2\mathsf{C}z + b = 0, \tag{6.135}$$

whose positive solution is:

$$z = \mathsf{C} + \sqrt{\frac{\alpha^2\beta^2}{8}\log\left(\frac{2}{\varepsilon}\right)}. \tag{6.136}$$

Thus the inequality in (6.132) is satisfied whenever:

$$\sqrt{N_{\mathsf{max}}} > \frac{1}{\mathscr{E}(\mathsf{R}^o)}\left(\mathsf{C} + \sqrt{\frac{\alpha^2\beta^2}{8}\log\left(\frac{2}{\varepsilon}\right)}\right), \tag{6.137}$$

or yet when:

$$N_{\mathsf{max}} > \left(\frac{\mathsf{C}}{\mathscr{E}(\mathsf{R}^o)}\right)^2\left(1 + \frac{\alpha\beta}{2\mathsf{C}}\sqrt{\frac{1}{2}\log\left(\frac{2}{\varepsilon}\right)}\right)^2. \tag{6.138}$$

The final result of the theorem is established, since the bound in (6.138) is more stringent than (6.131).

6.D Proof of Proposition 6.1

Before introducing the proof, in order to establish the complexity of class $\mathcal{F}^{\mathsf{NN}}$, we will resort to a set of known inequalities involving the Rademacher complexity operator [113], [117], summarized in Property 6.1 (Appendix 6.E). The proof follows an inductive argument similar to the one used in [106], where we establish an upper bound for the Rademacher complexity of the output of one layer with respect to the output of the previous layer, then this bound is iterated over the depth of the Multilayer Perceptron (MLP).

We wish to analyze the complexity of the class of functions $\mathcal{F}^{\mathsf{NN}}$, which is defined in (6.62) as the difference between the outputs of the neural network z_1 and z_2, for an input vector $x \in \mathbb{R}^{n_0}$. That is, function f^{NN} has the following form (as seen in (6.62)):

$$f^{\mathsf{NN}}(x) = \log\frac{p(+1|x;f)}{p(-1|x;f)} = z_1 - z_2, \tag{6.139}$$

where z_1, z_2 implement functions $g^{(L)} \in \mathcal{G}^{(L)}$ as defined in (6.59) for $\ell = L$. We thus say that $f^{\mathsf{NN}} \in \mathcal{F}^{\mathsf{NN}}$, with

$$f^{\mathsf{NN}}(x) = g_1^{(L)}(x) - g_2^{(L)}(x), \tag{6.140}$$

where $g_1^{(L)}, g_2^{(L)} \in \mathcal{G}^{(L)}$.

From items 1 and 2 in Property 6.1 (Appendix 6.E). choosing $c = -1$, the empirical Rademacher complexity of $\mathcal{F}^{\mathsf{NN}}(x)$ will satisfy:

$$\begin{aligned}\mathcal{R}\left(\mathcal{F}^{\mathsf{NN}}(x)\right) &\leq \mathcal{R}\left(\mathcal{G}^{(L)}(x)\right) + \mathcal{R}\left(\mathcal{G}^{(L)}(x)\right) \\ &= 2\mathcal{R}\left(\mathcal{G}^{(L)}(x)\right).\end{aligned} \tag{6.141}$$

From (6.59), the Rademacher complexity of $\mathcal{G}^{(\ell)}(x)$ can be expressed as:

$$\mathcal{R}\left(\mathcal{G}^{(\ell)}(x)\right) = \mathbb{E}_r \sup_{w_j, g_j^{(\ell-1)}} \left| \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{r}_i \sum_{j=1}^{m} w_j \sigma\left(g_j^{(\ell-1)}(x_i)\right) \right|, \tag{6.142}$$

where $\boldsymbol{r}_i$ are independent and identically distributed Rademacher random variables, with $\mathbb{P}(\boldsymbol{r}_i = +1) = \mathbb{P}(\boldsymbol{r}_i = -1) = 1/2$. The term on the RHS of (6.142) can be rewritten as:

$$\begin{aligned}
&\mathbb{E}_r \sup_{w, g^{(\ell-1)}} \left| \frac{1}{N} \sum_{j=1}^{m} w_j \sum_{i=1}^{N} \boldsymbol{r}_i \sigma\left(g_j^{(\ell-1)}(x_i)\right) \right| \\
&\overset{\text{(a)}}{\leq} \mathbb{E}_r \sup_{w, g^{(\ell-1)}} \|w\|_1 \max_j \left| \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{r}_i \sigma\left(g_j^{(\ell-1)}(x_i)\right) \right| \\
&\overset{\text{(b)}}{\leq} b\, \mathbb{E}_r \sup_{g^{(\ell-1)}} \max_j \left| \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{r}_i \sigma\left(g_j^{(\ell-1)}(x_i)\right) \right| \\
&\overset{\text{(c)}}{=} b\, \mathcal{R}\left(\sigma \circ \mathcal{G}^{(\ell-1)}(x)\right) \overset{\text{(d)}}{\leq} 2bL_\sigma \mathcal{R}\left(\mathcal{G}^{(\ell-1)}(x)\right),
\end{aligned} \tag{6.143}$$

where (a) follows from triangle inequality and taking the maximum w.r.t. j, (b) follows from the assumption that $\|w\|_1 \leq b$, (c) follows from the fact that $g_j^{(\ell-1)} \in \mathcal{G}^{(\ell-1)}$ for all $j = 1, 2, \ldots, m$, (c) and (d) follows from the contraction principle (item 3 in Property 6.1, Appendix 6.E) in association with the assumption that σ is a Lipschitz function with constant L_σ.

Replacing (6.143) into (6.142), we have the following recursion:

$$\mathcal{R}\left(\mathcal{G}^{(\ell)}(x)\right) \leq 2bL_\sigma \mathcal{R}\left(\mathcal{G}^{(\ell-1)}(x)\right). \tag{6.144}$$

We can develop the recursion above across all layers up to ℓ:

$$\mathcal{R}\left(\mathcal{G}^{(\ell)}(x)\right) \leq (2bL_\sigma)^{\ell-1} \mathcal{R}\left(\mathcal{G}^{(1)}(x)\right). \tag{6.145}$$

It remains to bound the Rademacher complexity relative to $\mathcal{G}^{(1)}(x)$ of the first layer, whose functions have the form of $g_m^{(1)}$ defined in (6.60). For this purpose, we can directly use the result in Lemma 15 of [106], which bounds the Rademacher complexity of a linear separator with bounded ℓ_p norm. Applying this lemma with $p = 1$, $\gamma = b$ and $\|x\|_\infty = \max_i |x_i| \leq c$, we have:

$$\begin{aligned}
\mathcal{R}\left(\mathcal{G}^{(1)}(x)\right) &= \mathbb{E}_r \sup_{w_j} \left| \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{r}_i \sum_{j=1}^{d} w_j x_{i,j} \right| \\
&\leq \frac{2bc\sqrt{\log(2n_0)}}{\sqrt{N}}.
\end{aligned} \tag{6.146}$$

Replacing (6.145) with $\ell = L$ and (6.146) into (6.141) yields the final result.

6.E Auxiliary Lemmas

We list three key properties of Rademacher complexity, which are used in some of our results. These properties are well known and therefore are reported here without proof, which can be found in [113], [117].

Property 6.1 (Inequalities involving Rademacher complexity [113]**).** Let $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_K$ be classes of real-valued functions, and x a sequence of samples $\{x_1, x_2, \ldots, x_N\}$. The Rademacher complexity defined in (6.47) satisfies the following properties:

1. Subadditivity:
$$\mathcal{R}\left(\sum_{k=1}^{K} \mathcal{F}_k(x)\right) \leq \sum_{k=1}^{K} \mathcal{R}(\mathcal{F}_k(x)), \tag{6.147}$$
with $\mathcal{F}_1(x) + \mathcal{F}_2(x) \triangleq \{[f_1(x_1) + f_2(x_1), f_1(x_2) + f_2(x_2), \ldots, f_1(x_N) + f_2(x_N)] : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$.

2. Scaling: For every $c \in \mathbb{R}$,
$$\mathcal{R}(c\mathcal{F}(x)) \leq |c|\mathcal{R}(\mathcal{F}(x)), \tag{6.148}$$
where $c\mathcal{F}(x) \triangleq \{[cf(x_1), cf(x_2), \ldots, cf(x_N)] : f \in \mathcal{F}\}$.

3. Contraction principle: Let $\phi : \mathbb{R} \mapsto \mathbb{R}_+$ be Lipschitz with constant L_ϕ and $\phi(0) = 0$. Then:
$$\mathcal{R}(\phi \circ \mathcal{F}(x)) \leq 2L_\phi \mathcal{R}(\mathcal{F}(x)), \tag{6.149}$$
with $\phi \circ \mathcal{F}(x) \triangleq \{[\phi(f(x_1)), \phi(f(x_2)), \ldots, \phi(f(x_N))] : f \in \mathcal{F}\}$.

The next three lemmas are important auxiliary results used in the proofs of Theorem 6.1 and Theorem 6.3.

Lemma 6.2 (Upper bound on the estimation error for the empirical risk). *From the definitions in* (6.28)–(6.29) *and* (6.31)*, for $x > 0$ we have that:*

$$\mathbb{P}\left(R(\widetilde{\boldsymbol{f}}) - \mathsf{R}^o \geq x\right) \leq \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left|\widetilde{\boldsymbol{R}}(f) - R(f)\right| \geq \frac{x}{2}\right). \tag{6.150}$$

Proof. From (6.28) and (6.29), we can verify that for all $k = 1, 2, \ldots, K$:

$$\widetilde{\boldsymbol{R}}_k(\widetilde{\boldsymbol{f}}_k) \leq \widetilde{\boldsymbol{R}}(f_k), \text{ for all } f_k \in \mathcal{F}_k, \tag{6.151}$$

which imply, from the definitions in (6.31), that

$$\widetilde{\boldsymbol{R}}(\widetilde{\boldsymbol{f}}) \leq \widetilde{\boldsymbol{R}}(f), \text{ for all } f \in \mathcal{F}, \tag{6.152}$$

where $\mathcal{F}$ is the global class of functions defined in (6.111). We can develop the expression of the

estimation error to obtain the following uniform bound:

$$\begin{aligned} R(\widetilde{\boldsymbol{f}}) - \mathsf{R}^o &\overset{\text{(a)}}{=} R(\widetilde{\boldsymbol{f}}) - \inf_{f\in\mathcal{F}} R(f) \\ &= R(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{R}}(\widetilde{\boldsymbol{f}}) + \widetilde{\boldsymbol{R}}(\widetilde{\boldsymbol{f}}) - \inf_{f\in\mathcal{F}} R(f) \\ &= R(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{R}}(\widetilde{\boldsymbol{f}}) + \sup_{f\in\mathcal{F}} \Big(\widetilde{\boldsymbol{R}}(\widetilde{\boldsymbol{f}}) - R(f)\Big) \\ &\overset{\text{(b)}}{\leq} R(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{R}}(\widetilde{\boldsymbol{f}}) + \sup_{f\in\mathcal{F}} \Big(\widetilde{\boldsymbol{R}}(f) - R(f)\Big) \\ &\leq 2\sup_{f\in\mathcal{F}} \Big|\widetilde{\boldsymbol{R}}(f) - R(f)\Big|, \end{aligned} \tag{6.153}$$

where (a) follows from the definition in (6.28) and (b) follows from (6.152).

Finally, from (6.153) and using the target risk notation in (6.28), we note that

$$R(\widetilde{\boldsymbol{f}}) - \mathsf{R}^o \geq x \Rightarrow \sup_{f\in\mathcal{F}} \Big|\widetilde{\boldsymbol{R}}(f) - R(f)\Big| \geq x/2, \tag{6.154}$$

for any $x > 0$, thus concluding the proof. □

Lemma 6.3 (Uniform upper bound for Lipschitz cost functions). *Assume that the pair of sequences $(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)$ is sampled independently from the same joint distribution for all $n = 1, 2, \ldots, N_{\mathsf{max}}$. Let $f_k : \mathcal{H}_k \mapsto \mathbb{R}$ be a function belonging to class $\mathcal{F}_k$, and let $\phi : \mathbb{R} \mapsto \mathbb{R}_+$ be a $L_\phi-$Lipschitz function. Then it follows that*

$$\mathbb{E}_{h,\gamma} \sup_{f\in\mathcal{F}} \left| \sum_{k=1}^{K} \pi_k \left[\chi_k(f_k) - \frac{1}{N_k} \sum_{n=1}^{N_k} \phi(-\boldsymbol{\gamma}_{k,n} f_k(\boldsymbol{h}_{k,n})) \right] \right| \leq 4L_\phi \rho, \tag{6.155}$$

with

$$\chi_k(f_k) \triangleq \mathbb{E}_{h_k,\gamma_k} \phi(-\boldsymbol{\gamma}_{k,n} f_k(\boldsymbol{h}_{k,n})). \tag{6.156}$$

Proof. Introduce the artificial pair $\boldsymbol{h}'_n, \boldsymbol{\gamma}'_n$, sampled independently with the same joint distribution of $\boldsymbol{h}_n, \boldsymbol{\gamma}_n$. We develop the following symmetrization argument, inspired by the ones used in [110], [113].

First, we use the triangle inequality and the subadditive property of the supremum operator:

$$\begin{aligned} &\mathbb{E}_{h,\gamma} \sup_{f\in\mathcal{F}} \left| \sum_{k=1}^{K} \pi_k \left[\chi_k(f_k) - \frac{1}{N_k} \sum_{n=1}^{N_k} \phi(-\boldsymbol{\gamma}_{k,n} f_k(\boldsymbol{h}_{k,n})) \right] \right| \\ &\leq \sum_{k=1}^{K} \pi_k \mathbb{E}_{h_k,\gamma_k} \sup_{f_k\in\mathcal{F}_k} \left| \chi_k(f_k) - \frac{1}{N_k} \sum_{n=1}^{N_k} \phi(-\boldsymbol{\gamma}_{k,n} f_k(\boldsymbol{h}_{k,n})) \right|, \end{aligned} \tag{6.157}$$

where we recall that the argument f represents the ensemble of functions $\{f_k\}$, with $f_k \in \mathcal{F}_k$, and $\mathcal{F}$ denotes the global space of functions defined in (6.111).

We focus on the individual elements of the summation on the RHS of (6.157), that is, on each term indexed by k. We drop subscript k everywhere to simplify the notation.

$$
\begin{aligned}
&\mathbb{E}_{h,\gamma} \sup_{f\in\mathcal{F}} \left| \chi(f) - \frac{1}{N}\sum_{n=1}^{N} \phi(-\boldsymbol{\gamma}_n f(\boldsymbol{h}_n)) \right| \\
&\overset{\text{(a)}}{=} \mathbb{E}_{h,\gamma} \sup_{f\in\mathcal{F}} \left| \mathbb{E}_{h',\gamma'} \frac{1}{N}\sum_{n=1}^{N} \left[\phi(-\boldsymbol{\gamma}'_n f(\boldsymbol{h}'_n)) - \sum_{n=1}^{N} \phi(-\boldsymbol{\gamma}_n f(\boldsymbol{h}_n)) \right] \right| \\
&\overset{\text{(b)}}{\leq} \mathbb{E}_{h,\gamma} \mathbb{E}_{h',\gamma'} \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{n=1}^{N} \left[\phi(-\boldsymbol{\gamma}'_n f(\boldsymbol{h}'_n)) - \phi(-\boldsymbol{\gamma}_n f(\boldsymbol{h}_n)) \right] \right| \\
&\overset{\text{(c)}}{=} \mathbb{E}_{h,\gamma} \mathbb{E}_{h',\gamma'} \mathbb{E}_{r} \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{r}_n \left[\phi(-\boldsymbol{\gamma}'_n f(\boldsymbol{h}'_n)) - \phi(-\boldsymbol{\gamma}_n f(\boldsymbol{h}_n)) \right] \right| \\
&\overset{\text{(d)}}{\leq} 2\mathbb{E}_{h,\gamma} \mathbb{E}_{r} \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{r}_n \phi(-\boldsymbol{\gamma}_n f(\boldsymbol{h}_n)) \right| \\
&\overset{\text{(e)}}{\leq} 4L_\phi \mathbb{E}_{h,\gamma} \mathbb{E}_{r} \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{r}_n \boldsymbol{\gamma}_n f(\boldsymbol{h}_n) \right| \\
&\overset{\text{(f)}}{\leq} 4L_\phi \mathbb{E}_{h} \mathbb{E}_{r} \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{r}_n f(\boldsymbol{h}_n) \right| \\
&= 4L_\phi \mathbb{E}_{h} \mathcal{R}\left(\mathcal{F}(\boldsymbol{h})\right).
\end{aligned}
\tag{6.158}
$$

We explain now each of the steps (a)–(f) performed in (6.158). In (a) we used the i.i.d. property of the artificial samples $(\boldsymbol{h}'_n, \boldsymbol{\gamma}'_n)$, (b) follows from the following two properties: **i)** $|\mathbb{E}\,\boldsymbol{x}| \leq \mathbb{E}\,|\boldsymbol{x}|$; **ii)** $\sup_{f\in\mathcal{F}} \mathbb{E}\,|\boldsymbol{y}(f)| \leq \mathbb{E}\,\sup_{f\in\mathcal{F}} |\boldsymbol{y}(f)|$.

In (c) we introduced the i.i.d. Rademacher random variables, i.e., $\boldsymbol{r}_n \in \{-1, +1\}$ with uniform probability, which are independent of samples $(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)$ and $(\boldsymbol{h}'_n, \boldsymbol{\gamma}'_n)$. Since $(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)$ and $(\boldsymbol{h}'_n, \boldsymbol{\gamma}'_n)$ are identically distributed and independently sampled, exchanging $(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)$ and $(\boldsymbol{h}'_n, \boldsymbol{\gamma}'_n)$ is immaterial and therefore we can safely introduce the Rademacher random variables $\boldsymbol{r}_n$ in the summation.

In (d), we used the triangle inequality for the absolute value and the fact that $(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)$ and $(\boldsymbol{h}'_n, \boldsymbol{\gamma}'_n)$ are identically distributed. In (e), we use the Lipschitz property of ϕ associated with the contraction principle of the Rademacher complexity (item 3 in Property 6.1) to conclude that:

$$
\mathbb{E}_{r} \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{r}_n \phi(-\boldsymbol{\gamma}_n f(\boldsymbol{h}_n)) \right| \leq 2L_\phi \mathbb{E}_{r} \sup_{f\in\mathcal{F}} \left| \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{r}_n \boldsymbol{\gamma}_n f(\boldsymbol{h}_n) \right|. \tag{6.159}
$$

Step (f) follows from similar symmetrization arguments, considering that $\boldsymbol{\gamma}_n$ assumes values ± 1 and that $\boldsymbol{r}_n$ and $-\boldsymbol{r}_n$ are equally distributed and independent form the samples and over i. Finally replacing (6.158) into (6.157) for each of the summands indexed by k, for all terms indexed by k, and recalling the definition of ρ in (6.50), we obtain (6.155). □

Lemma 6.4 (Auxiliary result for bounded differences). *Assume $S(f)$ and $T(f)$ are operators dependent on a real-valued function $f \in \mathcal{F}$, and consider the following quantities:*

$$s_1 = \sup_{f\in\mathcal{F}} \Big|S(f)\Big|, \quad s_2 = \sup_{f\in\mathcal{F}} \Big|S(f) + T(f)\Big|. \tag{6.160}$$

Then, we have that:

$$|s_1 - s_2| \leq \sup_{f\in\mathcal{F}} \Big|T(f)\Big|. \tag{6.161}$$

Proof. The proof is split in two cases.

a) Case $s_2 \geq s_1$:

$$\begin{aligned} s_2 - s_1 &= \sup_{f\in\mathcal{F}} \Big|S(f) + T(f)\Big| - \sup_{f\in\mathcal{F}} \Big|S(f)\Big| \\ &\leq \sup_{f\in\mathcal{F}} \Big|S(f)\Big| + \sup_{f\in\mathcal{F}} \Big|T(f)\Big| - \sup_{f\in\mathcal{F}} \Big|S(f)\Big| = \sup_{f\in\mathcal{F}} \Big|T(f)\Big|, \end{aligned} \tag{6.162}$$

where the inequality follows from the triangle inequality and the subadditive property of the supremum operator, i.e., $\sup_{f\in\mathcal{F}}[a(f) + b(f)] \leq \sup_{f\in\mathcal{F}} a(f) + \sup_{f\in\mathcal{F}} b(f)$.

b) Case $s_2 < s_1$:

$$\begin{aligned} s_1 - s_2 &= \sup_{f\in\mathcal{F}} \Big|S(f)\Big| - \sup_{f\in\mathcal{F}} \Big|S(f) + T(f)\Big| \\ &= \sup_{f\in\mathcal{F}} \left(\Big|S(f)\Big| - s_2\right) \\ &\overset{\text{(a)}}{\leq} \sup_{f\in\mathcal{F}} \left(\Big|S(f)\Big| - \Big|S(f) + T(f)\Big|\right) \\ &\leq \sup_{f\in\mathcal{F}} \Big|\,\Big|S(f)\Big| - \Big|S(f) + T(f)\Big|\,\Big| \\ &\overset{\text{(b)}}{\leq} \sup_{f\in\mathcal{F}} \Big|S(f) - S(f) + T(f)\Big| = \sup_{f\in\mathcal{F}} \Big|T(f)\Big|, \end{aligned} \tag{6.163}$$

where (a) follows from the definition of s_2, and (b) from the reverse triangle inequality, i.e., $|a - b| \geq |\,|a| - |b|\,|$.

Using (6.162) and (6.163), we obtain the desired result in (6.161). □

7 Conclusions

In this thesis, we studied multiple aspects of social learning strategies. Our investigations allowed us to examine questions regarding the topology learning problem and the effect of sharing partial information under stationary conditions—seen in Part I of the thesis. They also enabled us to overcome two critical assumptions in social learning, namely, that world conditions are stationary and that perfect statistical models are available. Both premises are neither realistic to model social dynamics since the world with which we interact is nonstationary, nor desirable in a decision-making system since the statistical models used by sensors in the system are usually the result of a training process. The solutions proposed in Part II of the thesis take us one step closer to making social learning strategies suitable for fully data-based applications.

7.1 Summary of Main Results

In Chapter 3, we considered the *network* aspect of social learning and addressed the reverse learning problem in weakly connected networks. The weakly connected network models the existence of influence dynamics in real social networks, wherein some *influential* (sending) subnetworks have control over the opinions at *influenced* (receiving) subnetworks. In Chapter 3, we propose to learn the amount of power exerted by each sending subnetwork on each receiving agent, from observing the belief evolution of the latter. The reverse learning task is formally posed as a topology learning problem. We show that a necessary condition for the problem to be feasible is that the number of hypotheses is greater or equal than the number of sending subnetworks. More specifically, we show that when the likelihood models across sending networks present *little diversity*, i.e., they all belong to the same family of Gaussian distributions, the topology learning problem is not feasible in general. This is mitigated when the models across sending networks have *greater diversity*, i.e., they do not follow a fixed family of distributions, under which the problem is almost always feasible.

In Chapter 4, we tapped into the *exchanged information* in social learning and addressed the problem of sharing partial information in strongly connected networks. Instead of allowing agents to share their full belief vectors with neighbors, we constrain the communication among agents even further. They can share only one *hypothesis of interest* at all times. The objective of the network is to assess the validity of such hypothesis, i.e., whether it corresponds to the true

state of the world or not. We show that the network learns the truth unequivocally when the true hypothesis is shared. When the shared hypothesis is not the truth but sufficiently close to it, the network can converge to a wrong conclusion. The exact limitations to this approach are analytically detailed in Chapter 4.

In Chapter 5, we consider nonstationary *world* conditions. Social learning is designed for stationary environments, and performs poorly when the environment changes, e.g., when there is a change in the underlying true state. We explain this phenomenon and propose an *adaptive* social learning strategy, which allows agents to adapt their opinion in view of changing world conditions. A step-size parameter is introduced to the formulation allowing agents to exploit a trade-off between *learning performance* and *adaptation*. We characterize the steady-state behavior and show that consistent learning occurs for small step sizes. In particular, we show that the steady-state error probability decreases exponentially with a decreasing step size. We then investigate the transient behavior and show that the adaptation time decreases with a growing step size, therefore characterizing how the aforementioned trade-off, between learning accuracy and adaptation, relies on the step-size parameter.

In Chapter 6, we consider imperfect likelihood *models*. Social learning assumes that likelihood models are exactly known. In this chapter, we propose instead a fully data-based strategy, in which these models are *trained* using a finite amount of data. We also cast the social learning problem into a distributed classification problem under streaming observations. We propose a machine learning framework, where, in a first stage, multiple classifiers, belonging to a fairly general functional class, are independently trained given *heterogeneous* features, and in a second stage, these classifiers collaborate to classify streaming unlabeled observations in a distributed manner. We show that this structure results in consistent learning with high probability, and characterize how the number of training samples should scale as a function of different parameters of the learning problem. Contrary to traditional boosting solutions, the proposed solution also enables agents to continually improve accuracy over time.

7.2 Future Directions

Partial Information

The results in Chapter 4 suggest interesting future directions of research. In contrast to the memoryless approach to partial information introduced in (4.3), we can also consider a *memory-aware* approach, namely, for $\theta \neq \theta_{\mathsf{TX}}$:

$$\widehat{\boldsymbol{\psi}}_{\ell k,i}(\theta) = \frac{\boldsymbol{\psi}_{k,i}(\theta)}{1-\boldsymbol{\psi}_{k,i}(\theta_{\mathsf{TX}})}(1-\boldsymbol{\psi}_{\ell,i}(\theta_{\mathsf{TX}})). \tag{7.1}$$

The memory-aware choice allows agent k to use local prior knowledge, in the form of their intermediate beliefs $\boldsymbol{\psi}_{k,i}$, to fill in the knowledge gap regarding the non-transmitted components received from its neighbor ℓ. Preliminary results using this strategy suggest that the mislearning scenarios of the memoryless approach can be completely avoided.

Another interesting extension is to consider that the *global* hypothesis of interest evolves

randomly over time. This extension finds motivation in the fact that real social networks do not always discuss the same topics or hypotheses. Instead, discussions follow trends depending on contemporary events. A first setup with randomized transmitted hypothesis is investigated in [118], where the hypothesis of interest is sampled from a fixed distribution at every time instant. The authors found that the randomization allows agents to overcome communication constraints and learn the truth in the traditional social learning sense. More involved scenarios would be to consider that this global hypothesis of interest evolves according to a Markov chain, mimicking the setting in which a topic of discussion evolves over time with a certain coherence with respect to the past. Another more intricate scenario would take into account that subsets of the network discuss different topics at the same time, in which case the transmitted hypothesis is a *localized* instead of a global random variable.

Privacy in Social Learning

In social interactions, a key human concern is to preserve privacy. People are usually unwilling to share information that is deemed too personal or that might make them feel disapproved of by other individuals. An example is the Bradley effect in political polls, where voters give inaccurate answers for fear of criticism from society. Similar privacy concerns can be found when information is exchanged in distributed engineering systems. Agents do not want to give away any information regarding their private observations, except the strict necessary for performing the distributed learning task. Therefore, an interesting future extension is to add a privacy mechanism to social learning.

For that purpose, the concept of differential privacy [119] is frequently used. Its main strategy is to add noise to the algorithm output, and it provides guarantees that the output does not carry information about private data samples. The strategy has been used in distributed learning [120], [121], with successful privacy-preserving results at the expense, however, of their learning performance. To mitigate the loss in performance, graph-homomorphic perturbations, as shown in [122], can be used to enable differential privacy.

Social Machine Learning

In Chapter 6, the social machine learning framework uses a set of classifiers distributed in space to perform social learning during the prediction phase over a growing data stream. We require, however, a large amount of streaming observations, i.e., $i \to \infty$, to guarantee consistent learning during prediction with high probability. In other words, the expression of consistent learning found Theorem 6.1 takes into account implicitly the steady-state performance during prediction.

A meaningful extension would be to characterize the learning performance assuming a finite number of unlabeled samples in the prediction phase. The expression of consistent learning found in Theorem 6.1 should be modified to incorporate an additional multiplicative probability term that should approach 1 as i grows. This important contribution would allow us to exploit, not only the number of training samples N_k for $k = 1, 2, \ldots, K$, but also the length of the prediction data stream as an additional dimension of the social learning problem.

Bibliography

[1] N. De Condorcet, *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Imprimerie Royale, 1785.

[2] F. Galton, "Vox populi (the wisdom of crowds)", *Nature*, vol. 75, no. 7, pp. 450–451, 1907.

[3] D. Acemoglu and A. Ozdaglar, "Opinion dynamics and learning in social networks", *Dynamic Games and Applications*, vol. 1, no. 1, pp. 3–49, 2011.

[4] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning", *Games and Economic Behavior*, vol. 76, no. 1, pp. 210–225, 2012.

[5] X. Zhao and A. H. Sayed, "Learning over social networks via diffusion adaptation", in *Proc. Asilomar Conference on Signals, Systems and Computers*, 2012, pp. 709–713.

[6] C. Chamley, A. Scaglione, and L. Li, "Models for the diffusion of beliefs in social networks: An overview", *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 16–29, 2013.

[7] B. Golub and E. Sadler, "Learning in social networks", *Available at SSRN: https://ssrn.com/abstract=2919146*, 2017.

[8] E. Mossel and O. Tamuz, "Opinion exchange dynamics", *Probability Surveys*, vol. 14, pp. 155–204, 2017.

[9] A. Zellner, "Optimal information processing and Bayes's theorem", *The American Statistician*, vol. 42, no. 4, pp. 278–280, 1988.

[10] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain", *Journal of physiology-Paris*, vol. 100, no. 1-3, pp. 70–87, 2006.

[11] M. Oaksford and N. Chater, *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press, 2007.

[12] R. H. Berk, "Limiting behavior of posterior distributions when the model is incorrect", *The Annals of Mathematical Statistics*, vol. 37, no. 1, pp. 51–58, 1966.

[13] J. L. Doob, "Application of the theory of martingales", *Le Calcul des Probabilités et ses Applications*, pp. 23–27, 1949.

[14] D. A. Freedman, "On the asymptotic behavior of Bayes' estimates in the discrete case", *The Annals of Mathematical Statistics*, vol. 34, no. 4, pp. 1386–1403, 1963.

[15] D. A. Freedman, "On the asymptotic behavior of Bayes' estimates in the discrete case ii", *The Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 454–456, 1965.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1991.

[17] M. H. DeGroot, *Optimal Statistical Decisions*. John Wiley & Sons, 2005.

[18] P. Billingsley, *Probability and Measure*. John Wiley & Sons, 2008.

[19] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions", in *Proc. Berkeley Symposium on Mathematical Statistics and Probability*, 1967, p. 221.

[20] H. Akaike, "Information theory and an extension of the maximum likelihood principle", in *Proc. International Symposium of Information Theory*, 1973, pp. 267–281.

[21] H. White, "Maximum likelihood estimation of misspecified models", *Econometrica*, pp. 1–25, 1982.

[22] J. Hązła, A. Jadbabaie, E. Mossel, and M. A. Rahimian, "Bayesian decision making in groups is hard", *Operations Research*, vol. 69, no. 2, pp. 632–654, 2021.

[23] A. V. Banerjee, "A simple model of herd behavior", *The Quarterly Journal of Economics*, vol. 107, no. 3, pp. 797–817, 1992.

[24] S. Bikhchandani, D. Hirshleifer, and I. Welch, "Learning from the behavior of others: Conformity, fads, and informational cascades", *Journal of Economic Perspectives*, vol. 12, no. 3, pp. 151–170, 1998.

[25] L. Smith and P. Sørensen, "Pathological outcomes of observational learning", *Econometrica*, vol. 68, no. 2, pp. 371–398, 2000.

[26] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks", *The Review of Economic Studies*, vol. 78, no. 4, pp. 1201–1236, 2011.

[27] Y. Kanoria and O. Tamuz, "Tractable Bayesian social learning on trees", in *Proc. IEEE International Symposium on Information Theory*, 2012, pp. 2721–2725.

[28] E. Mossel and O. Tamuz, "Making consensus tractable", *ACM Transactions on Economics and Computation (TEAC)*, vol. 1, no. 4, pp. 1–19, 2013.

[29] H. A. Simon, "Bounded rationality", in *Utility and Probability*, Springer, 1990, pp. 15–18.

[30] J. Conlisk, "Why bounded rationality?", *Journal of Economic Literature*, vol. 34, no. 2, pp. 669–700, 1996.

[31] M. H. DeGroot, "Reaching a consensus", *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[32] R. L. Berger, "A necessary and sufficient condition for reaching a consensus using DeGroot's method", *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 415–418, 1981.

[33] A. H. Sayed, "Diffusion adaptation over networks", in *Academic Press Library in Signal Processing*, vol. 3, Elsevier, 2014, pp. 323–454.

[34] L. G. Epstein, J. Noor, A. Sandroni, *et al.*, "Non-Bayesian learning", *The BE Journal of Theoretical Economics*, vol. 10, no. 1, pp. 1–20, 2010.

[35] A. G. Dimakis, S. Kar, J. M. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing", *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[36] A. H. Sayed, "Adaptive networks", *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, 2014.

[37] A. H. Sayed, "Adaptation, learning, and optimization over networks", *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[38] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, "Diffusion strategies for adaptation and learning over networks: An examination of distributed strategies and network behavior", *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 155–171, 2013.

[39] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed detection: Finite-time analysis and impact of network topology", *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3256–3268, 2015.

[40] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-Bayesian learning", *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5538–5553, 2017.

[41] A. Lalitha, T. Javidi, and A. D. Sarwate, "Social learning and distributed hypothesis testing", *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6161–6179, 2018.

[42] V. Matta, V. Bordignon, A. Santos, and A. H. Sayed, "Interplay between topology and social learning over weak graphs", *IEEE Open Journal of Signal Processing*, vol. 1, pp. 99–119, 2020.

[43] P. Molavi, A. Tahbaz-Salehi, and A. Jadbabaie, "A theory of non-Bayesian social learning", *Econometrica*, vol. 86, no. 2, pp. 445–490, 2018.

[44] C. A. Uribe, A. Olshevsky, and A. Nedich, "Non-asymptotic concentration rates in cooperative learning part I: Variational non-Bayesian social learning", *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 1–1, 2022.

[45] M. Kayaalp, Y. Inan, E. Telatar, and A. H. Sayed, "On the arithmetic and geometric fusion of beliefs for distributed inference", *arXiv:2204.13741*, Apr., 2022.

[46] Y. Inan, M. Kayaalp, E. Telatar, and A. H. Sayed, "Social learning under randomized collaborations", *arXiv:2201.10957*, Jan., 2022.

[47] B. Ying and A. H. Sayed, "Information exchange and learning dynamics over weakly connected adaptive networks", *IEEE Transactions on Information Theory*, vol. 62, no. 3, pp. 1396–1414, 2016.

[48] H. Salami, B. Ying, and A. H. Sayed, "Social learning over weakly connected graphs", *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 222–238, 2017.

[49] V. Matta, A. Santos, and A. H. Sayed, "Exponential collapse of social beliefs over weakly-connected heterogeneous networks", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5267–5271.

[50] C. A. Uribe, A. Olshevsky, and A. Nedich, "Non-asymptotic concentration rates in cooperative learning part II: Inference on compact hypothesis sets", *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 1–1, 2022.

[51] V. Matta, V. Bordignon, A. Santos, and A. H. Sayed, "Learning graph influence from social interactions", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 5535–5539.

[52] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms", *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[53] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging", *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[54] A. Speranzon, C. Fischione, and K. H. Johansson, "Distributed and collaborative estimation over wireless sensor networks", in *Proc. IEEE Conference on Decision and Control*, 2006, pp. 1025–1030.

[55] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation", *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1035–1048, 2009.

[56] S. Kar and J. M. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures", *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1383–1400, 2009.

[57] A. Nedić and A. Ozdaglar, "Cooperative distributed multi-agent optimization", in *Convex Optimization in Signal Processing and Communications*, Cambridge University Press, 2009, pp. 340–386.

[58] S. U. Pillai, T. Suel, and S. Cha, "The Perron-Frobenius theorem: Some of its applications", *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 62–75, 2005.

[59] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.

[60] H. Salami, B. Ying, and A. H. Sayed, "Belief control strategies for interactions over weakly-connected graphs", *IEEE Open Journal of Signal Processing*, vol. 2, pp. 265–279, 2021.

[61] V. Matta and A. H. Sayed, "Consistent tomography under partial observations over adaptive networks", *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 622–646, 2018.

[62] A. Santos, V. Matta, and A. H. Sayed, "Local tomography of large networks under the low-observability regime", *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 587–613, 2019.

[63] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing", *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 16–43, 2019.

[64] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean distance matrices: Essential theory, algorithms, and applications", *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12–30, 2015.

[65] P. Wedin, "Perturbation theory for pseudo-inverses", *BIT Numerical Mathematics*, vol. 13, no. 2, pp. 217–232, 1973.

[66] J. Baksalary and R Kala, "The matrix equation $AX - YB = C$", *Linear Algebra and its Applications*, vol. 25, pp. 41–43, 1979.

[67] J. C. Gower, "Properties of Euclidean and non-Euclidean distance matrices", *Linear algebra and its applications*, vol. 67, pp. 81–97, 1985.

[68] V. Bordignon, V. Matta, and A. H. Sayed, "Social learning with partial information sharing", *arXiv:2006.13659*, Jun., 2020.

[69] V. Bordignon, V. Matta, and A. H. Sayed, "Social learning with partial information sharing", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 5540–5544.

[70] J. Mills, J. Hu, and G. Min, "Communication-efficient federated learning for wireless edge intelligence in iot", *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 5986–5994, 2020.

[71] P. S. Bullen, *Handbook of Means and their Inequalities*. Springer, 2013, vol. 560.

[72] W. Rudin *et al.*, *Principles of Mathematical Analysis*. McGraw-Hill, 1976, vol. 3.

[73] J. Shao, *Mathematical Statistics*. Springer, 2003.

[74] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, 1993.

[75] M. Loève, "On almost sure convergence", in *Proc. Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, pp. 279–303.

[76] V. Bordignon, V. Matta, and A. H. Sayed, "Adaptation in online social learning", in *Proc. European Signal Processing Conference (EUSIPCO)*, 2020, pp. 2170–2174.

[77] V. Bordignon, V. Matta, and A. H. Sayed, "Adaptive social learning", *IEEE Transactions on Information Theory*, vol. 67, no. 9, pp. 6053–6081, 2021.

[78] V. Matta and A. H. Sayed, "Estimation and detection over adaptive networks", in *Cooperative and Graph Signal Processing*, Elsevier, 2018, pp. 69–106.

[79] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Diffusion-based adaptive distributed detection: Steady-state performance in the slow adaptation regime", *IEEE Transactions on Information Theory*, vol. 62, no. 8, pp. 4710–4732, 2016.

[80] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Springer, 2010, vol. 38.

[81] F. Den Hollander, *Large Deviations*. American Mathematical Society, 2000, vol. 14.

[82] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Distributed detection over adaptive networks: Refined asymptotics and the role of connectivity", *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 442–460, 2016.

[83] E. N. Gilbert, "Capacity of a burst-noise channel", *The Bell System Technical Journal*, vol. 39, no. 5, pp. 1253–1265, 1960.

[84] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels", *The Bell System Technical Journal*, vol. 42, no. 5, pp. 1977–1997, 1963.

[85] P. Hu, V. Bordignon, S. Vlaski, and A. H. Sayed, "Optimal aggregation strategies for social learning over graphs", *arXiv:2203.07065*, Mar., 2022.

[86] M. Kayaalp, V. Bordignon, S. Vlaski, and A. H. Sayed, "Hidden Markov modeling over graphs", *arXiv:2111.13626*, Nov., 2021.

[87] V. Shumovskaia, K. Ntemos, S. Vlaski, and A. H. Sayed, "Online graph learning from social interactions", in *Proc. Asilomar Conference on Signals, Systems, and Computers*, 2021, pp. 1263–1267.

[88] V. Shumovskaia, K. Ntemos, S. Vlaski, and A. H. Sayed, "Explainability and graph learning from social interactions", *arXiv:2203.07494*, Mar., 2022.

[89] P. Hu, V. Bordignon, S. Vlaski, and A. H. Saye, "Optimal combination policies for adaptive social learning", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 5842–5846.

[90] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, 2019, vol. 49.

[91] W. Feller, *An Introduction to Probability Theory and its Applications, vol 2*. John Wiley & Sons, 1971.

[92] V. Bordignon, S. Vlaski, V. Matta, and A. H. Sayed, "Network classifiers based on social learning", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5185–5189.

[93] V. Bordignon, S. Vlaski, V. Matta, and A. H. Sayed, "Learning from heterogeneous data based on social interactions over graphs", *arXiv:2112.09483*, Dec., 2021.

[94] J. Z. Hare, C. A. Uribe, L. Kaplan, and A. Jadbabaie, "Non-Bayesian social learning with uncertain models", *IEEE Transactions on Signal Processing*, vol. 68, pp. 4178–4193, 2020.

[95] J. Z. Hare, C. A. Uribe, L. Kaplan, and A. Jadbabaie, "A general framework for distributed inference with uncertain models", *IEEE Transactions on Signal and Information Processing over Networks*, vol. 7, pp. 392–405, 2021.

[96] T. G. Dietterich, "Ensemble methods in machine learning", in *Proc. International Workshop on Multiple Classifier Systems*, 2000, pp. 1–15.

[97] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges", *Information Fusion*, vol. 38, pp. 43–54, 2017.

[98] L. Breiman, "Bagging predictors", *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[99] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[100] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training", in *Proc. Conference on Computational Learning Theory*, 1998, pp. 92–100.

[101] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors I. Fundamentals", *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, 1997.

[102] R. S. Blum, S. A. Kassam, and H. V. Poor, "Distributed detection with multiple sensors II. Advanced topics", *Proceedings of the IEEE*, vol. 85, no. 1, pp. 64–79, 1997.

[103] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT Press, 2018.

[104] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[105] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[106] B. Neyshabur, R. Tomioka, and N. Srebro, "Norm-based capacity control in neural networks", in *Proc. Conference on Learning Theory*, 2015, pp. 1376–1401.

[107] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities", in *Measures of Complexity*, Springer, 2015, pp. 11–30.

[108] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 2013, vol. 31.

[109] V. Koltchinskii, "Rademacher penalties and structural risk minimization", *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1902–1914, 2001.

[110] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: A survey of some recent advances", *ESAIM: Probability and Statistics*, vol. 9, pp. 323–375, 2005.

[111] P. L. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation", *Machine Learning*, vol. 48, no. 1, pp. 85–113, 2002.

[112] C. Cortes, M. Mohri, and U. Syed, "Deep boosting", in *Proc. International Conference on Machine Learning*, 2014, pp. 1179–1187.

[113] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results", *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[114] Y. LeCun, C. Cortes, and C. J. Burges, *MNIST handwritten digit database*, ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2010.

[115] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting", *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[116] C. McDiarmid, "On the method of bounded differences", in *Surveys in Combinatorics*, vol. 141, Cambridge University Press, 1989, pp. 148–188.

[117] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2013.

[118] M. Kayaalp, V. Bordignon, and A. H. Sayed, "Random information sharing over social networks", *arXiv:2203.02466*, Mar., 2022.

[119] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy", *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[120] M. Pathak, S. Rane, and B. Raj, "Multiparty differential privacy via aggregation of locally trained classifiers", vol. 23, 2010, pp. 1–9.

[121] C. Li, P. Zhou, L. Xiong, Q. Wang, and T. Wang, "Differentially private distributed online learning", *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 8, pp. 1440–1453, 2018.

[122] S. Vlaski and A. H. Sayed, "Graph-homomorphic perturbations for private decentralized learning", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5240–5244.

Curriculum Vitae

Virginia Bordignon
virginia.bordignon@epfl.ch

Education

PhD in Electrical Engineering , Adaptive Systems Laboratory, Ecole Polytechnique Fédérale de Lausanne (EPFL) *Sep 2018-Sep 2022*

Masters in Electrical Engineering , Data-Driven Control Group, Universidade Federal do Rio Grande do Sul (UFRGS) *Sep 2016-Aug 2018*

Masters in Engineering , Ecole Centrale de Lyon, Double degree with Universidade Federal do Rio Grande do Sul (UFRGS) *Sep 2012-Aug 2016*

Bachelor in Electrical Engineering , Universidade Federal do Rio Grande do Sul (UFRGS): Double degree with Ecole Centrale de Lyon *Feb 2010-Aug 2016*

Research experience

Adaptive Systems Laboratory, EPFL, Doctoral thesis: "Opinion Formation over Adaptive Networks". Supervisor: Prof. Ali H. Sayed *Sep 2018-Sep 2022*

Data-Driven Control Group, UFRGS, Master thesis: "Optimal Criterion for Regulatory Data-Based Control Design". Supervisor: Prof. Luciola Campestrini *Sep 2016-Aug 2018*

Additional experience

AEL Sistemas, Porto Alegre, Brazil, Software Engineering Intern *Jun 2015-Jun 2016*

Volvo Group, Lyon, France, Electronic Vehicle Architecture Intern *Apr 2014-Nov 2014*

Awards

Research Scholarship, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil *Sep 2016*

Diploma Cum Laude, School of Engineering, UFRGS, Brazil *Aug 2016*

Eiffel Excellence Scholarship, Ministère des Affaires Etrangères, France *Sep 2012*

Journal publications or submissions

V. Bordignon, S. Vlaski, V. Matta, and A. H. Sayed, "Learning from heterogeneous data based on social interactions over graphs," *arXiv:2112.09483*, Dec. 2021. Submitted and under review.

V. Bordignon, V. Matta, and A. H. Sayed, "Adaptive social learning," in *IEEE Transactions on Information Theory*, vol. 67, no. 9, pp. 6053-6081, Sep. 2021.

V. Matta, V. Bordignon, A. Santos, and A. H. Sayed, "Interplay between topology and social learning over weak graphs," in *IEEE Open Journal of Signal Processing*, vol. 1, pp. 99-119, Jul. 2020.

V. Bordignon, V. Matta, and A. H. Sayed, "Social learning with partial information sharing," *arXiv:2006.13659*, Jun. 2020. Submitted and under review.

Conference publications or submissions

K. Ntemos, V. Bordignon, S. Vlaski, and A. H. Sayed, "Social learning with disparate hypotheses," in *Proc. European Signal Processing Conference (EUSIPCO)*, Sep. 2022.

M. Kayaalp, V. Bordignon, S. Vlaski, and A. H. Sayed, "Hidden Markov modeling over graphs," in *Proc. IEEE Data Science and Learning Workshop*, Jun. 2022.

R. Nassif, V. Bordignon, S. Vlaski, and A. H. Sayed, "Decentralized learning in the presence of low-rank noise," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2022, pp. 5667-5671.

P. Hu, V. Bordignon, S. Vlaski, and A. H. Sayed, "Optimal combination policies for adaptive social learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2022, pp. 5842-5846.

V. Bordignon, S. Vlaski, V. Matta, and A. H. Sayed, "Network classifiers based on social learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 5185-5189.

K. Ntemos, V. Bordignon, S. Vlaski and A. H. Sayed, "Social learning under inferential attacks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 5479-5483.

V. Bordignon, V. Matta, and A. H. Sayed, "Adaptation in online social learning," in *Proc. European Signal Processing Conference (EUSIPCO)*, Jan. 2021, pp. 2170-2174.

V. Matta, V. Bordignon, A. Santos, and A. H. Sayed, "Learning graph influence from social interactions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 5535-5539.

V. Bordignon, V. Matta, and A. H. Sayed, "Social learning with partial information sharing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 5540-5544.

Skills

Languages Coded: Python, C, Matlab (Simulink), Latex, …

Languages Spoken: English, French, German, Spanish, Portuguese