LEARNING FROM OPINIONS

BY

NOYAN CEM SEVÜKTEKİN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Doctoral Committee:

Professor Andrew Carl Singer, Chair
Associate Professor Maxim Raginsky
Associate Professor Mohamed-Ali Belabbas
Assistant Professor Alexander Gerhard Schwing

ABSTRACT

Modern engineering challenges motivate a transition from conventional systems that rely on measurements of physical quantities to systems that interpret and respond to subjective evaluations of the world. Different from engineering problems that have quantifiable objectives, such as controlling a system based on noisy measurements or transmitting information through a medium, sources that provide subjective information, which we will refer to as "experts", evaluate the world based on a potentially hidden rationale. Learning, inference, and decision making based on subjective evaluations, or opinions, are not only common aspects of human learning but they are fundamental engineering challenges due to the hidden uncertainty. The objective of this work is to establish fundamentals of learning from opinions by addressing key problems that rely on subjective information with hidden models. Specifically, Chapter 2 focuses on sequential consultation of experts, Chapter 3 investigates statistical methods for opinion aggregation, Chapter 4 addresses fidelity-based error detection and mitigation, and Chapter 5 studies the impact of high-dimensional uncertainty on networks.

Contextually, an opinion has associated costs. Consulting an expert incurs among others, time, resource, and opportunity costs. Particularly in engineering systems, such costs further manifest as circuit-area, system-complexity, runtime, or memory requirements. The conventional decision-making framework with pre-allocated resources might not necessarily capture the trade-off between the utility to be gained by consulting an expert and the associated costs. Sequential consultation of experts arises naturally in this context, where the objective is to decide whether to consult another expert or to make a decision based on the opinions received up to that time. In this context, the true utility of consulting another expert does not only depend on the cost associated with consulting or the individual expertise, but it also depends on the instantaneous decision strength based on the statistics hitherto. A fundamental challenge is to find a sequential strategy that addresses this trade-off. In Chapter 2, we show that the strategy achieving maximum expected reward is in the form of a sequential likelihood ratio test, where a unique threshold function depends on the cost-performance trade-off of all future experts to be consulted.

Reliable mathematical models for experts might be difficult to obtain or quantify, even in some cases impossible, due to the inherent subjectivity of a task, limited insight that training might yield for real-world encounters, or due to massively high-dimensional space from which an expert might build a rationale for decision making. However, difficulty of modeling does not necessarily render statistical inference implausible. It is often reasonable to accept experts as honest-but-fallible sources of information that do not purposefully deceive the decision maker. Populations comprising such experts are less subjective than their individual constituents and a natural understanding of correctness arises: When objective truth is not achievable, one might choose to accept the consensus of opinions as truth to the best of one's knowledge. This leads to an alternative notion of reliability, termed "pseudo competence", which in turn allows reliable statistical inference. In Chapter 3, we show that pseudo competences can be estimated empirically on test data by centralized computation or they can be estimated in distribution on strongly connected networks. We further show that opinion aggregation mechanisms that use pseudo competences can, in some cases, achieve performance comparable to decision rules that have reliable models for experts.

Experts as error-prone computational units are often subject to unknown, or high-dimensional, failure mechanisms. However, the robustness of a computational unit can be inferred relatively reliably from the corresponding system complexity, motivating fidelity-based safe-guarding mechanisms against what is often called, "black swan" events; failures that happen with low probability yet have high impact on the system. A method for jointly testing for failure and bypassing erroneous outcomes, called algorithmic noise tolerance, uses computational units that are robust yet of lower fidelity to safeguard the system against high-impact errors from high-fidelity computational units, without requiring exact models for operation or failure. In Chapter 4, we propose model-independent design principles for algorithmic noise tolerance and address fundamental limits of distributed error bypassing.

Networks comprising stochastic components is a consequence of the uncertainty inherent to the embedding and integration of systems into physical realizations and substrates. Due to the massive dimensionality of assembly, fabrication, and integration processes, stochastic modeling of such uncertainty can be prohibitive and current methods are exceedingly conservative, often leading to massive over-design. In Chapter 5, we investigate concentration properties of certain network quantities for linear resistive networks for topology-preserving uncertainty profiles without relying on exact mathematical models for componentwise or network uncertainty. Furthermore, we quantify the effects of Johnson-Nyquist noise and address inter-component dependence due to the integration processes.

Dedicated to my parents.

ACKNOWLEDGMENTS

I would like to express my gratitude to everyone who contributed to the forming, developing, and testing of the ideas that have eventually led to this work, which is no small feat. First and foremost, I would like to thank my boss, Professor Andy Singer, whose subtle yet persistent guidance allowed me to conduct my research freely. He introduced me to a wide spectrum of ideas and allowed me to explore my own principles for facilitating my own way of research. I have always considered it a privilege to have the autonomy that I had and I can only hope that our work achieves the objective quality that matches this freedom. Prof. Singer taught me how to interpret the broader picture of my research, motivated me to get the work done when I was caught up on the details, and helped me learn the most out of the challenges that I faced. He did so while leading the most awesome research team around (no bias there), speaking of which, a big shout out to my colleagues and friends: Sijung Yang, Mehmet Ali Dönmez, Ryan Corey, Gizem Tabak, Jae Won Choi, and Dariush Kari! Thank you, Andy, for your lessons in research and in life, I will carry them with me for the rest of my days.

I had the distinct privilege of collaborating with several members of the University of Illinois family. Most of the engineering ideas that resonate well with me go back to Prof. Belabbas' lectures and our later collaboration helped me truly understand and appreciate the impact of these ideas. Prof. Varshney brings his unique creativity that I can only describe as beyond inspiration to every discussion while writing with a surgical precision and it was a true privilege to collaborate with him. I have found it enjoyable besides enlightening to discuss, develop, and argue over mathematical ideas with Prof. Raginsky. Discussions with Prof. Schwing effectively laid the foundations of my research and introduced me to an altogether new research philosophy. Prof. Hanumolu's patient responses to my many systems questions helped me build an intuition that I could never achieve alone. I found that a thousand ten-minute discussions with Prof. Bresler, at almost completely random times of course, can make a huge difference. I owe my "words of mathematics" to Prof. Laugesen, who never hesitated to lend a helping hand to a random engineering student that took his class years ago. Finally, in my first draft, I wrote "John Buck is awesome," and planned to make my way toward a proper thank you to Prof. Buck, but here we are.

Friends and family define the true magnitude of one's challenges, thus mine was never beyond my bearing. I would like to express my most sincere gratitude to them. Not to reveal a superhero's secret identity, but I have seen it with my own eyes–Peggy Wells saves graduate students of CSL on a daily basis! Before Lucas Buccafusca says it wittier, smarter, and non-negligibly sassier, I would like to make it known that knocking out a twelve-foot mountain troll might be one of those things, but it does not compare to tackling 538 together. The adventure epic "Lost in LA", starring Mustafa Uğur Daloğlu, might not appear on a silver screen yet, but it will never be forgotten. Of all things, engineering fits Onur Berkay Gamgam like a glove and I am proud to call him my friend.

Finally, I would like to thank my parents, to whom this dissertation is dedicated. Without their unwavering faith in my ability to pursue my goals, none of this would be possible. *Var olasınız.*

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The universe is transformation: life is opinion.

Marcus Aurelius

Engineering principles designed for processing noisy measurements of physical quantities to achieve well-defined goals might not adapt well to problems based on subjective evaluations. Risk-conservative challenges including disaster detection [1], personalized medicine [2], and autonomous driving [3] do not only rely on the ability to interpret subjective information, but also require a robust response to the underlying subjectivity. Sources of subjective information, often termed *experts*, interpret the current state of the environment, often with a hidden rationale, and provide their subjective evaluations, or *opinions*, to the decision maker. Experts are immutable sources of information in the sense that the decision maker has limited to no initiative on how an expert generates opinions. Conceptually, the notion of experts having hidden and immutable rationale captures a wide range of human and engineering decision making scenarios, while raising several interesting engineering challenges. The objective of this work is to establish some fundamental limits of learning from opinions by addressing key engineering problems subject to different forms of system uncertainty.

Chapter 2 proposes the problem of determining the optimal stopping time for sequentially consulting experts. Conventional hypothesis testing with pre-allocated resources results in decision strategies that focus exclusively on using sources most effectively to maximize the target reward [4]. Similarly, sequential hypothesis testing leads to strategies that effectively make decisions based on resources gathered until the time of decision making [5]. Nonetheless, effective use of resources does not necessarily yield efficient use of resources when consulting an expert, since acquiring an opinion, has associated costs. Contextually, in human decision making, consulting an expert incurs, among others, time, resource, and opportunity costs, where in engineering applications, this cost further manifests as, circuit-area, system-complexity, runtime, or memory requirements, motivating a sequential consultation process. A dynamic programming framework successfully captures a wide range of sequential decision making problems when a known reward function exists and serves as a reliable reference point for decision making [6]. Specifically, scheduling tasks is addressed in [7], dynamic portfolio analysis, analogous to building subcommittees of experts, is given in [8, 9], an abstract asset selling problem appears as an example of a stopping problem with known

rewards [6, Section 3.4], [10]. In the sequential consultation problem, the uncertainty introduced in the form of opinions creates a trade-off between expected current reward and the expected value of future opinions as opposed to more commonly observed trade-off between exact current reward and expected future value [6]. We first propose a stopping rule for consulting experts that maximize the expected reward when fixed models for experts are known. We further show that the resulting rule takes the form of a likelihood ratio test with a threshold that depends exclusively on the reliability of future experts and the cost associated with consulting them. We extend the results to a Bayesian framework, where the reliability of each expert has a known probability law. The proposed strategy achieves the maximum expected reward even when the cost of consulting yields a diminishing reward to be acquired only upon correct decision making extending the use of conventional sequential hypothesis testing, [5], by use of dynamic programming. Furthermore, for equally reliable experts with unknown reliability, we show that optimal stopping time in the conventional framework almost never depends on the underlying reliability of experts, allowing reliability estimation for such cases, for instance [11], to be bypassed.

Chapter 3 addresses the opinion aggregation problem, where experts generate opinions for a set of tasks with a hidden-but-fixed probability law. The proposed model-unaware, or unsupervised, approach employs what is termed *pseudo-competence*, a notion of reliability based exclusively on the collection of opinions and hence can be estimated sequentially during operation, without needing additional training. We further propose distributed averaging techniques to allow inference of local expert reliability using pseudo competences [12]. Consensus-based distributed averaging rules have been investigated in [13] with the noisy observation case given in [14], consensus rules under communication delays and changing topologies have been addressed in [15], and a related least-mean-square adaptive diffusion rule is discussed in [16]. Pseudo-competences preserve certain ordering properties of the "committee" of experts and allow adaptive, block processing, and instantaneous opinion aggregation via decision rules that resemble the naïve Bayes decision rule [17], which achieves minimum probability of error when the underlying probability law is known [4]. Unsupervised opinion aggregation research in the context of social choice dates back to the essays of marquis de Condorcet (1785) motivating the idea of *vox populi* and ever since has been an active field of research, viz. [18–20]. Block-processing (off-line) decision aggregating rules date back at least to [21], where a maximum likelihood solution was found via the expectation maximization (EM) algorithm. Several variants of the EM approach have been proposed since then, including the GLAD algorithm [22], which estimates states while simultaneously learning source reliabilities and task difficulties, [23], which proposes a form of spectral estimation for what they term confusion matrices for each source to initialize EM, [24], which

uses EM to aggregate soft decisions, and [25], which extends the use of EM to an on-line setting. Alternative to EM-based approaches, [26] proposes a belief propagation (BP)-type algorithm, [27] proposes a variational inference model and uses belief propagation and mean field methods, [28] discusses the optimality of the belief propagation approach and [29] studies an accuracy vs. budget trade-off for the belief propagation approach. Furthermore, [30] proposes a spectral meta-learner (SML) that uses the dominant eigenvectors of a certain empirical covariance matrix of workers to find weights and [31] proposes the use of deep neural networks. Statistical approach to opinion aggregation brings about several advantages over the current methods that rely on off-line iterative solutions or computationally demanding mathematical techniques: First, pseudo-competence notion allows unsupervised ordering and inference of true competences through on-line or decentralized computation, [12]; a task which is often achieved by off-line centralized computation, as done by SML [30], or as side-products of iterative rules such as BP or EM [24–26]. Since pseudo competence can be estimated empirically in real time, the resulting rules have significant runtime improvements. Furthermore, proposed unsupervised rules achieve performance of the best supervised rule as the number of experts increase; such asymptotic optimality has only been achieved by BP before, [28], which in addition requires a good task assignment strategy to prevent loops in the underlying Trellis graph.

Chapter 4 investigates fault-resilient computational principles for employing error-prone computational units as experts. The existence of such units lead to system-wide failures with hidden statistics that can be difficult to model, owing to the dimensionality of the error-inducing physical factors including process, temperature and voltage variations, power-to-circuit-area ratio, electromagnetic interference among many others. At the subsystem level, this motivates safeguarding mechanisms to provide robust, low-power solutions to mitigate failures [32–36], and the generality of such rules has created interest in interpretation of these ideas in a more information- and decision-theoretic framework [37, 38]. We propose a fidelity-based testing of hypotheses with hidden models [39], which not only provides a near optimal fault mitigation improving upon the decision-theoretic framework but also operates at single-task regime not requiring the asymptotic guarantees similar to the information-theoretic framework. Furthermore, we explore fundamental limits of fault-resilient distributed computation in the context of discarding faulty outcomes to prevent propagation of random failures across the network.

Chapter 5 focuses on linear noisy resistive networks with stochastic components subject to topology-preserving uncertainty profiles. Information processing through networks with stochastic components is a natural consequence of fabrication, assembly, and integration uncertainties. The massive dimensionality of the underlying physical processes governing

component and network uncertainty has led to using a *minmax* design philosophy relying on Monte Carlo simulations to predict a subset of the failure space as the *de facto* standard. Such a standard is not only expensive and time-consuming, but it is inherently limited in the intuition it provides on the underlying stochasticity. Therefore, a reliable framework for linear noisy networks with stochastic components is necessary to accurately incorporate the impact of fabrication uncertainty into the design process. The theory of linear noisy networks with stochastic components has far-reaching roots: Construction of reliable computing circuits via unreliable components was addressed in [40]. A mathematical framework for circuits using unreliable relays with static failure statistics was proposed in [41]. The noise due to thermal agitation of individual circuit components was shown to exhibit Gaussian statistics in [42]. The impacts of thermal noise in linear networks with deterministic components was discussed in [43]. We investigate concentration properties of effective resistance, mean-square branch voltage, and expected power dissipation for topology-preserving uncertainty profiles, quantify the effects of Johnson-Nyquist noise and address inter-component dependence due to the integration processes [44].

Notation

Standard notation from analysis, algebra, probability, and graph theory are employed. Specifically, $\mathbb{C}$, $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$, $\mathbb{N}$ denote the set of complex, real, rational, integer, and natural numbers respectively. Sets are denoted by $\{\cdots\}$, vectors are by $[\cdots]$, and $(\cdot)^c$ is the set complement operator. Deterministic matrices $\underline{\mathbf{A}}, \underline{\mathbf{B}}, \ldots$ are denoted by bold, capital, underlined letters. Deterministic vectors are always defined as column vectors $\boldsymbol{a}, \boldsymbol{b}, \ldots$ denoted by bold, lowercase, italic letters. Range space and null space operators are denoted by $\mathcal{R}(\cdot)$ and $\mathcal{N}(\cdot)$, respectively. Furthermore, $\mathcal{R}(\cdot)$ denotes the range of a function, complementing the domain notation of $\mathcal{D}(\cdot)$. Transpose, inverse and pseudo inverse operators are denoted by $(\cdot)^\top$, $(\cdot)^{-1}$, $(\cdot)^\dagger$, respectively. Sign and indicator functions are denoted by $\operatorname{sign}(\cdot)$ and $\mathbb{1}(\cdot)$, respectively. Deterministic sequences a_n are denoted by lowercase letters with subscript indices. Absolute value of a number and the cardinality of sets and vectors are denoted by $|\cdot|$. On a Hilbert space $\mathcal{H}$, $\|\cdot\|_\ell$ denotes the corresponding ℓ-norm, in the context of matrices, $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_\sigma$ denotes the spectral norm. The gradient of a function f with respect to a vector $\boldsymbol{x}$ is denoted by $\nabla_{\boldsymbol{x}} f = [\partial f/\partial x_1 \cdots \partial f/\partial x_1]^\top$. We use the notation $\nabla_{\boldsymbol{x}} f(\boldsymbol{u})$ to denote the gradient evaluated at some point $\boldsymbol{u} \in \mathcal{D}(f)$.

Random variables defined on $(\Omega, \mathscr{F}, \mathbb{P})$ are denoted with uppercase letters $X, Y, \ldots$, where Ω is the event space, $\mathscr{F}$ is a σ-field defined on Ω and $\mathbb{P}(\cdot)$ is the probability measure.

Random vectors $\mathbf{X}, \mathbf{Y}, \ldots$ are defined as column vectors and denoted with boldface, capital letters. Samples from random variables and vectors are denoted by their lowercase counterparts $x, y, \ldots$ and $\mathbf{x}, \mathbf{y}$ respectively. A subtle convention that is more conceptual than formal is followed when the samples from random vectors are not italicized: $(\mathbf{X}, \mathbf{x})$ as opposed to $(\mathbf{X}, \boldsymbol{x})$. Probability density function of a random variable X is denoted by $p_X(x)$. Random processes $X(t), Y(t), \ldots$ are defined on $(\Omega, \mathscr{F}, \mathbb{P}, \mathbb{T})$ with $t \in \mathbb{T}$ being the index set. When it is clear from context, $\mathbb{E}[\cdot]$ (and $\mathbb{E}\cdot$ when dealing with expected norms) denotes the expectation operator otherwise, expectation with respect to a marginal distribution, for instance that of a random variable X, is denoted by $\mathbb{E}_X[\cdot]$. The conditional probability and expectation operators are denoted by $\mathbb{P}(\cdot \mid \cdot)$, $\mathbb{E}[\cdot \mid \cdot]$. When the intersection of events $\omega_1, \omega_2 \in \Omega$ are concerned, $\mathbb{P}(\omega_1, \omega_2) \equiv \mathbb{P}(\omega_1 \cap \omega_2)$. Gaussian distribution with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$.

The function $w(p) = \log {}^{p}/_{(1-p)} \equiv \log {}^{p}/_{q}$ for $q = 1 - p$ appears frequently. In all instances of its use, the domain of $w(\cdot)$ is allowed to be $[0, 1]$ with the convention that $\log {}^{1}/_{0} = \infty$ and $\log {}^{0}/_{1} = -\infty$. Furthermore, variables (p, q) always obey $p+q = 1$ with various extensions for instance, $p_i+q_i = 1$ for competences, $\tilde{p}_i+\tilde{q}_i = 1$ for pseudo competences, and $\mathcal{P}(\cdot)+\mathcal{Q}(\cdot) = 1$ for log-likelihood to probability of correctness and error mappings to name a few. However, for all such uses, we provide explicit in-text definitions.

Graphs are denoted by $\mathcal{G}$ with the corresponding vertex and edge sets $(\mathcal{V}, \mathcal{E})$. The adjacency matrix of a graph is denoted by $\underline{\mathbf{A}}$ and the Laplacian is denoted by $\underline{\mathbf{L}}$. When a quantity $r(\mathcal{G})$ (random or deterministic) between two vertices $i, j \in \mathcal{V}$ is needed, double subscript r_{ij} is used. When $i, j \in \mathcal{V}$ are connected, $i \leftrightarrow j \in \mathcal{E}$ is written. Neighborhood of a vertex i is denoted by $\mathcal{N}_i$.

Indexing

Elements of deterministic matrices are referred to $(\underline{\mathbf{A}})_{ij}$ in row-column form and elements of deterministic vectors vectors are written as $(\boldsymbol{a})_i$. Standard unit vectors are denoted by $\boldsymbol{e}_i$ such that $(\boldsymbol{e}_i)_j = \mathbb{1}(i = j)$ and the matrix $\mathbf{J}^{ij} \triangleq (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top$ is employed. In general, for any vector $\boldsymbol{x}$ and any subset $\mathcal{I}$ of an index set $\boldsymbol{x}_\mathcal{I} = \{x_i : i \in \mathcal{I}\}$, same convention applies to random vectors $\mathbf{X}_\mathcal{I}$. When a specific index, for instance i, is removed, the remaining vector is denoted by $\mathbf{X}_{\setminus i}$. A vector-valued random process is given as $\mathbf{X}(t) = [X_1(t), \ldots, X_N(t)]^\top$, $\forall t \in \mathbb{T}$. The matrix $\mathbf{X}_1^T \triangleq [\mathbf{X}(1), \ldots, \mathbf{X}(T)]$ denotes all outcomes from $t = 1$ up to $t = T$. We further denote the set: $[N] = \{1, \ldots, N\}$.

CHAPTER 2

OPTIMAL STOPPING TIME FOR CONSULTING EXPERTS

Stochastic experts are sources of information that provide subjective evaluations of the current state-of-the-world, or *opinions*, without any deliberate intent to deceive their recipients [45]. The rationale behind an opinion and the concomitant proficiency of an expert is often hidden, or difficult to model, leading to a *consultation* process, where a variety of opinions are gathered prior to decision making. Different from a straightforward engineering application of the idea *vox populi*, [46], with all experts fully committing to a cause, the operational cost of consulting each expert motivates a *sequential* consultation process that allows the decision maker to collect opinions until stopping upon necessity or confidence.

Many modern engineering applications, including disaster detection [1], personalized medicine [2], and autonomous driving [3], rely on experts as sources of extrinsic information that are accessible at an operational cost. For instance, *the internet of things* is conjectured to be particularly rich in highly localized processing units under power, circuit area, and latency constraints are designed to collectively address global computational tasks [47]. These networks comprise error-prone computational units that, when employed, incur operational costs in the form of communication overhead, memory requirements, or processing power. Furthermore, physico-chemical processes, such as genomic data sequencing, operate with remarkably high target efficiency under processing-time and material constraints [48]. Similarly, computer vision applications often employ tree-based classifiers, such as random forests, thanks to their versatility [11]. Tree-based classifiers introduce exponential growth of the number of available classifiers, making strategies for dynamic resource allocation necessary. Such applications share an inherent trade-off between the reward associated with achieving a task and the costs incurred in the process.

The classical framework for the Bayesian hypothesis testing often admits a probability law on *pre-allocated* resources, or experts, available to a decision maker that aims primarily to maximize a known reward [4]. Since the observation statistics is known a priori in this framework, the decision maker uses all available opinions to maximize the reward, often without an initiative to consult experts dynamically. In the event of model uncertainty, randomized decision rules, which still employ all available resources, are called to action

[4]. When the underlying probability law is unknown, or hidden, feedback-based decision-aggregation rules are often relied upon, leading to the ideas from the literature concerning so-called mixture of experts and boosting i.a. [7, 45]. Consequently, decision rules with pre-allocated resources often focuses on correct decision making, potentially under model uncertainty, with limited regard to costs associated with consulting experts.

Sequential resource allocation problems, on the other hand, often aim to maximize known rewards in the sequential hypothesis testing [5] or in the dynamic programming framework [6]. For instance, dynamic portfolio analysis investigates how a fixed budget can be allocated among assets, analogous to proficiency of experts, to maximize the expected return from such assets [8, 9]. Furthermore, scheduling arguments assert that a certain order of allocating resources, or an order of experts to be consulted, can prove to be more rewarding than arbitrary allocations [49]. In applications including random forests, experts are available at the nature's behest [11]. In such cases, the decision maker faces an optimal stopping problem, where the objective is to stop when the current reward is superior to all expected future rewards [6, Section 3.4]. Further properties of stopping times are addressed in [50, Sections 4.8, 7.3]. Sequential resource allocation techniques often enjoy reliable information on the current rewards to determine how to allocate resources with limited regard to the how the resources are used.

The fundamental challenge for sequentially consulting experts is to strike the balance between the cost of consulting experts and the reward to be a acquired from correctly aggregating their opinions. Therefore, in order to find the optimal stopping time for consulting experts, one needs to jointly consider the strength of a decision and the the cost associated with consulting another expert. Further challenges arise when experts are generated from a stochastic family or when they are of hidden statistics. Chapter 2 addresses these challenges: By introducing a reward function that diminishes in the numbers of experts consulted, we capture the trade-off between the cost consulting an expert and the decision strength to be gained from consulting. This, in comparison to sequential hypothesis testing, [5], incorporates the cost of consultation into decision making directly. Furthermore, dynamic programming framework allows us to conclude that in the classical, no-cost sequential decision making based on opinions from equally reliable experts, such as those addressed in [5, 11], unsupervised competence estimate is almost never relevant when deciding whether to consult more experts. Finally, when experts are generated from a stochastic family, we show that rather than estimating the underlying reliability of experts, one could iterative compute a unique stopping rule that achieves the maximum reward in expectation.

2.1 Background and Problem Definition

Let a binary random variable $Y \in \mathcal{Y} \triangleq \{-1, 1\}$ capture the true state of the world. A stochastic expert evaluates the current state of the world and produces an opinion $X_t \in \mathcal{Y}$. The proficiency of an expert, often called *competence*, is the probability with which the produced opinion X_t captures the current state Y:

$$p_t = \mathbb{P}\left(X_t = Y\right). \tag{2.1}$$

Experts that can be described sufficiently by the condition (2.1) are often called *stochastic experts* [45]. Conceptually, a stochastic expert might *fail* but not *deceive*. Furthermore, we consider experts that *generate opinions independently*, which does *not* imply that opinions X_t are independent but rather implies that X_t are *conditionally* independent *given the current state* Y, that is $\forall t_1 \neq t_2 \in [T]$:

$$X_{t_1} - Y - X_{t_2}. \tag{2.2}$$

Such experts neither collaborate nor purposefully mislead the decision maker as might happen in a game-theoretic framework. Non-stochastic, or adversarial, experts as they are often referred to, are addressed extensively in [45] from the mixture of experts perspective and in [51] from that of multi-armed bandits.

It is well-known that given opinions $X^T = x^T$ of experts with competences $\{p_1, \cdots, p_T\}$ the maximum a posteriori (MAP) decision rule:

$$\delta^*\left(x^T\right) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}\left(Y = y \;\middle|\; X^T = x^T\right) \tag{2.3}$$

minimizes the probability of error [4]:

$$\delta^*\left(X^T\right) = \arg\min_{\delta \in \mathcal{D}_T} \mathbb{P}\left(Y \neq \delta\left(X^T\right)\right). \tag{2.4}$$

Here $\mathcal{D}_T$ is the family of decision rules that map $\mathcal{Y}^T \to \mathcal{Y}$. Furthermore, the optimal decision rule (2.3) necessarily takes the form of a *likelihood ratio test* [4]:

$$\sum_{t=1}^{T} X_t \log \frac{p_t}{q_t} \gtrless \eta, \tag{2.5}$$

where $q_t = 1 - p_t$, $\forall t$ and threshold $\eta = \log {}^{\mathbb{P}(Y=1)}/_{\mathbb{P}(Y=-1)}$. The likelihood ratio test in (2.5), written in the log-likelihood form here, is also commonly called the *naïve Bayes* decision rule [17] with the understanding that $\delta^*\left(X^T\right) = 1$ when the log-likelihood exceeds the

threshold.

Observe that the decision rule (2.3) does not take into account the margin by which the log-likelihood might exceed the threshold. This is a direct consequence of resources being pre-allocated and the decision maker aiming to minimize probability of error (2.4) only. The decision maker is provided with a set of experts at no cost, or at a fixed cost that decision maker cannot control, therefore, there is no initiative for not consulting them. The sequential consultation of experts incorporates the cost of opinions into the decision-making framework.

2.1.1 Sequential Consultation of Experts

Conceptually, consulting an expert to obtain an opinion has associated costs, among others, time, circuit-area, memory, computational complexity, or opportunity cost, depending on the application. Therefore, one needs to consider not only the benefit from an additional opinion but also the associated cost. The following *reward* function captures this trade-off over the binary alphabet:

$$r\left(t; X^t\right) = \beta_t \mathbb{1}\left(Y = \delta\left(X^t\right)\right), \tag{2.6}$$

where the pay-off function β_t is monotonically decreasing in t; $\beta_t \geq \beta_{t+1}$, $\forall t \in [T]$, and is independent of the opinions X^T. Decision aggregation rule $\delta(\cdot) \in \mathcal{D}_t$, where:

$$\mathcal{D}_t = \left\{\delta : \mathcal{Y}^t \to \mathcal{Y}\right\}$$

is the set of rules that aggregate up to t opinions. Note that the decision rule $\delta(\cdot)$ is chosen at the discretion of the decision maker. Therefore, the reward function (2.6) captures a wide range of problem setups that do not yield a reward upon failure while penalizing excess use of resources.

Sequential consultation of experts aims to maximize the reward in expectation and hence, the problem is cast into a dynamic programming framework. Let the *value function* $V_t\left(x^t\right)$ be the maximum expected reward starting from a set of opinions $X^t = x^t$ at time t:

$$V_t\left(x^t\right) = \max_{\substack{\tau \geq t \\ \delta \in \mathcal{D}_\tau}} \mathbb{E}\left[r\left(\tau; X^\tau\right) \,\middle|\, X^t = x^t\right]. \tag{2.7}$$

The maximum is taken over all future times τ and all decision rules $\mathcal{D}_\tau$ attainable at those times, as shown in Appendix A.1. This is different from the standard dynamic programming setup, [6, Section 3.4] with fixed rewards that take the maximum over time alone. Further note that the reward $r\left(t; X^t\right)$ is earned only if the decision maker decides to stop *and* the aggregated opinions yield the correct outcome. The following lemma formulates the Bellman

equation that captures this trade-off:

Lemma 2.1. *The Bellman equation that correspond to* (2.7) *for sequentially consulting experts is as follows:*

$$V_t\left(x^t\right) = \max\left(\beta_t \max_{y\in\mathcal{Y}} \mathbb{P}\left(y \mid X^t = x^t\right), \mathbb{E}\left[V_{t+1}\left(X^{t+1}\right) \mid X^t = x^t\right]\right).$$

The proof is mostly technical and relies on the fact that a decision maker can aggregate opinions optimally upon stopping via the MAP decision rule (2.3) if competences p_t, $t \in [T]$ are known – given in Appendix A.1.

Albeit achievable when competences $\{p_1, \cdots, p_T\}$ are known, or when they are known to belong a family $\{P_1, \cdots, P_T\}$ with a probability law $f_{P_1,\cdots,P_T}(\cdot)$, decision maker might not always be able to construct the MAP decision rule based on the opinions alone. Notably, when experts have hidden competences, MAP rule cannot be constructed exactly, which might lead to fixing a decision rule such as majority voting and designing the stopping rule achieving the maximum reward constrained to that rule, or designing a minmax stopping rule if the experts are known to be equally reliable.

Next, we formulate the optimal stopping rule for consulting experts when the competences are known.

2.2 Optimal Rule for Joint Stopping and Opinion Aggregation

Let nature provide an *ordered* set of experts with competences $\{p_1, \cdots, p_T\}$ for consultation. At each time instance $t \in [T]$, upon receiving X_t, the decision maker may decide to stop and aggregate opinions X^t, or continue consulting. The pay-off function $\beta_t, \forall t \in [T]$, and the competences $\{p_1, \cdots, p_T\}$, along with their ordering, are known.

The main purpose of this section is to propose an instantaneously realizable function $f : \mathcal{Y}^t \to \mathbb{R}$ and a recursively realizable function η_t such that the optimal stopping time $\mathcal{T}^*$ for consulting experts takes the form:

$$\mathcal{T}^* = \left\{\min_t f\left(X^t; p^t\right) > \eta_t\left(p_{t+1}^T, \beta_t^T\right)\right\}. \tag{2.8}$$

Here, p^t denotes past competences and $\left(p_{t+1}^T, \beta_t^T\right)$ denotes future competences as well as current and future possible pay-offs. Note that the optimal stopping time $\mathcal{T}^*$ is a random variable, the function $f(\cdot)$ of opinions is past measurable, and the threshold function $\eta(\cdot)$ is a function of future competences and pay-offs alone; independent of any opinion sample

path x^T.

A stopping rule of the form (2.8) has conceptual advantages: First, a suitable Markov property-preserving function $f : \mathcal{Y}^t \to \mathbb{R}$ could quantify a notion of *decision strength* and hence, would not lead to an exhaustive search over the random walk X_{t+1}^{τ}, $\forall \tau \in [t+1, T]$ of opinions. Second, such a rule would allow the consultation process to stop as soon as the pre-computed threshold η_t is exceeded. Furthermore, albeit not obvious from (2.8), η_t could quantify the cost-performance trade-off for consulting more experts as well as the impact of their *ordering*. These ideas are formally addressed in Sections 2.2.1-2.2.2.

Certain properties of $f(\cdot)$ and $\eta_t(\cdot)$ should be clear conceptually: The more *past* opinions agree the larger $f(\cdot)$ should be. Moreover, the more competent *future* experts are or the slower β_t diminishes, the larger $\eta_t(\cdot)$ should be. Appendix A.2 provides a technical motivation based on Lemma 2.1 for the instantaneously realizable statistics that we propose next.

2.2.1 Likelihood Ratio as Sufficient Statistic

Let the random process L_t represent the log-likelihood ratio without the conventional emphasis on a specific state $y \in \mathcal{Y}$. Instead, let it be the *log-likelihood of correct decision making*:

$$L_t \triangleq \log \frac{\max\limits_{y \in \mathcal{Y}} \mathbb{P}\left(Y = y \mid X^t = x^t\right)}{\min\limits_{y \in \mathcal{Y}} \mathbb{P}\left(Y = y \mid X^t = x^t\right)}. \tag{2.9}$$

The random process L_t is well-defined over the index set $[T]$ as long as the *ordering* of the experts is fixed. As could be expected, it satisfies $L_t > 0$ almost surely. One should note that (2.9) does *not* assume any prior distribution on the underlying state Y: If such a prior were known, Bayes' rule would apply. The goal of this section is to show that L_t *is a Markov process* and that it provides sufficient statistics for joint stopping and decision aggregation.

Observe that given $L_t = \ell_t$, the maximum probability of correct decision making for *any* set of opinions $X^t = x^t$ that amount to ℓ_t can be written as:

$$\max_{y \in \mathcal{Y}} \mathbb{P}\left(Y = y \;\middle|\; X^t = x^t\right) = \frac{1}{1 + e^{-\ell_t}},$$

where the right-hand side is often referred to as *sigmoid, logistic, or logit* function. Indeed, it is useful to define separately the mapping from log-likelihood to probability of correctness:

$$\mathcal{P}(\ell_t) \triangleq \frac{1}{1 + e^{-\ell_t}}, \tag{2.10}$$

with $\mathcal{Q}(\cdot) = 1 - \mathcal{P}(\cdot)$ being the corresponding mapping from log-likelihood to probability of error. Note that $\mathcal{P}(L_t) \geq 1/2$ almost surely. We deviate from the standard $\sigma(\cdot)$ notation for the sigmoid function, [52, Section 1.4.6] in order to use $\mathcal{P}(\cdot)$ as the probability of correct decision making.

Similarly, define the log-likelihood ratio of an individual expert being correct:

$$\theta_t \triangleq \log \frac{\max(p_t, q_t)}{\min(p_t, q_t)}, \tag{2.11}$$

where, different from (2.9), the set $\{\theta_1, \cdots, \theta_T\}$ is deterministic. Further note that θ_t is *not* the likelihood of correct decision making based on X_t as it does not take into account the prior on Y. The next lemma addresses the Markov property of L_t.

Lemma 2.2. *Let a finite set of experts with log-likelihood of correctness $(\theta_1, \cdots, \theta_T)$ generate opinions $(X_1, \cdots, X_T)$ independently $(X_i - Y - X_j, \forall i \neq j \in [T])$ upon being consulted. The log-likelihood process:*

$$L_t = \log \frac{\max_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X^t = x^t)}{\min_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X^t = x^t)}$$

is a Markov process that given $L_t = \ell_t$ evolves with:

$$L_{t+1} = \begin{cases} \ell_t + \theta_{t+1}, & w.p. \quad \tilde{p}_{t+1}, \\ \left|\ell_t - \theta_{t+1}\right| & w.p. \quad \tilde{q}_{t+1}, \end{cases} \tag{2.12}$$

where the transition probabilities are given by:

$$\tilde{p}_{t+1} = \mathcal{P}(\theta_t)\,\mathcal{P}(\ell_t) + \mathcal{Q}(\theta_t)\,\mathcal{Q}(\ell_t), \tag{2.13}$$

where $\tilde{q}_{t+1} = 1 - \tilde{p}_{t+1}$ and $\mathcal{P}(\cdot) = 1 - \mathcal{Q}(\cdot)$ is the sigmoid function (2.10).

Proof is given in Appendix A.3. Conceptually, one can interpret $\mathcal{P}(\ell_t)$ as the *instantaneous competence* of the decision maker and hence, ℓ_t as the instantaneous *strength* of the decision to be made. The state transition (2.12) quantifies this notion by establishing the probability distribution of the strength to be attained by consulting another expert. Furthermore, the transition probabilities (2.13) indicate that the strength of decision maker increases when the new expert agrees with the current consensus, which not only happens when both the consensus and the new expert are correct but happens when both are *incorrect* as well.

We now set $f(X^t; p^t) \equiv L_t$ and next, show that the optimal stopping time for consulting experts takes the form in (2.8) by finding the threshold $\eta_t\left(p_{t+1}^T, \beta_t^T\right)$.

2.2.2 A Likelihood Ratio Test for the Optimal Stopping Time

A direct consequence of Lemma 2.2 and (A.3) in Appendix A.1 is that one can write the Bellman equation given in Lemma 2.1 in terms of the log-likelihood process L_t:

$$V_t\left(\ell_t\right) = \max\left(\beta_t \mathcal{P}\left(\ell_t\right), \mathbb{E}_{L_{t+1}}\left[V_{t+1}\left(L_{t+1}\right) \mid L_t = \ell_t\right]\right), \tag{2.14}$$

which leads to an optimal stopping time for consulting experts that takes the form:

$$\mathcal{T}^* = \left\{\min_t : \beta_t \mathcal{P}\left(\ell_t\right) \geq \mathbb{E}_{L_{t+1}}\left[V_{t+1}\left(L_{t+1}\right) \mid L_t = \ell_t\right]\right\}.$$

Observe that, different from the opinion process of Lemma 2.1, the log-likelihood process of Lemma 2.2 admits the expected maximum future reward to be computed $\forall t \in [T-1]$ via:

$$\mathbb{E}_{L_{t+1}}\left[V_{t+1}\left(L_{t+1}\right) \mid L_t = \ell_t\right] = \tilde{p}_{t+1} V_{t+1}\left(\ell_t + \theta_{t+1}\right) + \tilde{q}_{t+1} V_{t+1}\left(\left|\ell_t - \theta_{t+1}\right|\right).$$

Further note that when there exist finitely many experts ($T < \infty$) the value function, upon consulting the last expert, amounts to receiving the minimal payoff β_T with a probability determined by all available opinions. Equivalently:

$$V_T\left(\ell_T\right) = \beta_T \mathcal{P}\left(\ell_T\right) = \beta_T \mathcal{P}\left(\max\left(\ell_T, 0\right)\right), \tag{2.15}$$

which leads to the main result of Section 2.2.2. Note that the notation $(X_1, \cdots, X_T)$ rather than $\{X_1, \cdots, X_T\}$ is used to emphasize that experts are ordered in time.

Theorem 2.1. *Let a finite set of experts with log-likelihood of correctness $(\theta_1, \cdots, \theta_T)$ generate opinions $(X_1, \cdots, X_T)$ independently $(X_i - Y - X_j, \forall i \neq j \in [T])$ upon being consulted. For a non-increasing pay-off function β_t, starting at $X^t = x^t$, the stopping time that maximizes the reward:*

$$r\left(t; x^t\right) = \beta_t \mathbb{1}\left(Y = \delta\left(x^t\right)\right),$$

in expectation is called the optimal stopping time for consulting experts. A decision maker that stops consulting upon the first occurrence of the event:

$$L_t \geq \eta_t$$

and uses the maximum a priori decision rule:

$$\delta^*\left(x^t\right) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}\left(Y = y \;\middle|\; X^t = x^t\right)$$

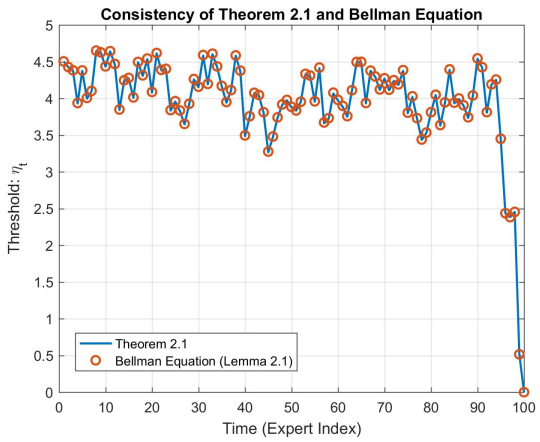


Figure 2.1: Comparison of Thresholds Acquired from Theorem 2.1 and Bellman Equation in Lemma 2.1

for aggregation opinions achieves the maximum expected reward. Here, L_t is the log-likelihood ratio process:

$$L_t \triangleq \log \frac{\max_{y \in \mathcal{Y}} \mathbb{P}\left(Y = y \mid X^t = x^t\right)}{\min_{y \in \mathcal{Y}} \mathbb{P}\left(Y = y \mid X^t = x^t\right)}$$

and starting at $\eta_T = 0$, the threshold function η_t is defined recursively via:

$$\eta_t = \log \max \left(\frac{\mathcal{P}\left(\theta_{t+1}\right) \mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right) \mathcal{Q}\left(\eta_{t+1}\right)}, \frac{\mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\eta_{t+1}\right)}, \frac{\mathcal{P}\left(\theta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right)} \right), \tag{2.16}$$

where, $\delta_{t+1} = \frac{\beta_t - \beta_{t+1}}{\beta_{t+1}}$ and $\mathcal{P}(\cdot) = 1 - \mathcal{Q}(\cdot)$ is the sigmoid function (2.10)*. Therefore, the optimal stopping time for consulting experts is said to be given by:*

$$\mathcal{T}^* = \min \left\{ t : L_t \geq \eta_t \right\}.$$

Theorem 2.1 proposes a stopping rule of desired form (2.8) with decision statistics $f\left(X^t; p^t\right) \equiv L_t$ and threshold function $\eta_t\left(p_{t+1}^T, \beta_t^T\right)$ given in (2.16). Note that η_t is not necessarily non-negative as it captures, among others, the trade-off between instantaneous cost for per consultation δ_{t+1} and the likelihood θ_{t+1} of that expert being correct. The threshold changes monotonically in its parameters: It increases in individual likelihoods θ_τ, $\forall \tau > t$, which corresponds to $|p_\tau - 1/2|$ increasing, and it decreases in rate δ_τ, $\forall \tau > t$, with which the pay-off function diminishes. A proof of Theorem 2.1 and those of some of the properties of η_t are given in Appendix A.4. Figure 2.1 illustrates that Bellman equation in Lemma 2.1 yields numerically the threshold function η_t from Theorem 2.1, as it should.

It is of interest to observe that the value function $V_t(\ell_t)$ that underlies Theorem 2.1 takes the form:

$$V_t(\ell_t) = \beta_t \mathcal{P}(\max(\ell_t, \eta_t)), \forall t \in [T], \tag{2.17}$$

for the unique threshold function η_t. Conceptually, η_t is the *discounted future likelihood of correctness*; where the discounting factor follows directly from the pay-off diminishing and the concomitant δ_{t+1} factor in (2.16). This complements *the principle optimality* as it should: optimal stopping happens when the current likelihood of correctness exceeds the discounted future likelihood.

Diminishing pay-off function β_t is a fundamental aspect of the problem setup for determining an optimal stopping time for consulting experts. Through dynamic programming, Theorem 2.1 concludes that the proposed version of the sequential likelihood ratio test achieves maximum expected reward for arbitrary pay-off functions beyond the standard, constant-payoff formulation in [5]. Nevertheless, when the pay-off is constant and hence, when the decision maker is allowed to consult until achieving confidence, Theorem 2.1 yields interesting insights, as discussed next.

Constant Pay-off while Consulting Non-Identical Experts

If the pay-off does not diminish, $\beta_t = \beta_{t+1} = 1$, $\forall t \in [T]$, equivalently, when opinions can be acquired at no extra cost, the threshold function η_t is given by:

$$\eta_t = \log \frac{\mathcal{P}(\theta_{t+1})\,\mathcal{P}(\eta_{t+1})}{\mathcal{Q}(\theta_{t+1})\,\mathcal{Q}(\eta_{t+1})} = \sum_{j=t+1}^{T} \log \frac{\mathcal{P}(\theta_j)}{\mathcal{Q}(\theta_j)}.$$

Recall that $\eta_T = 0$ and $L_T \geq 0$ almost surely, which yields that decision-maker necessarily stops and decides whichever choice is more likely. Let Y have the uniform prior for ease of exposition. Then, the optimal stopping time attains the simple form:

$$\mathcal{T}^* = \min \left\{ t : \left| \sum_{i=1}^{t} X_i \log \frac{\mathcal{P}(\theta_i)}{\mathcal{Q}(\theta_i)} \right| \geq \sum_{j=t+1}^{T} \log \frac{\mathcal{P}(\theta_j)}{\mathcal{Q}(\theta_j)} \right\}. \tag{2.18}$$

Conceptually, the stopping time in (2.18) exhibits a key phenomenon: In the constant pay-off setup, to maximize the probability of correct decision making, one should consult experts sequentially until the remaining experts *in unanimity* are unable to change the current decision. Equivalently, decision maker should compare the current decision strength to the combined *strength* of the future experts. Furthermore, the decision maker employs the naïve Bayes rule, shown in (2.5), to aggregate the available opinions. More generally, the *magnitude*

of the gathered statistics determines whether the decision maker should stop consulting or not and the corresponding *sign* determines which decision to make upon stopping.

Constant Pay-off while Consulting Identical Experts

If the experts are identical in their competences $p_i = p_j = p$, $\forall i, j \in [T]$, yet still generating opinions independently, the threshold η_t takes the form:

$$\eta_t = (T-t) \log \frac{\mathcal{P}(\theta)}{\mathcal{Q}(\theta)} = (T-t) \log \frac{\max(p,q)}{\min(p,q)}.$$

The corresponding optimal stopping time takes the form:

$$\mathcal{T}^* = \min \left\{ t : \left| \left(\sum_{i=1}^{t} X_i \right) \log \frac{\max(p,q)}{\min(p,q)} \right| \geq (T-t) \log \frac{\max(p,q)}{\min(p,q)} \right\},$$

which further simplifies to:

$$\mathcal{T}^* = \begin{cases} 1 & \text{if} \quad p \in \{0,1\}, \\ 0 & \text{if} \quad p = 1/2, \\ \min\{t : M_t \geq T-t\} & \text{otherwise}, \end{cases} \tag{2.19}$$

where the random variable M_t denotes the *margin* between votes, $M_t = \left|\sum_{i=1}^{t} X_i\right|$. It important to note that the stopping rule in (2.19) is an *unsupervised* stopping rule in the sense that it does *not* depend on the value of the competence p but on experts being identical *and* whether $p \in \{1/2, 0, 1\}$ or not.

Even though the stopping rule takes an intuitive form when the competences are known, in practice it is a demanding constraint to meet. Next, we address the optimal stopping time for consulting experts with hidden competences subject to a known probability law.

2.3 Bayesian Stopping Times for Consulting Experts

Let a probability law f_{Θ^T} governing the log-likelihood $\Theta^T \equiv \{\Theta_1, \cdots, \Theta_T\}$ of correctness for the experts be given:

$$\mathbb{P}(X_t = Y) = \mathcal{P}(\Theta_t).$$

Further allow that $\Theta_{t_1} \perp \Theta_{t_2}$, $\forall t_1 \neq t_2 \in [T]$ and that $\Theta_t > 0$, $\forall i \in [T]$ almost surely. Unless the consultation process was stopped previously, at each time $t \in [T]$ an expert of

competence $\mathcal{P}(\Theta_t)$ generates an opinion X_t, which is then revealed to the decision maker *without* the underlying likelihood $\Theta_t = \theta_t$. In this case, the law of total probability yields that log-likelihood process L_t takes the form:

$$L_t = \log \frac{\max\limits_{y \in \mathcal{Y}} \mathbb{E}_{\Theta^t} \left[\mathbb{P}\left(Y = y \mid X^t, \Theta^t\right) \right]}{\min\limits_{y \in \mathcal{Y}} \mathbb{E}_{\Theta^t} \left[\mathbb{P}\left(Y = y \mid X^t, \Theta^t\right) \right]}, \tag{2.20}$$

where the conditional should be understood as ordered pairs $(X_1, \theta_1), \cdots, (X_t, \theta_t)$. The purpose of this section is to show that for any such probability law f_{Θ^T}, a unique threshold for optimal stopping exists.

Theorem 2.1 establishes the existence of a unique intersection point between $\beta_{t-1}\mathcal{P}\left(\ell_{t-1}\right)$ and $\mathbb{E}_{L_t}\left[V_t\left(L_t\right) \mid \ell_{t-1}, \theta_t\right]$ over the domain of ℓ_{t-1} when the value function is given by (2.17), as shown in Appendix A.4. Unlike the discussion in Section 2.2, in the Bayesian framework, $\{\theta_1, \cdots, \theta_T\}$ are not revealed to the decision maker, which leads to the expected future value function taking the form of an ensemble average:

$$\mathbb{E}_{L_t}\left[V_t\left(L_t\right) \mid \ell_{t-1}\right] = \mathbb{E}_{\Theta_t}\left[\mathbb{E}_{L_t}\left[V_t\left(L_t\right) \mid \ell_{t-1}, \Theta_t\right]\right]. \tag{2.21}$$

This is an immediate consequence of the law of total probability. The following lemma helps establish whether there exists a unique intersection point between $\beta_{t-1}\mathcal{P}\left(\ell_{t-1}\right)$ and $\mathbb{E}_{L_t}\left[V_t\left(L_t\right) \mid \ell_{t-1}\right]$, which would yield the optimal stopping time in the form of (2.8).

Lemma 2.3. *Let U be a random variable with a probability law $f_U(u)$. If $\forall u : f_U(u) > 0$ there exists a unique intersection point $x(u)$ satisfying $h(x(u)) = g(x(u), u)$ then, there exists a unique intersection point x_0 such that:*

$$h\left(x_0\right) = \mathbb{E}_U\left[g\left(x_0, U\right)\right].$$

Proof is given in Appendix A.5. The following result is a corollary of Theorem 2.1 and it establishes the optimal stopping time for consulting experts under a probability law f_{Θ^T}.

Corollary 2.1. *For a given probability law f_{Θ^T} governing Θ^T, there exists a unique threshold function η_t such that:*

$$\mathcal{T}^* = \min\left\{t : L_t \geq \eta_t\right\},$$

yields the optimal stopping for consulting experts.

Let us motivate that a closed-form solution similar to that in (2.16) might not be attainable for an arbitrary probability law over Θ^T. Observe that value function $V_T\left(\ell_T\right)$ still

attains the form in (2.15), where the expected future reward at time $t = T - 1$ follows from (2.21). The Bellman equation (2.14) yields that the consulting process stops at time $T - 1$ if $\ell_{T-1} \geq \eta_{T-1}$, where η_{T-1} is the solution to:

$$\eta_{T-1} = \log \frac{\int_{\eta_{T-1}}^{\infty} \mathcal{P}(\Theta_T)\, df_{\Theta_T}}{\delta_T + \int_{\eta_{T-1}}^{\infty} \mathcal{Q}(\Theta_T)\, df_{\Theta_T}}.$$

One should note the similarity with (2.16) and the fact that such a threshold exists for all Θ^T. Nonetheless, η_t might not be finite as evidenced by allowing $\mathcal{P}(\Theta_t)$ to be uniform over $[1/2, 1]$, $\forall t$ for the no-cost formulation.

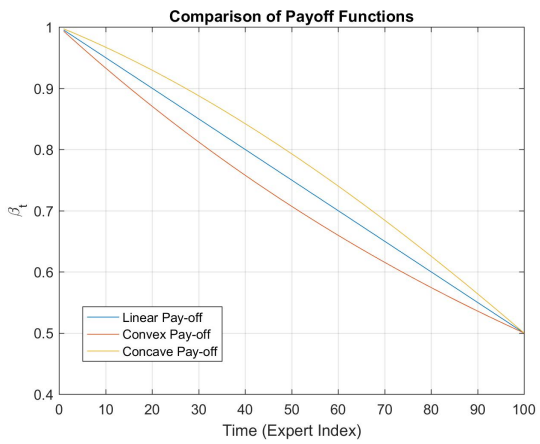
2.4 Experiments

The experiments consider $T = 100$ experts being consulted sequentially and the decision maker receiving rewards that correspond to different pay-off functions β_t. We consider the following pay-off functions:

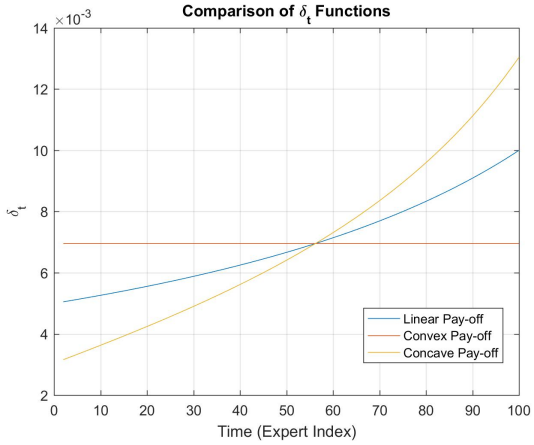
1. Linear pay-off: $\beta_t = 1 - t/2T$.
2. Convex pay-off: $\beta_t = \gamma^t$, where $\gamma = \sqrt[T]{t}$.
3. Concave pay-off: $\beta_t = 2 - \gamma^t - t/T$.

Note that all cost functions are normalized to ensure $\beta_0 = 1$ and $\beta_T = 1/2$. Figure 2.2 illustrates these pay-off functions and the corresponding δ_t functions, which might be considered *relative cost per consultation*. Figure 2.2b shows that δ_t functions intersect at $t = 57$, and $\forall t < 57$, convex pay-off has the highest δ_t, followed by that that of linear pay-off and that of concave pay-off. This carries significance as the threshold η_t is monotonic in δ_t.

Figure 2.3 illustrates the properties of the threshold function η_t. As stated in Section 2.2.2, the threshold function increases in θ_t and decreases in δ_t. Figure 2.3a shows that for equally competent experts with competence $p \in \{0.5, \cdots, 1\}$ (thus monotonically ordered θ), higher competence yields higher thresholds for stopping. One should note that Figure 2.3a illustrates the case for convex pay-off however, the behavior persists for any cost function. Figure 2.3b, on the other hand, illustrates the impact of δ_t when competences are chosen at random: Recall from Figure 2.2b that up to time $t = 57$, the δ_t of convex pay-off is greater than that of linear pay-off and that of concave pay-off. Figure 2.3b illustrates that the corresponding thresholds have the correct ordering.

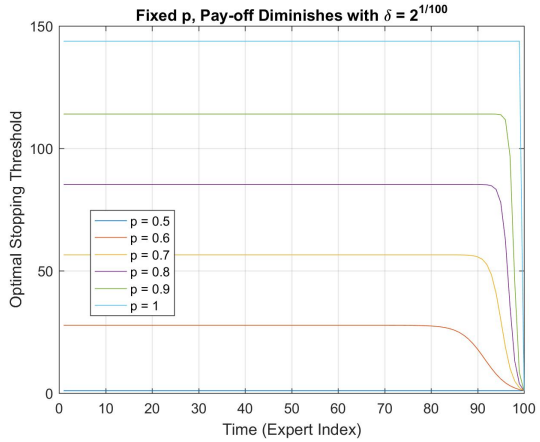


(a) Pay-off Functions Used for Experiments

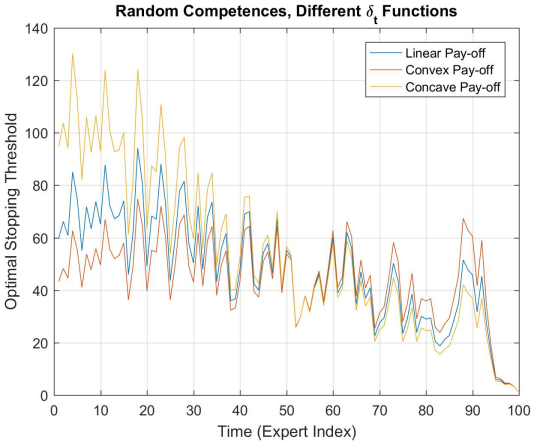


(b) Functions δ_t that Correspond to Pay-off Functions

Figure 2.2: Linear, Convex, and Concave Pay-off Functions Used for Experiments and the Corresponding δ_t Functions



(a) Pay-off Functions Used for Experiments



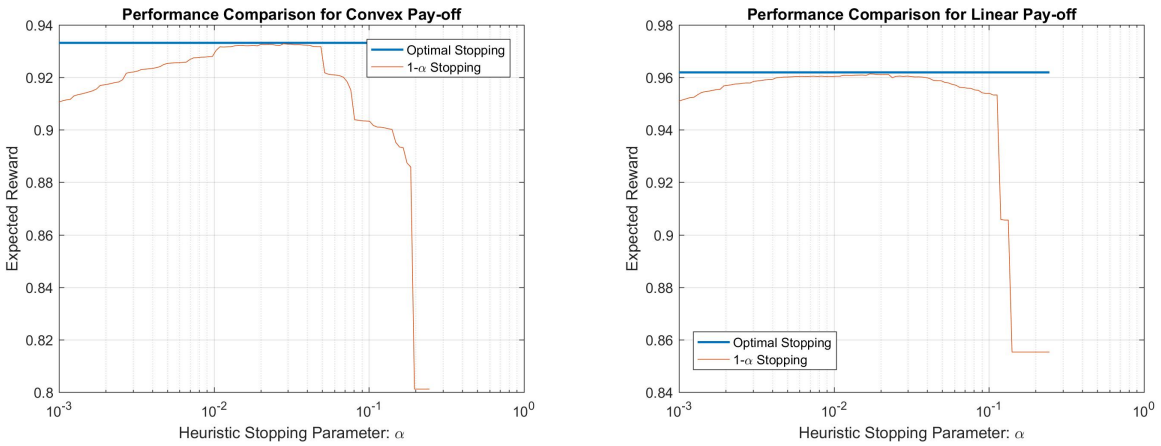
(b) Functions δ_t that Correspond to Pay-off Functions

Figure 2.3: Linear, Convex, and Concave Pay-off Functions Used for Experiments and the Corresponding δ_t Functions

2.4.1 The Optimal Stopping Time for Known Competences

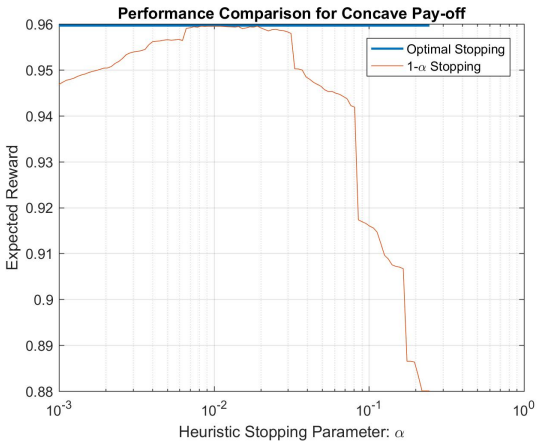
We first address the problem setup, where competence of every expert (along with their ordering) is known to the decision maker. The optimal stopping rule from Theorem 2.1 is compared against a heuristic $1 - \alpha$ rule that stops upon the first occurrence of the event:

$$\max_{y \in \mathcal{Y}} \mathbb{P}\left(Y = y \,\middle|\, X^t = x^t\right) \geq 1 - \alpha.$$



(a) Convex Pay-off

(b) Linear Pay-off

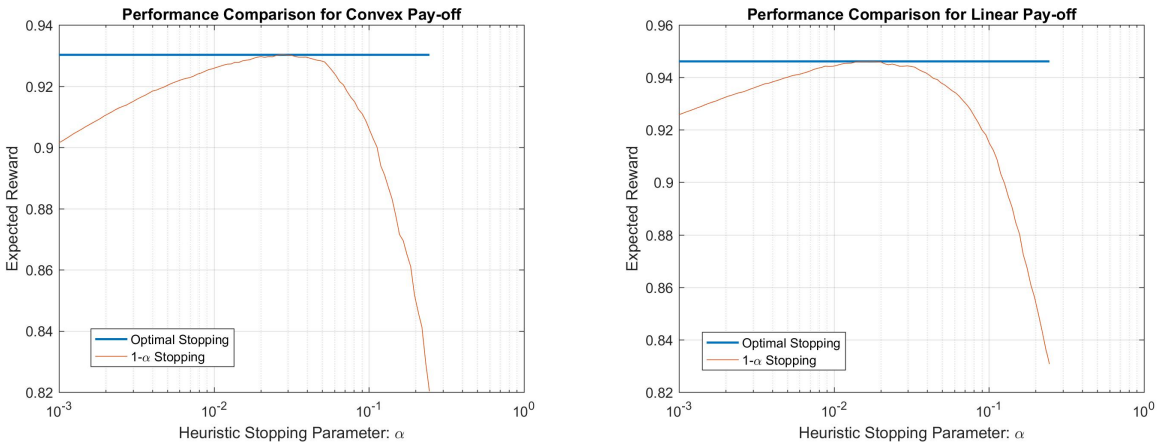


(c) Concave Pay-off

Figure 2.4: Comparison of the Optimal Stopping Rule with Heuristic $(1-\alpha)$ Rules When Competences Are Randomly Chosen

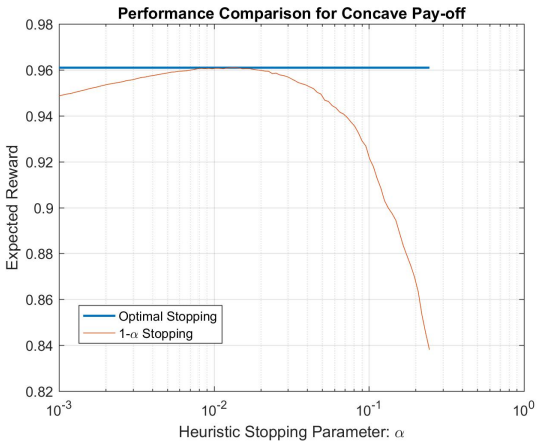
Figure 2.4 illustrates performance comparison between the expected reward of the optimal stopping time and those from $1-\alpha$ rule with α values taken from the interval $[0.001, 0.25)$. The optimal stopping rule outperforms all heuristic rules as expected. Importantly, the results are shown for an arbitrary set of competences randomly drawn from the interval $[0.5, 0.9]$. The performance is achieved over an average of 1000 trials, where 1000 different tasks were chosen at random but the competences remained fixed (picked at random in the beginning of the experiment). It appears that there exist α values that ensure comparable performance even though it is not known how to compute them before the experiment.

Figure 2.4 shows that there are discontinuities in the performance of $(1-\alpha)$ rules. This is due to competences being picked at random in the beginning of the experiment. Even though the optimal stopping rule should (and does) outperform other stopping rules for any set of fixed competences, one could repeat this experiment over again to capture the



(a) Convex Pay-off

(b) Linear Pay-off



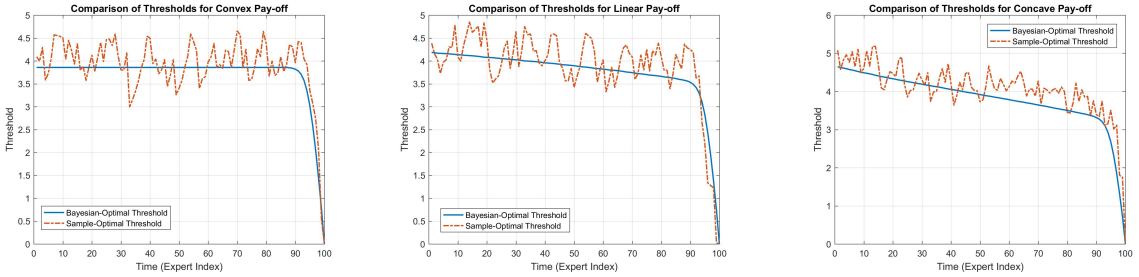
(c) Concave Pay-off

Figure 2.5: Comparison of the Optimal Stopping Rule with Heuristic $(1-\alpha)$ Rules when Averaged over Competences

behavior of other rules more accurately. Figure 2.5 illustrates when the previous experiment with 100 tasks is repeated 100 times, for each new trial competences are picked at random. As expected, the performance of heuristic rules can be observed more clearly, where the expected reward of the optimal rule still exceed those of other, heuristic, rules.

2.4.2 The Optimal Stopping Time for Bayesian Competences

In this experiment, we allow each expert to have a competence P_i drawn from the uniform distribution over $[0.5, 0.9]$, the competences are independent and identically distributed (i.i.d.). The unique threshold for optimal stopping in Corollary 2.1 computed numerically. Figure 2.6 illustrates a comparison between optimal stopping threshold for known competences from (2.16) and that for the Bayesian case. Similar to the experiment for known competences,



(a) Convex Pay-off (b) Linear Pay-off (c) Concave Pay-off

Figure 2.6: Comparison of the Optimal Stopping Thresholds for Bayesian and Known Competences

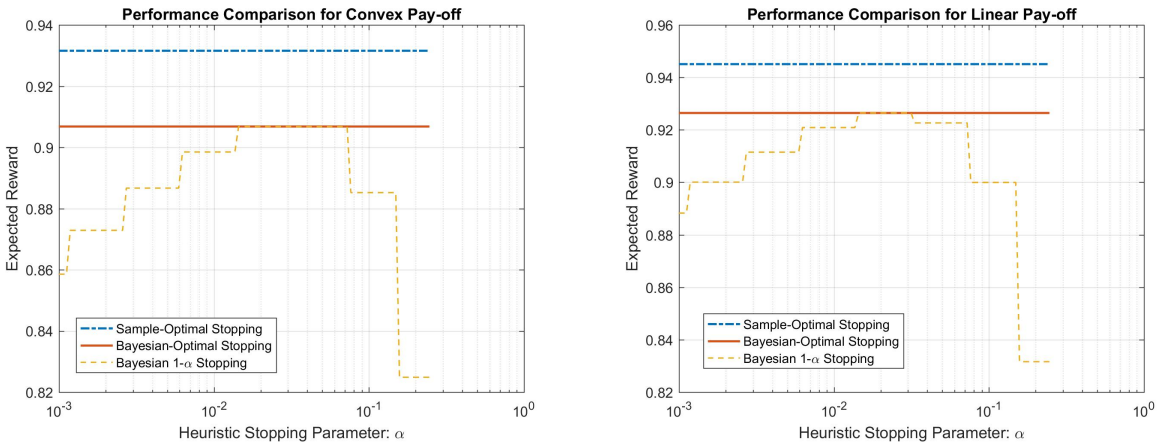
experts with random competences are consulted for 100 tasks and process is repeated for 100 times with competences being generated randomly at the beginning of each experiment. It is important to note that when competences are chosen i.i.d. with log-likelihood process Θ, the likelihood ratio for the Bayesian case (2.20) takes the form:

$$L_t = \left| \sum_{i=1}^{t} X_i \right| \log \frac{\mathbb{E}\left[\mathcal{P}\left(\Theta\right)\right]}{\mathbb{E}\left[\mathcal{Q}\left(\Theta\right)\right]}. \tag{2.22}$$

The $(1-\alpha)$ rules, as well as the optimal stopping rule uses (2.22) for stopping and opinion aggregation. Figure 2.7 illustrates that performance of the Bayesian optimal stopping rule outperforms the heuristic rules while suffering from a performance degradation with respect to the optimal rule for known competences, termed "sample-optimal", which has direct access to competences.

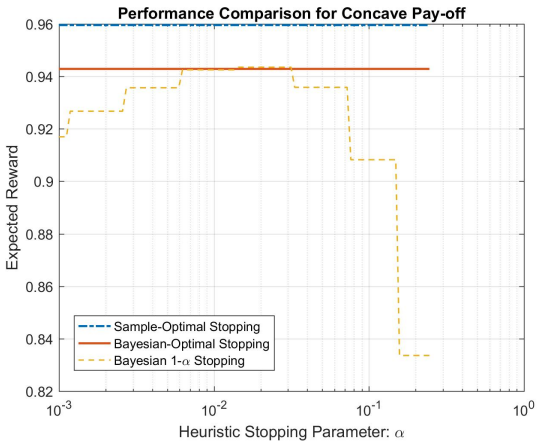
2.4.3 Constant Pay-off while Consulting Identical Experts

Experiments so far rely on information on competences either directly or through Bayesian prior. Here we address equally reliable experts being consulted at no cost, which leads to a constant pay-off for all times ($\beta_t = 1, \forall t \in [T]$), hence aims to sequentially minimize the probability of error. Figure 2.8 illustrates that the optimal stopping rule (2.19) outperforms $(1-\alpha)$ rules *that use true competences for decision making*. On one hand, this indicates that the minimum probability of error achieving rule does not rely on the true competence value and thus, is an unsupervised rule (competences $p \in \{0, 1/2, 1\}$ are exceptions, as noted in Section 2.2.2). On the other hand, it indicates the knowing competences can not improve performance in this framework.



(a) Convex Pay-off

(b) Linear Pay-off



(c) Concave Pay-off

Figure 2.7: Comparison of the Optimal Stopping Rule with Heuristic $(1-\alpha)$ Rules when Average over Competences

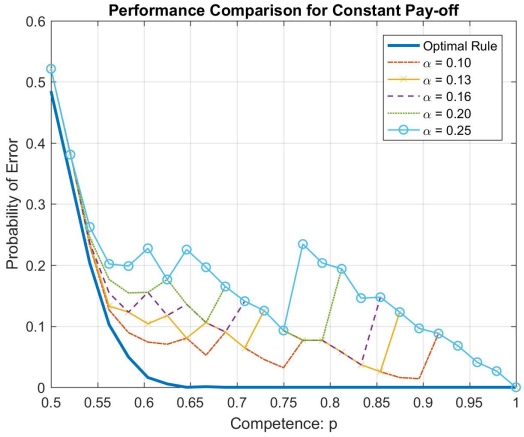


Figure 2.8: Constant Pay-off for Equally Competent Experts

CHAPTER 3

UNSUPERVISED OPINION AGGREGATION

As seen in Chapter 2, the performance of decision making after consulting experts, opinion aggregation is a key factor in applications that rely on subjective information. Even though optimal opinion aggregation rules in various problem formulations are well-known when the probability law governing the competence of each expert is known, it is of interest to pursue opinion aggregation rules that do not rely on such prior information.

The opinion aggregation, has far-reaching roots: Bayesian hypothesis testing sets the fundamental limits of opinion aggregation provided that the complete probability law governing the experts exists and is known to the decision maker [4]. Nonetheless, it is difficult, if at all possible, to model reliably, over the entire application space, the probability law that underlies highly specialized processing units, which are often trained on limited data. The modeling error and the concomitant performance degradation in opinion aggregation might prove to be uncontrollable in such cases. The uncertainty in modeling, however, does not necessarily render statistical inference of the underlying probability law implausible. Dempster-Schafer theory addresses the Bayesian inference problem under model uncertainty by incorporating *belief* and *plausibility* functions to substitute direct application of the probability law [53, 54]. In general, use of approximate models for reasoning is called *fuzzy logic* [55] and has wide range of applications in control theory [56].

In the absence of approximate or complete probability laws, *feedback* is often used to learn and mitigate the reliability of each source and the concomitant probability distribution over opinions. The mixture of experts setup successfully aggregates opinions from potentially adversarial experts by the use of reliable past information [45]. Similarly, boosting and other associated meta-learning concepts use feedback, often at the expense of additional training data, to learn and use the reliability of different strategies, or classifiers [7]. The use of feedback is both the strength and the fundamental limitation of such strategies, as feedback is often expensive even when it is feasible to generate. The Bayesian ideas are called to action and a prior distribution on the underlying probability law is often *assumed* when feedback is not feasible. The Bayesian approach enables the use of a rather powerful toolbox of iterative algorithms including belief propagation, expectation-maximization, and mean-

field methods [19, 21, 23, 25, 28]. Bayesian decision-aggregation methods are powerful to the extent with which the prior successfully captures reality.

The existence of reliable or approximate models, feedback, or prior information fundamentally defines the application space on which the associated ideas can be employed, and thus, are limited to such applications. As such, we may consider opinion-aggregation techniques that use some form of side information as *supervised* rules. Conceptually, the presence of such side information renders an otherwise vital fact obsolete: Experts aim to achieve a common task, not to fail it, as long as a fixed probability governing how experts produce opinions exists. The opinion-aggregation rules that rely solely on the existence of such a probability law can therefore be considered *unsupervised* rules. Often a community of experts is less subjective than its individual constituents, motivating statistical inference to compensate the lack of supervision. The fundamental challenge is to design unsupervised decision rules that reliably aggregate opinions without using side information on the underlying, hidden, and fixed probability law governing the experts.

Chapter 3 addresses this challenge from an inherently statistical perspective. Section 3.1 provides the formal background for opinion aggregation and identifies different regimes of operation. Section 3.2 introduces a novel technique for estimating, dynamically in real-time, the reliability of each expert from a set of opinions and discusses the properties of the proposed method. The purpose of such reliability estimation is made clear by using these estimates to infer the unknown probability law. Section 3.4 introduces a sharp upper-bound on the minimum probability of error achievable when the probability law is known, improving upon the state-of-the-art. Section 3.5 proposes an unsupervised opinion aggregation rule based on the minimum probability of error rule that uses the unsupervised reliability estimates and investigate its fundamental limits. Section 3.6 addresses empirical extensions that aggregate a fixed block of opinions as well as doing so adaptively in real-time. Experiments are given in Section 3.7. The proofs are deferred to Appendix B.

3.1 Background and Problem Definition

Let a set of *tasks* $\mathbb{T}$ for identifying hidden binary *states* $Y(t) \in \{-1, 1\}$, $\forall t \in \mathbb{T}$ (in the context of classification, $Y(t)$ is often called *label or ground-truth* instead) be generated independently, $Y(t) \perp Y(\tau)$, $\forall t \neq \tau$, with a uniform prior:

$$\mathbb{P}\left(Y(t) = 1\right) = \mathbb{P}\left(Y(t) = -1\right) = 1/2, \forall t \in \mathbb{T}. \tag{3.1}$$

Even though there exist applications that do not admit the uniform prior, such as group testing [57], where populations are severely skewed, (3.1) is a common assumption in opinion aggregation problems [17]. If a non-uniform prior on Y exists, standard Bayesian decision-making principles can be used to adapt the results, see, for instance, [4].

Experts generate binary opinions $X_i \in \{-1, 1\}$, $\forall i \in [N]$, that identify the true state $Y(t)$ with some probability:

$$p_i \triangleq \mathbb{P}\left(X_i(t) = Y(t)\right), \forall t \in \mathbb{T}. \tag{3.2}$$

The *true competence* p_i of an expert is considered *fixed* across tasks. Let there be N experts and assume that opinions are *generated* independently, which amounts to:

$$X_i(t) - Y(t) - X_j(t), \forall i \neq j \in [N], \tag{3.3}$$

for every task $t \in \mathbb{T}$. One should note that (3.3) does *not* indicate that opinions are independent. Indeed, it is conceptually clear that opinions should be dependent for meaningful inference as an opinion is a subjective evaluation of the current state, not an arbitrary input.

Experts that can be reliably defined by a stochastic law, as done here via $(3.1) - (3.3)$, are sometimes called *stochastic experts* to separate them from the more game-theoretic framework of *adversarial experts* [45], a similar distinction exists for multi-armed bandits [51]. Furthermore, when p_i is a function of the underlying task space $\mathbb{T}$, unlike how it is defined here, experts exhibit task-dependent competence. The Neyman-Pearson formulation of binary hypothesis testing [58] and the two-coin Dawid-Skene model [21], are examples of expert competences changing over the task space $\mathbb{T}$.

Opinion-based systems are sometimes referred to as *semi-supervised* systems due to the availability of experts and the concomitant subjective information [52]. However, in the context of opinion aggregation, one may consider supervision as *any* form of side-information that yields inference of the underlying probability law beyond the extent that opinions alone would allow. Formally, a *supervised* opinion-aggregation rule refers to a function $f(\cdot)$ mapping a set of opinions $\boldsymbol{X}(t)$ to an estimate of the true state $Y(t)$:

$$\hat{Y}(t) = f\left(\boldsymbol{X}(t); \mathcal{S}\right), \tag{3.4}$$

where $\hat{Y}(t) \in \{-1, 1\}$ and $\mathcal{S}$ denotes some form of side information. Supervised opinion-aggregation methods require different forms of side-information $\mathcal{S}$: $\{p_1, \cdots, p_N\}$ for binary hypothesis testing [4], known subsets of $\{y_1, \cdots, y_{t-1}\}$ for boosting and mixture of experts [7, 45], or a priori distribution $p_{P_1,\cdots,P_N}(\cdot)$ on competences for Bayesian techniques; [19, 21, 23,

25, 28] are some of the prominent examples.

Unsupervised opinion aggregation, on the other hand, refers to functions that only rely on the existence of the underlying probability law, here, for instance, characterized by (3.1)-(3.3). As such, they assume the form:

$$\hat{Y}(t) = f(\boldsymbol{X}(t)). \tag{3.5}$$

The most prominent example of unsupervised decision aggregation is *majority voting*, which is commonly accepted as the *baseline* for unsupervised techniques. A modern unsupervised technique, called spectral meta learner, relies on the singular values of the empirical covariance matrix of a collection of opinions [30].

Binary opinion-aggregation techniques often aim to minimize the probability of error:

$$\min_{f \in \mathcal{D}} \mathbb{P}(f(\boldsymbol{X}(t)) \neq Y(t)), \tag{3.6}$$

where $\mathcal{D}$ is the family of admissible opinion-aggregation rules that depend on the problem setup. For instance, the admissible rules comprise all, potentially randomized, decision rules for known competences, past-measurable decision rules for feedback available from hidden models, and functions that directly map available opinions to decisions on the corresponding tasks in the absence of all side information.

Opinion-aggregation rules (3.4)-(3.5), and the corresponding performance metric (3.6), are written in the form of single-task opinion aggregation, where a set of opinions are used to make a decision $\hat{Y}(t)$. We next discuss several modes of operation for unsupervised opinion aggregation.

3.1.1 Unsupervised Opinion Aggregation

Unsupervised opinion aggregation refers to employing a function $f(\cdot)$ of opinions $\{X_i(t)\}$ for $(i,t) \in \mathcal{N} \times \mathcal{T}$, where $\mathcal{N} \times \mathcal{T} \subset \mathbb{T} \times [N]$ to identify a set of hidden states $\{Y(t) : t \in \mathcal{T}\}$, *directly*:

$$\left\{\hat{Y}(t) : t \in \mathcal{T}\right\} = f(\{X_i(t) : (i,t) \in \mathcal{N} \times \mathcal{T}\}). \tag{3.7}$$

Note that there is no side information $\mathcal{S}$ as an input to the decision rule and that the subset $\mathcal{N} \times \mathcal{T}$ determines the *operational meaning* of the opinion-aggregation rule: The function $f(\cdot)$ might be fixed for all tasks $\mathbb{T}$, such as majority voting [46], or it might change adaptively in tasks. Furthermore, $f(\cdot)$ might process blocks of opinions (often iteratively and non-adaptively) [19, 21, 23, 25, 28]. Formal definitions of unsupervised opinion-aggregation rules

are discussed next.

Instantaneous Opinion Aggregation

A fixed function directly applicable to opinions $\{X_i(t) : i \in [N]\}$ on any task $t \in \mathbb{T}$, is an instantaneous opinion-aggregation strategy. Conceptually, these rules do not require additional memory to store past opinions. Formally, they take form $\hat{Y}(t) = f(\boldsymbol{X}(t))$, as illustrated in Figure 3.1.

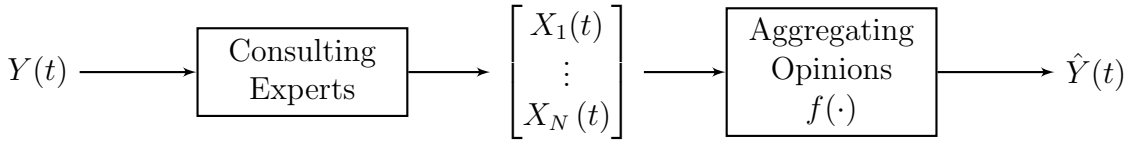


Figure 3.1: Instantaneous Opinion Aggregation

It is often difficult to find meaningful unsupervised rules that are instantaneous, with a notable exception: Majority voting, denoted by $f^{MV}(\boldsymbol{X}(t))$, is a commonly-accepted baseline for unsupervised rules:

$$f^{MV}(\boldsymbol{X}(t)) = \text{sign}\left(\sum_{i=1}^{N} X_i(t)\right), \tag{3.8}$$

where ties are broken arbitrarily. It is often taken for granted that ties being broken arbitrarily is a direct consequence of (3.1). Indeed, if a prior $p_Y(y)$ on $Y(t)$ were known,

$$\hat{Y} = \underset{y \in \{-1,1\}}{\arg\max}\, \mathbb{P}(Y = y)$$

should be chosen in the event of a tie [4].

Conjectures of marquis de Condorcet, [46], have long been debated in the social choice literature i.a. [18, 20, 59], revealing that majority voting is not reliable when heterogeneous ($p_i \in [0, 1]$) or arbitrarily weak ($p_i \to 1/2$) populations of experts are concerned.

Block(-Iterative) Opinion Aggregation

A strategy that processes a *collection* of opinions $\{X_i(t) : (i, t) \in \mathcal{N} \times \mathcal{T}\}$ to estimate the corresponding states $\left\{\hat{Y}(t) : t \in \mathcal{T}\right\}$ is a block opinion-aggregation rule. Conceptually, these rules, often iteratively, process past opinions to decide for the respective block of tasks. We

focus on block rules that leverage *all* the available opinions with $\mathcal{N} = [N]$ and $\mathcal{T} = \mathbb{T}$, as illustrated in Figure 3.2.

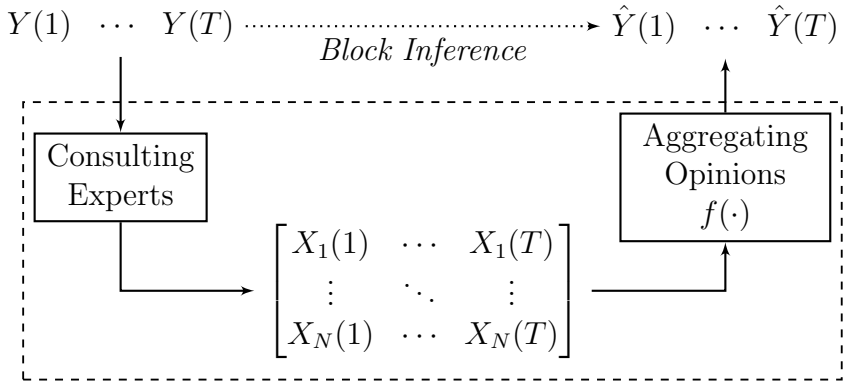


Figure 3.2: Block(-iterative) Opinion Aggregation

Often, off-line techniques such as expectation maximization, belief propagation, or spectral decomposition methods are used iteratively on a collection of opinions [19, 21, 23, 25, 28]. Specifically, belief propagation has been shown to asymptotically minimize the probability of error for specific sparse subsets $\mathcal{N} \subset [N]$, often referred to as task-assignment [28]. The singular value distribution of the empirical covariance matrix of opinions has also been investigated for opinion aggregation [23, 30]. Adaptations of expectation maximization have been proposed for adaptive block-processing and task-dependent modeling of competences [21, 25]. These methods are generally computationally expensive and they seldom yield provable guarantees for their performance.

Adaptive Opinion Aggregation

A strategy that infers the underlying probability law sequentially from past observations is an adaptive opinion-aggregation strategy. Formally, it has the form $\hat{Y}(t) = f_t(\mathbf{X}(t))$, as illustrated in Figure 3.3.

These strategies often estimate and employ empirical competence estimates $\{\hat{p}_1, \cdots, \hat{p}_N\}$ in the decision making process and, traditionally, they are almost exclusively formulated as supervised opinion-aggregation strategies that have access to state feedback or additional "meta"-training [7, 17, 45].

We propose a set of non-iterative, unsupervised decision-aggregation strategies with quantifiable performance guarantees. Inspired from the naïve Bayes decision rule, which follows from the likelihood ratio test for known competences $\{p_1, \cdots, p_N\}$, [4, 17], we propose block-decision aggregation rules and discuss their adaptive extension. The proposed rules employ biased estimates of expert competences, called *pseudo competences*, directly.

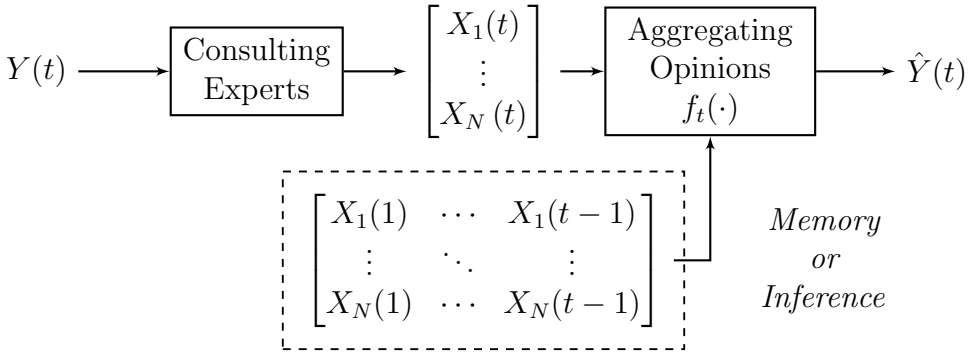


Figure 3.3: Adaptive Opinion Aggregation

3.2 Unsupervised Estimation of Competences

Let the experts $\{X_1, \cdots, X_N\}$ be characterized by the probability law (3.1)-(3.3). Then, given knowledge of the true states $\{Y(1), \cdots, Y(T)\}$ the true competence p_i of an expert can objectively be measured by the frequency with which the expert successfully identifies the true state:

$$\hat{p}_i(T) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(X_i(t) = Y(t)\right). \tag{3.9}$$

The ergodicity of the process $\mathbb{1}\left(X_i(t) = Y(t)\right)$, which follows from (3.1)-(3.3), yields that $\hat{p}_i(T) \to p_i$ as the number of tasks increases. However, an unsupervised decision maker does not have access to $Y\left(t\right)$ and therefore, cannot make use of these reliable competence estimates $\hat{p}_i(T)$ for decision making.

Conceptually, measuring the quality of an opinion without knowing the true state, or the ground-truth, is a commonly encountered challenge in human decision-making: One might accept the consensus of extrinsic opinions on a task as a proxy for the truth to the best of one's knowledge. We define a form of opinion-based reliability, or the *pseudo competence* of an expert, as the likelihood of an expert agreeing with independently-generated opinions from other experts:

$$\tilde{p}_i \triangleq \mathbb{P}\left(X_i(t) = f^{MV}\left(\mathbf{X}_{\setminus i}(t)\right)\right), \forall t \in \mathbb{T}. \tag{3.10}$$

As discussed in Section 3.1, majority vote is an intuitive decision rule that is often accepted as a baseline, and it leads to the notion of *agreeing with peers*. Formally, the subset $\mathbf{X}_{\setminus i}$ are the *peers* of the expert X_i and their collective competence under majority vote is denoted by:

$$p_{\setminus i} \triangleq \mathbb{P}\left(Y(t) = f^{MV}\left(\mathbf{X}_{\setminus i}(t)\right)\right), \forall t \in \mathbb{T}. \tag{3.11}$$

We refer to $p_{\setminus i}$ as *self-excluding majority* and use it as a reference point for measuring the

pseudo competence of the corresponding expert. It follows from the law of total probability that pseudo competence is a function of the entire committee $\{p_1, \cdots, p_N\}$ [60]:

$$\tilde{p}_i = p_i p_{\setminus i} + q_i q_{\setminus i}, \tag{3.12}$$

where, $q_{\setminus i} = 1 - p_{\setminus i}$. If one *chooses* to accept the frequency with which an expert agrees with peers as the empirical competence estimate for that expert:

$$\hat{\rho}_i(T) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(x_i(t) = f^{MV}\left(\mathbf{x}_{\setminus i}(t)\right)\right), \tag{3.13}$$

then it is possible to infer competences, in the form of pseudo competences, in real-time. Similar to the true competence estimates, the ergodicity of the process:

$$A_i(t) = \mathbb{1}\left(X_i(t) = f^{MV}\left(\mathbf{X}_{\setminus i}(t)\right)\right),$$

which follows (3.1)-(3.3), yields that $\hat{\rho}_i(T) \to \tilde{p}_i$ as the number of tasks increases. Furthermore, (3.13) enables distributed estimation and ranking of competences on connected networks, as discussed in Section 3.3, [12].

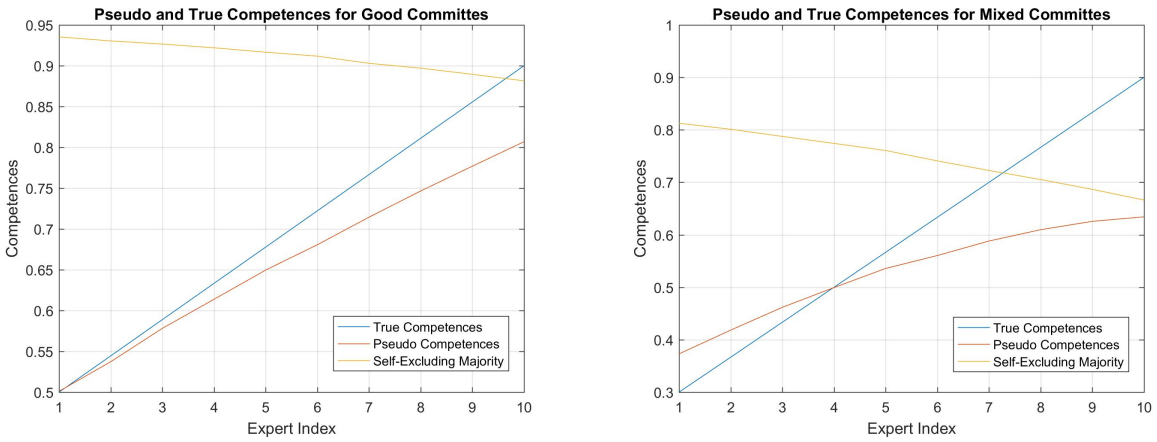
A key aspect of the pseudo competence is the exclusion of self-opinions $\mathbf{X}_{\setminus i}$ and the concomitant competence of the peers $p_{\setminus i}$. Exclusion of self-opinions is rather intuitive as an expert always agrees with itself, hence including self-opinions would bias the pseudo competence toward5 the expert that is being measured. The collective expertise of the peers on the other hand, is critical for pseudo competence to make sense. Conceptually, as one should not measure the competence of an expert with the likeliness of failure, one should not measure it by the likelihood of agreement with those who fail often.

We next address what we call good committees ($p_i > 1/2, \forall i \in [N]$) and mixed committees ($p_i \in [0,1], \forall i \in [N]$) to explore the conditions that would yield reliable and meaningful inference using pseudo competences instead of true competences.

3.2.1 Properties of Pseudo Competence for Good Committees

We call a committee *good* if $p_i > 1/2$, $\forall i \in [N]$ for a finite $N > 2$, or $p_i > 1/2 + \varepsilon$, $\forall i$, for some $\varepsilon \in (0, 1/2)$ when the committee is countably infinite. Contextually, a good committee consists of experts that generate correct opinions *in general*. Formally, it follows that for a good committee, the self-excluding majority satisfies $p_{\setminus i} > 1/2$, $\forall i$.

Proposition 3.1 summarizes the key properties of pseudo competence for a good commit-



(a) A Good Committee (b) A Mixed Committee

Figure 3.4: A Comparison of Pseudo Competences $\tilde{p}_i$, (3.10), to True Competences p_i, (2.1), Self-Excluding Majority $p_{\setminus i}$, (3.11), as Reference

tee.

Proposition 3.1. *For every finite committee* $\{p_1, \cdots, p_N\}$ *with* $p_i > 1/2$, $\forall i \in [N]$, *or for every countably infinite committee* $\{p_1, \cdots, p_i, \cdots\}$ *with* $p_i > 1/2 + \varepsilon$, $\forall i \in \mathbb{N}$ *for some* $\varepsilon \in (0, 1/2)$, *the pseudo competence satisfies:*

1. *Ordering:* $p_i > p_j \iff \tilde{p}_i > \tilde{p}_j$.

2. *Under-estimation:* $1/2 < \tilde{p}_i < p_i$.

The pseudo competences preserve the ordering of the true competences and they are strictly greater than $1/2$ for good committees – proof is given in Appendix B.1. On the other hand, pseudo competence penalizes the most competent experts while evaluating lower-competence experts more accurately. Figure 3.4a illustrates this phenomenon for $N = 10$ experts with competences uniformly spaced over $[0.5, 0.9]$. A good committee ensures that $\big(p_{\setminus i} > 1/2, \forall i\big)$, however, it is not necessary for a committee to be *good* to guarantee the same condition. We next discuss mixed committees that ensure reliable peers for every expert.

3.2.2 Properties of Pseudo Competence for Mixed Committees

Pseudo competence not only relies on the notion that the committee, as a whole, is a sufficiently competent reference point to measure the competence of each expert but also relies on committee being robust to the absence of any individual expert in the sense that $p_{\setminus i} > 1/2$ for

every expert. Certainly, this notion is not valid for all mixed committees: Neither reliable inference of pseudo competence and nor reliable unsupervised opinion aggregation is unattainable when, for instance, a significant portion of the committee is unreliable ($|i : p_i \leq 1/2|$ is large), or the committee as a whole is indecisive $\left(\mathbb{P}\left(f^{MV}\left(\mathbf{X}\right) = Y\right) \approx 1/2\right)$. It is rather difficult to characterize the mixed committees that ensure reliable unsupervised inference. Several works have proposed application-specific conditions on the mixed committees, for instance, [18, 19, 25, 26, 28].

It is of practical interest to investigate mixed committees that are competent as a whole: Often, a notion of *consistency* arises in various supervised or unsupervised learning, inference, and decision rules, i.a. [45]. Consistency under majority vote, also known as *Condorcet's Jury Theorem* [18, 46, 59], is defined as follows.

Definition 3.1 (Consistency)**.** *A committee of experts with competences $\{p_i : i \in \mathbb{N}\}$ is consistent under majority voting if:*

$$\lim_{N\to\infty} \mathbb{P}\left(f^{MV}\left(X_1, \cdots, X_N\right) = Y\right) = 1.$$

An explicit characterization of consistent mixed committees is non-trivial and addressed in [18, 20, 59]. Nonetheless, for every consistent committee, there exists a monotonically increasing, tight lower bound a_N that we call the *rate* of consistency:

$$a_N = \inf_{n\geq N} \mathbb{P}\left(f^{MV}\left(X_1, \cdots, X_n\right) = Y\right). \tag{3.14}$$

The rate of consistency determines the asymptotic performance of decision aggregation rules that use pseudo competences. Proposition 3.2 extends the notion of under-estimation in Proposition 3.1 to what-we-call *pessimistic* estimation for mixed committees.

Proposition 3.2. *For every consistent committee $\{p_i\}_{i\in\mathbb{N}}$, there exists a committee size n^* such that $\forall N > n^*$ pseudo competences satisfy:*

1. *Ordering:* $p_i \geq p_j \iff \tilde{p}_i \geq \tilde{p}_j$.
2. *Pessimism:* $\min\{\tilde{p}_i, 1-\tilde{p}_i\} \geq \min\{p_i, 1-p_i\}$.

Conceptually, pessimism property indicates that good experts ($p_i > 1/2$) are underestimated and bad ones ($p_j < 1/2$) are overestimated – proof given in Appendix B.2. Figure 3.4b illustrates Proposition 3.2 for $N = 10$ experts with competences uniformly spaced over $[0.3, 0.9]$. A condition for *finite* mixed committees to satisfy Proposition 3.2 is given by (B.1) in Appendix B.1.

The pseudo competence (3.10) provides a metric that can be estimated in real-time via (3.13) and that requires no side-information beyond the existence of a probability law characterized by (3.1)-(3.3). Committees that satisfy $p_{\setminus i} > 1/2$, $\forall i$ ensure that pseudo competence preserves the ordering of the true competences and it neither mistakes a bad expert for a good one nor it does the opposite. Therefore, it provides a reliable unsupervised metric for estimating the underlying probability law directly. Another advantage of pseudo competence metric is that it admits distributed estimation, which is addressed next.

3.3 Distributed Unsupervised Estimation of Competences

Network models of opinion sources include locally generated opinions of varying reliability. In power-constrained, low-bandwidth applications, distributed learning of nodal competences is of interest. Consider a network modeled by a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ comprising a finite set of experts $(X_1, \ldots, X_N)$ sitting at vertices, where $N = |\mathcal{V}|$. Experts are characterized by (3.1)-(3.3), and the edge set $\mathcal{E}$ defines the interconnections among experts. Let $\underline{\mathbf{A}}$ be the adjacency matrix of the network, that is, $\forall i, j \in \mathcal{V}$:

$$(\underline{\mathbf{A}})_{ij} = \mathbb{1}\left(i \leftrightarrow j \in \mathcal{E}\right).$$

Recall that for every pair of vertices $i, j \in \mathcal{V}$, $i \leftrightarrow j \in \mathcal{E}$ denotes being connected by an edge and define diagonal "edge-degree" matrix $\underline{\mathbf{\Lambda}}$ such that $\forall i \in \mathcal{V}$, $(\underline{\mathbf{\Lambda}})_{ii} = \sum_{k=1}^{N} (\underline{\mathbf{A}})_{ik}$. Using the edge-degree matrix, define the Laplacian of the network:

$$\underline{\mathbf{L}} = \underline{\mathbf{\Lambda}} - \underline{\mathbf{A}}.$$

Observe that for a connected graph, as is of interest here, there exists a walk from every node to every other node and hence, $\underline{\mathbf{L}}$ has exactly one zero-eigenvalue.

The goal is to locally aggregate opinions on tasks $\mathbb{T}$ and reach a consensus on the ordering of competences when the network comprises heterogeneous experts ($p_i \neq p_j$) and reach a consensus on an estimate when it comprises homogeneous experts ($p_i = p$, $\forall i \in \mathcal{V}$). Formally, let $\mu(\cdot)$ denote the *network average*; $\forall \mathbf{a} \in \mathbb{R}^N$:

$$\mu(\mathbf{a}) = \frac{1}{N} \sum_{i=1}^{N} a_i. \tag{3.15}$$

Two regimes of operation are investigated here. The first is the low task-frequency, or "sequential regime", where the network has time to diffuse information in-between tasks,

hence global estimates of competence can be attained. The second is the high task-frequency, or "batch regime", where the network is compelled to infer competences locally. Formally, these regimes are defined as follows:

$$\{\mathbf{X}(t) \longrightarrow_c \mathbf{1}\mu(\mathbf{X}(t)) \longrightarrow_l \tilde{\mathbf{p}}^g(t)\}_{t\in\mathbb{T}} \longrightarrow_t \tilde{\mathbf{p}}^g(|\mathbb{T}|), \tag{3.16}$$

$$\left\{\mathbf{X}(t) \longrightarrow_l \tilde{\mathbf{p}}^\ell(t)\right\}_{t\in\mathbb{T}} \longrightarrow_t \tilde{\mathbf{p}}^\ell(|\mathbb{T}|) \longrightarrow_c \mathbf{1}\mu\left(\tilde{\mathbf{p}}^\ell(|\mathbb{T}|)\right). \tag{3.17}$$

Here, $\longrightarrow_c$ denotes distributed averaging via consensus, $\longrightarrow_l$ denotes local information processing, and $\longrightarrow_t$ denotes exhausting the available tasks. An empirical competence estimates based on global information in the sequential regime (3.16) and local information in the batch regime (3.17) *up to* task t are denoted by $\tilde{\mathbf{p}}^g(t)$ and $\tilde{\mathbf{p}}^\ell(t)$ respectively.

Any consensus rule can be employed in this setup: Let $\underline{\mathbf{D}} \in \mathbb{R}^{N\times N}$ be a doubly stochastic matrix on the network: $\sum_j(\underline{\mathbf{D}})_{ij} = 1$, $\sum_i(\underline{\mathbf{D}})_{ij} = 1$ with $(\underline{\mathbf{D}})_{ij} \geq 0$, $\forall i, j \in \mathscr{V}$ and $\forall i \neq j$, $(\underline{\mathbf{D}})_{ij} = 0$ whenever $(\underline{\mathbf{A}})_{ij} = 0$. Formally, let $\mathbf{1}$ denote the all-one column vector and note that $\forall t \in \mathbb{T}$, $\mathbf{x}(t) \longrightarrow_c \mu(\mathbf{x}(t))\mathbf{1}$, denotes the process:

$$\mathbf{x}(n;t) = \underline{\mathbf{D}}\mathbf{x}(n-1;t),$$

with $\mathbf{x}(0;t) = \mathbf{x}(t)$ being the network state, or opinion pool, on a task and $\lim_{n\to\infty}\mathbf{x}(n;t) = \mu(\mathbf{x}(t))\mathbf{1}$ being the average network opinion. Furthermore, such a consensus rule guarantees strong, or almost sure (a.s.), convergence: Observe that random process $\mathbf{X}(n;t) = \underline{\mathbf{D}}^n\mathbf{X}(t)$ and $\lim\limits_{n\to\infty} x(n;t) = \mu(\mathbf{x}(t))$ for every sample $\mathbf{x}(t)$, which yields that:

$$\lim_{n\to\infty}\mathbf{X}(n;t) = \lim_{n\to\infty}\underline{\mathbf{D}}^n\mathbf{X}(t) = \mu(\mathbf{X}(t)) = \frac{\mathbf{1}}{N}\sum_{i=1}^{N} X_i(t) \quad \text{a.s.} \tag{3.18}$$

An analysis of the fastest converging consensus rules is given in [13] if other consensus rules are to be considered.

As discussed in Section 3.2, in the absence of true labels, it is often difficult, if at all possible, to obtain an unbiased estimator for the competence of an expert. Therefore, we propose local and global extensions of pseudo competence (3.10), which can be locally and instantaneously estimated:

$$\tilde{p}_i^g = \mathbb{P}\left(X_i = f^{MV}\left(\mathbf{X}_{\setminus i}\right)\right), \tag{3.19}$$

$$\tilde{p}_i^\ell = \mathbb{P}\left(X_i = f^{MV}\left(\mathbf{X}_{\mathcal{N}_i}\right)\right). \tag{3.20}$$

Here, $\mathcal{N}_i = \{j \neq i : i \leftrightarrow j \in \mathcal{E}\}$ denotes the neighborhood of the node i excluding the node

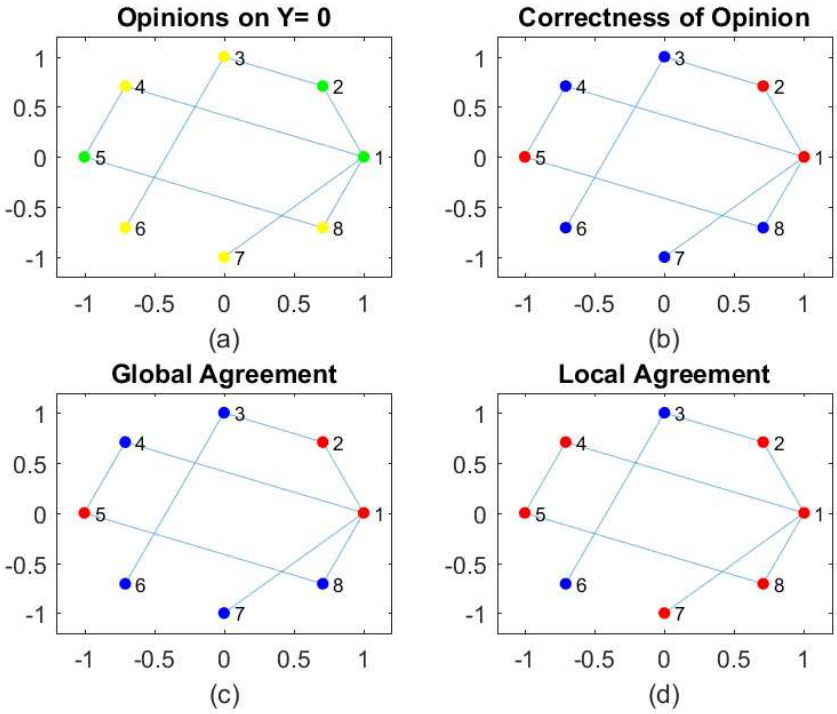


Figure 3.5: Correctness, Local, and Global Agreement: (a) Opinions Provided Through a Network, (b) Correctness of Each Opinion, Blue Nodes Are Correct, (c) Agreement to Network Consensus, Blue Nodes Agree to Global Decision, (d) Local Agreement as Defined in (3.21), Blue Nodes Agree with Local Decision

itself, we further allow $\mathbf{X}_{\mathcal{N}_i} = \{X_j : j \in \mathcal{N}_i\}$ and $\mathbf{X}_{\setminus i} = \{X_j : j \in \mathcal{V}, j \neq i\}$. Observe that $\tilde{p}_i^g$ is the frequency with which an expert agrees with the global majority (when its own vote is excluded), similarly, $\tilde{p}_i^\ell$ is that with local majority. We note that $\tilde{p}_i^g$ can be estimated via (3.13) and similarly, $\tilde{P}_i^\ell(t) = {}^1\!/\!_t \sum_{\tau=1}^t \mathbb{1}\left(X_i(\tau) = f^{MV}\left(\mathbf{X}_{\mathcal{N}_i}(\tau)\right)\right)$ can be used to estimate $\tilde{p}_i^\ell$. Here, the convergence in the number tasks follows from ergodicity of the agreement processes $\mathbb{1}\left(X_i(t) = f^{MV}\left(\mathbf{X}_{\setminus i}(t)\right)\right)$ and $\mathbb{1}\left(X_i(t) = f^{MV}\left(\mathbf{X}_{\mathcal{N}_i}(t)\right)\right)$. Figure 3.5 illustrates how the local and global agreement might instantaneously differ.

In addition to ease of local and instantaneous estimation, both $\tilde{p}_i^\ell$ and $\tilde{p}_i^g$ have properties that allow their use in a distributed setup, despite their bias: First, for finite graphs $\tilde{p}_i^\ell \leq \tilde{p}_i^g \leq p_i$, $\forall i \in \mathcal{V}$ as long as $p_i > {}^1\!/\!_2$, $\forall i \in \mathcal{V}$. In words, measuring the competence of an expert by its agreement frequency to some other group of experts yields *under*-estimation of its true competence. Naturally, if $\exists i \in \mathcal{V}$ such that $p_i = {}^1\!/\!_2$, then $p_i = \tilde{p}_i^\ell = \tilde{p}_i^g$. Furthermore, $\tilde{p}^g$ exhibits the ordering property of Proposal 3.1: $p_i > p_j \iff \tilde{p}_i^g > \tilde{p}_j^g$. Section 3.3.1 employs this property to order competences of experts in the sequential regime. Observe that given the agent-excluding majority vote, $f^{MV}\left(\mathbf{X}_{\setminus i}(t)\right)$ can be represented in terms of

the average opinion $\mu(\mathbf{X}(t))$ via $f^{MV}\left(\mathbf{X}_{\setminus i}(t)\right) = \mathbb{1}\left(N\mu(\mathbf{X}(t)) - X_i(t) > {}^{(N-1)}/_{2}\right)$. Therefore, similar to local estimates $\tilde{p}^{\ell}(t)$, global estimates $\tilde{p}^{g}(t)$ can be updated with each incoming task provided that the network can converge between tasks.

3.3.1 Distributed Ordering of Experts

Consider the following procedure on a heterogeneous network in the sequential regime:

$$\begin{aligned}
\mathbf{X}(n;t) &= \underline{\mathbf{D}}^{n}\mathbf{X}(t),\\
\lim_{n\to\infty}\mathbf{X}(n;t) &= \mathbf{1}\mu(\mathbf{X}(t)),\\
\mathbf{V}(t) &= \mathbb{1}\left(N\mathbf{1}\mu(\mathbf{X}(t)) - \mathbf{X}(t) > {}^{(N-1)}/_{2}\right),\\
\tilde{\mathbf{P}}^{g}(t) &= \frac{t-1}{t}\tilde{\mathbf{P}}^{g}(t-1) + \frac{1}{t}\mathbb{1}\left(\mathbf{X}(t) = \mathbf{V}(t)\right).
\end{aligned}$$

The first step is the consensus step and the proceeding voting and agreement checks are carried out nodally (index-wise). Observe that:

$$\lim_{t\to\infty}\tilde{\mathbf{P}}^{g}(t) = \tilde{\mathbf{p}}^{g} = \begin{bmatrix}\tilde{p}_1^{g}\\ \vdots \\ \tilde{p}_N^{g}\end{bmatrix}.$$

Therefore, for every network with nodal experts, as a committee, preserving ordering, that is: $\forall i, j \in \mathcal{V}$ such that $p_i > p_j$, the following holds:

$$\lim_{t\to\infty}\mathbb{P}\left(\tilde{P}_j^{g}(t) > \tilde{P}_i^{g}(t)\right) = 0.$$

Therefore, Proposition 3.1 concludes that ordering of experts can be attained via consensus for good committees and Proposition 3.2 concludes its extension for mixed committees self-excluding majorities satisfy $p_{\setminus i} > {}^{1}/_{2}$, $\forall i \in \mathcal{V}$. Next, distributed estimation of competences on homogeneous networks ($p_i = p$, $\forall i \in \mathcal{V}$) are addressed in the batch and sequential regimes.

3.3.2 Distributed Estimation via Local Updates

Conceptually, representing the true competence p by $\tilde{p}_i^{\ell} = \mathbb{P}\left(X_i = f^{MV}\left(\mathbf{X}_{\mathcal{N}_i}\right)\right)$ is equivalent to measuring the competence of each expert at a level defined by the competence of its neighbors. Formally,

$$\tilde{p}_i^{\ell} = p_{\mathcal{N}_i}p + q_{\mathcal{N}_i}q.$$

Here, $p_{\mathcal{N}_i} \triangleq \mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\mathcal{N}_i}\right) = Y\right)$ denotes the neighborhood competence and $q_{\mathcal{N}_i} = 1 - p_{\mathcal{N}_i}$. Note that $\tilde{p}_i^{\ell}$ always *under*-estimates the true competence with a bias:

$$p - \tilde{p}_i^{\ell} = (p-q)q_{\mathcal{N}_i} > 0, \forall i \in \mathcal{V}. \tag{3.21}$$

The inequality follows from $p > 1/2 > q$. As (3.21) indicates, local estimates vary in their reliabilities as functions of the true competence and their nodal degrees $|\mathcal{N}_i|$. Consider the batch setup updates:

$$\begin{aligned}
V_i(t) &= \mathbb{1}\left(|\mathbf{X}_{\mathcal{N}_i}(t)| > {}^{|\mathcal{N}_i|}/_2\right), \forall i \in \mathcal{V}, \\
\tilde{\mathbf{P}}^{\ell}(t) &= \frac{t-1}{t}\tilde{\mathbf{P}}^{\ell}(t-1) + \frac{1}{t}\mathbb{1}\left(\mathbf{X}(t) = \mathbf{V}(t)\right), t \in \{1, \cdots, |\mathbb{T}|\}, \\
\tilde{\mathbf{P}}^{\ell}(n; |\mathbb{T}|) &= \underline{\mathbf{D}}^n \tilde{\mathbf{P}}^{\ell}(|\mathbb{T}|), \\
\lim_{n\to\infty} \underline{\mathbf{D}}^n \tilde{\mathbf{P}}^{\ell}(|\mathbb{T}|) &= \mathbf{1}\mu(\tilde{\mathbf{P}}^{\ell}(|\mathbb{T}|)).
\end{aligned}$$

The variance of estimator the $\tilde{\mathbf{P}}^{\ell}$ depends on $|\mathbb{T}|$, its bias, on the other hand, depends on the local connectivity. The difference between the true competence and the network average of local expected agreements $\mu(\tilde{\mathbf{p}}^{\ell}) = \mu\left(\lim_{t\to\infty} \tilde{\mathbf{P}}^{\ell}(t)\right)$, or the network estimation bias, is given as follows:

$$p - \mu(\tilde{\mathbf{p}}^{\ell}) = (p-q)\frac{1}{N}\sum_{i=1}^{N} q_{\mathcal{N}_i} = (p-q)\mu(\mathbf{q}_{\mathcal{N}}). \tag{3.22}$$

Here $(\mathbf{q}_{\mathcal{N}})_i = q_{\mathcal{N}_i}$. Since $p > 1/2$, Cramér's theorem [60] yields that $q_{\mathcal{N}_i} \leq \exp\left(-|\mathcal{N}_i| \, l\left(1/2\right)\right) = (4pq)^{\frac{|\mathcal{N}_i|}{2}}$, where $l(a) \triangleq a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}$, $\forall a \in (0,1)$. Therefore, we can write that:

$$\frac{p - \mu(\tilde{\mathbf{p}}^{\ell})}{p-q} \leq \frac{1}{N}\sum_{i=1}^{N}(4pq)^{|\mathcal{N}_i|/2} = \frac{1}{N}\sum_{j=1}^{N}\alpha_j(4pq)^{j/2}.$$

Here, $\alpha_j = |\{i \in \mathcal{V} : |\mathcal{N}_i| = j\}|$. The Cauchy-Schwarz inequality and bounding the geometric terms yield that:

$$p - \mu(\tilde{\mathbf{p}}^{\ell}) \leq \frac{p-q}{\sqrt{1-4pq}}\left(\frac{1}{N}\sqrt{\sum_{j=1}^{N}\alpha_j^2}\right). \tag{3.23}$$

The second term of (3.23) quantifies the impact of network connectivity on the network estimation bias $p - \mu(\tilde{\mathbf{p}})$. Naturally, on a fully connected network, the bias is minimized. Alternatively, in the low-frequency regime the minimum bias is attainable.

3.3.3 Distributed Estimation via Global Updates

On a network of experts with identical competences, $p_i = p$, note that:

$$\tilde{p}_i^g = p_{\setminus i} p + q_{\setminus i} q,$$

where $p_{\setminus i} = \mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus i}\right) = Y\right)$ is the self-excluding majority and since experts have identical competences $\tilde{p}_i^g = \tilde{p}_j^g = \tilde{p}^g$, $\forall i, j \in \mathcal{V}$ for some $\tilde{p}^g$. Therefore,

$$p - \mu(\tilde{\mathbf{p}}^g) = p - \tilde{p}^g = (p - q) q_{\setminus i}. \tag{3.24}$$

Observe that $p_{\setminus i}$ is a monotonically increasing function of the number of available expert N, as long as $p > 1/2$. Hence, $p_{\setminus i} > p_{\mathcal{N}_i}$, $\forall i \in \mathcal{V}$. Consequently,

$$\tilde{p}_i^\ell < \tilde{p}^g < p, \forall i \in \mathcal{V}.$$

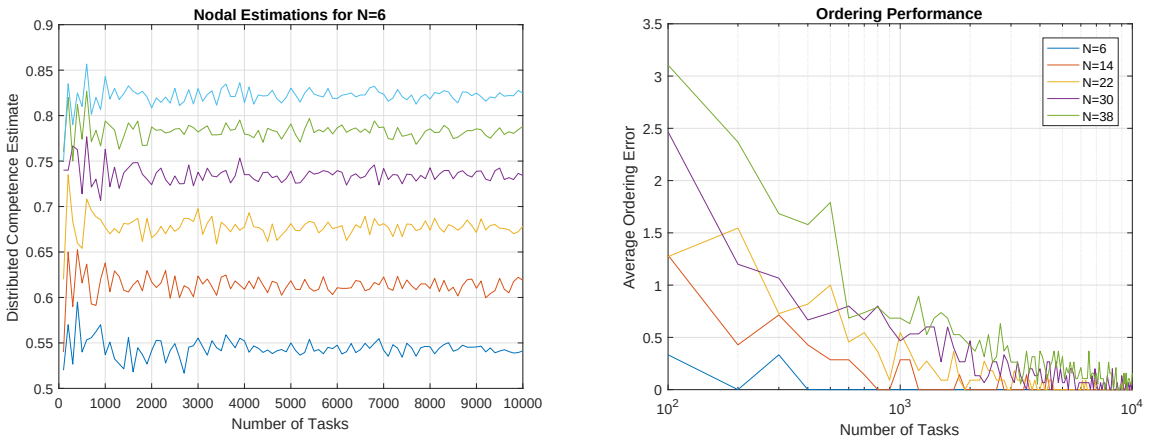
When the estimation rule (C.2) in Appendix C.2 is employed on an homogeneous network, the competence of each expert is estimated within an identical bias that effectively depends on the size of the network; when employed on a heterogeneous network, competences are estimated within varying biases, yet the ordering of competences is preserved asymptotically.

3.3.4 Experiments

We have investigated the impacts of the number of tasks $|\mathbb{T}|$, true competence profile $\mathbf{p}$, number of experts N, and local connectivity, defined by the network topology $\mathcal{G}$ on the performance of distributed competence estimation.

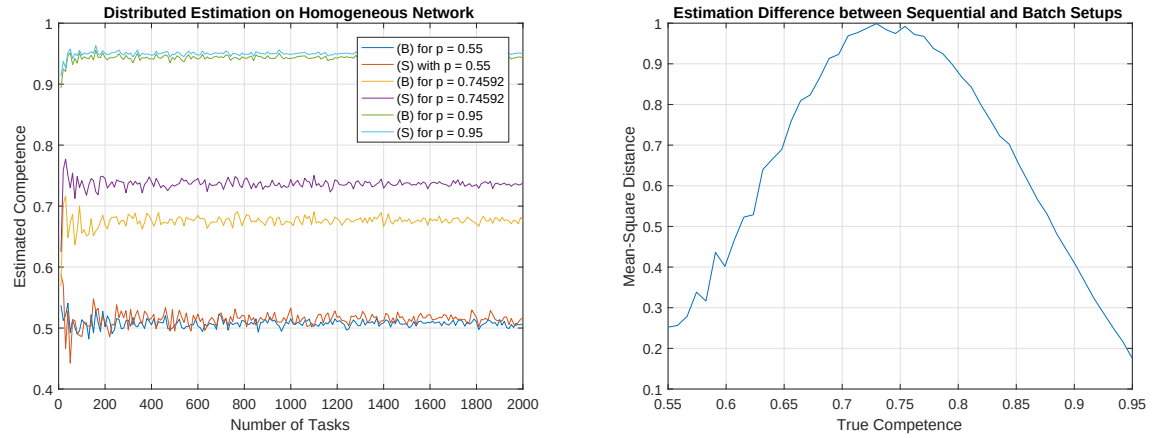
Figure 3.6 demonstrates the ordering strategy given in (C.2). We allowed $\mathbf{p} \in [0.55, 0.95]$ to be equally spaced with $p_{i+1} - p_i = 0.4/N$. As seen in Figure 3.6a, competence estimators $\tilde{p}^g$ converge to distinct estimates when true competences are sufficiently separated. We observed the impact of closely chosen true competences by increasing the number of experts on the fixed interval: Figure 3.6b illustrates an average ordering error, which is the average absolute distance between true and network-estimated ordering indices, as functions of N and $|\mathbb{T}|$.

Figure 3.7 illustrates distributed estimation on homogeneous networks of fixed topology. We demonstrated the difference $\tilde{p}^g - \tilde{p}^\ell$ as a function of true competence p: Figure 3.7a illustrates a set of sample paths, where the network that employs nodes of competence $p \approx 0.75$ exhibits the largest deviation between rules (C.2) and (C.3) setups. Figure 3.7b illustrates a behavior that is not obvious from (3.21)-(3.24), albeit intuitive: As the connectivity of the



(a) (b)

Figure 3.6: Distributed Ordering of Experts on Heterogeneous Networks: (a) An Illustration of Sample Paths vs. Number of Tasks (b) Average Ordering Error vs. Number of Tasks for Different Numbers of Nodes



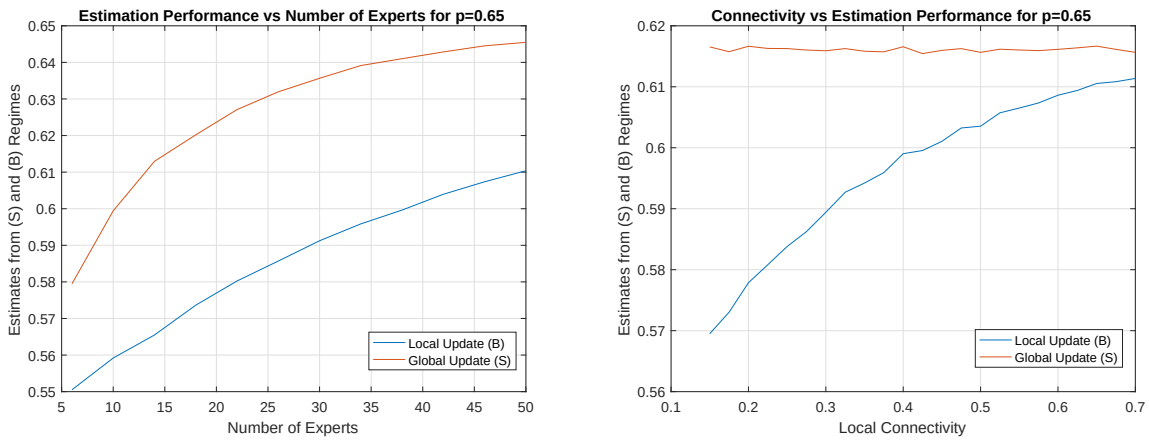
(a) (b)

Figure 3.7: Distributed Competence Estimation on Homogeneous Networks: (a) Sample Paths over Number of Tasks (b) Mean-Square Distance Between $\tilde{p}^{\ell}$ and $\tilde{p}^{g}$ vs. True Competence p

graph increases, the bias between local and global estimates diminish.

We further investigated the impact of connectivity and number of nodes on the estimation performance. Figure 3.8a verifies that the bias of $\tilde{p}^{\ell}$ is mostly constrained by local connectivity as the bias of $\tilde{p}^{g}$ diminishes faster than that of $\tilde{p}^{\ell}$ with the number of experts. Figure 3.8b further verifies this notion; $\tilde{p}^{g}$ remains unaffected by the local connectivity while the bias of $\tilde{p}^{\ell}$ monotonically decreases.

Section 3.4 investigates the minimum probability of error achieving opinion-aggregation



(a) (b)

Figure 3.8: Impact of The Number of Experts and Local Connectivity on Estimation Performance (a) Increasing Number of Experts on Networks of Fixed Average Connectivity (b) Changing Local Connectivity on Networks of Fixed Number of Nodes

strategy, then Section 3.5 substitutes the true probability law with in this strategy with the pseudo probability law.

3.4 Naïve Bayes Decision Rule and Its Performance Guarantees

Consider the family $\mathcal{D}$ of possibly randomized decision rules that aggregate N opinions $\mathbf{X}(t)$ to decide on a state $Y(t)$. The minimum probability of error (MPE)-achieving strategy is an instantaneous opinion-aggregation rule:

$$f^{MPE} = \arg\min_{f \in \mathcal{D}} \mathbb{P}\left(f\left(\mathbf{X}\right) \neq Y\right).$$

Note that the competences $\{p_1, \cdots, p_N\}$ being fixed across all tasks $t \in \mathbb{T}$ allows task-dependency of opinions $\mathbf{X}(t)$ and states $Y(t)$ to be dropped. It is well known that the maximum a posteriori rule achieves the minimum probability of error [4]. Therefore, the MPE-achieving supervised opinion-aggregation strategy is given by:

$$f^{MPE}\left(\mathbf{X}\right) = \arg\max_{y \in \{\pm 1\}} \mathbb{P}\left(y \mid \mathbf{X}\right),$$

with the corresponding likelihood-ratio test:

$$f^{MPE}(\mathbf{X}) = \operatorname{sign}\left(\sum_{i=1}^{N} X_i \log \frac{p_i}{q_i}\right) \equiv f^{NB}(\mathbf{X}). \tag{3.25}$$

The decision rule in (3.25) is often referred to as the *naïve Bayes decision rule* (NB) [17]. The existence of the probability law (3.1)-(3.3) ensure that the NB decision rule is instantaneous and that it has no *bias term* in the sign function.

As a direct consequence of being the maximum a posteriori rule, the probability of error of the naïve Bayes decision rule can be written explicitly as:

$$\mathbb{P}\left(f^{NB}(\mathbf{X}) \neq Y\right) = \frac{1}{2} \sum_{\boldsymbol{x} \in \{\pm 1\}^N} \min_{y \in \{\pm 1\}} \mathbb{P}\left(\mathbf{X} = \boldsymbol{x} \mid Y = y\right).$$

Even though $\min_{y\in\{\pm1\}} \mathbb{P}\left(\mathbf{X} = \boldsymbol{x} \mid Y = y\right)$ can be readily computed $\forall \boldsymbol{x}$ when $\{p_1, \cdots, p_N\}$ is known, the sum is still intractable for large N, which motivates research for lower and upper bounds. It has been shown that a set of lower and upper bounds can be found in the form [17, Theorem 1]:

$$-\log \mathbb{P}\left(f^{NB}(\mathbf{X}) \neq Y\right) \asymp \Phi,$$

where $\asymp$ denotes upper and lower bounds within a constant factor. As a function of the true competences $\{p_1, \cdots, p_N\}$, Φ is called the *committee potential*, [17, 61, 62] and it is given by:

$$\Phi\left(p_1, \cdots, p_N\right) = \sum_{i=1}^{N} \left(p_i - \frac{1}{2}\right) \log \frac{p_i}{q_i}. \tag{3.26}$$

The upper-bound given in [17, Theorem 1(i)] makes a use of the Chernoff bounding technique, see, for instance, [63, Section 2.2.1], and the Kearns-Saul inequality [62, Lemma 1]. A detailed discussion of Kearns-Saul and Berend-Kontorovich concentration inequalities for mixtures of independent, bounded random variables is given in [63, Section 2.2.4].

Interestingly, in the case of the naïve Bayes decision rule, the subsequent use of Kearns-Saul inequality appears unnecessary. As shown in Appendix B.3, a direct consequence of the Chernoff bounding technique is as follows.

Theorem 3.1. *Let $Y \in \{\pm 1\}$ be uniformly distributed, experts with competences $\{p_1, \cdots, p_N\}$, where $p_i = \mathbb{P}\left(X_i = Y\right)$, generate opinions independently: $X_i - Y - X_j$, $\forall i \neq j \in [N]$. Con-*

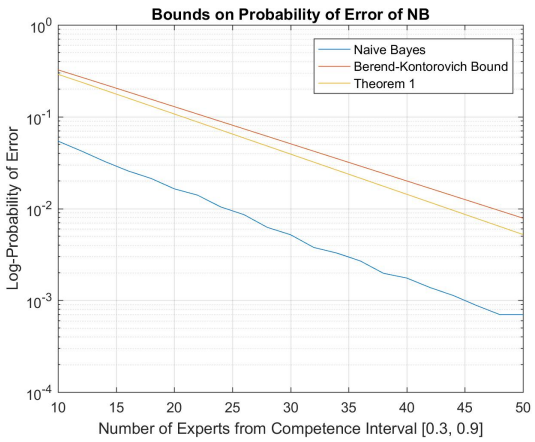


Figure 3.9: A Comparison of Upper-Bounds for Naïve Bayes Probability of Error

sider the naïve Bayes decision rule:

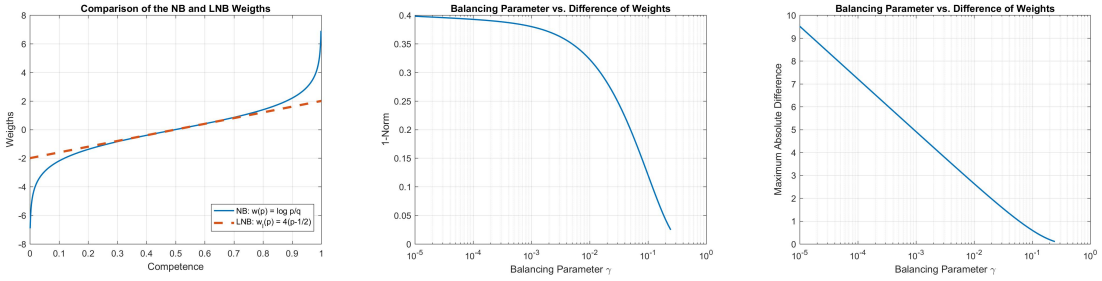
$$f^{NB}(\mathbf{X}) = \operatorname{sign}\left(\sum_{i=1}^{N} X_i \log \frac{p_i}{1-p_i}\right),$$

where $\log \frac{1}{0} = \infty$ *by convention. Then, the following tight upper-bound on the probability of error holds:*

$$\mathbb{P}\left(f^{NB}(\mathbf{X}) \neq Y\right) \leq \prod_{i=1}^{N} \sqrt{4p_i q_i}.$$

Theorem 3.1 is the sharpest bound attainable by the use of Chernoff bounding technique and it is tight, as evidenced when $\exists i : p_i = 1$, in addition to the regimes discussed in [17]. Figure 3.9 provides a comparison of upper bounds between that given in Theorem 3.1 and that in [17, Theorem 1 (i)] for $N \in [10, 50]$ experts with equally spaced competences chosen from the interval $[0.3, 0.9]$.

Observe that $w(p) = \log p/q \in (-\infty, \infty)$ is an unbounded function that monotonically increases over $p \in [0, 1]$. On one hand, when competences are known reliably, $w(p)$ ensures that NB relies almost exclusively on highly competent experts with $p_i \approx \{0, 1\}$ (note that when competences are known, vote of an expert with $p < 1/2$ is necessarily flipped). On the other hand, in empirical setups, where competences are to be estimated on set of tasks, direct use of $w(\hat{p})$ could jeopardize the robustness of opinion-aggregation rule: An expert with arbitrary competence $p_i < 1$ could be assigned a competence estimate $\hat{p}_i \approx 1$ with non-negligible probability and thus, an unbounded weight. Although such events become increasingly improbable as the number of labeled tasks increases, for robustness, the *linearized*



(a) (b) (c)

Figure 3.10: Comparison of the NB and the LNB Weighs and the Impact of the Balancing Parameter

naïve Bayes (LNB) decision rule could be considered:

$$f^{LNB}(\mathbf{X}) = \text{sign}\left(\sum_{i=1}^{N} X_i \left(p_i - 1/2\right)\right). \tag{3.27}$$

Note that the term "linearized" in LNB refers to the linear estimation of the weights as the first-order Taylor series expansion of $w(p)$ around $p = 1/2$ is given by $4\left(p - 1/2\right)$ (the positive constant factor is dropped since it does not contribute to decision making).

Alternatively, one might focus on a practical subset of committees, one that we call *absolutely balanced* committees, that excludes almost-all-knowing ($p \approx 1$) and almost-ever-lying ($p \approx 0$) experts.

Definition 3.2 (Absolute Balance)**.** *A committee of experts with competences* $\{p_i : i \in \mathbb{N}\}$ *is called absolutely balanced if* $\exists \gamma \in (0, 1/2)$ *such that* $p_i \in [\gamma, 1-\gamma]$, $\forall i$.

Figure 3.10 illustrates the difference between weights and the impact of the balancing parameter γ on ℓ_1-norm as well as the absolute maximum of the difference between weight vectors (when p_i are equally spaced in $[\gamma, 1-\gamma]$). For an absolutely balanced committee, not only do the true competence estimates (3.9), yield robust empirical implementation of the NB rule through direct substitution into (3.25) but they also enable the empirical use of the LNB rule by substitution into (3.27) that exhibits similar performance to the NB rule. In fact, for absolutely balanced committees with modest balancing parameters, the performance difference between the NB rule and the LNB rule appears insignificant. Section 3.7 illustrates this phenomenon empirically.

Section 3.2 shows that a set of opinions can be used to estimate true competences $\{p_1, \cdots, p_N\}$ in the form of pseudo competences $\{\tilde{p}_1, \cdots, \tilde{p}_N\}$. Section 3.5 indeed shows that for an absolutely balanced committee, one could achieve an unsupervised opinion-aggregation performance that scales with the committee potential, therefore, with the performance of

the MPE-achieving NB rule, while suffering from error due to bias between pseudo and true competences that *diminishes* in the number of experts consulted.

3.5 The Pseudo Naïve Bayes Decision Rule

In the unsupervised problem setup, true competence estimates $\{\hat{p}_1, \cdots, \hat{p}_N\}$ are not available, where the pseudo competence estimates $\{\hat{\rho}_1(T), \cdots, \hat{\rho}_N(T)\}$ can be over (unlabeled) T tasks. An opinion-aggregation rule that use pseudo competence estimates not only suffers performance degradation due to empirical estimation error but also suffers degradation due to the bias $|p_i - \tilde{p}_i|$ as well. The goal of this section is to first propose fundamental limits for a decision rule that has access to pseudo competences $\{\tilde{p}_1, \cdots, \tilde{p}_N\}$ *directly*.

Let us call the following opinion aggregation rule the *pseudo naïve Bayes* (PNB) rule:

$$f^{PNB}(\mathbf{X}) = \operatorname{sign}\left(\sum_{i=1}^{N} X_i \log \frac{\tilde{p}_i}{\tilde{q}_i}\right). \tag{3.28}$$

The PNB rule corresponds to *assuming* that the underlying probability law is characterized by the pseudo competences $\{\tilde{p}_1, \cdots, \tilde{p}_N\}$, where it is actually characterized by the true competences $\{p_1, \cdots, p_N\}$. The PNB rule, similar to the NB rule, is an instantaneous decision rule. Unlike the NB rule, the PNB rule is empirically achievable without supervision.

Formally, we show that the performance of PNB rule scales with the underlying true committee potential $\Phi(p_1, \cdots, p_N)$ as defined in (3.26), and the performance degradation due to the use of pseudo competences is quantified as follows:

$$-\log \mathbb{P}\left(f^{PNB}(\mathbf{X}) \neq Y\right) \asymp (1 - \delta(a_N, \gamma))\Phi.$$

Here, $\delta(\cdot)$ represents the performance degradation due to lack of supervision and is a bounded function of the rate of consistency a_N from (3.14) and the balancing parameter γ from Definition 3.2. The variable $\delta(\cdot)$ diminishes both in a_N and γ due to the difference between pseudo competences and true competences diminishing in a_N and γ limiting the maximum difference between pseudo competences and true competences.

Let us note that the performance of the PNB decision rule exhibits a similar scaling to that of NB, as long as exclusion of any expert leaves committee reliable ($p_{\setminus i} > 1/2$, $\forall i$):

$$-\log \mathbb{P}\left(f^{PNB}(\mathbf{X}) \neq Y\right) \asymp \tilde{\Phi}. \tag{3.29}$$

Here, we call the term $\tilde{\Phi}$ the pseudo committee potential:

$$\tilde{\Phi} = \sum_{i=1}^{N} \left(p_i - \frac{1}{2} \right) \log \frac{\tilde{p}_i}{1 - \tilde{p}_i}.$$

The relation in (3.29) follows algebraically from the proof of [17, Theorem 1] by the use of Property 3.2, as provided in the Appendix B.4.

The pseudo committee potential $\tilde{\Phi}$ and true committee potential Φ converge at a rate determined by the rate at which the underlying committee becomes consistent in the majority vote. Theorem 3.2 quantifies the rate at which the performance of the PNB decision rule scales with that of NB.

Theorem 3.2. *Every absolutely balanced, consistent committee with rate a_N and balancing parameter γ satisfies:*

$$\frac{\tilde{\Phi}}{\Phi} \geq 1 - C(1/2 - \gamma)\rho_N \longrightarrow_N 1.$$

Here, $\rho_N = {}^{(1-a_N)}/_{(a_N - 1/2)}$ and $C(x)$ is a positive function supported on $x \in [-1/2, 1/2]$.

Theorem 3.2 indicates that the PNB rule is not only asymptotically optimal, but it approaches to the performance of the optimal (supervised) decision rule that becomes consistent at a rate that is faster than that of majority vote, a_N. A corollary of this result can be formulated as follows.

Corollary 3.1. *A sufficient condition for any absolutely balanced committee of size N to ensure that $1 - \delta \leq \frac{\tilde{\Phi}}{\Phi}$ is as follows:*

$$\frac{1 - p_{\setminus i}}{p_{\setminus i} - 1/2} \leq \frac{\delta}{C(1/2 - \gamma)}, \forall i \in [N].$$

Decision rules that employ some relevant statistics in a functional form that is known to perform well are often called "plug-in" decision rules. Plug-in rules are often difficult to analyze and are often suboptimal [64, Chapter 1]. Interestingly, despite being a plug-in rule, the PNB rule achieves minimum probability of error asymptotically, thanks to the construction of the pseudo competences.

The PNB rule, as defined in (3.28), is instantaneous because it uses pseudo competences directly, which, in practice, can be estimated from opinions empirically. The manner in which pseudo competence estimates are updated gives rise to the operational meaning of the corresponding decision aggregation rule, as addressed in Section 3.6.

3.6 Empirical Rules That Use Pseudo-Competences

Empirical unsupervised decision aggregation rules that use pseudo competences to process a block of tasks are of the form:

$$f^B(\mathbf{X}_1^T) = \text{sign}\left(\sum_{i=1}^{N} \boldsymbol{X}_i w(\tilde{p}_i(T))\right). \tag{3.30}$$

Here, T is the number of available tasks, $w(\cdot)$ is a weight function operating on empirical competences estimated $\tilde{p}_i(t)$, as defined in (3.13). Note that the adaptive extension of (3.30) is as follows:

$$f_t^A(\mathbf{X}(t)) = \text{sign}\left(\sum_{i=1}^{N} X_i(t) w(\tilde{p}_i(t))\right). \tag{3.31}$$

Asymptotically, both of the empirical rules should achieve optimality at a rate close to that of NB rule, as Theorem 3.2 indicates. The objective of Section 3.6 is to quantify the impact of empirical estimation error.

3.6.1 Unsupervised Block Decision Aggregation Rules

Pseudo competences can be estimated dynamically in real-time or over a block of opinions. When a block of opinions is to be processed, the performance of the corresponding empirical PNB decision rule is determined by two factors:

1. *Pseudo Competence Bias:* $\|\tilde{\boldsymbol{p}} - \boldsymbol{p}\|_1 = \sum_{i=1}^{N} |\tilde{p}_i - p_i|$.

2. *Empirical Estimation Error:* $\|\hat{\boldsymbol{\rho}}(T) - \tilde{\boldsymbol{p}}\|_1 = \sum_{i=1}^{N} |\hat{\rho}_i(T) - \tilde{p}_i|$.

Note that pseudo competence bias is the cost of operating in an unsupervised setup and it is a hidden function of $\boldsymbol{p}$ that is fixed for a given committee. Empirical estimation error, on the other hand, introduces a nonlinear distortion that propagates through the weights of the decision rule.

When each expert can only be consulted for a small number T of tasks, rather coarse, high-variance, estimates $\{\hat{\rho}_1(T), \cdots, \hat{\rho}_N(T)\}$ are achievable, and hence the corresponding weights might be arbitrarily larger than what they are supposed to be. In order to rectify this non-robust behavior, a *linearized* version of the weights that follows from the first-order

Taylor series expansion of $\log {}^{x}/_{1-x}$ around $x = 1/2$, can be used:

$$w(x) = \log \frac{x}{1-x} = 4\left(x - \frac{1}{2}\right) + \frac{16}{3}\left(x - \frac{1}{2}\right)^3 + \cdots.$$

The corresponding opinion aggregation rule for small number of tasks is as follows:

$$f^{LPNB}(\mathbf{X}) = \text{sign}\left(\sum_{i=1}^{N} X_i(\tilde{p}_i(T) - 1/2)\right). \tag{3.32}$$

It is clear that (3.32) should not suffer from estimation error as much as the empirical PNB rule with $w(\tilde{p}) = \log {}^{\tilde{p}}/_{\tilde{q}}$, as given in (3.30). The main challenge for such a rule is to achieve consistency, as addressed next.

Proposition 3.3. *If a committee satisfies:*

1. *(Naïve Bayes)* $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (p_i - 1/2)^2 = \infty$,
2. *(Majority)* $\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (p_i - 1/2) \geq \sqrt{\frac{\log 2}{8}}$,

then $f^L(\cdot)$ *is consistent:* $\lim\limits_{N\to\infty} \mathbb{P}\left(f^L(\mathbf{X}) \neq Y\right) = 0$.

Proposition 3.3 shows that when arbitrarily large number of experts are consulted for a small number of tasks each, the unsupervised empirical PNB rule becomes reliable. The first condition of Proposition 3.3 is a direct consequence of the Hoeffding's inequality and it is sufficient for consistency when each expert is *tested* with a small number of *labeled* tasks and competences are estimated via (3.9) [17, Theorem 7]. Interestingly, pseudo competences can facilitate this consistency without the need for labeled data but at the expense of the second condition, which is not restrictive: As discussed in Appendix B.5, it amounts to $\lim_{N\to\infty} \mathbb{P}\left(f^{MV}(\boldsymbol{X}) = Y\right) > 1/2$. Similar to the previous discussion on the performance similarity between the NB and the LNB rules, Section 3.7 empirically demonstrates that pseudo naïve Bayes and its linearized version perform similarly.

When there are sufficiently many tasks to be processed, the error due to empirically estimating $\tilde{\boldsymbol{p}}$ diminishes and pseudo competence bias becomes the dominant factor of performance degradation. Thus, the PNB rule given in (3.28) naturally extends to a decision rule (3.30) that empirically estimates the pseudo competence of each expert over T tasks and applies to $w(\tilde{p}_i) = \log {}^{\tilde{p}_i}/_{\tilde{q}_i}$. For an arbitrary committee $\boldsymbol{p} \in [0,1]^N$, the difference between pseudo competences and true competences become unbounded for experts with $p_i \approx \{0, 1\}$. However, for absolutely balanced committees, pseudo competence bias is necessarily bounded. Lemma 3.1 quantifies the committees that limit the difference between PNB and NB weights.

Lemma 3.1. *Let* $R(\gamma) = \frac{2\gamma(1-\gamma)}{1-2\gamma}$*, if an absolutely balanced committee satisfies for some* $\epsilon > 0$

$$\min_{i\in[N]} p_{\setminus i} \geq \frac{1}{2} + \frac{1}{2 + \epsilon R(\gamma)},$$

then $\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|_1 \leq \frac{\epsilon N}{2}$.

Observe $R(\gamma)$ increases in γ, equivalently, committees that concentrate toward the center of the cube $[0,1]^N$ yield closer weights. Conceptually, this amounts to discussion on weak classifiers; it is often easier to boost weak classifiers to form a stronger one [7]. The next theorem jointly addresses the empirical estimation error and pseudo competence bias:

Theorem 3.3. *Let a committee be consistent with rate* a_N*,* $\forall \delta \in (0,1)$ *define* $C(\delta; N, T) \triangleq \frac{12}{T} \log \frac{8N}{\delta}$*. Then,*

$$\forall \epsilon \in \left(\left(\rho_N C(\delta; N, T)\right)^{1/3}, \min\left\{5, \frac{2\Phi}{N}\right\} \right),$$

and for all absolutely balanced committees with parameter $\gamma > C(\delta; N, T) \left(\frac{2}{\sqrt{4\epsilon+1}-1}\right)^2$*:*

$$\mathbb{P}\left(f^B(\mathbf{X}) \neq \mathbf{Y}\right) \leq \delta + \exp\left[-\frac{\left(2\Phi - \epsilon N\right)^2}{8\Phi}\right].$$

Property 3.2 along with Lemma 3.1 allows $f^B(\cdot)$ to scale similar to an empirical NB as long as the underlying worker committee is sufficiently strong, which is captured by δ. Theorem 3.3 borrows its empirical analysis from that of [17, Theorem 11], which quantifies the performance of empirical NB under sufficiently long training. Albeit insightful, Proposition 3.3 and Theorem 3.3 analyze the performance of empirical PNB decisions rules for a *block* of opinions. An adaptive and instantaneous extension is addressed next.

3.6.2 An Unsupervised Adaptive Decision Aggregation Rule

Let $f^A_\tau(\cdot)$ be an empirical pseudo naïve Bayes decision rule:

$$f^A_\tau(\mathbf{X}(\tau)) = \operatorname{sign}\left(\sum_{i=1}^{N} X_i(\tau) \log \frac{\hat{\rho}_i(\tau)}{1 - \hat{\rho}_i(\tau)}\right).$$

We call the probability that the decision rule $f^A_\tau(\cdot)$ makes the correct decision based on $X(t)$, that is, $\mathbb{P}\left(f^A_\tau(\mathbf{X}(t)) = Y(t)\right)$, for some $t > \tau$, the *confidence* of the adaptive decision rule. Theorem 3.4 characterizes this notion of confidence.

Theorem 3.4. *Let* $\delta \geq \sum_{i=1}^{N} |p_i - \tilde{p}_i| + \frac{N}{\sqrt{\tau}}$ *and define the event* $R(\tau)$ *as follows:*

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left(\tilde{p}_i(\tau)-\frac{1}{2}\right)\log\frac{\tilde{p}_i(\tau)}{1-\tilde{p}_i(\tau)}\right) \leq \frac{\delta}{2}.$$

Then $\forall t > \tau$, $\mathbb{P}\left(R(\tau) \cap \left\{f_\tau^A\left(\mathbf{X}(t)\right) \neq Y\right\}\right) \leq \delta$.

The term $\sum_{i=1}^{N} |p_i - \tilde{p}_i| \leq \sum_{i=1}^{N}(1-p_{\setminus i})$, which diminishes with the committee potential. The proof is given in Appendix B.7 and it borrows the analysis of [17, Theorem 13] on the adaptive empirical naïve Bayes decision rule, which is based on the committee potential being empirically estimated from some labeled data to control the worst case performance on the test data. Theorem 3.4 extends this analysis to use empirical pseudo competences instead, resulting in a real-time algorithm where the player builds an empirical pseudo committee potential and makes decisions with dynamic confidence.

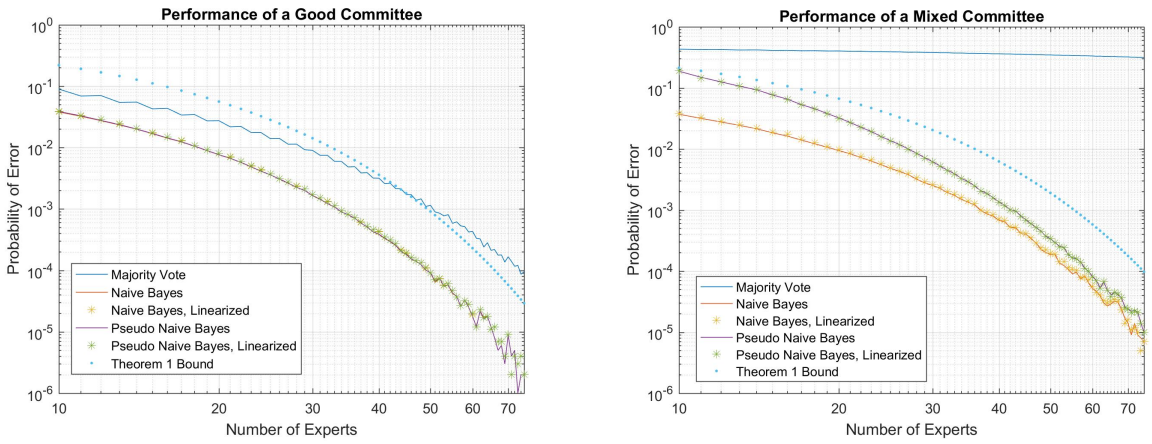
The adaptive decision aggregation rule $f_\tau^A(\cdot)$ is a sequential decision making mechanism: at any given time $t \in [T]$ the algorithm makes a decision with confidence δ if $R(\tau)$ has happened for some $\tau < t$ otherwise, it assigns the majority vote. This allows the algorithm to a make decisions with dynamic confidence; once a confidence level is achieved at $t = \tau$, there is no need to keep updating the weights as the decision rule is finalized and algorithm uses that fixed decision rule on the incoming data.

3.7 Experiments

First, we examine the performance of pseudo naïve Bayes decision rule in different operational regimes including, mixed/good committees of varying size and linearized/true decision rules. Then, we compare the performance of PNB decision rule to spectral meta learner (SML) [30], expectation maximization (EM) [19, 25], and belief propagation (BP) [26].

3.7.1 Comparison between the PNB and NB Rules

In the first set of experiments, pseudo-random tasks and experts that satisfy (3.1)-(3.3) are generated in MATLAB. In order to compare the performance of naïve Bayes decision rule to its linearized and unsupervised counterparts, we consider a committee of experts with sizes varying from $N = 10$ to $N = 75$ with competences equally spaced in the intervals $[0.15, 0.9]$ for the mixed-committee case and from $[0.5, 0.9]$ for the good-committee case. Figure 3.11 compares the performance of NB to that of PNB, as defined in (3.28) and converged over



(a) A Good Committee (b) A Mixed Committee

Figure 3.11: A Performance Comparison between Naïve Bayes Rules and Their Unsupervised Counterparts

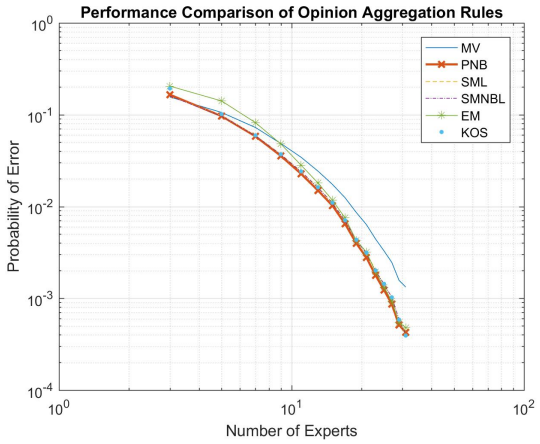


Figure 3.12: A Performance Comparison between the PNB Rule and Other Opinion-Aggregation Rules

$T = 1e + 6$ tasks. The objective of such large of tasks is to observe the fundamental operational tendencies: Figure 3.11a indicates that in the case good committees, where the probability of error of majority vote starts around 0.1, the performance difference between PNB and NB are negligible, which is to be expected as $p_i > 1/2$ for every expert. In the case of a mixed committee with majority vote error varying in the interval $(0.4, 0.37)$ for every committee size, the performance difference with PNB and NB rules is evident. However, linearization does not seem to introduce a significant performance degradation, as expected. Furthermore, one should note that performance improvement in the mixed case is due to majority voting performance becoming more robust to individual perturbations, formally, it tends to $p_{\setminus i} \approx \mathbb{P}\left(f^{MV}\left(\mathbf{X}\right) = Y\right)$ for all experts.

3.7.2 Comparison with Other Opinion-Aggregation Rules

We consider 1000 iterations of the following setup: 150 tasks are evaluated by experts of pseudo-random competences from the range $[0.5, 1]$. The number of experts range between $[3, 31]$. PNB rule is compared against the SML, EM, and BP algorithms. Figure 3.12 illustrates that PNB rule shows similar performance with other, more computationally expensive, opinion-aggregation rules.

CHAPTER 4

FAULT-TOLERANT COMPUTATION

Often what is called soft information (non-binary valued opinions) is available from computational sources that are of high performance, yet prone to difficult-to-model failure statistics, such as those due to component degradation, process/temperature and voltage variations, and similar power-reliability trade-offs. Computational units prone to random failures with unknown statistics capture a wide array of such models. For fault-detection or fault-mitigation circuitry, unknown operation and failure statistics creates the challenge of testing hypotheses (fault/no fault) with unknown statistics to detect and bypass faulty readings. A system-level idea, one that is often called "algorithmic noise tolerance" (ANT), is a robust framework to test such hypotheses through the use of low-fidelity, robust estimation units that safeguard the potentially faulty main unit [33]. Section 4.1 extends the use of ANT to detect and bypass failures for arbitrary main, failure, and calibration statistics based on a contextual fidelity-ordering.

The objective of ANT-like statistical error compensation techniques is to jointly detect and bypass failures of error-prone units. As such, such techniques result in the use of an *additional* low-resolution calibration unit. Conceptually, each source having its own safeguarding mechanism, such as ANT, is equivalent to each expert from Chapter 3 rethinking its opinions based on a compass, much like Captain Jack Sparrow's, that tilts slightly toward to the true state. On the other hand, a distributed system provides a reliable framework to locally detect faults based on soft information in a manner similar to local pseudo competence estimation. Section 4.2 employs this notion to propose a network averaging technique, through discarding faulty sources locally.

4.1 Fidelity-Based Testing of Hidden Hypotheses

Chapters 2-3 provided the interpretation of experts as sources of binary opinions. However, power-to-area-constrained computational units often produce non-binary readings while suffering random failures that are only partially captured by additive noise models. Furthermore, such failures are often functions of operational dynamics and hence, difficult to model

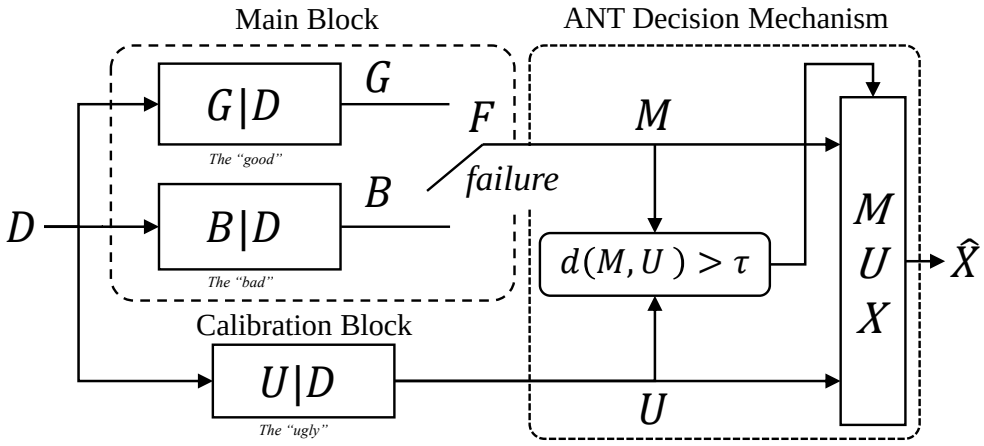


Figure 4.1: Generalized Model for ANT Architecture

reliably. Unless mitigated, such errors diffuse through the network reducing the overall performance, as addressed in Section 4.2. Therefore, it is of interest to develop safeguarding mechanisms of such non-binary experts that do not only detect failures but also *bypass* the corresponding faulty outcomes.The system-level idea called algorithmic noise tolerance employs a low-resolution yet robust calibration unit to detect and bypass randomly occurring failures and has been shown to be analogous to Gaussian CEO problem for certain noise profiles [37, 38]. There has been a set of heuristic rules established in practice that can be explained with the fidelity-based framework as illustrated in Figure 4.1.

4.1.1 Problem Definition

Let data $D(X)$ depend on a random variable X, and let the purpose of the computation be to generate $\hat{X} = f(D)$ with a *loss function* $\ell : \mathbb{R}^2 \to \mathbb{R}^+$ measuring the performance. Here, we propose a mixture model that incorporates fidelity-ordered main, calibration and failure statistics to generalize the additive noise model of [34, 37, 38].

The ANT architecture is modeled as consisting of *three* computational units, henceforth called blocks: a main block that comprises a *good* block producing an outcome G and a bad one that produces B, which models intermittent failure of the main computational path, and a calibration, or so-called ugly, block producing U, modeling the lower-fidelity alternative, should it be decided that the main block has failed, as seen in Figure 4.1. Intuitively, the main block produces $M = G$, when there is no hardware failure, and produces $M = B$ when there is a failure. The ANT decision mechanism can use the calibration block U to detect hardware failures and bypass them by switching between M and U.

Formally, the generalized model for ANT allows the main block to *switch modes of operation* back and forth between the good block and the bad block. Let an independent Bernoulli random variable F of parameter $p \in (0,1)$, determine the hardware failure, the following mixture model characterizes the main block:

$$M = G\bar{F} + BF. \tag{4.1}$$

Here, $\bar{F} = 1 - F$ and a failure happens when $F = 1$. We allow each block to be statistically independent, that is, given the hidden variable X, outcomes G, B, U are conditionally independent from one another:

$$\begin{aligned} G - X - B, \\ G - X - U, \\ B - X - U, \end{aligned}$$

equivalently, $p_{GBU|X} = p_{G|X} p_{B|X} p_{U|X}$, [60]. This assumption is similar to the independence of estimation error and hardware error in [34, 38] and to the independence of noise in different branches in [37]. The triplet (G, B, U) is conditioned on the hidden variable X and the conditional probability density functions $p_{G|X}$, $p_{B|X}$ and $p_{U|X}$ represent the statistical characteristics of the main block, the main block under hardware failure and the calibration block respectively. These distributions represent the computational properties of respective blocks under uncertainties due to process, temperature and voltage variations, which are commonly unknown or too costly to model [33].

The ANT decision rule, denoted by $\delta^{ANT}(\cdot)$, operates on the branch outcomes, (M, U), to make a decision between them: $\delta^{ANT}(M, U) \in \{M, U\} \equiv \{G, B, U\}$. The block with *lower expected loss* is understood to have *higher fidelity*:

$$\mathbb{E}\ell\left(G, X\right) < \mathbb{E}\ell\left(U, X\right) < \mathbb{E}\ell\left(B, X\right). \tag{4.2}$$

The purpose of ANT is to use the calibration random variable U to determine whether a failure ($F = 1$) has happened, or not ($F = 0$), and bypass the main block with the calibration block when it does. If $p_{G|X}$ and $p_{B|X}$, were known a priori, likelihood ratio would provide a sufficient statistic for testing whether a failure has occurred or not, and the Neyman-Pearson rule could be built upon it [4]. Instead, the ANT architecture *builds* statistics using the pair $(M, U) \equiv (\{G, B\}, U)$. In Section 4.1.2, we introduce the ANT decision rule on an arbitrary metric space and define a measure of performance as the regret with respect to the optimal-yet-unattainable oracle decision rule.

4.1.2 Performance Criterion for ANT

ANT builds decision statistics from the pair (M, U) to test whether $M \sim G$ or $M \sim B$ and bypasses the main block when $M = B$ appears more likely. Let $d(\cdot, \cdot)$ be a distance measure defined on $\mathbb{R}$ satisfying the axioms in [65]. A general ANT decision rule has the following form:

$$\delta^{ANT}(M, U) = \begin{cases} M & \text{if} \quad d(M, U) \leq \tau, \\ U & \text{if} \quad d(M, U) > \tau. \end{cases}$$

Intuitively, the ANT decision rule "favors" the main computational unit when it passes a "calibration check", otherwise it uses the calibration unit to bypass the main unit that is "flagged" with hardware failure.

Now consider an *oracle* that has access to reliable information on when a hardware failure, F, occurs. Such an oracle minimizes its loss via the following decision rule:

$$\delta^{O}(M, U; F = f) = \begin{cases} M & \text{if} \quad f = 0, \\ U & \text{if} \quad f = 1. \end{cases}$$

In practice, information on F is not easily, if at all, available. However, it serves as a useful benchmark for measuring the performance of ANT. We propose a conservative measure of performance by defining the *regret* of ANT as the expected loss suffered from using the ANT decision rule, δ^{ANT}, against that of the oracle decision rule δ^{O}:

$$R^{ANT}(\tau) = \mathbb{E}\ell\left(\delta^{ANT}(M, U), X\right) - \mathbb{E}\ell\left(\delta^{O}(M, U; F), X\right).$$

A more intuitive form for $R^{ANT}(\tau)$, follows from the independence of F and the total law of probability, [60].

Proposition 4.1. *For any triplet (G, B, U) of computational units, the regret of ANT satisfies:*

$$R^{ANT}(\tau) = \bar{p} R_{UG} \Phi_d^{GU}(\tau) + p R_{BU} F_d^{BU}(\tau). \tag{4.3}$$

Here, $\bar{p} = 1 - p$, $R_{\alpha\beta} \triangleq \mathbb{E}\ell(\alpha, X) - \mathbb{E}\ell(\beta, X)$, where $(\alpha, \beta) \subset \{G, B, U\}$, $\Phi_d^{GU} \triangleq \mathbb{P}(d(G, U) > \tau)$ and $F_d^{BU} \triangleq \mathbb{P}(d(B, U) \leq \tau)$, when the distance metric is known from context, we drop the subscript. The regret in (4.3) shows that when a hardware failure happens with probability p, the ANT "misses" it with probability $F_d^{BU}(\tau)$, and allows performance degradation R^{BU} and hence, suffers a regret of $pR_{BU}F_d^{BU}(\tau)$. Similarly, with probability $\Phi_d^{GU}(\tau)$, ANT raises a "false alarm" during normal operation that happens with probability $(1 - p)$, and switches back to U, degrading performance by R^{GU}, and thus, suffering a

regret $\bar{p}R_{UG}\Phi_d^{GU}(\tau)$. Note that R^{UG} and R^{BU} are functionals of $\ell(\cdot,\cdot)$, where $\Phi_d^{GU}(\tau)$ and $F_d^{BU}(\tau)$ are functionals of $d(\cdot,\cdot)$. In Section 4.1.3, we explore the connection between these functionals and quantify a fidelity-based characterization of the regret.

4.1.3 Fidelity-Based Characterization

The regret of ANT is a mixed functional of the distance measure $d(\cdot,\cdot)$ used to build the decision statistics and the loss function $\ell(\cdot,\cdot)$ that determines the fidelity of a computational unit. On a Hilbert space, $\mathcal{H}$, the distance measure is given by:

$$d(M,C) = \|M - C\|,$$

where $\|\cdot\|$ is the norm associated with $\mathcal{H}$, [65]. This section explores the fundamental limits for the regret of ANT for fidelities defined by any $\mathcal{C}$-bi-Lipschitz loss function on a Hilbert space, $\mathcal{H}$. That is, $\forall\{M,C\} \subset \{G,B,U\}$:

$$\frac{1}{\mathcal{C}}\|M - C\| \leq |\ell(M,X) - \ell(C,X)| \leq \mathcal{C}\|M - C\|. \tag{4.4}$$

If $\exists\tau : R^{ANT}(\tau) = 0$, then ANT is regret-optimal, that is, it operates with no regret. The fidelity ordering in (4.2) yields that:

$$R^{ANT}(\tau) = 0 \iff \Phi_d^{GU}(\tau) = F_d^{BU}(\tau) = 0. \tag{4.5}$$

This follows from positivity of p, $\bar{p}$, R^{GU} and R^{BU} and it yields the following statistical necessary condition.

Proposition 4.2. *A necessary condition for* (4.5) *is* $\mathbb{E}\|U - G\| < \mathbb{E}\|B - U\|$. *Furthermore,* $\mathbb{E}\ell(G,X) < \mathbb{E}\ell(U,X) < \frac{1}{2\mathcal{C}^2+1}\mathbb{E}\ell(B,X)$ *implies* $\mathbb{E}\|U - G\| < \mathbb{E}\|B - U\|$.

The relation $\mathbb{E}\ell(G,X) < \mathbb{E}\ell(U,X) < \frac{1}{2\mathcal{C}^2+1}\mathbb{E}\ell(B,X)$ allows $\Phi_d^{GU}(\tau)$ and $F_d^{BU}(\tau)$ to be bounded in terms of *fidelities* rather than expected distances.

Proposition 4.3. *Any ANT rule inheriting its distance measure from the Hilbert space* $\mathcal{H}$ *on which* (G,B,U) *are defined satisfy the following properties for any* $\mathcal{C}$ *bi-Lipschitz loss function:*

1. *Distance to Regret:* $\mathbb{E}\|B - U\| \geq \frac{1}{\mathcal{C}}R^{BU}$ *and* $\mathbb{E}\|G - U\| \leq \mathcal{C}\Sigma^{GU}$, *where* $\Sigma^{GU} \triangleq \mathbb{E}\ell(G,X) + \mathbb{E}\ell(U,X)$.

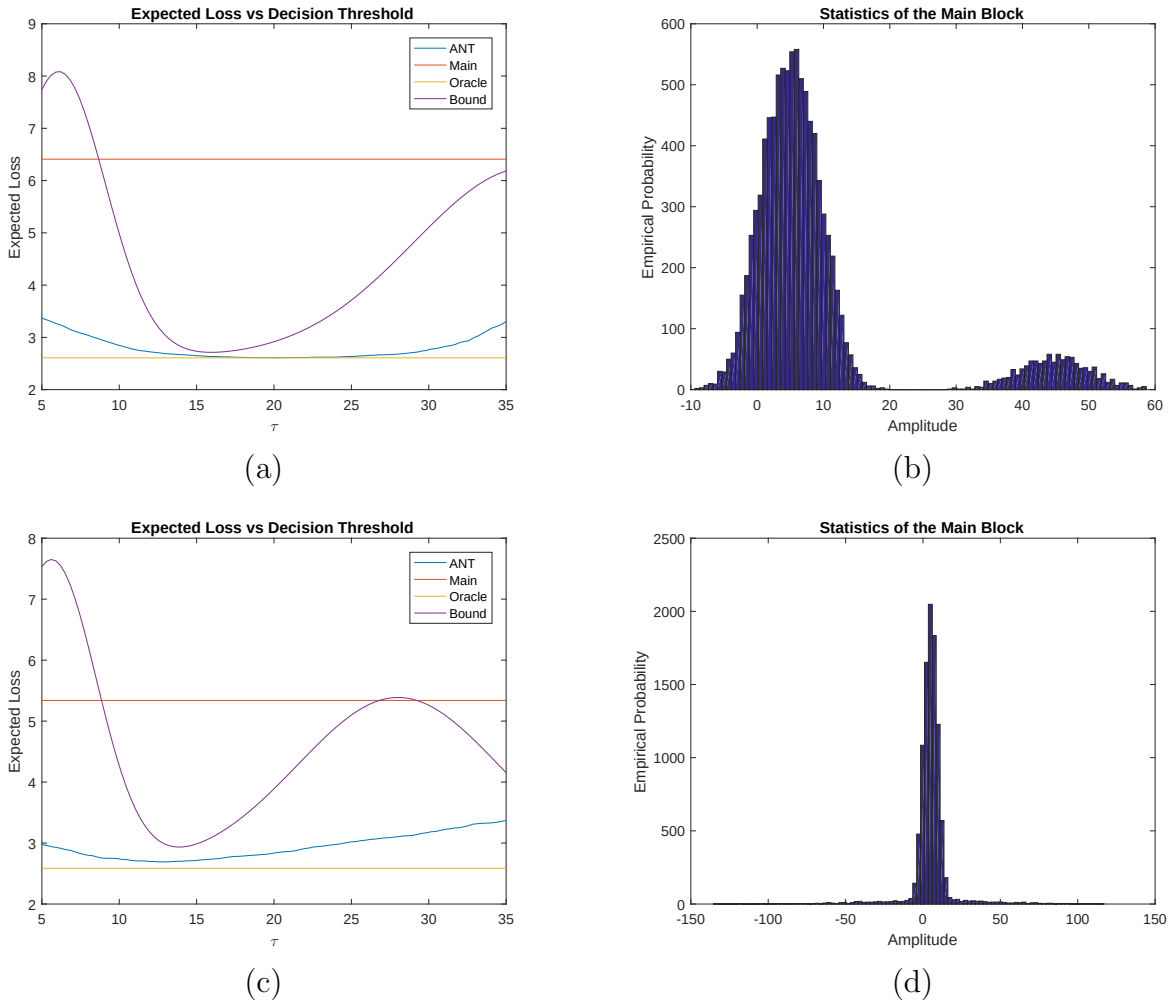


Figure 4.2: Universal Bounds vs. Performance of ANT for Different Error Statistics: (a-b) Bias-Introducing Hardware Failure, (c-d) "Burying Noise" (High-Variance) Hardware Failure

2. *Chernoff Bound:* $\log \Phi^{GU}(\tau) \leq -\frac{\left(\tau - \mathcal{C}\Sigma^{GU}\right)^2}{\tau + \mathcal{C}\Sigma^{GU}}$ *and* $\log F^{BU}(\tau) \leq -\frac{\left(\tau - R^{BU}/\mathcal{C}\right)^2}{2R^{BU}/\mathcal{C}}$, $\forall \tau \in \left(\mathcal{C}\Sigma^{GU}, R^{BU}/\mathcal{C}\right)$.

Here, $\Sigma^{GU} = \mathbb{E}\ell(G, X) + \mathbb{E}\ell(U, X)$. We note that the Chernoff bound is not generally sharp, however, it lends the necessary tool for our fidelity-based analysis to be universal, which we demonstrate in Section 4.1.4.

4.1.4 Experiments

We propose a Gaussian mixture with $G \sim \mathcal{N}(X, \sigma_G)$, $U \sim \mathcal{N}(X, \sigma_U)$ and the bad-block either introducing a bias $B \sim \mathcal{N}(X + \mu_B, \sigma_B)$ or introducing a large-variance noise. We

compare the performance of ANT to that of the *oracle* over the range of τ values and demonstrate that the Chernoff bounds that we propose in observation 2, indicate accurately when the performance of ANT is optimal.

The experiment specifications are as follows: Figure 4.2 (a)-(b) demonstrate the setup, where the bias that the "bad" block introduces is the main source of distortion as, X is distributed uniformly on $(0, 10)$, $\sigma_G = 10$, $\sigma_B = \sigma_U = 20$ with $\mu_B = 40$. Figure 4.2 (c)-(d) illustrates the case, where the "bad" block has large variance: $\sigma_G = 10$, $\sigma_U = 15$, yet $\sigma_B = 1.5e + 3$. Figure 4.2 (a)-(c) indicate that as σ_B increases, the performance of ANT deteriorates as the performance of the ANT no longer achieves that of oracle, as expected from Proposition 4.2.

The use of calibration unit U in the ANT architecture not only allows fault-detection, but also bypassing of such erroneous outcomes by the output from the calibration unit. We next investigate the performance of an oracle rule that has no access to a calibration unit and thus, chooses to *discard* a faulty outcome rather than bypassing it.

4.2 Fault-Rejecting Averaging

Let a set of error-prone computational units $\{M_i : i \in [n]\}$ provide opinions of the form:

$$M_i = G_i \bar{F}_i + B_i F_i,$$

where, similar to Section 4.1, the failure F_i is a Bernoulli random variable independent from (G_i, B_i). We allow F_i to be independent and identically distributed across different units. The goal is to discover the fundamental limits for fault-rejecting averaging, denoted by $\mu(\mathbf{M}; \delta)$, where upon detecting a failure, say $\hat{F}_i = \delta(\mathbf{M}) = 1$, via some decision rule $\delta(\cdot)$, output M_i from the unit marked as faulty, is discarded. Formally:

$$\mu(\mathbf{M}; \delta) = \frac{1}{\left|\left\{i : \hat{F}_i = 0\right\}\right|} \sum_{i:\hat{F}_i=0} M_i. \tag{4.6}$$

Note that (4.6) is a random average of random variables and our focus is the expectation and the variance of $\mu(\mathbf{M}; \delta)$. First note that mean and variance of a computational unit determines the mean-square error since,

$$\mathbb{E}\left[|X - \mu(\mathbf{X})|^2\right] = \left(\mathbb{E}\left[X - \mu(\mathbf{X})\right]\right)^2 + \mathrm{Var}\left(\mu(\mathbf{X})\right).$$

We first explore the performance of the oracle decision rule: $\delta^O(M_i) = F_i$ in order to explore the impact of *fault rejecting* directly.

4.2.1 Network Average with Oracle Fault Detector

Observe that as long as an oracle detects the failures via δ^O, the following network average can be achieved:

$$\mu^O(\mathbf{M}) = \frac{1}{|\{i : F_i = 0\}|} \sum_{i:F_i=0} G_i. \tag{4.7}$$

Since F_i is modeled as independent from (G_i, B_i), the network average in (4.7) is a random average of random variables, where the number of random variables averaged, the random variable $N = |\{i : F_i = 0\}|$ is independent of $\mathbf{G}$. It is important, however, to note that when $N = 0$, there does not exist any $\mu^O(\mathbf{M})$. Therefore, we first address an abstraction, where N independent and identically distributed reliable sources (G_i) are mixed, where $N \geq 1$ almost surely.

Proposition 4.4. *Let G_i be independent identically distributed sources and let $N \geq 1$ a.s. and N be independent of $\{G_1, \cdots, G_N\}$. Then, the mean and variance of the random average of random variables $A = \frac{1}{N}\sum_{i=1}^{N} G_i$ are:*

$$\begin{aligned} \mathbb{E}[A] &= \mathbb{E}[G], \\ \mathrm{Var}(A) &= \mathbb{E}[1/N]\,\mathrm{Var}(G). \end{aligned}$$

Observe that Proposition 4.4 is consistent with averaging over a constant number of sources. The purpose of Proposition 4.4 is mostly to build intuition; random number of random variables can be averaged and the variance of the result is determined by the expected averaging size $\mathbb{E}[1/N]$. However, Proposition 4.4 does not yet capture averaging of outcomes from faulty sources M_i in form (4.7). Since F_i is a Bernoulli random variable $\forall i \in [n]$, there is non-zero probability that the average is not defined. If one chooses to abstain in the event of all sources being faulty, one can ensure that fault-rejecting averaging is well-defined.

Proposition 4.5. *Let $\{G_1, \cdots, G_n\}$ be independent and let each source be subjected to an independent failure F_i with probability $p_F = 1 - q_F$ (identical for every source). Define $N = \left\|\bar{F}\right\|_1$, which is a Binomial random variable with parameters (q_F, n). Refusing to answer in the case of total network failure ($N = 0$) yields:*

$$\tilde{\mathbb{E}}\left[\mu^O(\mathbf{M})\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[G_i].$$

Furthermore, if $\mathbb{E}[G_i] = 0$, $\forall i$, *then:*

$$\tilde{\mathrm{Var}}\left(\mu^O(\mathbf{M})\right) = \tilde{\mathbb{E}}\left[1/N\right]\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(G_i\right)\right),$$

where $\tilde{\mathbb{E}}[\cdot]$ *and* $\tilde{\mathrm{Var}}(\cdot)$ *denotes expectation and variance with respect to the* 0*-rejecting modified distribution* (C.4)*.*

The exact expression for $\tilde{\mathrm{Var}}\left(\mu^O(\mathbf{M})\right)$ in the case of biased sources ($\mathbb{E}[G_i] \neq 0$) is given in (C.6). As it should, mean and variance of the average $\mu^O(\mathbf{M})$ is consistent with a deterministic average.

CHAPTER 5

NETWORKS WITH STOCHASTIC COMPONENTS

All circuit components have uncertainties inherent to the underlying fabrication process that, in large networks, change the overall circuit response unpredictably. In the absence of a robust and general model to incorporate individual component uncertainties and the concomitant stochastic thermal noise characteristics, under-simulating from the massively high-dimensional experiment space via Monte Carlo techniques and over-designing the final product against potentially undiscovered faults have become *de facto* standard. Such practices not only cost simulation time, excess circuit area and power, but they also provide a partial understanding of the underlying uncertainty and of ways to exploit it. This chapter investigates the impact of component uncertainties in linear resistive networks, where individual elements are subject to Johnson-Nyquist noise.

Fabrication of integrated circuits as well as assembly of discrete circuit components are subject to variability inherent to the physics of the underlying production techniques. The resulting stochasticity is observable at the ensemble level as final products with the same initial design exhibit varying performance characteristics, potentially outside the original design specifications. Unpredictable impacts of such variability and the consequential cost incurred in mass production due to yield motivate the search for a robust and general framework that incorporates fabrication stochasticity as a part of the design process. The fabrication process is massively high-dimensional, comprising the complete set of factors that govern product variability. Hence, it is often difficult, if at all feasible, to build quantitative relationships between fabrication processes and proceeding ensemble of products.

In the absence of a robust and reliable framework to incorporate fabrication stochasticity, a *minmax* design philosophy is established in practice: Monte Carlo simulations are employed, often ad hoc, to explore a subspace of the massively high-dimensional fabrication space. Design principles based on fault tolerance and robustness to process variations are relied upon to compensate potentially undiscovered variabilities at the expense of increased circuit area and power consumption. These practices increase the simulation and testing time significantly, contributing to the aforementioned production costs. Under-simulating and over-designing devices provide limited insights into the underlying sources of uncertainty

and ways to compensate for them.

The massive dimensionality of fabrication uncertainty and topology-dependent propagation of resulting doubly-stochastic thermal noise processes are fundamental challenges of modern circuit design. State-of-the-art circuit simulators, on the other hand, commonly employ *static models* for circuit elements to simulate an abstraction of the network response [66]. Their use of fast matrix methods and *piecewise linearization*, allows them to capture a large family of non-linear circuit characteristics, while operating within the principles of linear network theory. This work addresses the impact of component stochasticity on linear noisy networks to build a framework that incorporates fabrication stochasticity into design process.

We discuss stochasticity in linear resistive networks due to the variations in individual components and the Johnson-Nyquist noise exhibited in each resistive element. The analysis applies to both circuit topologies constructed from discrete components on a fabricated circuit board, as well as resistive networks within an integrated circuit. We follow a graph theoretic framework to investigate the concentration of the overall network response around its designed mean as well as its ensemble mean in terms of individual component stochasticity and network topology. In particular, we discuss effective resistance, power dissipation, and mean-squared branch voltages on arbitrary fixed circuit topologies.

5.1 Linear Resistive Networks with Stochastic Components

Given any linear resistive network, there exists an equivalent weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{g})$, where $\mathcal{V}$ represents nodes of the circuit and $\mathcal{E}$ represents edges with weights being conductances $\boldsymbol{g}$. We assume that the every subgraph of edges is connected, equivalently, there is a single circuit of interest. Furthermore, without loss of generality, we allow a *unique* conductance g_{ij} to exist $\forall i \leftrightarrow j \in \mathcal{E}$. When there are multiple components connecting vertices $i, j \in \mathcal{V}$, they can be replaced by a single resistor with equivalent conductance. Appendix D.4 demonstrates how branch statistics are modified in such a case.

Our purpose is to investigate how properties of a linear resistive network of fixed topology concentrates around its circuit-ensemble and designed mean when subjected to component stochasticity and the Johnson-Nyquist noise. We use the term circuit property, and denote it by $h(\mathbf{G}; \mathcal{G})$, to collectively refer to quantities such as effective resistance r_{ij}^{eff} (for any $i, j \in \mathcal{V}$), total effective resistance, σ, average power dissipation, $\mathbb{E}[P]$, mean-squared branch voltages, $\mathbb{E}\left[V_{ij}^2\right]$ (for $i \leftrightarrow j \in \mathcal{E}$), and so on. Conceptually, the designed mean of a circuit property, $\mathbb{E}_{\mathcal{T}}[h(\boldsymbol{g}; \mathcal{G}) \mid \boldsymbol{g}]$, is the expected network response when there is no fabrication uncertainty ($\boldsymbol{g} = \mathbb{E}_{\mathbf{G}}[\mathbf{G}]$, being the designed circuit parametrization, is decided during the

design process). Ensemble mean of a network, $\mathbb{E}_{\mathbf{G}}\left[\mathbb{E}_{\mathcal{T}}\left[h(\mathbf{G};\mathcal{G}) \mid \mathbf{G}\right]\right]$, on the other hand, is the expected response over the fabricated circuits. Formally, we pursue bounds of the form:

$$\mathbb{P}\left(h(\mathbf{G};\mathcal{G}) - \mathbb{E}_{\mathcal{T}}\left[h(\boldsymbol{g};\mathcal{G}) \mid \boldsymbol{g}\right] > \varepsilon\right) \leq f(\varepsilon;\mathcal{G}), \tag{5.1}$$

$$\mathbb{P}\left(h(\mathbf{G};\mathcal{G}) - \mathbb{E}_{\mathbf{G}}\left[\mathbb{E}_{\mathcal{T}}\left[h(\mathbf{G};\mathcal{G}) \mid \mathbf{G}\right]\right] > \varepsilon\right) \leq f(\varepsilon;\mathcal{G}). \tag{5.2}$$

When a concentration result of form (5.1) is shown, we say that the circuit property concentrates around its *designed mean*, when a result of form (5.2) is shown we say that concentrates around its circuit ensemble mean, or simply, *ensemble mean*. We do not pursue concentration results of form $f(\varepsilon;\mathcal{G},p_{\mathbf{G}})$ as statistics of the fabrication process is seldom, if at all, known. Instead, we seek results of form $f(\varepsilon;\mathcal{G})$ for bounded fabrication variabilities that preserve topology.

The analysis presented here extends from assembly of discrete components subject to statistically independent fabrication variability to integration through processes such as lithography, chemical deposition and so on, that result in spatially correlated network components. To that end, we introduce mixing matrices $\mathbf{\Phi}$ for resistances $\mathbf{R}$ ($\mathbf{\Gamma}$ for conductances $\mathbf{G}$): Mixing matrices $\mathbf{\Phi},\mathbf{\Gamma} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ contain coefficients that measure the dependence between random variables $\{R_{ij}\}_{i\leftrightarrow j\in\mathcal{E}}$ and $\{G_{ij}\}_{i\leftrightarrow j\in\mathcal{E}}$ respectively. For integrated circuits, mixing matrices should be interpreted as the spatial correlation between components as a result of the integration process. When a mixing matrix, say $\mathbf{\Phi}$, is available, we pursue concentration results of form $f(\varepsilon;\mathcal{G},\mathbf{\Phi})$. Johnson-Nyquist noise is the response of a conductor to thermal agitation and it has been shown to exhibit Gaussian statistics with variance depending on the operation temperature and the conductor under excitement [42]. Formally, a branch current I_{ij} becomes a Gaussian random variable, $I_{ij} \sim \mathcal{N}\left(g_{ij}, 2ktg_{ij}^2\right)$, where t is the operation temperature in degrees Kelvin and k is the Boltzmann constant. Hierarchical structure between Johnson-Nyquist noise and the fabrication process is noteworthy: *after* the circuit is fabricated, the thermal noise takes over the resulting circuit. In other words, the thermal noise process is conditioned on the fabrication process. Section 5.2 quantifies the impact of the thermal noise. Next, we introduce the graph theoretic techniques that we employ.

Consider a fixed circuit topology $\mathcal{G} = (\mathcal{V},\mathcal{E})$ of a linear resistive network with the corresponding weighted adjacency matrix $\underline{\mathbf{C}}$ containing *conductances*: $\left(\underline{\mathbf{C}}\right)_{ij} \triangleq g_{ij}\mathbb{1}\left(i \leftrightarrow j \in \mathcal{E}\right)$, $\forall i,j \in \mathcal{V}$. Here, $g_{ij} = {}^1\!/_{r_{ij}}$ corresponds to branch resistances r_{ij}. Consider $\underline{\mathbf{\Lambda}}$, a diagonal matrix with entries $\left(\underline{\mathbf{\Lambda}}\right)_{ii} = \sum_{k=1}^{|\mathcal{V}|}\left(\underline{\mathbf{C}}\right)_{ik}$ and define the Laplacian $\underline{\mathbf{L}}$ of the network:

$$\underline{\mathbf{L}} = \underline{\mathbf{\Lambda}} - \underline{\mathbf{C}}.$$

It has been shown that the *effective resistance* between any pairs of nodes in the circuit,

which is the resistance shown to a potential applied these nodes, is a distance metric on such a graph [67, 68]. The effective resistance r_{ij}^{eff} between any pairs of nodes $i, j \in \mathcal{V}$, has the following form:

$$r_{ij}^{\mathit{eff}} = (\boldsymbol{e}i - \mathbf{e}_j)^\top \underline{\mathbf{L}}\dagger(\boldsymbol{e}i - \boldsymbol{e}j) = (\underline{\mathbf{L}}\dagger)_{ii} + (\underline{\mathbf{L}}\dagger)_{jj} - 2(\underline{\mathbf{L}}\dagger)_{ij}. \tag{5.3}$$

When necessary, we explicitly emphasize that effective resistance is a function of the branch conductances by writing $r_{ij}^{\mathit{eff}}(\boldsymbol{g})$ (when branch conductances are random, $r_{ij}^{\mathit{eff}}(\mathbf{G})$). We write $r_{ij}^{\mathit{eff}}(\boldsymbol{g})$ as a function of *conductances*, although we take derivatives with respect to resistances when necessary.

Furthermore, the *total effective resistance*, the sum of effective resistances between all pairs of nodes, characterizes power dissipation and mean-square voltages across branch resistances [68, 69]. We denote the total effective resistance by Σ and write it as a function of conductances when necessary:

$$\Sigma = \sum_{i<j} r_{ij}^{\mathit{eff}}.$$

When the branch conductances are random variables, we use $\Sigma(\mathbf{G})$ to denote the resulting random variable. Total effective resistance is related to the spectrum of the graph that represents the circuit, expected power dissipation, mean-square voltages across branches, and network criticality [68, 69]. Notably, Σ can be represented as follows:

$$\frac{1}{|\mathcal{V}|}\Sigma = \operatorname{tr}(\underline{\mathbf{L}}\dagger) = \sum_{i=2}^{|\mathcal{V}|} \frac{1}{\lambda_i}. \tag{5.4}$$

Here, $\{\lambda_i\}$ are eigenvalues of the Laplacian $\underline{\mathbf{L}}$. A proof of (5.4) can be found in, for instance, [68, 69]. A comprehensive survey on the extensions and applications of algebraic graph theory for circuit analysis appear in [67].

In this section, we investigate linear resistive networks with stochastic components in the absence of Johnson-Nyquist noise and propose concentration results around the designed mean $h(\mathbb{E}[\mathbf{G}]; \mathcal{G})$ and ensemble mean $\mathbb{E}[h(\mathbf{G}; \mathcal{G})]$ (whether the expectation $\mathbb{E}[\cdot]$ is with respect to $\mathbf{R}$ or to $\mathbf{G}$ is stated explicitly below). We allow topology-preserving bounded perturbations on the branch conductances: $\mathbb{P}(G_{ij} \in [\ell_{ij}, u_{ij}]) = 1$ for some $u_{ij} > \ell_{ij} > 0$.

We define two diagonal matrices of use in hindsight:

$$(\mathbf{P})_{i\leftrightarrow j\in\mathcal{E}} = \frac{u_{ij}-\ell_{ij}}{\ell_{ij}}, \tag{5.5}$$

$$(\mathbf{D})_{i\leftrightarrow j\in\mathcal{E}} = \frac{u_{ij}-\ell_{ij}}{u_{ij}\ell_{ij}}. \tag{5.6}$$

When we refer to the fabrication process, we assume that support $(\boldsymbol{u},\boldsymbol{\ell})$ and mixing matrices $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$ are known.

A dimension-dependent bound the concentration of the effective resistance between any two nodes of a linear resistive network with resistances of bounded support follows from the monotonicity of the effective resistance.

Proposition 5.1. *Let branch conductances G_{ij}, be statistically independent, bounded random variables, then, $\forall a,b\in\mathcal{V}$:*

$$\mathbb{P}\left(\left|r_{ab}^{\mathit{eff}}(\mathbf{G})-\mathbb{E}_{\mathbf{G}}\left[r_{ab}^{\mathit{eff}}(\mathbf{G})\right]\right|>\varepsilon\right)\leq 2\exp\left(-\frac{2\varepsilon^2}{|\mathcal{E}|\left|r_{ab}^{\mathit{eff}}(\boldsymbol{\ell})-r_{ab}^{\mathit{eff}}(\boldsymbol{u})\right|^2}\right).$$

Dimensionality dependence in Proposition 5.1 is undesirable. Instead, one can exploit the relation between the circuit topology and the support of the resistance process to propose a sharper bound. The matrix $\mathbf{P}$ is defined in (5.5).

Theorem 5.1. *Let branch conductances G_{ij}, be statistically independent bounded random variables, then, effective resistance concentrates around its ensemble mean: $\forall a,b\in\mathcal{V}$:*

$$\mathbb{P}\left(\left|r_{ab}^{\mathit{eff}}(\mathbf{G})-\mathbb{E}_{\mathbf{G}}\left[r_{ab}^{\mathit{eff}}(\mathbf{G})\right]\right|>\varepsilon\right)\leq 2\exp\left(-\frac{2\varepsilon^2}{\left|\boldsymbol{g}^{\mathit{eff}}(\boldsymbol{\ell})^{\top}\mathbf{P}\nabla_{\boldsymbol{g}}r_{ab}^{\mathit{eff}}(\boldsymbol{\ell})\right|^2}\right).$$

A corollary of Theorem 5.1 has a weaker form that connects it back to Proposition 5.1.

Corollary 5.1. *The following (weaker) bound follows:*

$$\mathbb{P}\left(\left|r_{ab}^{\mathit{eff}}(\mathbf{G})-\mathbb{E}_{\mathbf{G}}\left[r_{ab}^{\mathit{eff}}(\mathbf{G})\right]\right|>\varepsilon\right)\leq 2\exp\left(-\frac{2\varepsilon^2}{|\mathbf{P}\boldsymbol{r}^{\mathit{eff}}(\boldsymbol{\ell})|^2}\right).$$

Here, $\boldsymbol{r}^{\mathit{eff}}(\boldsymbol{\ell})$ takes values over the edges of the graph. A second corollary characterizes how the effective resistance concentrates around its designed mean as well.

Corollary 5.2. *Let branch resistances R_{ij}, be statistically independent bounded random variables, then, effective resistance concentrates around its designed mean: $\forall a, b \in \mathcal{V}$:*

$$\mathbb{P}\left(r_{ab}^{\mathit{eff}}(\mathbf{G}) - r_{ab}^{\mathit{eff}}(\mathbb{E}_{\mathbf{R}}[\mathbf{G}]) > \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\left|\boldsymbol{r}^{\mathit{eff}}(\boldsymbol{\ell})^{\top}\mathbf{P}\nabla_{\boldsymbol{r}} r_{ab}^{\mathit{eff}}(\boldsymbol{\ell})\right|^2}\right).$$

Theorem 5.1 considers a fabrication process, where discrete components are assembled by assuming independent conductance statistics. When spatial dependence due to an integration process is taken into account, one should consider the statistical dependence among components. Formally, let the matrix $\mathbf{D}$ be as defined in (5.6):

Theorem 5.2. *Let branch resistances R_{ij}, $\forall i \leftrightarrow j \in \mathcal{E}$ be bounded random variables, with the mixing matrix $\boldsymbol{\Phi}$. Then, $\forall a, b \in \mathcal{V}$:*

$$\mathbb{P}\left(r_{ab}^{\mathit{eff}}(\mathbf{G}) - \mathbb{E}_{\mathbf{R}}\left[r_{ab}^{\mathit{eff}}(\mathbf{G})\right] > \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\left\|\boldsymbol{\Phi}\right\|^2 \mathbb{E}_{\mathbf{R}}\left|\mathbf{D}\nabla_{\boldsymbol{r}} r_{ab}^{\mathit{eff}}(\mathbf{R})\right|^2}\right).$$

The proof follows from modifying [70, Eqn. 2.15] for an arbitrary compact domain. We provide an outline in the Appendix D.2. The concentration around its designed mean $r_{ab}^{\mathit{eff}}(\mathbb{E}_{\mathbf{R}}[\mathbf{G}])$ follows from the concavity of $r_{ab}^{\mathit{eff}}(\boldsymbol{g})$ with respect to $\boldsymbol{r}$, [69], similar to Corollary 5.2. Formally:

$$\mathbb{P}\left(r_{ab}^{\mathit{eff}}(\mathbf{G}) - r_{ab}^{\mathit{eff}}(\mathbb{E}_{\mathbf{R}}[\mathbf{G}]) > \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\left\|\boldsymbol{\Phi}\right\|^2 \mathbb{E}_{\mathbf{R}}\left|\mathbf{D}\nabla_{\boldsymbol{r}} r_{ab}^{\mathit{eff}}(\mathbf{R})\right|^2}\right).$$

Next, we investigate the concentration of the total effective resistance around its ensemble mean:

Theorem 5.3. *Let branch resistances R_{ij}, $\forall i \leftrightarrow j \in \mathcal{E}$ be bounded random variables, with the mixing matrix $\boldsymbol{\Phi}$. Then, $\forall a, b \in \mathcal{V}$, the total effective resistance of the circuit concentrates around its ensemble mean:*

$$\mathbb{P}\left(\Sigma(\mathbf{G}) - \mathbb{E}_{\mathbf{R}}[\Sigma(\mathbf{G})] > \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\left\|\boldsymbol{\Phi}\right\|^2 \mathbb{E}_{\mathbf{R}}\left|\mathbf{D}\nabla_{\boldsymbol{r}}\Sigma(\mathbf{R})\right|^2}\right).$$

A corollary of Theorem 5.3, characterizes the average power consumption of the circuit when a random current $\mathbf{J}$ is injected to the network. Following the example in [69], we allow $\mathbb{E}[\mathbf{J} \mid \mathbf{G} = \boldsymbol{g}] = 0$ and $\mathbb{E}[J_i J_j \mid \mathbf{G} = \boldsymbol{g}] = \mathbb{1}(i \neq j)$, $\forall i, j \in \mathcal{V}$ to be injected to the network,

which results in power dissipation $P(\boldsymbol{g}) = \mathbf{J}^\top \underline{\mathbf{L}}\dagger(\boldsymbol{g})\mathbf{J}$ that obeys the following corollary:

Corollary 5.3. *The expected dissipated power in a resistor network concentrates around its designed mean:*

$$\mathbb{P}\left(\mathbb{E}\left[P(\mathbf{G})\right] - \mathbb{E}_{\mathbf{R}}\left[\mathbb{E}\left[P \mid \mathbf{G}\right]\right] > \varepsilon\right) \leq \exp\left(-\frac{\left(\varepsilon\left|\mathcal{E}\right|\right)^2}{2\left\|\boldsymbol{\Phi}\right\|^2 \mathbb{E}_{\mathbf{R}}\left|\mathbf{D}\nabla_{\boldsymbol{r}}\Sigma(\mathbf{R})\right|^2}\right).$$

Theorem 5.3 and Corollary 5.3 can be modified, using Jensen's inequality similar to what is done in Corollary 5.2, to incorporate designed means $\Sigma\left(\mathbb{E}_{\mathbf{R}}\left[G\right]\right)$ and $\mathbb{E}\left[P(\mathbb{E}_{\mathbf{R}}\left[\mathbf{G}\right])\right]$ instead of their ensemble counterparts.

Finally, we investigate the mean-square voltages appearing across each resistor. Unlike Theorems 5.2-5.3, the next result employs concavity of $\nabla_{\boldsymbol{g}} r_{ij}^{\mathit{eff}}(\boldsymbol{g})$, $\forall i \leftrightarrow j \in \mathcal{E}$.

Theorem 5.4. *Let branch conductances G_{ij}, $\forall i \leftrightarrow j \in \mathcal{E}$ be bounded random variables, with mixing matrix $\boldsymbol{\Gamma}$. Then, $\forall i \leftrightarrow j \in \mathcal{E}$, the mean-square voltage $\mathbb{E}\left[V_{ij}^2(\mathbf{G})\right]$ appearing across the branch resistor concentrates around its ensemble mean:*

$$\mathbb{P}\left(\mathbb{E}\left[V_{ij}^2(\mathbf{G})\right] - \mathbb{E}_{\mathbf{G}}\left[\mathbb{E}\left[V_{ij}^2 \mid \mathbf{G}\right]\right] > \varepsilon\right) \leq \exp\left(-\frac{\left(\varepsilon\left|\mathcal{E}\right|\right)^2}{2\left\|\boldsymbol{\Gamma}\right\|^2 \mathbb{E}_{\mathbf{G}}\left|\boldsymbol{\ell}^\top \mathbf{P}\nabla_{\boldsymbol{g}}\Sigma(\mathbf{G})\right|^2}\right).$$

Linear resistive networks exhibit concentration phenomena over varying circuit properties when they are subjected to bounded fabrication uncertainty. In the next section, we incorporate Johnson-Nyquist noise into the concentration results on effective resistance.

5.2 Linear Noisy Networks with Stochastic Components

Post-fabrication, circuits are subject to Johnson-Nyquist noise and the impacts of thermal noise on the concentration properties of the effective resistances around their designed and ensemble means are investigated here.

A key fact about Johnson-Nyquist noise is of use: When all resistances in a linear resistive network (of deterministic components) are subject to thermal noise, the effective resistance between any two nodes is subject to the thermal noise that would operate on the equivalent resistance [43]. Formally, $\forall i, j \in \mathcal{V}$, when a fixed current I flows into node j and flows out of node i, the potential has Gaussian statistics with

$$V_{ij} \sim \mathcal{N}\left(Ir_{ij}^{\mathit{eff}}, 2kt\right). \tag{5.7}$$

Let a Gaussian random variable $T \sim \mathcal{N}(0, 2kt)$. The distribution in (5.7) is useful for characterizing the impact of fabrication stochasticity: For any (sample) circuit parametrized with $\mathbf{G} = \boldsymbol{g}$, the potential observed from the sample circuit under thermal noise, denoted by $V_{ij}(\boldsymbol{g}; \mathcal{T})$, obeys:

$$V_{ij}(\boldsymbol{g}; \mathcal{T}) = V_{ij}(\boldsymbol{g}) + T.$$

Over the ensemble of the circuit fabrication process, $V_{ij}(\mathbf{G}; \mathcal{T})$ has a joint distribution over $\mathbf{G}$ and $\mathcal{T}$ and it obeys $V_{ij}(\mathbf{G}; \mathcal{T}) = V_{ij}(\mathbf{G}) + T$. Hence, for every sample circuit $\boldsymbol{g}$, $\mathbb{E}_{\mathcal{T}}[V_{ij}(\mathbf{G}; \mathcal{T}) \mid \mathbf{G} = \boldsymbol{g}] = V_{ij}(\boldsymbol{g})$, yielding that:

$$\mathbb{E}_{\mathbf{G}}\left[\mathbb{E}_{\mathcal{T}}[V_{ij}(\mathbf{G}; \mathcal{T}) \mid \mathbf{G}]\right] = \mathbb{E}_{\mathbf{G}}\left[V_{ij}(\mathbf{G})\right].$$

Therefore, deviation of $V_{ij}(\mathbf{G}; \mathcal{T})$ from its designed and ensemble means, collectively denoted by μ are of the form:

$$|V_{ij}(\mathbf{G}; \mathcal{T}) - \mu| = |Vij(\mathbf{G}) + T - \mu|.$$

Let us allow a unit current to be applied so that we can investigate $r_{ij}^{\mathit{eff}}(\mathbf{G}; \mathcal{T})$ rather than Vij.

$$\begin{aligned}
\mathbb{P}\left(\left|r_{ij}^{\mathit{eff}}(\mathbf{G}; \mathcal{T}) - \mu\right| \geq \varepsilon\right) &= \mathbb{P}\left(\left|r_{ij}^{\mathit{eff}}(\mathbf{G}) + T - \mu\right| \geq \varepsilon\right), \\
&= \int_{-\infty}^{\infty} \mathbb{P}\left(\left|r_{ij}^{\mathit{eff}}(\mathbf{G}) + \alpha - \mu\right| \geq \varepsilon\right) dP_T(\alpha), \\
&\leq \mathbb{P}(|T| \geq \varepsilon) + \int_{-\varepsilon}^{\varepsilon} \mathbb{P}\left(\left|r_{ij}^{\mathit{eff}}(\mathbf{G}) - \mu\right| \geq \varepsilon - \alpha\right) dP_T(\alpha).
\end{aligned}$$

The argument in the integral above can be bounded using the results from Proposition 5.1, Theorems 5.1-5.2, and their corollaries, depending on whether μ is the ensemble mean or designed mean. One-sided results are handled similarly.

Chapter 5 discusses the impact of fabrication variability and thermal noise on linear resistive networks. Due to the massive dimensionality of the fabrication process, we addressed fabrication processes that preserve circuit topology by introducing variability on a compact support. We proposed concentration bound for circuit properties such as effective resistance, total effective resistance, average power dissipation, and mean-square branch voltage around the corresponding circuit-ensemble mean and designed mean values. We further addressed the impact of Johnson-Nyquist noise, using the hierarchical structure between fabrication and thermal noise processes. We leave topology altering fabrication uncertainty as an exciting open problem.

CHAPTER 6

CONCLUSION

This dissertation addressed the problems of sequential consultation, opinion aggregation, error detection and mitigation, and quantifying the impact of component uncertainty in the absence of error model, observation statistics, or feedback. The problems of interest spanned from component uncertainty at the circuit-level to computational uncertainty at the systems-level, and eventually to uncertainty due to a large, and potentially, unknown set of factors at the architecture-level. The use of a variety of subjective evaluations of the world to deduce objective information emerged as the overarching theme of this work.

The sequential consultation problem was addressed in the dynamic programming framework, allowing cost-dependent stopping times that are not easily attained in the conventional sequential-probability-ratio formulation. Specifically, the magnitude of the log-likelihood process proved sufficient for stopping the consultation process, complementing the use of its sign for decision making. We provided a closed-form expression for the optimal stopping threshold, quantifying the impact of the cost function. Furthermore, the results from model-dependent optimal stopping extended naturally to the Bayesian framework with the theoretical guarantees on the existence of a unique Bayesian threshold supported the numerical computation of the Bayesian optimal stopping threshold. Interestingly, the cost-free consultation of equally-reliable experts, which aims to minimize the probability of error directly, admitted an unsupervised consultation strategy. Overall, the agreement among experts was quantified and used in a manner to ensure cost-efficient sequential acquisition of subjective information.

The unsupervised opinion aggregation problem was addressed from a statistical perspective that allowed instantaneous, computationally efficient, and distributed inference. The notion of agreement among experts emerged as a key tool for statistical inference of reliability in the absence of opinion-generation models, opinion statistics, or feedback. In the pursuit of an effective use of what-was-called "pseudo" competences, an interesting concentration inequality emerged, guaranteeing the sharpest attainability by the use of Chernoff bounding technique. Furthermore, pseudo competence, as a measure of reliability, which can be inferred in an unsupervised framework, was shown to preserve the true ordering of

the competences, allowing block and adaptive aggregation of opinions. Therefore, the agreement among experts were used to facilitate reliable and computationally efficient opinion aggregation.

A set of fault-tolerant computational principles were addressed to limit the impact of low-probability yet high-impact failures, also commonly known as the "black swan" events, at the systems level. The decision-theoretic interpretation of what is called "algorithmic noise tolerance" was shown to provide model-independent framework for failure detection and error mitigation. Specifically, in the presence of a robust computational unit, fidelity ordering proved to be an effective way model-independent inference. We further addressed, failure-rejected averaging and discovered the fundamental limits of averaging reliable outputs from an error-prone computational unit. The notion of agreement among computational units allowed model-independent, or unsupervised in that sense, detection of failures.

The impact of component uncertainty due to fabrication, assembly, and integration processes was quantified by the use of concentration of measure inequalities. These processes were considered to have hidden statistics due to the difficulty of gathering relevant data in practice. When discrete component uncertainty with no statistical inter-component dependence was concerned, the effective resistance was shown to have bounded differences in terms of the individual resistances. The inter-component dependence, spatial dependence in particular, was addressed by using the concavity of the effective resistance. Furthermore, the impact of the Johnson-Nyquist noise in the proposed framework was quantified over the ensemble of circuits.

Several new challenges arose. First, a conceptually meaningful question remains: How one would sequentially consult experts if one could choose the competence of the next expert? Next, although it was shown that pseudo competence is sufficient to identify the most competent expert in the centralized and distributed frameworks, it was not shown here precisely how one would use pseudo competences for unsupervised exploration and exploitation. Finally, a spectral approach for linear noisy circuits with stochastic components is worth considering to complement to finite-circuit-size analysis presented here.

Information is often available indirectly and the reliability of information depends on a large set of factors that cannot always be modeled reasonably. However, the subjectivity of information does not necessarily render objective inference implausible. This dissertation established a set of ideas that allow such inference.

APPENDIX A

PROOFS FOR CHAPTER 2

A.1 Proof of Lemma 2.1

Proof. The expected future reward given that opinions $X^t = x^t$ have been observed is called the value function (2.7), and it is given as follows:

$$V_t\left(x^t\right) = \max_{\substack{\tau \geq t \\ \delta \in \mathcal{D}_\tau}} \mathbb{E}\left[r\left(\tau; X^\tau\right) \,\middle|\, X^t = x^t\right] = \max_{\tau \geq t} \max_{\delta \in \mathcal{D}_\tau} \mathbb{E}\left[\beta_\tau \mathbb{1}\left(\delta\left(X^\tau\right) = Y\right) \,\middle|\, x^t\right]. \tag{A.1}$$

The separation (A.1) of the joint maximization over stopping times $\tau > t$ and opinion aggregation rules $\delta \in \mathcal{D}_\tau$ follows the family decision aggregation rules being well-defined under the filtration $\mathscr{F}_t = \sigma\left(\mathcal{Y}^t\right)$, the smallest σ-algebra containing $\mathcal{Y}^t$. Equivalently, it follows since the decision maker cannot employ $\delta \in \mathcal{D}_\tau \setminus \mathcal{D}_t$ at any time $t < \tau$. Observe that:

$$\begin{aligned}\max_{\delta \in \mathcal{D}_\tau} \mathbb{E}\left[\beta_\tau \mathbb{1}\left(\delta\left(X^\tau\right) = Y\right) \,\middle|\, X^t = x^t\right] &= \beta_\tau \max_{\delta \in \mathcal{D}_\tau} \mathbb{E}_{X_{t+1}^\tau}\left[\mathbb{E}\left[\mathbb{1}\left(\delta\left(X^\tau\right) = Y\right) \,\middle|\, x^t, X_{t+1}^\tau\right] \,\middle|\, x^t\right], \\ &= \beta_\tau \max_{\delta \in \mathcal{D}_\tau} \mathbb{E}_{X_{t+1}^\tau}\left[\mathbb{P}\left(\delta\left(X^\tau\right) = Y \mid X^\tau\right) \,\middle|\, x^t\right].\end{aligned}$$

The first equality follows from what is sometimes called *tower property* [60], which is marginalization with respect to X_{t+1}^τ, followed by the total law of probability. Recall that the pay-off function β_τ is independent of the opinions $\forall \tau \in [T]$. Then, the following equality holds:

$$\max_{\delta \in \mathcal{D}_\tau} \mathbb{E}_{X_{t+1}^\tau}\left[\mathbb{P}\left(\delta\left(X^\tau\right) = Y \mid X^\tau\right) \,\middle|\, x^t\right] = \mathbb{E}_{X_{t+1}^\tau}\left[\max_{\delta \in \mathcal{D}_\tau} \mathbb{P}\left(\delta\left(X^\tau\right) = Y \mid X^\tau\right) \,\middle|\, x^t\right]. \tag{A.2}$$

The first inequality ($\leq$) follows since maximum of the average is dominated by the average of maxima:

$$\max_{\delta \in \mathcal{D}_\tau} \mathbb{E}_{X_{t+1}^\tau}\left[\mathbb{P}\left(\delta\left(X^\tau\right) = Y \mid X^\tau\right) \,\middle|\, x^t\right] \leq \mathbb{E}_{X_{t+1}^\tau}\left[\max_{\delta \in \mathcal{D}_\tau} \mathbb{P}\left(\delta\left(X^\tau\right) = Y \mid X^\tau\right) \,\middle|\, x^t\right].$$

For the second equality ($\geq$), observe that the following holds:

$$\max_{\delta\in\mathcal{D}_\tau} \mathbb{E}_{X_{t+1}^\tau}\left[\mathbb{P}\left(\delta\left(X^\tau\right)=Y \mid X^\tau\right) \,\middle|\, x^t\right] \geq \mathbb{E}_{X_{t+1}^\tau}\left[\mathbb{P}\left(\delta\left(X^\tau\right)=Y \mid X^\tau\right) \,\middle|\, x^t\right], \forall \delta\in\mathcal{D}_\tau.$$

Allowing $\delta_\tau^* = \arg\max_{\delta\in\mathcal{D}_\tau} \mathbb{P}\left(\delta\left(X^\tau\right)=Y \mid X^\tau=x^\tau\right)$, $\forall \tau \geq t$ yields (A.2). Conceptually, this amount to a simple rationale: when the optimal opinion-aggregation rule is achievable $\forall \tau \in [T]$, it must be applied at all times for optimal stopping. Further note that since δ_τ^* is indeed the MAP rule:

$$\begin{aligned}\max_{\delta\in\mathcal{D}_\tau} \mathbb{E}_{X_{t+1}^\tau}\left[\mathbb{P}\left(\delta\left(X^\tau\right)=Y \mid X^\tau\right) \,\middle|\, x^t\right] &= \mathbb{E}_{X_{t+1}^\tau}\left[\max_{\delta\in\mathcal{D}_\tau}\mathbb{P}\left(\delta\left(X^\tau\right)=Y \mid X^\tau\right) \,\middle|\, x^t\right], \\ &= \mathbb{E}_{X_{t+1}^\tau}\left[\max_{y\in\mathcal{Y}}\mathbb{P}\left(y \mid X^\tau\right) \,\middle|\, x^t\right].\end{aligned}$$

Note that we write $\mathbb{P}\left(Y=y \mid X^\tau\right) \equiv \mathbb{P}\left(y \mid X^t\right)$ for brevity and for emphasizing the random variables X_{t+1}^τ over which the expectation is taken. Then, the value function can be written as:

$$V_t\left(x^t\right) = \max_{\tau\geq t} \mathbb{E}_{X_{t+1}^\tau}\left[\beta_\tau \max_{y\in\mathcal{Y}}\mathbb{P}\left(y \mid X^\tau\right) \,\middle|\, x^t\right]. \tag{A.3}$$

When competences are known, $\beta_\tau \max_{y\in\mathcal{Y}}\mathbb{P}\left(y \mid x^\tau\right)$ is a deterministic function of x^t and since for a fixed ordering of experts X^t is a Markov process due to (2.2), the asserted Bellman equation follows. Nonetheless, in order to briefly explain the equality between (A.3) and the Bellman equation, note that:

$$\begin{aligned}&\mathbb{E}_{X_{t+1}}\left[V_{t+1}\left(X^{t+1}\right) \mid x^t\right] \\ &\quad= \mathbb{E}_{X_{t+1}}\left[\max_{\tau\geq t+1}\mathbb{E}_{X_{t+2}^\tau}\left[\beta_\tau \max_{y\in\mathcal{Y}}\mathbb{P}\left(y \mid X^\tau\right) \,\middle|\, x^{t+1}\right] \,\middle|\, x^t\right] \\ &\quad= \max_{\tau\geq t+1}\mathbb{E}_{X_{t+1}^\tau}\left[\beta_\tau \max_{y\in\mathcal{Y}}\mathbb{P}\left(y \mid X^\tau\right) \,\middle|\, x^t\right].\end{aligned}$$

The first equality follows directly from the value function (A.3) being evaluated at time $t+1$. For the second equality, observe that ($\geq$) follows from expectation of maxima dominating the maximum of the expectation and ($\leq$) follows from maximum of expected future rewards dominating all future rewards and hence, by letting:

$$\tau^* = \arg\max_{\tau\geq t+1}\mathbb{E}_{X_{t+1}^\tau}\left[\beta_\tau \max_{y\in\mathcal{Y}}\mathbb{P}\left(y \mid X^\tau\right) \,\middle|\, x^{t+1}\right].$$

A more detailed treatment for the Bellman equation for stopping a Markov process under known rewards can be found in [6, Section 3.4]. □

A.2 Motivation for Instantaneously-Realizable Statistics

Let us first remark that the notation X_t has been chosen deliberately to denote a *new* expert being consulted at a discrete time instance $t \in [T]$. As such, we refer to $t \in [T]$ as *time* rather than the *next expert consulted* for ease of presentation.

Observe that Lemma 2.1 yields that optimal stopping happens at time t given the opinions $X^t = x^t$ if:

$$\beta_t \max_{y \in \mathcal{Y}} \mathbb{P}\left(y \mid x^t\right) \geq \mathbb{E}\left[V_{t+1}\left(X^{t+1}\right) \mid x^t\right] = \max_{\tau \geq t} \mathbb{E}_{X_{t+1}^\tau}\left[\beta_\tau \max_{y \in \mathcal{Y}} \mathbb{P}\left(y \mid X^\tau\right) \,\middle|\, x^t\right]. \tag{A.4}$$

Note that $\mathbb{E}_{X_{t+1}^\tau}\left[\beta_\tau \max_{y \in \mathcal{Y}} \mathbb{P}\left(y \mid X^\tau\right) \mid x^t\right]$ is defined over the opinion process X_{t+1}^τ, the probability distribution of which coincides with that of a weighted random walk with up to $2^{\tau - t}$ possible outcomes at each time $\tau > t$. Therefore, computing (A.4) over all sample paths $X_{t+1}^\tau = x_{t+1}^\tau$ is challenging, and in its current form it yields limited insight into how optimal stopping and opinion aggregation problems are coupled. Here, we motivate the log-likelihood process L_t to address these computational and conceptual issues.

The expected *reward* at time τ, $\forall \tau > t$, given that $X^t = x^t$ is as follows:

$$\mathbb{E}_{X_{t+1}^\tau}\left[\beta_\tau \max_{y \in \mathcal{Y}} \mathbb{P}\left(Y = y \mid X^\tau\right) \,\middle|\, X^t = x^t\right] = \beta_\tau \mathbb{E}_{X_{t+1}^\tau}\left[\max_{y \in \mathcal{Y}} \frac{\mathbb{P}\left(X_{t+1}^\tau \mid y, x^t\right) \mathbb{P}\left(y \mid x^t\right)}{\mathbb{P}\left(X_{t+1}^\tau \mid x^t\right)} \,\middle|\, x^t\right].$$

Let $P_{X_{t+1}^\tau | X^t = x^t}$ be the conditional probability distribution of X_{t+1}^τ given $X^t = x^t$ and write the expectation explicitly:

$$\begin{aligned}
\mathbb{E}_{X_{t+1}^\tau}\left[\max_{y \in \mathcal{Y}} \mathbb{P}\left(y \mid X^\tau\right) \,\middle|\, x^t\right] &= \int \max_{y \in \mathcal{Y}} \frac{\mathbb{P}\left(X_{t+1}^\tau \mid y\right) \mathbb{P}\left(y \mid x^t\right)}{\mathbb{P}\left(X_{t+1}^\tau \mid x^t\right)} \, dP_{X_{t+1}^\tau | X^t = x^t}, \\
&= \sum_{x_{t+1}^\tau \in \mathcal{Y}^{\tau - t}} \max_{y \in \mathcal{Y}} \frac{\mathbb{P}\left(x_{t+1}^\tau \mid y\right) \mathbb{P}\left(y \mid x^t\right)}{\mathbb{P}\left(x_\tau^t \mid x^t\right)} \mathbb{P}\left(x_\tau^t \mid x^t\right), \\
&= \sum_{x_{t+1}^\tau \in \mathcal{Y}^{\tau - t}} \max_{y \in \mathcal{Y}} \mathbb{P}\left(x_{t+1}^\tau \mid y\right) \mathbb{P}\left(y \mid x^t\right).
\end{aligned}$$

Note that the number of sample paths x_{t+1}^τ increases (up to) exponentially in $\tau - t$. Since $\mathbb{P}\left(y \mid x^t\right)$ is not a function of future sample paths x_{t+1}^τ, one might interpret the summation above as the superposition of future probabilities of correct decision making defined with respect to the current probability of correct decision making. In order to observe this relation,

recall that the MAP decision (2.3) and note that:

$$\delta^*(x^\tau) = \arg\max_{y\in\mathcal{Y}} \mathbb{P}\left(x_{t+1}^\tau \;\middle|\; y\right)\mathbb{P}\left(y \;\middle|\; x^t\right) = \begin{cases} \delta^*(x^t) & \text{if } \frac{\mathbb{P}\left(x_{t+1}^\tau \;\middle|\; Y=\delta^*\left(x^t\right)\right)}{\mathbb{P}\left(x_{t+1}^\tau \;\middle|\; Y\neq\delta^*(x^t)\right)} > \frac{\min_{y\in\mathcal{Y}}\mathbb{P}\left(y \;\middle|\; x^t\right)}{\max_{y\in\mathcal{Y}}\mathbb{P}(y \mid x^t)}, \\ -\delta^*(x^t) & \text{otherwise,} \end{cases} \tag{A.5}$$

where the term $-\delta^*(x^t)$ follows from $\delta^*(x^t) \in \mathcal{Y} = \{\pm 1\}$. The relation in (A.5) yields that whether a decision y_t^* is changed or not is determined by a likelihood ratio test between the current and future likelihoods. This motivates the random process L_t that captures log-likelihood of correctness in (2.9). Appendices A.3-A.4 use (A.5) to formulate L_t and discuss its properties in detail.

A.3 Proof of Lemma 2.2

Proof. Since opinions are generated independently on a given task $(X_i - Y - X_j)$, one could note that an incoming opinion is conditionally independent from all of the past opinions. Formally, X_{t+1} is conditionally independent from X^t, $\left(X_{t+1} - Y - X^t\right)$, which leads to:

$$L_{t+1} = \log \frac{\max_{y\in\mathcal{Y}} \mathbb{P}\left(X_{t+1} \;\middle|\; y\right)\mathbb{P}\left(y \mid X^t\right)}{\max_{y\in\mathcal{Y}} \mathbb{P}\left(X_{t+1} \;\middle|\; y\right)\mathbb{P}\left(y \mid X^t\right)}.$$

Note that this equation follows from the Bayes' rule and it holds for any prior on Y. Therefore, L_{t+1} can be written as:

$$L_{t+1} = \begin{cases} L_t + \log \frac{\mathbb{P}\left(X_{t+1} \;\middle|\; Y=\delta^*\left(X^t\right)\right)}{\mathbb{P}\left(X_{t+1} \;\middle|\; Y\neq\delta^*\left(X^t\right)\right)} & \text{if} \quad \delta^*(X^{t+1}) = \delta^*(X^t)\,, \\ \log \frac{\mathbb{P}\left(X_{t+1} \;\middle|\; Y\neq\delta^*\left(X^t\right)\right)}{\mathbb{P}\left(X_{t+1} \;\middle|\; Y=\delta^*\left(X^t\right)\right)} - L_t & \text{if} \quad \delta^*(X^{t+1}) \neq \delta^*(X^t)\,. \end{cases}$$

Observe that the event of whether a decision is changed by an incoming opinion is characterized by (A.5). Given that $L_t = \ell_t$, $\delta^*(X^{t+1}) = \delta^*(X^t)$ if:

$$\log \frac{\mathbb{P}\left(X_{t+1} \;\middle|\; Y = \delta^*(X^t)\,,\ \ell_t\right)}{\mathbb{P}\left(X_{t+1} \;\middle|\; Y \neq \delta^*(X^t)\,,\ \ell_t\right)} > -\ell_t,$$

which directly follows from substituting $\tau = t+1$ in (A.5), conditioned on $L_t = \ell_t$, and writing it for random variables X^{t+1} rather than sample opinions x^{t+1}.

More importantly, observe that $\log \frac{\mathbb{P}\left(X_{t+1} \;\middle|\; Y=\delta^*\left(X^t\right),\ \ell_t\right)}{\mathbb{P}\left(X_{t+1} \;\middle|\; Y\neq\delta^*(X^t),\ \ell_t\right)}$ is a *random variable* with distri-

bution:

$$\begin{aligned} \log \frac{p_{t+1}}{q_{t+1}} \quad & \text{w.p.} \quad \mathbb{P}\left(X_{t+1} = \delta^*\big(X^t\big) \;\middle|\; L_t = \ell_t\right), \\ -\log \frac{p_{t+1}}{q_{t+1}} \quad & \text{w.p.} \quad \mathbb{P}\left(X_{t+1} \neq \delta^*\big(X^t\big) \;\middle|\; L_t = \ell_t\right). \end{aligned}$$

Let us denote the probability of an incoming opinion X_{t+1} agreeing with the current decision $\delta^*(X^t)$ given that $L_t = \ell_t$ by:

$$\alpha_{t+1} \triangleq \mathbb{P}\left(X_{t+1} = \delta^*\big(X^t\big) \;\middle|\; L_t = \ell_t\right). \tag{A.6}$$

Hence, $\bar{\alpha}_{t+1} \triangleq \mathbb{P}\left(X_{t+1} = \delta^*(X^t) \mid L_t \neq \ell_t\right) = 1 - \alpha_{t+1}$ is the probability of disagreeing. Consequently, the random process L_{t+1} can be written as:

$$L_{t+1} = \begin{cases} \ell_t + \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \alpha_{t+1} \mathbb{1}\left(\log \frac{p_{t+1}}{q_{t+1}} > -\ell_t\right), \\ \ell_t - \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \bar{\alpha}_{t+1} \mathbb{1}\left(\log \frac{p_{t+1}}{q_{t+1}} < \ell_t\right), \\ -\ell_t - \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \alpha_{t+1} \mathbb{1}\left(\log \frac{p_{t+1}}{q_{t+1}} < -\ell_t\right), \\ -\ell_t + \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \bar{\alpha}_{t+1} \mathbb{1}\left(\log \frac{p_{t+1}}{q_{t+1}} > \ell_t\right). \end{cases}$$

Depending on whether $p_{t+1} > 1/2$, L_t simplifies further. Specifically, when $p_{t+1} > 1/2$:

$$L_{t+1} = \begin{cases} \ell_t + \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \alpha_{t+1}, \\ \ell_t - \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \bar{\alpha}_{t+1} \mathbb{1}\left(\log \frac{p_{t+1}}{q_{t+1}} < \ell_t\right), \\ -\ell_t + \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \bar{\alpha}_{t+1} \mathbb{1}\left(\log \frac{p_{t+1}}{q_{t+1}} > \ell_t\right). \end{cases}$$

Similarly, when $p_{t+1} < 1/2$:

$$L_{t+1} = \begin{cases} \ell_t + \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \alpha_{t+1} \mathbb{1}\left(\log \frac{p_{t+1}}{q_{t+1}} > -\ell_t\right), \\ \ell_t - \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \bar{\alpha}_{t+1} \\ -\ell_t - \log \frac{p_{t+1}}{q_{t+1}} & \text{w.p.} \quad \alpha_{t+1} \mathbb{1}\left(\log \frac{p_{t+1}}{q_{t+1}} < -\ell_t\right). \end{cases}$$

Recall the definition of θ_{t+1} in (2.11) observe that both cases can be written in the form:

$$L_{t+1} = \begin{cases} \ell_t + \theta_{t+1}, & \text{w.p.} \quad \tilde{p}_{t+1}, \\ \left|\ell_t - \theta_{t+1}\right| & \text{w.p.} \quad \tilde{q}_{t+1}, \end{cases}$$

where $(\tilde{p}_{t+1}, \tilde{q}_{t+1})$ are *some* transition probabilities (not yet proven to be in the form (2.13)).

It is sufficient to show that:

$$\tilde{p}_{t+1} = \max\left(\alpha_{t+1}, \bar{\alpha}_{t+1}\right),$$
$$\tilde{q}_{t+1} = \min\left(\alpha_{t+1}, \bar{\alpha}_{t+1}\right),$$

and that α_{t+1} is a deterministic function of $\left(\ell_t, \theta_{t+1}\right)$ to conclude that L_t is a Markov process. In order to formulate α_{t+1} explicitly, first note that for any sample path x^t that satisfies $L_t = \ell_t$ the following holds:

$$\ell_t = \frac{\max\limits_{y\in\mathcal{Y}} \mathbb{P}\left(y \mid x^t\right)}{\min\limits_{y\in\mathcal{Y}} \mathbb{P}\left(y \mid x^t\right)} = \frac{\mathbb{P}\left(Y = \delta^*(x^t)\right)}{\mathbb{P}\left(Y \neq \delta^*(x^t)\right)}.$$

Therefore, one can conclude:

$$\mathbb{P}\left(Y = \delta^*\left(x^t\right)\right) = \mathbb{P}\left(Y = \delta^*\left(X^t\right) \;\middle|\; L_t = \ell_t\right),$$

which yields by (2.10) that for all such set of opinions x^t:

$$\mathcal{P}\left(\ell_t\right) = \mathbb{P}\left(Y = \delta^*\left(X^t\right) \;\middle|\; \ell_t\right) = \max_{y\in\mathcal{Y}} \mathbb{P}\left(y \;\middle|\; x^t\right).$$

The rest follows from the law of total probability:

$$\begin{aligned}
\alpha_{t+1} &= \mathbb{P}\left(X_{t+1} = \delta^*\left(X^t\right) \;\middle|\; L_t = \ell_t\right), \\
&= \mathbb{P}\left(X_{t+1} = \delta^*\left(X^t\right) \;\middle|\; Y = \delta^*\left(X^t\right), \ell_t\right) \mathbb{P}\left(Y = \delta^*\left(X^t\right) \;\middle|\; \ell_t\right) \\
&\quad + \mathbb{P}\left(X_{t+1} = \delta^*\left(X^t\right) \;\middle|\; Y \neq \delta^*\left(X^t\right), \ell_t\right) \mathbb{P}\left(Y \neq \delta^*\left(X^t\right) \;\middle|\; \ell_t\right), \\
&= \mathbb{P}\left(X_{t+1} = Y\right) \mathbb{P}\left(Y = \delta^*\left(X^t\right) \;\middle|\; \ell_t\right) + \mathbb{P}\left(X_{t+1} \neq Y\right) \mathbb{P}\left(Y \neq \delta^*\left(X^t\right) \;\middle|\; \ell_t\right), \\
&= p_{t+1}\mathcal{P}\left(\ell_t\right) + q_{t+1}\mathcal{Q}\left(\ell_t\right).
\end{aligned}$$

In order to conclude that the transition probabilities have the form given in (2.13), first note that the following relation holds:

$$p_{t+1} \gtrless 1/2 \iff \alpha_{t+1} \gtrless \bar{\alpha}_{t+1},$$

which follows from $\mathcal{P}\left(\ell_t\right) > \mathcal{Q}\left(\ell_t\right)$, see (2.9)-(2.10), via:

$$\begin{aligned}
\alpha_{t+1} &\gtrless \bar{\alpha}_{t+1}, \\
p_{t+1}\mathcal{P}\left(\ell_t\right) + q_{t+1}\mathcal{Q}\left(\ell_t\right) &\gtrless q_{t+1}\mathcal{P}\left(\ell_t\right) + p_{t+1}\mathcal{Q}\left(\ell_t\right), \\
\left(p_{t+1} - q_{t+1}\right)\mathcal{P}\left(\ell_t\right) &\gtrless \left(p_{t+1} - q_{t+1}\right)\mathcal{Q}\left(\ell_t\right).
\end{aligned}$$

Further note that when $p_{t+1} > 1/2$, $\mathcal{P}\left(\theta_{t+1}\right) = p_{t+1}$ and hence, the probability of *agreement* is:

$$\alpha_{t+1} = \mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right) + \mathcal{Q}\left(\theta_{t+1}\right)\mathcal{Q}\left(\ell_t\right).$$

Similarly, when $p_{t+1} < 1/2$, $\mathcal{P}\left(\theta_{t+1}\right) = q_{t+1}$ and hence, the probability of *disagreement* is:

$$\bar{\alpha}_{t+1} = \mathcal{P}\left(\theta_{t+1}\right)\mathcal{Q}\left(\ell_t\right) + \mathcal{Q}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right),$$

which concludes that $\tilde{p}_{t+1} = \max\left(\alpha_{t+1}, \bar{\alpha}_{t+1}\right)$ takes the from in (2.13). Finally, given that $L_t = \ell_t$, L_{t+1} is independent of all past likelihoods L_τ, $\tau < t$ therefore, it is a Markov process with its state-transition given in (2.12). □

A.4 Proof of Theorem 2.1

Let us start with a preliminary lemma about the relation between the sigmoid function $\mathcal{P}(\cdot)$, transition probabilities $(\tilde{p}_{t+1}, \tilde{q}_{t+1})$ and states $\ell_{t+1} \in \left\{\ell_t + \theta_{t+1}, \left|\ell_t - \theta_{t+1}\right|\right\}$.

Lemma A.1. *The following equalities hold:*

$$\begin{aligned}
\tilde{p}_{t+1}\mathcal{P}\left(\ell_t + \theta_{t+1}\right) &= \mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right), \\
\tilde{q}_{t+1}\mathcal{P}\left(\ell_t - \theta_{t+1}\right) &= \mathcal{Q}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right), \\
\tilde{q}_{t+1}\mathcal{P}\left(\theta_{t+1} - \ell_t\right) &= \mathcal{P}\left(\theta_{t+1}\right)\mathcal{Q}\left(\ell_t\right).
\end{aligned}$$

Proof of Lemma A.1. Direct substitution of (2.10) into $\mathcal{P}(\cdot)$ terms in (2.10) yields that:

$$\tilde{p}_{t+1} = \left[\mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right) + \mathcal{Q}\left(\theta_{t+1}\right)\mathcal{Q}\left(\ell_t\right)\right] = \frac{1 + e^{-\left(\ell_t + \theta_{t+1}\right)}}{\left(1 + e^{-\ell_t}\right)\left(1 + e^{-\theta_{t+1}}\right)}. \tag{A.7}$$

Similarly, the following holds:

$$\tilde{q}_{t+1} = \left[\mathcal{Q}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right) + \mathcal{P}\left(\theta_{t+1}\right)\mathcal{Q}\left(\ell_t\right)\right] = \frac{e^{-\ell_t} + e^{-\theta_{t+1}}}{\left(1 + e^{-\ell_t}\right)\left(1 + e^{-\theta_{t+1}}\right)}. \tag{A.8}$$

Therefore, the expressions (A.7)-(A.8) yield that:

$$\tilde{p}_{t+1}\mathcal{P}\left(\ell_t+\theta_{t+1}\right)=\tilde{p}_{t+1}\frac{1}{1+e^{-\left(\ell_t+\theta_{t+1}\right)}}=\frac{1}{\left(1+e^{-\ell_t}\right)\left(1+e^{-\theta_{t+1}}\right)}=\mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right),$$

$$\tilde{q}_{t+1}\mathcal{P}\left(\ell_t-\theta_{t+1}\right)=\tilde{q}_{t+1}\frac{1}{1+e^{-\left(\ell_t-\theta_{t+1}\right)}}=\frac{e^{-\theta_{t+1}}}{\left(1+e^{-\ell_t}\right)\left(1+e^{-\theta_{t+1}}\right)}=\mathcal{Q}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right),$$

$$\tilde{q}_{t+1}\mathcal{P}\left(\theta_{t+1}-\ell_t\right)=\tilde{q}_{t+1}\frac{1}{1+e^{-\left(\theta_{t+1}-\ell_t\right)}}=\frac{e^{-\ell_t}}{\left(1+e^{-\ell_t}\right)\left(1+e^{-\theta_{t+1}}\right)}=\mathcal{P}\left(\theta_{t+1}\right)\mathcal{Q}\left(\ell_t\right).$$

□

Next, we prove Theorem 2.1 by backwards induction from $t=T$. The proof relies on Lemma 2.2, and makes use of Lemma A.1.

Proof of Theorem 2.1. First, observe that the value function in (A.3) can be written as:

$$V_t\left(x^t\right)=\max_{\tau\geq t}\mathbb{E}_{X_{t+1}^{\tau}}\left[\beta_\tau\max_{y\in\mathcal{Y}}\mathbb{P}\left(y\mid X^\tau\right)\,\middle|\,x^t\right]=\max_{\tau\geq t}\mathbb{E}_{L_t}\left[\beta_\tau\mathcal{P}\left(L_\tau\right)\mid L_t=\ell_t\right]\equiv V_t\left(\ell_t\right).$$

The second equality follows from the definition of L_τ in (2.9) and the expectation being taken with respect to the likelihood process $\mathbb{E}\left[L_\tau\right]\cdot$ is a direct consequence of Lemma 2.2. The corresponding Bellman equation is written as:

$$V_t\left(\ell_t\right)=\max\left(\beta_t\mathcal{P}\left(\ell_t\right),\mathbb{E}_{L_{t+1}}\left[V_{t+1}\left(L_{t+1}\right)\mid L_t=\ell_t\right]\right).$$

It is sufficient to show that the value function is of the form:

$$V_t\left(\ell_t\right)=\beta_t\mathcal{P}\left(\max\left(\ell_t,\eta_t\right)\right),\forall t\in[T],$$

where η_t is given in (2.16). We argue by mathematical induction.

Note that at time $t=T$, the value function yields the expected payoff:

$$V_T\left(\ell_T\right)=\beta_T\mathcal{P}\left(\ell_T\right)=\beta_T\mathcal{P}\left(\max\left(\ell_T,0\right)\right),$$

which follows since $L_T\geq 0$ almost surely. Hence, the base step of the induction follows trivially.

Inductive argument is constructed backwards starting from $t=T$. Assume that:

$$V_{t+1}\left(\ell_{t+1}\right)=\beta_{t+1}\mathcal{P}\left(\max\left(\ell_{t+1},\eta_{t+1}\right)\right), \tag{A.9}$$

for some $\eta_{t+1} \in \mathbb{R}$ and observe that the expected future value at time t can be written as:

$$\begin{aligned}\mathbb{E}_{L_{t+1}}\left[V_{t+1}\left(L_{t+1}\right) \mid \ell_t\right] &= \beta_{t+1}\left[\tilde{p}_{t+1}V_{t+1}\left(\ell_t+\theta_{t+1}\right)+\tilde{q}_{t+1}V_{t+1}\left(\left|\ell_t-\theta_{t+1}\right|\right)\right], \\ &= \beta_{t+1}\left[\tilde{p}_{t+1}\mathcal{P}\left(\max\left(\ell_t+\theta_{t+1},\eta_{t+1}\right)\right)+\tilde{q}_{t+1}\mathcal{P}\left(\max\left(\left|\ell_t-\theta_{t+1}\right|,\eta_{t+1}\right)\right)\right].\end{aligned}$$

It is useful to address the cases where $\eta_{t+1} \gtrless \theta_{t+1}$ separately. If $-\infty < \eta_{t+1} < \theta_{t+1}$, one can write:

$$\begin{aligned}\mathbb{E}_{L_{t+1}}\left[V_{t+1}\left(L_{t+1}\right) \mid \ell_t\right] &= \beta_{t+1}\left[\begin{array}{l}\tilde{p}_{t+1}\mathcal{P}\left(\ell_t+\theta_{t+1}\right) \\ +\tilde{q}_{t+1}\mathcal{P}\left(\ell_t-\theta_{t+1}\right)\mathbb{1}\left(\ell_t \geq \theta_{t+1}+\eta_{t+1}\right) \\ +\tilde{q}_{t+1}\mathcal{P}\left(\eta_{t+1}\right)\mathbb{1}\left(\theta_{t+1}-\eta_{t+1} \leq \ell_t < \theta_{t+1}+\eta_{t+1}\right) \\ +\tilde{q}_{t+1}\mathcal{P}\left(\theta_{t+1}-\ell_t\right)\mathbb{1}\left(0 \leq \ell_t < \theta_{t+1}-\eta_{t+1}\right)\end{array}\right], \\ &= \beta_{t+1}\left[\begin{array}{l}\mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right) \\ +\mathcal{Q}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right)\mathbb{1}\left(\ell_t \geq \theta_{t+1}+\eta_{t+1}\right) \\ +\tilde{q}_{t+1}\mathcal{P}\left(\eta_{t+1}\right)\mathbb{1}\left(\theta_{t+1}-\eta_{t+1} \leq \ell_t < \theta_{t+1}+\eta_{t+1}\right) \\ +\mathcal{P}\left(\theta_{t+1}\right)\mathcal{Q}\left(\ell_t\right)\mathbb{1}\left(1 \leq \ell_t < \theta_{t+1}-\eta_{t+1}\right)\end{array}\right].\end{aligned}$$

The first equality follows from (A.9) and the second equality follows from Lemma A.1. Therefore, the value function can be written as:

$$V_t\left(\ell_t\right) = \max\left(\begin{array}{l}\beta_t\mathcal{P}\left(\ell_t\right), \beta_{t+1}\left[\begin{array}{l}\mathcal{P}\left(\ell_t\right)\mathbb{1}\left(\ell_t \geq \theta_{t+1}+\eta_{t+1}\right) \\ +\left[\mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right)+\tilde{q}_{t+1}\mathcal{P}\left(\eta_{t+1}\right)\right] \\ \quad \times\mathbb{1}\left(\theta_{t+1}-\eta_{t+1} \leq \ell_t < \theta_{t+1}+\eta_{t+1}\right) \\ +\mathcal{P}\left(\theta_{t+1}\right)\mathbb{1}\left(0 \leq \ell_t < \theta_{t+1}-\eta_{t+1}\right)\end{array}\right]\end{array}\right).$$

Clearly, $\forall \ell_t \geq \theta_{t+1}+\eta_{t+1}$, the value function $V_t\left(\ell_t\right) = \beta_t\mathcal{P}\left(\ell_t\right)$ as $\beta_t \geq \beta_{t+1}$. For other values of ℓ_t, first observe that:

$$\beta_t\mathcal{P}\left(\ell_t\right) \gtrless \beta_{t+1}\left[\mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\ell_t\right)+\tilde{q}_{t+1}\mathcal{P}\left(\eta_{t+1}\right)\right] \iff \ell_t \gtrless \log\frac{\mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\eta_{t+1}\right)}{\frac{\beta_t-\beta_{t+1}}{\beta_{t+1}}+\mathcal{Q}\left(\theta_{t+1}\right)\mathcal{Q}\left(\eta_{t+1}\right)}.$$

Define the relative cost for consulting per expert:

$$\delta_{t+1} \triangleq \frac{\beta_t-\beta_{t+1}}{\beta_{t+1}} \geq 0$$

and note that:

$$\log\frac{\mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1}+\mathcal{Q}\left(\theta_{t+1}\right)\mathcal{Q}\left(\eta_{t+1}\right)} \leq \theta_{t+1}+\eta_{t+1},$$

with equality if and only if $\beta_t = \beta_{t+1}$. Finally,

$$\beta_t \mathcal{P}\left(\ell_t\right) \gtrless \beta_{t+1} \mathcal{P}\left(\theta_{t+1}\right) \iff \ell_t \gtrless \log \frac{\mathcal{P}\left(\theta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right)}.$$

It appears that $V_t\left(\ell_t\right) = \beta_t \mathcal{P}\left(\ell_t\right)$ when either of the following conditions are satisfied:

$$\begin{aligned} \ell_t &> \log \frac{\mathcal{P}\left(\theta_{t+1}\right) \mathcal{P}\left(\eta_{t+1}\right)}{\frac{\beta_t - \beta_{t+1}}{\beta_{t+1}} + \mathcal{Q}\left(\theta_{t+1}\right) \mathcal{Q}\left(\eta_{t+1}\right)}, \\ \theta_{t+1} - \eta_{t+1} > \ell_t &> \log \frac{\mathcal{P}\left(\theta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right)}. \end{aligned}$$

A key observation yields that these conditions yield a unique threshold:

$$\begin{aligned} \frac{\mathcal{P}\left(\theta_{t+1}\right) \mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right) \mathcal{Q}\left(\eta_{t+1}\right)} \gtrless \frac{\mathcal{P}\left(\theta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right)} &\iff \eta_{t+1} \gtrless \log\left(1 + \frac{\delta_{t+1}}{\mathcal{Q}\left(\theta_{t+1}\right)}\right), \\ &\iff \log \frac{p_{t+1} \mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + q_{t+1} \mathcal{Q}\left(\eta_{t+1}\right)} \gtrless \theta_{t+1} - \eta_{t+1}. \end{aligned}$$

In words, the equivalence above indicates that the likelihood is compared to the dominant threshold in each interval, which formally yields that when $-\infty < \eta_{t+1} < \theta_{t+1}$:

$$\eta_t = \log \max\left(\frac{\mathcal{P}\left(\theta_{t+1}\right) \mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right) \mathcal{Q}\left(\eta_{t+1}\right)}, \frac{\mathcal{P}\left(\theta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right)}\right).$$

A similar behavior manifests itself when $\eta_{t+1} > \theta_{t+1}$. Observe that:

$$\begin{aligned} \mathbb{E}_{L_{t+1}}\left[V_{t+1}\left(L_{t+1}\right) \mid \ell_t\right] &= \beta_{t+1} \left[\begin{array}{l} \tilde{p}_{t+1} \mathcal{P}\left(\max\left(\ell_t + \theta_{t+1}, \eta_{t+1}\right)\right) \\ +\tilde{q}_{t+1} \mathcal{P}\left(\max\left(\ell_t - \theta_{t+1}, \eta_{t+1}\right)\right) \mathbb{1}\left(\ell_t \geq \theta_{t+1}\right) \\ +\tilde{q}_{t+1} \mathcal{P}\left(\eta_{t+1}\right) \mathbb{1}\left(1 \leq \ell_t < \theta_{t+1}\right) \end{array}\right], \\ &= \beta_{t+1} \left[\begin{array}{l} \tilde{p}_{t+1} \mathcal{P}\left(\ell_t + \theta_{t+1}\right) \mathbb{1}\left(\ell_t \geq \eta_{t+1} - \theta_{t+1}\right) \\ +\tilde{p}_{t+1} \mathcal{P}\left(\eta_{t+1}\right) \mathbb{1}\left(\ell_t < \eta_{t+1} - \theta_{t+1}\right) \\ +\tilde{q}_{t+1} \mathcal{P}\left(\ell_t - \theta_{t+1}\right) \mathbb{1}\left(\ell_t \geq \eta_{t+1} + \theta_{t+1}\right) \\ +\tilde{q}_{t+1} \mathcal{P}\left(\eta_{t+1}\right) \mathbb{1}\left(\theta_{t+1} \leq \ell_t < \eta_{t+1} + \theta_{t+1}\right) \\ +\tilde{q}_{t+1} \mathcal{P}\left(\eta_{t+1}\right) \mathbb{1}\left(\ell_t < +\theta_{t+1}\right) \end{array}\right]. \end{aligned}$$

Lemma A.1 yields that the value function in this case attains the following form:

$$V_t(\ell_t) = \max\left(\beta_t \mathcal{P}(\ell_t), \beta_{t+1}\left[\begin{array}{l} \mathcal{P}(\ell_t)\mathbb{1}\left(\ell_t \geq \eta_{t+1} + \theta_{t+1}\right) \\ + \left[\tilde{p}_{t+1}\mathcal{P}\left(\eta_{t+1}\right) + p_{t+1}\mathcal{P}(\ell_t)\right] \\ \times \mathbb{1}\left(\eta_{t+1} - \theta_{t+1} \leq \ell_t < \eta_{t+1} + \theta_{t+1}\right) \\ + \mathcal{P}\left(\eta_{t+1}\right)\mathbb{1}\left(1 \leq \ell_t < \eta_{t+1} - \theta_{t+1}\right) \end{array}\right]\right).$$

Similar to the previous discussion, observe that:

$$\beta_t \mathcal{P}(\ell_t) \gtrless \beta_{t+1}\mathcal{P}\left(\eta_{t+1}\right) \iff \ell_t \gtrless \log \frac{\mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\eta_{t+1}\right)}.$$

Furthermore, the following equivalence holds:

$$\begin{aligned} \frac{\mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right)\mathcal{Q}\left(\eta_{t+1}\right)} \gtrless \frac{\mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\eta_{t+1}\right)} &\iff \theta_{t+1} \gtrless \log\left(1 + \frac{\delta_{t+1}}{\mathcal{Q}\left(\eta_{t+1}\right)}\right) \\ &\iff \frac{\mathcal{P}\left(\theta_{t+1}\right)\mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\theta_{t+1}\right)\mathcal{Q}\left(\eta_{t+1}\right)} \gtrless \eta_{t+1} - \theta_{t+1}, \end{aligned}$$

which indicates that as long as $\eta_{t+1} \geq \theta_{t+1}$:

$$\eta_t = \max\left(\frac{p_{t+1}\mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + q_{t+1}\mathcal{Q}\left(\eta_{t+1}\right)}, \frac{\mathcal{P}\left(\eta_{t+1}\right)}{\delta_{t+1} + \mathcal{Q}\left(\eta_{t+1}\right)}\right).$$

Consequently, at any time $t \in [T]$, for all competences and competence orderings $\{\theta_1, \cdots, \theta_T\}$, and non-increasing pay-off functions β_t, (2.16) establishes a recursively computable fixed (with respect to the likelihood) threshold and it determines the optimal stopping time. □

A.5 Proof of Lemma 2.3

Proof of Existence. Formally one aims to prove that if $\forall u$: $f_U(u) > 0$, $\exists x$ such that $h(x) = g(x,u)$ then, $\exists x_0 : h(x_0) = \mathbb{E}[U]\, g(x_0, U)$. We argue by the contrapositive of the statement: If $\forall x$, $h(x) > \mathbb{E}[U]\, g(x,U)$ then, $h(x) > g(x,u)$, $\forall u$: $f_U(u) > 0$. Let $\mathcal{A}$ be a set with $\mathbb{P}(U \in \mathcal{A}) = \varepsilon > 0$ and note that $\forall x$:

$$\mathbb{E}[U]\, g(x,U) = \int_{\mathcal{A}} g(x,U)\, df_U + \int_{\mathcal{A}^c} g(x,U)\, df_U.$$

Here, $\mathcal{A}^c$ is the set complement of $\mathcal{A}$. Similarly,

$$h(x) = h(x)\mathbb{P}\left(U \in \mathcal{A}\right) + h(x)\mathbb{P}\left(U \in \mathcal{A}^c\right).$$

Since $h\left(x\right) > \mathbb{E}\left[U\right] g\left(x, U\right)$, $\forall x$, it follows that:

$$\frac{h(x) - \frac{1}{\mathbb{P}(U \in \mathcal{A})} \int\limits_{\mathcal{A}} g\left(x, U\right)\, df_U}{-h(x) + \frac{1}{\mathbb{P}(U \in \mathcal{A}^c)} \int\limits_{\mathcal{A}^c} g\left(x, U\right)\, df_U} > \frac{\mathbb{P}\left(U \in \mathcal{A}^c\right)}{\mathbb{P}\left(U \in \mathcal{A}\right)} > 0,$$

over the set $\left\{x : h(x) \neq \frac{1}{\mathbb{P}(U \in \mathcal{A}^c)} \int\limits_{\mathcal{A}^c} g\left(x_0, U\right)\, df_U\right\}$. Therefore, $\forall x$ the following holds:

$$h(x) > \frac{1}{\mathbb{P}\left(U \in \mathcal{A}\right)} \int\limits_{\mathcal{A}} g\left(x, U\right)\, df_U,$$

which yields that $h(x) > g(x, u)$, $\forall u : \; f_U(u) > 0$ by allowing the probability of $\mathcal{A}$ to be arbitrary. □

Proof of Uniqueness. **Argue uniqueness** – We argue by contradiction. Assume that $\exists x_1 \neq x_2$ such that $h(x_i) = \mathbb{E}\left[U\right] g(x_i, U)$ for $i \in \{1, 2\}$. □

APPENDIX B

PROOFS FOR CHAPTER 3

B.1 Proof of Proposition 3.1

Proof of Part (1)-Ordering. Consider any pairs of experts (X_i, X_j) for $i \neq j$, and allow $\eta_i = \mathbb{1}(X_i = Y)$. Observe that $\eta_i \perp \eta_j, \forall i \neq j$ are Bernoulli random variables with parameter p_i, denoted by $\mathcal{B}(p_i)$. Successive application of the law of total probability yields that:

$$\begin{aligned}\tilde{p}_i = \mathbb{P}\left(X_i = f^{MV}\left(\mathbf{X}_{\setminus i}\right)\right) &= \sum_{\eta_i} \mathbb{P}\left(X_i = f^{MV}\left(\mathbf{X}_{\setminus i}\right) \mid \eta_i\right) \mathbb{P}(\eta_i) \\ &= \sum_{\eta_i, \eta_j} \mathbb{P}\left(X_i = f^{MV}\left(\mathbf{X}_{\setminus i}\right) \mid \eta_i, \eta_j\right) \mathbb{P}(\eta_j \mid \eta_i) \mathbb{P}(\eta_i).\end{aligned}$$

Observe that $\mathbb{P}(\eta_j \mid \eta_i) = \mathbb{P}(\eta_j)$ due to the conditional independence of opinions (hence the independence of opinion *generation* η_i). A similar extension of $\tilde{p}_j$ yields that:

$$\tilde{p}_j = \sum_{\eta_i, \eta_j} \mathbb{P}\left(X_j = f^{MV}\left(\mathbf{X}_{\setminus j}\right) \mid \eta_i, \eta_j\right) \mathbb{P}(\eta_i) \mathbb{P}(\eta_j).$$

Since the rest of the committee is arbitrary, yet fixed, the following conditional probabilities are equal:

$$\begin{aligned}\mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus i}\right) = Y \mid X_j = Y\right) &= \mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus j}\right) = Y \mid X_i = Y\right), \\ \mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus i}\right) \neq Y \mid X_j \neq Y\right) &= \mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus j}\right) \neq Y \mid X_i \neq Y\right),\end{aligned}$$

and $p_i q_j - p_j q_i = (p_i - p_j)$. Therefore, the ratio between the differences between pseudo competences and true competences are given by:

$$\begin{aligned}&{}^{\tilde{p}_i - \tilde{p}_j}/_{p_i - p_j} = \\ &\left[\mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus j}\right) = Y \mid X_i \neq Y\right) - \mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus j}\right) \neq Y \mid X_i = Y\right)\right].\end{aligned} \tag{B.1}$$

As long as (B.1) is positive, pseudo competence preserves ordering. We now show that (B.1) is monotonically increasing in $p_i \in (1/2, 1)$, $\forall i$ and the minimum is zero. Observe that total law of probability yields that $\forall n \in [N]$:

$$\begin{aligned}\frac{\partial}{\partial p_k}\mathbb{P}\left(\sum_{i=1}^{N}\eta_i \geq n\right) &= \frac{\partial}{\partial p_k}\left[p_k\mathbb{P}\left(\sum_{i\neq k}\eta_i \geq n-1\right) + q_k\mathbb{P}\left(\sum_{i\neq k}\eta_i \geq n\right)\right] \\ &= \mathbb{P}\left(\sum_{i\neq k}\eta_i \geq n-1\right) - \mathbb{P}\left(\sum_{i\neq k}\eta_i \geq n\right). \end{aligned} \tag{B.2}$$

Therefore,

$$\frac{\partial}{\partial p_k}\mathbb{P}\left(\sum_{i=1}^{N}\eta_i \geq n\right) = \mathbb{P}\left(\sum_{i\neq k}\eta_i = n-1\right) \geq 0, \tag{B.3}$$

$$\frac{\partial}{\partial p_k}\mathbb{P}\left(\sum_{i=1}^{N}\eta_i \leq n\right) = -\mathbb{P}\left(\sum_{i\neq k}\eta_i = n\right) \leq 0. \tag{B.4}$$

Consequently, $\mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus j}\right) = Y \;\middle|\; X_i \neq Y\right)$ decreases in p_i for any expert, where the probability $\mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus j}\right) \neq Y \;\middle|\; X_i = Y\right)$ increases. Therefore,

$$\min_{\substack{\mathbf{p}\\ p_i > 1/2}} \frac{\tilde{p}_i - \tilde{p}_j}{p_i - p_j} = \lim_{\mathbf{p}\to \mathbf{1}/2} \frac{\tilde{p}_i - \tilde{p}_j}{p_i - p_j} = 0,$$

which yields that $\tilde{p}_i > \tilde{p}_j \iff p_i > p_j$, if $p_i > 1/2$, $\forall i$. □

Conceptually, (B.3)-(B.4) dictate that increasing the competence of an expert necessarily decreases the probability of error for majority vote. One should note that this does not contradict the discussion in Section 3.2 as the consistency of majority vote is concerned with *adding* a new expert, which does *not* ensure any monotonicity, instead of *increasing* the competence of an expert, which, as shown, *does* ensure monotonicity.

Proof of Part (2). Observe that $p_{\setminus i} > 1/2$ for all finite good committees with $p_i > 1/2$, $\forall i$. Then, equation (3.12) yields that:

$$\begin{aligned}\tilde{p}_i - p_i &= p_i p_{\setminus i} + (1-p_i)(1-p_{\setminus i}) - p_i \\ &= (1-2p_i)(1-p_{\setminus i}) < 0, \end{aligned} \tag{B.5}$$

$$\begin{aligned}\tilde{p}_i - 1/2 &= (p_i - 1/2)p_{\setminus i} + (1/2 - p_i)(1-p_{\setminus i}) \\ &= (p_i - 1/2)(2p_{\setminus i} - 1) > 0. \end{aligned} \tag{B.6}$$

Hence, if $p_i > 1/2$, $\forall i$, then $1/2 < \tilde{p}_i < p_i$. □

B.2 Proof of Proposition 3.2

Proof of Part (1). Equation (B.1) indicates that any committee satisfying

$$\mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus j}\right) = Y \mid X_i \neq Y\right) > \mathbb{P}\left(f^{MV}\left(\mathbf{X}_{\setminus j}\right) \neq Y \mid X_i = Y\right)$$

preserves ordering. For every consistent committee, $\exists N^*$ such that $\forall N > N^*$

$$\mathbb{P}\left(\sum_{k\neq i,j} \eta_k > \left\lfloor \frac{N-2}{2} \right\rfloor\right) > \mathbb{P}\left(\sum_{k\neq i,j} \eta_k < \left\lceil \frac{N-2}{2} \right\rceil\right),$$

which yields that pseudo competence preserves ordering for consistent mixed committees. □

Proof of Part (2). For every consistent committee, $\exists N^*$ such that $\forall N > N^*$, $p_{\setminus i} > 1/2$. Recall equations (B.5)-(B.6), which yield that:

$$\begin{aligned} 1/2 < \tilde{p}_i < p_i \quad &\text{if} \quad p_i > 1/2, \\ 1/2 > \tilde{p}_i > p_i \quad &\text{if} \quad p_i < 1/2, \\ 1/2 = p_i = \tilde{p}_i \quad &\text{if} \quad p_i = 1/2. \end{aligned}$$

Therefore, $\min\{\tilde{p}_i, 1-\tilde{p}_i\} \geq \min\{p_i, 1-p_i\}$ for consistent mixed committees. □

B.3 Proof Theorem 3.1

Proof Theorem 3.1. Let $w_i = \log p_i/q_i$. Chernoff bounding technique yields that [63, Section 2.2.1]:

$$\mathbb{P}\left(f^{NB}\left(\mathbf{X}\right) \neq Y\right) \leq e^{-t\Phi}\mathbb{E}\left[\exp\left(-t\sum_{i=1}^{N} w_i\left(\eta_i - p_i\right)\right)\right].$$

Observe that the expectation is with respect to $\eta_i \sim \mathcal{B}\left(p_i\right)$:

$$\mathbb{E}\left[e^{-t\sum_i w_i(\eta_i - p_i)}\right] = p_i e^{-q_i w_i t} + q_i e^{p_i w_i t}.$$

Therefore, the probability of error for the NB rule:

$$\begin{aligned}\mathbb{P}\left(f^{NB}\left(\mathbf{X}\right)\neq Y\right) &\leq e^{-t\Phi}\prod_i\left(p_i e^{-q_i w_i t}+q_i e^{p_i w_i t}\right)\\ &= \exp\underbrace{\left(-t\Phi+\textstyle\sum_i \log\left(p_i e^{-q_i w_i t}+q_i e^{p_i w_i t}\right)\right)}_{\triangleq -t\Phi+\phi(t;\boldsymbol{p})}. \end{aligned} \tag{B.7}$$

The derivative of $-t\Phi+\phi(t;\boldsymbol{p})$ is given as follows:

$$\begin{aligned}\frac{\partial}{\partial t}\left(\phi(t;\boldsymbol{p})-t\Phi\right) &= \sum_i w_i\left[\frac{-p_i q_i e^{-q_i w_i t}+p_i q_i e^{p_i w_i t}}{p_i e^{-q_i w_i t}+q_i e^{p_i w_i t}}-\left(p_i-1/2\right)\right]\\ &= \sum_i \frac{w_i}{2}\left[\frac{-p_i e^{-q_i w_i t}+q_i e^{p_i w_i t}}{p_i e^{-q_i w_i t}+q_i e^{p_i w_i t}}\right],\end{aligned}$$

which yields that (B.7) is minimized when $t=1$ since:

$$-p_i e^{-q_i\log\frac{p_i}{q_i}}+q_i e^{p_i\log\frac{p_i}{q_i}} = -p_i\left(\frac{p_i}{q_i}\right)^{-q_i}+q_i\left(\frac{p_i}{q_i}\right)^{p_i} = -{q_i}^{q_i}{p_i}^{p_i}+{p_i}^{p_i}{q_i}^{q_i}=0.$$

Hence,

$$\begin{aligned}\mathbb{P}\left(f^{NB}\left(\mathbf{X}\right)\neq Y\right) &\leq \exp\left(-\Phi+\phi(1;\boldsymbol{p})\right)\\ &= \exp\left(-\Phi+\sum_i \log 2q_i\left(\frac{p_i}{q_i}\right)^{p_i}\right)\\ &= \exp\left(\sum_i \log 2\sqrt{q_i p_i}\right),\end{aligned}$$

yielding the asserted bound. □

It is important to note that this bound is the sharpest possible using Chernoff bounding technique and it is a direct consequence of the weight function $w(p)=\log {}^{p}/_{q}$.

B.4 Proof of Theorem 3.2

Proof of Theorem 3.2. Let $\varepsilon_i = p_i - 1/2$, hence $\tilde{p}_i = 1/2 + 2\varepsilon_i\varepsilon_{\setminus i}$ and observe that

$$\frac{\tilde{p}_i}{1-\tilde{p}_i} = \frac{4\varepsilon_i\varepsilon_{\setminus i}+1}{1-4\varepsilon_i\varepsilon_{\setminus i}} = \frac{1+2\varepsilon_i+\left({}^{1}/_{2\varepsilon_{\setminus i}}-1\right)}{1-2\varepsilon_i+\left({}^{1}/_{2\varepsilon_{\setminus i}}-1\right)}.$$

Then, the ratio of the weights of PNB and NB decision rules are as follows:

$$\begin{aligned} 1 \geq \frac{\log \frac{4\varepsilon_i\varepsilon_{\setminus i}+1}{1-4\varepsilon_i\varepsilon_{\setminus i}}}{\log \frac{2\varepsilon_i+1}{1-2\varepsilon_i}} &= \frac{\log \frac{1+2\varepsilon_i+\left(1/2\varepsilon_{\setminus i}-1\right)}{1-2\varepsilon_i+\left(1/2\varepsilon_{\setminus i}-1\right)}}{\log \frac{1+2\varepsilon_i}{1-2\varepsilon_i}}, \\ &\geq 1-\left(1/2\varepsilon_{\setminus i}-1\right) \underbrace{\frac{4\varepsilon_i}{\left(1-4\varepsilon_i^2\right)\log \frac{1+2\varepsilon_i}{1-2\varepsilon_i}}}_{C(|\varepsilon_i|)}. \end{aligned}$$

The inequality follows from the Taylor series expansion of $\log \frac{1+a+x}{1-a+x}$ with respect to variable x around $x=0$, which corresponds to $\varepsilon_{\setminus i} \approx 1/2$. Observing that $\varepsilon_{\setminus i} \geq a(N) - 1/2$, $\forall N, i$ and $C(\varepsilon_i) \equiv C(|\varepsilon_i|)$ is monotonic in $|\varepsilon_i|$ yield that:

$$\frac{\tilde{\Phi}}{\Phi} = \frac{\sum_{i=1}^{N}\left(\varepsilon_i \log \frac{4\varepsilon_i\varepsilon_{\setminus i}+1}{1-4\varepsilon_i\varepsilon_{\setminus i}}\right)}{\sum_{i=1}^{N}\varepsilon_i \log \frac{2\varepsilon_i+1}{1-2\varepsilon_i}} \geq 1 - C(1/2-\gamma)\frac{1-a(n)}{a(n)-1/2}. \tag{B.8}$$

□

B.4.1 Proof of Corollary 3.1

Proof of Corollary 3.1. Similar to the proof of Theorem 3.2, consider the ratio in (B.8) and observe that:

$$C(1/2-\gamma)\left(1/2\varepsilon_{\setminus i}-1\right) \leq \delta, \forall i,$$

ensures that $\frac{\tilde{\Phi}}{\Phi} \geq 1-\delta$. Change of variables $\varepsilon_{\setminus i} = p_{\setminus i} - 1/2$ concludes proof. □

B.4.2 Proof of Equation (3.29)

The following proof is based on [17, Theorem 1]. We go over the algebraic manipulations necessary in order to prove (3.29). We first show that:

$$\mathbb{P}\left(f^{PNB}(\mathbf{X}) \neq Y\right) \leq \exp\left(-\frac{\tilde{\Phi}}{2}\right). \tag{B.9}$$

Observe that allowing $w(\tilde{p}_i) \triangleq \frac{\tilde{p}_i}{1-\tilde{p}_i}$ and $\xi \triangleq \mathbb{1}\left(X_i = Y\right) \sim \mathcal{B}(p_i)$:

$$\mathbb{P}\left(f^{PNB}(\mathbf{X}) \neq Y\right) = \mathbb{P}\left(\sum \xi_i w(\tilde{p}_i) - \mathbb{E}\left[\sum \xi_i w(\tilde{p}_i)\right] \leq -\sum\left(\frac{1}{2}-p_i\right)w(\tilde{p}_i)\right).$$

Subsequent application of Kearns-Saul inequality yields (B.9). The use of Kearns-Saul inequality yields sufficiently sharp bounds for the performance of NB (and PNB) decision rules and it is discussed in detail [17]. Next, a lower bound is needed:

$$\mathbb{P}\left(f^{PNB}(\mathbf{X}) \neq Y\right) \geq \frac{3/4}{1+\exp\left(2\tilde{\Phi}+4\sqrt{\tilde{\Phi}}\right)}.$$

Let $\eta_i \triangleq 2\mathbb{1}\left(X_i = Y\right) - 1$ and observe that:

$$\begin{aligned}\mathbb{E}\left[\sum \eta_i w(\tilde{p}_i)\right] &= \sum (p_i - q_i) w(\tilde{p}_i) = 2\tilde{\Phi},\\ \mathrm{Var}\left(\sum \eta_i w(\tilde{p}_i)\right) &= \sum p_i q_i w(\tilde{p}_i)^2 \leq 4\tilde{\Phi}.\end{aligned}$$

The upper bound on the variance is not straightforward. It follows from $w(\tilde{p}_i) \leq w(p_i)$ (which holds $\forall i$ such that $p_{\setminus i} > 1/2$ and from consistency $\forall N > N^*$ for some N^* it holds $\forall i$) and [17, Lemma 4]. The rest follows from observing that:

$$\exp(\sum \eta_i w(\tilde{p}_i)) = \prod_{i:\eta_i=1} \frac{\tilde{p}_i}{1-\tilde{p}_i} \prod_{i:\eta_i=-1} \frac{1-\tilde{p}_i}{\tilde{p}_i},$$

and repeating the exact same steps as the proof of [17, Theorem 1(ii)], which we will not repeat here (the fact that $\min\{\tilde{p}_i, 1-\tilde{p}_i\} \geq \min\{p_i, 1-p_i\}$ is useful).

B.5 Proof of Proposition 3.3

Proof of Proposition 3.3. The proof follows from the Hoeffding's inequality, since for any weighted mixture:

$$\mathbb{P}\left(\sum_{i=1}^{N} w_i \eta_i < 0\right) \leq \exp\left(-\frac{8}{N}\left(\sum_{i=1}^{N} \varepsilon_i w_i\right)^2\right).$$

The definition of the pseudo-competence (3.10) and $w_i = (\tilde{p}_i - 1/2)$ yield that the following is sufficient for the consistency of the rule (3.32):

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} 2\varepsilon_i^2 \varepsilon_{\setminus i} \to \infty \Rightarrow \mathbb{P}\left(f^L(\mathbf{X}) \neq Y\right) \to 0.$$

Similarly by allowing $w_i = 1$, $\forall i$, the following is sufficient for a committee to be not asymptotically weak under majority vote:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i \geq \sqrt{\frac{\log 2}{8}} \Rightarrow \exp\left(-\frac{8}{N}\left(\sum_{i=1}^{N}\varepsilon_i\right)^2\right) \leq \frac{1}{2},$$

which ensures that $\exists N^*$ such that $\forall N > N^*$, $\varepsilon_{\setminus i} \geq \delta > 0$, $\forall i$, yielding that:

$$\sum_{i=1}^{N} 2\varepsilon_i^2\varepsilon_{\setminus i} \geq \delta\sum_{i=1}^{N} 2\varepsilon_i^2 \to \infty,$$

as long as $\lim_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(p_i - 1/2)^2 = \infty$, which concludes that empirical PNB decision rule is consistent. □

B.6 Proof of Theorem 3.3

Proof of Lemma 3.1. Consider $|w_i - \tilde{w}_i|$ for an absolutely balanced committee; the following holds $\forall p_i \in (\gamma, 1-\gamma)$:

$$\left|\log\frac{1+2\varepsilon_i}{1-2\varepsilon_i} - \log\frac{4\varepsilon_i\varepsilon_{\setminus i}+1}{1-4\varepsilon_i\varepsilon_{\setminus i}}\right| \leq (1/2\varepsilon_{\setminus i} - 1)\frac{4\varepsilon_i}{(1-4\varepsilon_i^2)} \leq (1/2\varepsilon_{\setminus i} - 1)\frac{1/2-\gamma}{\gamma(1-\gamma)}.$$

As long as the right hand side is upper bounded by $\frac{\epsilon}{2}$, $\|\mathbf{w} - \tilde{\mathbf{w}}\|_1 < \frac{\epsilon N}{2}$. □

Proof of Theorem 3.3. This proof is an extension of [17, Theorem 11]. Consider the following:

$$\begin{aligned}
|\mathbf{w}\cdot\boldsymbol{\eta} - \tilde{\mathbf{w}}(T)\cdot\boldsymbol{\eta}| &= |\mathbf{w}\cdot\boldsymbol{\eta} - \tilde{\mathbf{w}}\cdot\boldsymbol{\eta} + \tilde{\mathbf{w}}\cdot\boldsymbol{\eta} - \tilde{\mathbf{w}}(T)\cdot\boldsymbol{\eta}|\,,\\
&\leq |\mathbf{w}\cdot\boldsymbol{\eta} - \tilde{\mathbf{w}}\cdot\boldsymbol{\eta}| + |\tilde{\mathbf{w}}\cdot\boldsymbol{\eta} - \tilde{\mathbf{w}}(T)\cdot\boldsymbol{\eta}|\,,\\
&\leq \|\mathbf{w}\cdot\boldsymbol{\eta} - \tilde{\mathbf{w}}\cdot\boldsymbol{\eta}\|_1 + \|\tilde{\mathbf{w}}\cdot\boldsymbol{\eta} - \tilde{\mathbf{w}}(T)\cdot\boldsymbol{\eta}\|_1\,.
\end{aligned}$$

The first inequality follows from the triangle inequality and the second inequality follows from the Hölder's inequality, then, [17, eqn. (41)] yields that:

$$\mathbb{P}\left(\tilde{\mathbf{w}}(T)\cdot\boldsymbol{\eta} \leq 0\right) \leq \mathbb{P}\left(\mathbf{w}\cdot\boldsymbol{\eta} \leq \epsilon N\right) + \mathbb{P}\left(\|\mathbf{w}\cdot\boldsymbol{\eta} - \tilde{\mathbf{w}}\cdot\boldsymbol{\eta}\|_1 + \|\tilde{\mathbf{w}}\cdot\boldsymbol{\eta} - \tilde{\mathbf{w}}(T)\cdot\boldsymbol{\eta}\|_1 > \epsilon N\right).$$

As long as a committee satisfies the condition in Lemma 3.1, this upper-bound boils down

to:

$$\mathbb{P}\left(\tilde{\mathbf{w}}(T)\cdot\boldsymbol{\eta}\leq 0\right)\leq\mathbb{P}\left(\mathbf{w}\cdot\boldsymbol{\eta}\leq\epsilon N\right)+\mathbb{P}\left(\left\|\tilde{\mathbf{w}}\cdot\boldsymbol{\eta}-\tilde{\mathbf{w}}(T)\cdot\boldsymbol{\eta}\right\|_1>{}^{\epsilon N}/_2\right).$$

Now, [17, Corollary 10] yields that $\forall\delta\in(0,1)$ and $\forall i\in[N]$ as long as

$$T\min\left\{\tilde{p}_i,(1-\tilde{p}_i)\right\}\geq 3\left(\frac{4}{\sqrt{4\epsilon+1}-1}\right)^2\log\frac{8N}{\delta},\tag{B.10}$$

the probability that empirical pseudo weights deviate from pseudo weights are bounded:

$$\mathbb{P}\left(\left\|\tilde{\mathbf{w}}\cdot\boldsymbol{\eta}-\tilde{\mathbf{w}}(T)\cdot\boldsymbol{\eta}\right\|_1>\frac{\epsilon N}{2}\right)<\delta.$$

Finally, by Property 3.2, $\min\left\{p_i,(1-p_i)\right\}$ satisfying (B.10) yields that:

$$\mathbb{P}\left(f^H(\mathbf{X})\neq Y\right)\leq\delta+\exp\left[-\frac{\left(2\Phi-\epsilon N\right)^2}{8\Phi}\right].$$

Observe that Lemma 3.1 and eqn. (B.10) are connected to consistency and absolute balance conditions respectively. Therefore, consider a consistent committee with rate $a(N)$ and observe that Lemma 3.1 holds as long as:

$$\frac{\epsilon R(\gamma)}{2}>\rho(N).\tag{B.11}$$

Observing that (B.10) is merely absolute balance condition with $\gamma=\frac{3}{T}\log\frac{8N}{\delta}\left(\frac{4}{\sqrt{4\epsilon+1}-1}\right)^2$, plugging into (B.11), and taking Taylor series expansion yields (a long algebraic manipulation that we skip here) yields that Lemma 3.1 holds as long as:

$$\epsilon>\left(\rho(N)\frac{12}{T}\log\frac{8N}{\delta}\right)^{1/3}.$$

Defining $C(\delta;N,T)=\frac{12}{T}\log\frac{8N}{\delta}$ concludes the proof. □

B.7 Proof of Theorem 3.4

Proof of Theorem 3.4. Due inter-worker and inter-task independence, the empirical pseudo naïve Bayes decision rule at any given time, $\tau\in[T]$ evolves in a well-defined filtration.

Hence, f_τ^H, the empirical decision rule using weights $\tilde{w}(\tau)$, obeys $\forall t \in (\tau+1, ..., T)$:

$$\begin{aligned}\mathbb{P}\left(R(\tau) \cap \left\{f^H(\mathbf{X}(t>\tau) \neq Y)\right\}\right) &= \mathbb{P}_{\mathbf{X}_1^\tau, \boldsymbol{\eta}}\left(R(\tau) \cap \{\tilde{\mathbf{w}}(\tau) \cdot \boldsymbol{\eta} \leq 0\}\right), \\ &= \mathbb{E}_{\mathbf{X}_1^\tau}\left[\mathbb{1}\left(R(\tau)\right) \mathbb{P}_{\boldsymbol{\eta}}\left(\tilde{\mathbf{w}}(\tau) \cdot \boldsymbol{\eta} \leq 0\right)\right].\end{aligned}$$

It is important to observe that as long as the tasks are static, these probabilities are a function of τ and the committee profile $\mathbf{p}$. Let $\tilde{\boldsymbol{\eta}}_\tau$ be a random vector with elements distributed independently with Bernoulli $\tilde{p}_i(\tau)$ and denote $\Delta(\tau) \triangleq \sum_{i=1}^N |p_i - \tilde{p}_i(\tau)|$ In other words, it is a random vector with a pseudo committee profile. A standard tensorization result from [17, 71] yields:

$$|\mathbb{P}_{\boldsymbol{\eta}}\left(\tilde{\mathbf{w}}(\tau) \cdot \boldsymbol{\eta} \leq 0\right) - \mathbb{P}_{\tilde{\boldsymbol{\eta}}_\tau}\left(\tilde{\mathbf{w}}(\tau) \cdot \tilde{\boldsymbol{\eta}}_\tau \leq 0\right)| \leq \Delta(\tau).$$

Then $\forall \tau \in [T]$, $\mathbb{P}_{\tilde{\boldsymbol{\eta}}_\tau}\left(\tilde{\mathbf{w}}(\tau) \cdot \tilde{\boldsymbol{\eta}}_\tau \leq 0\right)$ is the probability of error for the naïve Bayes decision rule of committee strength $\tilde{\Phi}(\tau) \triangleq \sum_{i=1}^N \left(\tilde{p}_i(\tau) - \frac{1}{2}\right) \log \frac{\tilde{p}_i(\tau)}{1-\tilde{p}_i(\tau)}$. Therefore, from [17]:

$$\mathbb{P}_{\tilde{\boldsymbol{\eta}}_\tau}\left(\tilde{\mathbf{w}}(\tau) \cdot \tilde{\boldsymbol{\eta}}_\tau \leq 0\right) \leq \exp\left(-\frac{1}{2}\tilde{\Phi}\left(\tau\right)\right).$$

Hence, $\mathbb{P}_{\boldsymbol{\eta}}\left(\tilde{\mathbf{w}}^{HS}(\tau) \cdot \boldsymbol{\eta} \leq 0\right) \leq \Delta(\tau) + \exp\left(-\frac{1}{2}\tilde{\Phi}\left(\tau\right)\right)$. Then, by the triangle inequality with a mean absolute deviation estimate from [17, 61], we see that:

$$\mathbb{E}_{\mathbf{X}_1^\tau}\left[\Delta(\tau)\right] \leq \sum_{i=1}^N |p_i - \tilde{p}_i| + \frac{N}{\sqrt{T}}.$$

This concludes the proof. □

APPENDIX C

PROOFS FOR CHAPTER 4

C.1 Proof of Proposition 4.1

Proof of Proposition 4.1. The proof is a direct calculation of the terms $\mathbb{E}\ell\left(\delta^{O}(M,U;F),X\right)$ and $\mathbb{E}\ell\left(\delta^{ANT}(M,U),X\right)$. Observe that:

$$\mathbb{E}\ell\left(\delta^{O}(M,U;F),X\right)=\int\mathbb{E}\left[\ell\left(\delta^{O}(M,U;f),X\right)\mid F=f\right]\,dF=\bar{p}\mathbb{E}\ell\left(G,X\right)+p\mathbb{E}\ell\left(U,X\right),$$

which is a direct consequence of genie decision rule having access to reliable failure information. The ANT decision rule on the other hand:

$$\begin{aligned}\mathbb{E}\ell\left(\delta^{ANT}(M,U),X\right)&=\int\mathbb{E}\left[\ell\left(\delta^{ANT}(M,U),X\right)\mid F\right]\,dF\\&=\bar{p}\mathbb{E}\left[\ell\left(\delta^{ANT}(M,U),X\right)\mid F=0\right]+p\mathbb{E}\left[\ell\left(\delta^{ANT}(M,U),X\right)\mid F=1\right]\\&=\bar{p}\mathbb{E}\ell\left(\delta^{ANT}(G,U),X\right)+p\mathbb{E}\ell\left(\delta^{ANT}(B,U),X\right)\\&=\bar{p}\left[\mathbb{E}\ell\left(G,X\right)\mathbb{P}\left(d(G,U)\leq\tau\right)+\mathbb{E}\ell\left(U,X\right)\mathbb{P}\left(d(G,U)>\tau\right)\right]\\&+p\left[\mathbb{E}\ell\left(U,X\right)\mathbb{P}\left(d(B,U)>\tau\right)+\mathbb{E}\ell\left(B,X\right)\mathbb{P}\left(d(B,U)\leq\tau\right)\right].\end{aligned}$$

Therefore, the regret of the ANT rule is given by:

$$\begin{aligned}R^{ANT}(\tau)&=\mathbb{E}\ell\left(\delta^{ANT}(M,U),X\right)-\mathbb{E}\ell\left(\delta^{O}(M,U;F),X\right)\\&=\bar{p}\underbrace{\left(\mathbb{E}\ell\left(U,X\right)-\mathbb{E}\ell\left(G,X\right)\right)}_{R^{UG}}\underbrace{\mathbb{P}\left(d(G,U)>\tau\right)}_{\Phi_{d}^{GU}(\tau)}\\&+p\underbrace{\left(\mathbb{E}\ell\left(B,X\right)-\mathbb{E}\ell\left(U,X\right)\right)}_{R^{BU}}\underbrace{\mathbb{P}\left(d(B,U)\leq\tau\right)}_{F_{d}^{BU}(\tau)}.\end{aligned}$$

□

C.2 Proof of Proposition 4.2

Proof of Proposition 4.2. Recall that the necessary and sufficient condition for $R^{ANT}(\tau) = 0$ is:

$$\Phi^{GU}(\tau) = \mathbb{P}\left(\|G - U\| > \tau\right) = 0 = \mathbb{P}\left(\|B - U\| \leq \tau\right) = F^{BU}(\tau). \tag{C.1}$$

A necessary condition for (C.1) is that $\exists\tau$ such that $\mathbb{E}\|U - G\| \leq \tau \leq \mathbb{E}\|B - U\|$. This is because if $\exists\tau$:

$$\begin{aligned}
\mathbb{P}\left(\|G - U\| > \tau\right) = 0 \Rightarrow \mathbb{E}\|U - G\| &= \int_0^\infty \mathbb{P}\left(\|G - U\| > t\right)\, dt \\
&= \int_0^\tau \mathbb{P}\left(\|G - U\| > t\right)\, dt \leq \tau, \\
\mathbb{P}\left(\|B - U\| \leq \tau\right) = 0 \Rightarrow \mathbb{E}\|U - B\| &= \int_0^\infty \left(1 - \mathbb{P}\left(\|B - U\| \leq t\right)\right)\, dt \\
&= \int_0^\tau dt + \int_\tau^\infty \mathbb{P}\left(\|B - U\| > t\right)\, dt \geq \tau.
\end{aligned}$$

Therefore, if $\exists\tau$ such that $\Phi^{GU}(\tau) = F^{BU}(\tau) = 0$, then $\mathbb{E}\|U - G\| \leq \tau \leq \mathbb{E}\|B - U\|$. The sufficient condition follows as:

$$\mathbb{E}\|B - U\| \geq \frac{1}{\mathcal{C}}\mathbb{E}\left|\ell(B, X) - \ell(U, X)\right| \geq \frac{1}{\mathcal{C}}\mathbb{E}\ell(B, X) - \frac{1}{\mathcal{C}}\mathbb{E}\ell(U, X), \tag{C.2}$$

$$\mathbb{E}\|G - U\| \leq \mathcal{C}\mathbb{E}\left|\ell(G, X) - \ell(U, X)\right| \leq \mathcal{C}\mathbb{E}\ell(G, X) + \mathcal{C}\mathbb{E}\ell(U, X), \tag{C.3}$$

where, the first inequalities of (C.2)-(C.3) follow from the definition of $\mathcal{C}$ bi-Lipschitz loss functions (4.4) and the second inequalities follow from the triangle inequality. Then,

$$\begin{aligned}
\mathbb{E}\ell\left(G, X\right) < \mathbb{E}\ell\left(U, X\right) &\Rightarrow \mathbb{E}\|G - U\| < 2\mathcal{C}\mathbb{E}\ell(U, X), \\
\frac{1}{2\mathcal{C}^2 + 1}\mathbb{E}\ell\left(B, X\right) > \mathbb{E}\ell\left(U, X\right) &\Rightarrow \mathbb{E}\|B - U\| \geq 2\mathcal{C}\mathbb{E}\ell(U, X),
\end{aligned}$$

yield that as long as $\mathbb{E}\ell\left(G, X\right) < \mathbb{E}\ell\left(U, X\right) < \frac{1}{2\mathcal{C}^2+1}\mathbb{E}\ell\left(B, X\right)$. □

C.3 Proof of Proposition 4.3

Proof of Proposition 4.3. Distance to regret relations follow from (C.2)-(C.3). Recall the Chernoff bounds:

$$\mathbb{P}\left(X \geq (1+\delta)\mathbb{E}\left[X\right]\right) \leq \exp\left(-\frac{\delta^2}{2+\delta}\mathbb{E}\left[X\right]\right),$$
$$\mathbb{P}\left(X \leq (1-\delta)\mathbb{E}\left[X\right]\right) \leq \exp\left(-\frac{\delta^2}{2}\mathbb{E}\left[X\right]\right).$$

The rest algebraically follows from observing that:

$$\mathbb{E}\left[X\right] < a < \tau \Rightarrow \frac{\left(\tau - \mathbb{E}\left[X\right]\right)^2}{\tau + \mathbb{E}\left[X\right]} > \frac{(\tau - a)^2}{\tau + a},$$
$$\mathbb{E}\left[X\right] > b > \tau \Rightarrow \frac{\left(\tau - \mathbb{E}\left[X\right]\right)^2}{2\mathbb{E}\left[X\right]} > \frac{(\tau - b)^2}{2b}.$$

□

C.4 Proof of Proposition 4.4

Proof of Proposition 4.4. Consider the s-transform of $F_A(s)$ of the random average A:

$$F_A(s) = \mathbb{E}\left[e^{-s\frac{1}{N}\sum_{i=1}^{N} G_i}\right] = \mathbb{E}_N\left[\mathbb{E}\left[e^{-s\frac{1}{N}\sum_{i=1}^{N} G_i} \;\middle|\; N = n\right]\right]$$
$$= \mathbb{E}_N\left[\left[F_G\left(s/N\right)\right]^N\right] = \sum_{n\geq 1}\left[F_G(s/n)\right]^n p_N\left(n\right).$$

The moments of A can be acquired by differentiation [72], the mean is calculated as follows:

$$\mathbb{E}\left[A\right] = -\frac{d}{ds}F_A(s)\bigg|_{s=0} = -\sum_{n\geq 1}\frac{d}{ds}\left[F_G(s/n)\right]^n\bigg|_{s=0} p_N\left(n\right).$$

Further note that:

$$\frac{d}{ds}\left[F_G(s/n)\right]^n\bigg|_{s=0} = n\underbrace{\left[F_G(s/n)\right]^{n-1}\big|_{s=0}}_{=1}\underbrace{\frac{d}{ds}F_G(s/n)\bigg|_{s=0}}_{=-\mathbb{E}[G]/n} = -\mathbb{E}\left[G\right],$$

yielding that $\mathbb{E}[A] = \mathbb{E}[G]$. The second moment is calculated next, but first observe that:

$$\begin{aligned}\frac{d^2}{ds^2}\left[F_G(s/n)\right]^n\Big|_{s=0} &= (n(n-1))\left[F_G(s/n)\right]^{n-2}\big|_{s=0}\underbrace{\left(\frac{d}{ds}F_G(s/n)\right)^2\Big|_{s=0}}_{=(\mathbb{E}[G]/n)^2}\\ &\quad + n\left[F_G(s/n)\right]^{n-1}\big|_{s=0}\underbrace{\frac{d^2}{ds^2}F_G(s/n)\Big|_{s=0}}_{\mathbb{E}\left[G^2\right]/n^2}\\ &= \mathbb{E}[G]^2 + \frac{\mathbb{E}[G^2]}{n} - \frac{\mathbb{E}[G]^2}{n} = \mathbb{E}[G]^2 + \frac{\mathrm{Var}(G)}{n},\end{aligned}$$

which yields by [72, page 101] that:

$$\begin{aligned}\mathrm{Var}(A) &= \frac{d^2}{ds^2}F_A(s)\Big|_{s=0} - \left[\frac{d}{ds}F_A(s)\right]^2\Big|_{s=0}\\ &= \sum_{n\geq 1}\left[\underbrace{\frac{d^2}{ds^2}\left[F_G(s/n)\right]^n\Big|_{s=0}}_{\mathbb{E}[G]^2+\frac{\mathrm{Var}(G)}{n}} - \underbrace{\left[\frac{d}{ds}\left[F_G(s/n)\right]^n\right]^2\Big|_{s=0}}_{\mathbb{E}[G]^2}\right]p_N(n).\end{aligned}$$

Consequently, $\mathrm{Var}(A) = \mathrm{Var}(G)\,\mathbb{E}[1/N]$. □

C.5 Proof of Proposition 4.5

Proof of Proposition 4.5 (First Moment). First, let $\mathcal{V} = [N]$ and observe that $\mathcal{I} = \{i : F_i = 0\}$ is the set active agents and $|\mathcal{I}| \sim \mathcal{F}(q_F; |\mathcal{V}|)$. With probability $p_F^{|\mathcal{V}|}$, $|\mathcal{I}| = 0$, in which case the random average, and hence its s-transform is not defined. Define the probability measure derived from $p_{\mathcal{I}}(l)$:

$$\tilde{p}_{\mathcal{I}}(l) = \frac{1}{1-p_F^{|\mathcal{V}|}}p_{\mathcal{I}}(l), \forall l \geq 1, \tag{C.4}$$

which allows a well-defined average almost always. For the rest of the proof, we will allow that $\tilde{\mathbb{E}}[\cdot] \equiv \mathbb{E}[\cdot]$ and $\tilde{\mathrm{Var}}(\cdot) \equiv \mathrm{Var}(\cdot)$ for notational clarity. Even though G_i are not i.i.d.,

s-transform of $\mu^O(\mathbf{M}) \equiv A$ (this is also for notational clarity) has a closed form:

$$F_A(s) = \mathbb{E}\left[\exp\left(-s\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} G_i\right)\right] = \mathbb{E}_{\mathcal{I}}\left[\mathbb{E}\left[\exp\left(-s\frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0} G_i\right)\,\middle|\,\mathcal{I}=\mathcal{I}_0\right]\right]$$
$$= \mathbb{E}\left[\prod_{i\in\mathcal{I}} F_{G_i}\left(s/|\mathcal{I}|\right)\right]. \tag{C.5}$$

Observe that:

$$\frac{d}{ds}\prod_{i\in\mathcal{I}_0} F_{G_i}\left(s/|\mathcal{I}_0|\right)\Big|_{s=0} = \sum_{i\in\mathcal{I}_0} \underbrace{\frac{d}{ds}F_{G_i}\left(s/|\mathcal{I}_0|\right)\Big|_{s=0}}_{=-\frac{\mathbb{E}[G_i]}{|\mathcal{I}_0|}} \underbrace{\prod_{j\in\mathcal{I}_0\setminus i} F_{G_j}\left(s/|\mathcal{I}_0|\right)\Big|_{s=0}}_{=1}.$$

Furthermore,

$$\frac{d}{ds}F_A(s)\Big|_{s=0} = \sum_{\mathcal{I}_0:|\mathcal{I}_0|>0} \frac{d}{ds}\prod_{i\in\mathcal{I}_0} F_{G_i}\left(s/|\mathcal{I}_0|\right)\Big|_{s=0} \tilde{p}_{\mathcal{I}}(\mathcal{I}_0) = -\sum_{\mathcal{I}_0:|\mathcal{I}_0|>0} \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\mathbb{E}\left[G_i\right]\tilde{p}_{\mathcal{I}}(\mathcal{I}_0),$$

which yields for $\tilde{p}_{\mathcal{I}}(\mathcal{I}_0)$ that:

$$\mathbb{E}\left[A\right] = \sum_{\mathcal{I}_0:|\mathcal{I}_0|>0} \tilde{p}_{\mathcal{I}}(\mathcal{I}_0)\frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\mathbb{E}\left[G_i\right].$$

Observe that by the law of total probability:

$$\tilde{p}_{\mathcal{I}}(\mathcal{I}_0) = \sum_{l=1}^{|\mathcal{V}|} p_{\,\mathcal{I}\,|\,|\mathcal{I}|}\left(\,\mathcal{I}_0 \mid l\right)\tilde{p}_{|\mathcal{I}|}\left(l\right) = \sum_{l=1}^{|\mathcal{V}|}\frac{1}{\binom{|\mathcal{V}|}{l}}\tilde{p}_{|\mathcal{I}|}\left(l\right).$$

The fact that $p_{\,\mathcal{I}\,|\,|\mathcal{I}|}\left(\,\mathcal{I}_0 \mid l\right) = 1/\binom{|\mathcal{V}|}{l}$ is a property of the binomial distribution; every random vector $\bar{\mathbf{F}}$ is equally likely conditioned on its norm $\left\|\bar{\mathbf{F}}\right\|_1 = |\{i : F_i = 0\}|$, which extends to heterogeneous failure probabilities non-trivially. As long as $\left\|\bar{F}\right\|_1 \sim \mathcal{F}(q_F; |\mathcal{V}|)$,

$$\mathbb{E}\left[A\right] = \sum_{l=1}^{|\mathcal{V}|}\frac{1}{\binom{|\mathcal{V}|}{l}}\tilde{p}_{|\mathcal{I}|}\left(l\right)\frac{1}{l}\underbrace{\sum_{\mathcal{I}_0:|\mathcal{I}_0|=l}\sum_{i\in\mathcal{I}_0}\mathbb{E}\left[G_i\right]}_{\binom{|\mathcal{V}|-1}{l-1}\sum_{i\in\mathcal{V}}\mathbb{E}[G_i]} = \underbrace{\sum_{l=1}^{|\mathcal{V}|}\tilde{p}_{|\mathcal{I}|}\left(l\right)}_{=1}\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i\right] = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i\right].$$

It is worth noting that the same derivation for probability distributions on $\mathcal{I}$ that are *not* induced from a Binomial random variable and almost always non-empty yields that

$$\mathbb{E}[A] = \sum_{\mathcal{I}_0} \frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} \mathbb{E}[G_i] \, p_{\mathcal{I}}(\mathcal{I}_0).$$

□

Proof of Proposition 4.5 (Second Moment). The second moment calculations follow similarly to those of first moment calculations from (C.5). Observe that:

$$\begin{aligned}
\frac{d^2}{ds^2} \prod_{i \in \mathcal{I}_0} F_{G_i}(s/|\mathcal{I}_0|)\Big|_{s=0} &= \sum_{i \in \mathcal{I}_0} \underbrace{\frac{d^2}{ds^2} F_{G_i}(s/|\mathcal{I}_0|)\Big|_{s=0}}_{\frac{\mathbb{E}[G_i^2]}{|\mathcal{I}_0|^2}} \underbrace{\prod_{j \in \mathcal{I}_0 \setminus i} F_{G_j}(s/|\mathcal{I}_0|)\Big|_{s=0}}_{=1} \\
&\quad + \mathbb{1}(|\mathcal{I}_0| > 1) \sum_{i \neq j \in \mathcal{I}_0} \underbrace{\frac{d}{ds} F_{G_i}(s/|\mathcal{I}_0|) \frac{d}{ds} F_{G_j}(s/|\mathcal{I}_0|)\Big|_{s=0}}_{\frac{\mathbb{E}[G_i]}{|\mathcal{I}_0|} \frac{\mathbb{E}[G_j]}{|\mathcal{I}_0|}} \underbrace{\prod_{k \in \mathcal{I}_0 \setminus i,j} F_{G_k}(s/|\mathcal{I}_0|)\Big|_{s=0}}_{=1} \\
&= \sum_{i \in \mathcal{I}_0} \frac{\mathbb{E}[G_i^2]}{|\mathcal{I}_0|^2} + \mathbb{1}(|\mathcal{I}_0| > 1) \sum_{\substack{i \neq j \in \mathcal{I}_0 \\ |\mathcal{I}_0| > 1}} \frac{\mathbb{E}[G_i] \, \mathbb{E}[G_j]}{|\mathcal{I}_0|^2},
\end{aligned}$$

which yields that:

$$\begin{aligned}
\mathbb{E}\left[A^2\right] &= \sum_{\mathcal{I}_0:|\mathcal{I}_0|>0} \tilde{p}_{\mathcal{I}}(\mathcal{I}_0)\left[\sum_{i\in\mathcal{I}_0}\frac{\mathbb{E}\left[G_i^2\right]}{|\mathcal{I}_0|^2}+\sum_{\substack{i\neq j\in\mathcal{I}_0\\|\mathcal{I}_0|>1}}\frac{\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right]}{|\mathcal{I}_0|^2}\right]\\
&=\sum_{l=1}^{|\mathcal{V}|}\frac{1}{\binom{|\mathcal{V}|}{l}}\tilde{p}_{|\mathcal{I}|}\left(l\right)\sum_{\mathcal{I}_0:|\mathcal{I}_0|=l}\left[\sum_{i\in\mathcal{I}_0}\frac{\mathbb{E}\left[G_i^2\right]}{l^2}+\mathbb{1}\left(l>1\right)\sum_{\substack{i\neq j\in\mathcal{I}_0\\l>1}}\frac{\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right]}{l^2}\right]\\
&=\sum_{l=1}^{|\mathcal{V}|}\frac{1}{\binom{|\mathcal{V}|}{l}}\tilde{p}_{|\mathcal{I}|}\left(l\right)\underbrace{\sum_{\mathcal{I}_0:|\mathcal{I}_0|=l}\left[\sum_{i\in\mathcal{I}_0}\frac{\mathbb{E}\left[G_i^2\right]}{l^2}\right]}_{\binom{|\mathcal{V}|-1}{l-1}\frac{1}{l^2}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i^2\right]}\\
&\qquad+\sum_{l=1}^{|\mathcal{V}|}\frac{1}{\binom{|\mathcal{V}|}{l}}\tilde{p}_{|\mathcal{I}|}\left(l\right)\mathbb{1}\left(l>1\right)\underbrace{\sum_{\mathcal{I}_0:|\mathcal{I}_0|=l}\left[\sum_{\substack{i\neq j\in\mathcal{I}_0\\l>1}}\frac{\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right]}{l^2}\right]}_{\binom{|\mathcal{V}|-2}{l-2}\frac{1}{l^2}\sum_{i\neq j\in\mathcal{V}}\mathbb{E}[G_i]\mathbb{E}[G_j]}\\
&=\tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right]\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i^2\right]+\sum_{l=1}^{|\mathcal{V}|}\frac{l-1}{l}\tilde{p}_{|\mathcal{I}|}\left(l\right)\mathbb{1}\left(l>1\right)\underbrace{\frac{1}{|\mathcal{V}|\left(|\mathcal{V}|-1\right)}\sum_{i\neq j\in\mathcal{V}}\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right]}_{\equiv\mathcal{C}}.
\end{aligned}$$

Observe that:

$$\begin{aligned}
\sum_{l=1}^{|\mathcal{V}|}\frac{l-1}{l}\tilde{p}_{|\mathcal{I}|}\left(l\right)\mathbb{1}\left(l>1\right)\mathcal{C},&=\sum_{l=1}^{|\mathcal{V}|}\tilde{p}_{|\mathcal{I}|}\left(l\right)\mathbb{1}\left(l>1\right)\mathcal{C}-\sum_{l=1}^{|\mathcal{V}|}\frac{1}{l}\tilde{p}_{|\mathcal{I}|}\left(l\right)\mathbb{1}\left(l>1\right)\mathcal{C},\\
&=\sum_{l=1}^{|\mathcal{V}|}\tilde{p}_{|\mathcal{I}|}\left(l\right)\mathbb{1}\left(l>1\right)\mathcal{C}\\
&\qquad-\sum_{l=1}^{|\mathcal{V}|}\frac{1}{l}\tilde{p}_{|\mathcal{I}|}\left(l\right)\mathbb{1}\left(l>1\right)\mathcal{C}+\mathcal{C}\tilde{p}_{|\mathcal{I}|}(1)-\mathcal{C}\tilde{p}_{|\mathcal{I}|}(1),\\
&=\sum_{l=1}^{|\mathcal{V}|}\tilde{p}_{|\mathcal{I}|}\left(l\right)\mathcal{C}-\sum_{l=1}^{|\mathcal{V}|}\frac{1}{l}\tilde{p}_{|\mathcal{I}|}\left(l\right)\mathcal{C}\\
&=\mathcal{C}-\mathcal{C}\tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[A^2\right] &= \tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right]\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i^2\right] + (1-\tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right])\frac{1}{|\mathcal{V}|\left(|\mathcal{V}|-1\right)}\sum_{i\neq j\in\mathcal{V}}\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right], \\ &= \tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right]\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathrm{Var}(G_i) + \tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right]\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i\right]^2 \\ &\quad + (1-\tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right])\frac{1}{|\mathcal{V}|\left(|\mathcal{V}|-1\right)}\sum_{i\neq j\in\mathcal{V}}\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right].\end{aligned}$$

And hence,

$$\begin{aligned}\mathrm{Var}\left(A\right) &= \mathbb{E}\left[A^2\right] - \mathbb{E}\left[A\right]^2 \\ &= \tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right]\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathrm{Var}(G_i) + \tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right]\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i\right]^2 \\ &\quad + (1-\tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right])\frac{1}{|\mathcal{V}|\left(|\mathcal{V}|-1\right)}\sum_{i\neq j\in\mathcal{V}}\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right] - \frac{1}{|\mathcal{V}|^2}\sum_{i,j\in\mathcal{V}}\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right], \\ &= \tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right]\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathrm{Var}(G_i) + \tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right]\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i\right]^2 - \frac{1/|\mathcal{V}|}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right] \\ &\quad + (1-\tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right])\frac{1}{|\mathcal{V}|\left(|\mathcal{V}|-1\right)}\sum_{i\neq j\in\mathcal{V}}\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right] + \frac{1-1/|\mathcal{V}|}{|\mathcal{V}|\left(|\mathcal{V}|-1\right)}\sum_{i\neq j\in\mathcal{V}}\mathbb{E}\left[G_i\right]\mathbb{E}\left[G_j\right].\end{aligned}$$

Consequently, the variance of the average takes the following form:

$$\begin{aligned}\mathrm{Var}\left(A\right) &= \tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right]\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathrm{Var}(G_i) \\ &\quad + \left(\tilde{\mathbb{E}}\left[1/|\mathcal{I}|\right] - 1/|\mathcal{V}|\right)\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\mathbb{E}\left[G_i\right]\left(\mathbb{E}\left[G_i\right] - \frac{1}{|\mathcal{V}|-1}\sum_{j\in\mathcal{V}\setminus i}\mathbb{E}\left[G_i\right]\right).\end{aligned} \tag{C.6}$$

□

APPENDIX D

PROOFS FOR CHAPTER 5

D.1 Proof of Theorem 5.1 and Its Corollaries

Proof of Theorem 5.1 employs McDiarmid's inequality, hence the first step is to show that r_{ab}^{eff} has bounded differences [63]. First, the proof of Proposition 5.1 is given below:

Proof of Proposition 5.1. This proposition is a consequence of the compact support of the fabrication process. Formally, let $\boldsymbol{g}_{\setminus ij}$ be the conductance vector when only the branch conductance g_{ij} for some $i \leftrightarrow j \in \mathcal{E}$ is replaced by some g'_{ij}. Then, $\forall \boldsymbol{g}, \boldsymbol{g}_{\setminus ij} \in \prod_{k \leftrightarrow l \in \mathcal{E}} [\ell_{kl}, u_{kl}]$, the following bound holds:

$$\left| r_{ab}^{\mathit{eff}}(\boldsymbol{g}_{\setminus ij}) - r_{ab}^{\mathit{eff}}(\boldsymbol{g}) \right| \leq r_{ab}^{\mathit{eff}}(\boldsymbol{\ell}) - r_{ab}^{\mathit{eff}}(\boldsymbol{u}).$$

Hence, the proposition follows. □

For Theorem 5.1, we first employ Meyer's relation, a version of Woodbury's identity for pseudo-inverse:

Lemma D.1 (Meyer's Relation, [73])**.** *Let* $\boldsymbol{c} \in \mathcal{R}(\underline{\mathbf{A}})$, $\boldsymbol{c} \in \mathcal{R}(\underline{\mathbf{A}}^\top)$ *and* $(1 + \boldsymbol{d}^T \underline{\mathbf{A}}^\dagger \boldsymbol{c}) \neq 0$. *Then,*

$$\left(\underline{\mathbf{A}} + \boldsymbol{c}\boldsymbol{d}^\top\right)^\dagger = \underline{\mathbf{A}}^\dagger - \left(1 + \boldsymbol{d}^\top \underline{\mathbf{A}}^\dagger \boldsymbol{c}\right)^{-1} \underline{\mathbf{A}}^\dagger \boldsymbol{c}\boldsymbol{d}^\top \underline{\mathbf{A}}^\dagger.$$

Proof and extensions of Lemma D.1 are given in [73]. As it is stated here, Lemma D.1 is only valid for changes at the branch conductances that preserve circuit topology, not for addition and or severance of edges since $\forall i, j \in \mathcal{V}$ such that $i \leftrightarrow j \notin \mathcal{E}$, $(\boldsymbol{e}_i - \boldsymbol{e}_j) \notin \mathcal{R}(\underline{\mathbf{L}})$.

Lemma D.2 (Derivative of Effective Resistance)**.** *For a given circuit topology* $\mathcal{G}$, $\forall a, b \in \mathcal{V}$, *and* $\forall i \leftrightarrow j \in \mathcal{E}$, *the derivative of the effective resistance is given by:*

$$\frac{\partial}{\partial g_{ij}} r_{ab}^{\mathit{eff}} = -\left(\Delta_{ij}^{ab}\right)^2 \quad \textit{and} \quad \frac{\partial}{\partial r_{ij}} r_{ab}^{\mathit{eff}} = \left(\frac{\Delta_{ij}^{ab}}{r_{ij}}\right)^2. \tag{D.1}$$

where, Δ_{ij}^{ab} is a function of the original circuit. Explicitly:

$$\Delta_{ij}^{ab} \triangleq \frac{r_{aj}^{\mathit{eff}} + r_{bi}^{\mathit{eff}} - r_{ai}^{\mathit{eff}} - r_{bj}^{\mathit{eff}}}{2}, \tag{D.2}$$

which is bounded and symmetric over the connected nodes: $\Delta_{ij}^{ab} \leq \min\left\{r_{ij}^{\mathit{eff}}, r_{ab}^{\mathit{eff}}\right\}$ and $\Delta_{ij}^{ab} = \Delta_{ab}^{ij}, \forall i \leftrightarrow j, a \leftrightarrow b \in \mathcal{E}$.

Proof of Lemma D.2. First, let us show that the definition in (D.2) is not arbitrary; observe that:

$$\begin{aligned}\Delta_{ij}^{ab} &= \left[(\underline{\mathbf{L}}^{\dagger})_{ai} + (\underline{\mathbf{L}}^{\dagger})_{bj} - (\underline{\mathbf{L}}^{\dagger})_{aj} - (\underline{\mathbf{L}}^{\dagger})_{bi} + 0 \cdot \left((\underline{\mathbf{L}}^{\dagger})_{aa} + (\underline{\mathbf{L}}^{\dagger})_{bb} + (\underline{\mathbf{L}}^{\dagger})_{jj} + (\underline{\mathbf{L}}^{\dagger})_{ii}\right)\right] \\ &= \frac{1}{2}\left[r_{aj}^{\mathit{eff}} + r_{bi}^{\mathit{eff}} - r_{ai}^{\mathit{eff}} - r_{bj}^{\mathit{eff}}\right].\end{aligned} \tag{D.3}$$

Therefore, we define:

$$\Delta_{ij}^{ab} = (\boldsymbol{e}_a - \boldsymbol{e}_b)^{\top}\underline{\mathbf{L}}^{\dagger}(\boldsymbol{e}_i - \boldsymbol{e}_j),$$

which yields that $\Delta_{ij}^{ab} = \Delta_{ab}^{ij}$ provided $(\boldsymbol{e}_i - \boldsymbol{e}_j) \in \mathcal{R}(\underline{\mathbf{L}})$ and $(\boldsymbol{e}_a - \boldsymbol{e}_b) \in \mathcal{R}(\underline{\mathbf{L}})$, equivalently, $i \leftrightarrow j, a \leftrightarrow b \in \mathcal{E}$. Since the effective resistance is a distance on a graph, the triangle inequality on (D.2) yields that $\Delta_{ij}^{ab} \leq r_{ab}^{\mathit{eff}}$ and $\Delta_{ij}^{ab} \leq r_{ij}^{\mathit{eff}}$.

Let $\underline{\mathbf{L}}(\rho_{ij}) \triangleq \underline{\mathbf{L}} + \rho_{ij}(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^{\top}$ denote the Laplacian of a circuit acquired by changing a branch conductance without changing the circuit topology: $\boldsymbol{g}_{ij} \rightarrow \boldsymbol{g}_{ij} + \rho_{ij}$. Formally, $(\boldsymbol{e}_i - \boldsymbol{e}_j) \in \mathcal{R}(\underline{\mathbf{A}})$, therefore, by Lemma D.1:

$$\underline{\mathbf{L}}(\rho_{ij})^{\dagger} = \underline{\mathbf{L}}^{\dagger} - \frac{\rho_{ij}}{1 + \rho_{ij} r_{ij}^{\mathit{eff}}}\underline{\mathbf{L}}^{\dagger}(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^{\top}\underline{\mathbf{L}}^{\dagger}.$$

Using (5.3), we can deduce that:

$$\begin{aligned} r_{ab}^{\mathit{eff}}(\rho_{ij}) - r_{ab}^{\mathit{eff}} &= (\boldsymbol{e}_a - \boldsymbol{e}_b)^{\top}\left(\underline{\mathbf{L}}^{\dagger}(\rho_{ij}) - \underline{\mathbf{L}}^{\dagger}\right)(\boldsymbol{e}_a - \boldsymbol{e}_b) \\ &= \frac{-\rho_{ij}}{1 + \rho_{ij} r_{ij}^{\mathit{eff}}}(\boldsymbol{e}_a - \boldsymbol{e}_b)^{\top}\underline{\mathbf{L}}^{\dagger}\mathbf{J}^{ij}\underline{\mathbf{L}}^{\dagger}(\boldsymbol{e}_a - \boldsymbol{e}_b).\end{aligned} \tag{D.4}$$

From the definition of the derivative, we conclude that:

$$\frac{\partial}{\partial g_{ij}} r_{ab}^{\mathit{eff}} = \lim_{\rho_{ij} \to 0} \frac{r_{ab}^{\mathit{eff}}(\rho_{ij}) - r_{ab}^{\mathit{eff}}}{\rho_{ij}} = -\left(\Delta_{ij}^{ab}\right)^2.$$

For the derivative with respect to branch resistance r_{ij}, change of variables from $r_{ij} \rightarrow r_{ij} + h$ to $g_{ij} \rightarrow g_{ij} - \frac{h g_{ij}^2}{1 + h g_{ij}}$ yields (D.1). □

Lemmata D.1-D.2 are sufficient to prove that the effective resistance r_{ab}^{eff} between any two nodes $a, b \in \mathcal{V}$ has bounded differences over topology-preserving changes in the branch resistances/conductances.

Lemma D.3 (Bounded Difference Property)**.** *Let $r_{ab}^{\mathit{eff}}(\boldsymbol{g}_{\setminus ij})$ be the effective resistance between nodes $a, b \in \mathcal{V}$ when only the branch conductance g_{ij} for some $i \leftrightarrow j \in \mathcal{E}$ is replaced by some g'_{ij}. Then,*

$$\sup_{\substack{g'_{ij} \in [1/u_{ij},\, 1/\ell_{ij}] \\ \boldsymbol{g} \in \prod_{k \leftrightarrow l \in \mathcal{E}} [1/u_{kl},\, 1/\ell_{kl}]}} \left| r_{ab}^{\mathit{eff}}(\boldsymbol{g}_{\setminus ij}) - r_{ab}^{\mathit{eff}}(\boldsymbol{g}) \right| \leq p_{ij}.$$

Here, $p_{ij} \triangleq \frac{u_{ij} - \ell_{ij}}{\ell_{ij}} \left| g_{ij}^{\mathit{eff}}(\boldsymbol{\ell}) \frac{\partial}{\partial g_{ij}} r_{ab}^{\mathit{eff}}(\boldsymbol{\ell}) \right|$.

Proof of Lemma D.3. From (D.2)-(D.4), we can deduce that

$$\begin{aligned} \left| r_{ab}^{\mathit{eff}}(\boldsymbol{g}_{\setminus ij}) - r_{ab}^{\mathit{eff}}(\boldsymbol{g}) \right| &= \left| \frac{g'_{ij} - g_{ij}}{1 + \left(g'_{ij} - g_{ij}\right) r_{ij}^{\mathit{eff}}(\boldsymbol{g})} \right| \left| \frac{\partial}{\partial g_{ij}} r_{ab}^{\mathit{eff}}(\boldsymbol{g}) \right| \\ &= \left| \frac{g'_{ij} - g_{ij}}{g_{ij}^{\mathit{eff}}(\boldsymbol{g}) + g'_{ij} - g_{ij}} \right| \left| g_{ij}^{\mathit{eff}}(\boldsymbol{g}) \frac{\partial}{\partial g_{ij}} r_{ab}^{\mathit{eff}}(\boldsymbol{g}) \right|. \end{aligned}$$

Now observe that for every branch $i \leftrightarrow j \in \mathcal{E}$ with a conductance g_{ij}, $g_{ij}^{\mathit{eff}}(\boldsymbol{g}) = g_{ij} + n_{ij}$ for some $n_{ij} \geq 0$, which follows reducing the remaining circuit down to its effective conductance. Therefore,

$$\left| \frac{g'_{ij} - g_{ij}}{g_{ij}^{\mathit{eff}}(\boldsymbol{g}) + g'_{ij} - g_{ij}} \right| = \left| \frac{g'_{ij} - g_{ij}}{g'_{ij} + n_{ij}} \right| \leq \frac{u_{ij} - \ell_{ij}}{\ell_{ij}}.$$

The last inequality follows from $n_{ij} \geq 0$. Finally, using the fact that $g_{ij}^{\mathit{eff}}(\boldsymbol{g})$ and $\frac{\partial}{\partial g_{ij}} r_{ab}^{\mathit{eff}}(\boldsymbol{g})$ are nondecreasing in $\boldsymbol{g}$, we conclude the effective resistance is of bounded difference. □

Proof of Theorem 5.1. Given the bounded differences, p_{ij}, McDiarmid's inequality follows. □

Corollary 5.1 follows from Theorem 5.1 and the upper bound from Lemma D.2:

Proof of Corollary 5.1. Using $\Delta_{ab}^{ij} \leq \left(r_{ij}^{\mathit{eff}}\right)^2$ yields that

$$\left| g_{ij}^{\mathit{eff}}(\boldsymbol{g}) \frac{\partial}{\partial g_{ij}} r_{ab}^{\mathit{eff}}(\boldsymbol{g}) \right| \leq \left| -g_{ij}^{\mathit{eff}} \left(r_{ij}^{\mathit{eff}}\right)^2 \right| = r_{ij}^{\mathit{eff}}.$$

The rest follows from McDiarmid's inequality. □

Corollary 5.2 shows that how the end-products of the fabrication processes concentrate around their designed mean $r_{ab}^{\mathit{eff}}\left(\mathbb{E}\left[\mathbf{G}\right]\right)$.

Proof of Corollary 5.2. Observe that $r_{ab}^{\mathit{eff}}(\boldsymbol{g})$ is a concave function of resistance $\boldsymbol{r}$, which follows from (D.1) and also stated in [69]. By Jensen's inequality, it follows that $\forall a, b \in \mathcal{V}$:

$$r_{ab}^{\mathit{eff}}\left(\mathbb{E}_{\mathbf{R}}\left[\mathbf{G}\right]\right) \geq \mathbb{E}_{\mathbf{R}}\left[r_{ab}^{\mathit{eff}}(\mathbf{G})\right].$$

Therefore, for any fabrication *event*, the following holds:

$$r_{ab}^{\mathit{eff}}(\mathbf{G}) - r_{ab}^{\mathit{eff}}\left(\mathbb{E}_{\mathbf{R}}\left[\mathbf{G}\right]\right) > \varepsilon \Rightarrow r_{ab}^{\mathit{eff}}(\mathbf{G}) - \mathbb{E}_{\mathbf{R}}\left[r_{ab}^{\mathit{eff}}(\mathbf{G})\right] > \varepsilon.$$

Hence, Corollary 5.2 follows with the modification of the denominator follows from (D.1) with $r_{ij}\frac{\partial}{\partial r_{ij}} r_{ab}^{\mathit{eff}} = g_{ij}\frac{\partial}{\partial g_{ij}} r_{ab}^{\mathit{eff}}$, $\forall a, b \in \mathcal{V}$. □

D.2 Proof of Theorem 5.2

Proofs of Theorems 5.2-5.4 utilize [70, Corollary 3]. We have modified their result to incorporate arbitrary compact-support concave functions. The modified result is as follows:

Corollary D.1 (Modified Corollary 3, [70])**.** *For any smooth concave function f defined on $\prod_{i=1}^{n}[\ell_i, u_i]$ for $u_i > \ell_i$, for any random vector $\mathbf{X}$ supported on that domain with mixing matrix $\mathbf{\Gamma}$, and for $\varepsilon > 0$,*

$$\mathbb{P}\left(f(\mathbf{X}) - \mathbb{E}_{\mathbf{X}}\left[f(\mathbf{X})\right] \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\left\|\mathbf{\Gamma}\right\|^2 \mathbb{E}_{\mathbf{X}}\left|\mathbf{D}\nabla_{\boldsymbol{x}} f\right|^2}\right).$$

Modifications for the Proof of Corollary 3, [70]. Employing the natural upper bound for the absolute index-wise distance, $|x_i - y_i| \leq \mathbb{1}\left(x_i = y_i\right)$ for $\mathbf{x}, \mathbf{y} \in [0,1]^n$ yields [70, Corollary 1, Equation (2.14)]. Instead, one can use $|x_i - y_i| \leq |u_i - \ell_i|\,\mathbb{1}\left(x_i = y_i\right)$ on an arbitrary compact set and propose a modified version of [70, Equation (2.25)] for any smooth concave function:

$$\int f\, dQ - \int f\, dP \leq d_2\left(Q, P\right)\left(\int \left|\mathbf{B}\nabla f\right|^2 dP\right)^{1/2}. \tag{D.5}$$

The matrix $\mathbf{B}$ is a diagonal matrix with $\left(\mathbf{B}\right)_{ii} = (u_i - \ell_i)$. Given (D.5), the proof in [70] follows directly. □

Proof of Theorem 5.2. Concavity of effective resistance with respect to branch resistances allows direct use of Corollary D.1 that yields Theorem 5.2. □

D.3 Proofs for Theorems 5.3-5.4

Theorems 5.3-5.4 are consequences of total effective resistance and its rather interesting properties. We first prove Theorem 5.3. Followed by results from [69] about total effective resistance, without proof, we conclude Corollary 5.3 and Theorem 5.4.

Proof of Theorem 5.3. Since the total effective resistance is a concave function of the resistances $\boldsymbol{r}$, [69] (also a direct consequence of (D.1)), Theorem 5.3 follows from Corollary D.1. □

Lemma D.4 (Power Dissipation, [69])**.** *Let a random current be injected* $\mathbf{J}$ *into a (deterministic) circuit* $\mathcal{G}$*, then, the expected dissipated power is* $\mathbb{E}[P] \triangleq \mathbb{E}\left[\mathbf{J}\underline{\mathbf{L}}^{\dagger}\mathbf{J}\right] = \frac{1}{|\mathcal{V}|}\sigma$.

Proof of Corollary 5.3. Direct application of Lemma D.4 and Theorem 5.3 yields Corollary 5.3. □

Lemma D.5 (Mean-square Branch Voltages, [69])**.** *A deterministic circuit* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{g})$ *excited by a randomly injected current as in Lemma D.4, satisfies* $\forall i \leftrightarrow j \in \mathcal{E}$*:*

$$\frac{\partial}{\partial g_{ij}}\Sigma(\boldsymbol{g}) = -|\mathcal{V}|\,\mathbb{E}\left[V_{ij}^2\right].$$

The proof of Theorem 5.4. The proof employs concavity of $\nabla_{\boldsymbol{g}}\sigma$ in $\boldsymbol{g}$, which follows algebraically from:

$$(\boldsymbol{e}_a - \boldsymbol{e}_b)^{\top}\frac{\partial \underline{\mathbf{L}}^{\dagger}/2}{\partial g_{ij}}(\boldsymbol{e}_a - \boldsymbol{e}_b) = (\boldsymbol{e}_a - \boldsymbol{e}_b)^{\top}\underline{\mathbf{L}}^{\dagger}\mathbf{J}^{ij}\underline{\mathbf{L}}^{\dagger}(\boldsymbol{e}_a - \boldsymbol{e}_b).$$

Repeated application of this formula yields that $\nabla_{\boldsymbol{g}}\Sigma$ is concave in $\boldsymbol{g}$. The rest follows from Corollary D.1 with a modification of the denominator: $\left(\boldsymbol{\ell}^{\top}\mathbf{P}\right)_{i\leftrightarrow j} = u_{ij} - \ell_{ij}$ is the length of the support for the conductance g_{ij}. □

D.4 Multiple Components on a Single Branch

In a linear noisy resistive network with deterministic components, multiple resistors between any two nodes can be replaced with their equivalent resistance. As a proof of concept, let a fabrication process yield statistically independent resistances R_1 and R_2 with known densities $p_{R_1}(r_1)$ and $p_{R_2}(r_2)$ respectively. When a fixed potential V is applied to the nodes that these resistances connect, the ensemble average currents $\mathbb{E}_{\mathcal{F}}[I_s]$ (for series connection)

and $\mathbb{E}_{\mathcal{F}}[I_p]$ (for parallel connection) exhibits Gaussian statistics with $\varsigma = \sqrt{2kt}$,

$$\mathbb{E}[I_s] \sim \mathcal{N}\left(V\left\|\frac{p_S(s)}{s}\right\|_1, \left(\varsigma\left\|\frac{p_S(s)}{s}\right\|_2\right)^2\right),$$

$$\mathbb{E}[I_p] \sim \mathcal{N}\left(V\sum_{i\in\{1,2\}}\left\|\frac{p_{R_i}(r_i)}{r_i}\right\|_1, \varsigma^2\sum_{i\in\{1,2\}}\left\|\frac{p_{R_i}(r_i)}{r_i}\right\|_2^2\right),$$

where $p_S(\cdot) = p_{R_1}(\cdot) * p_{R_2}(\cdot)$, linear convolution of the individual densities. Even when component stochasticity is present, any two nodes might be considered to be connected by a unique component.

REFERENCES

[1] M. Bahrepour, N. Meratnia, M. Poel, Z. Taghikhaki, and P. J. Havinga, "Distributed event detection in wireless sensor networks for disaster management," in *2010 International Conference on Intelligent Networking and Collaborative Systems*. IEEE, 2010, pp. 507–512.

[2] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *New England Journal of Medicine*, vol. 363, no. 4, pp. 301–304, 2010.

[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI Vision Benchmark Suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[4] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Springer Science & Business Media, 2013.

[5] A. Wald, *Sequential Analysis*. John Wiley & Sons, 1947.

[6] D. P. Bertsekas, *Dynamic Programming and Optimal Control 4th Edition*. Belmont, MA: Athena Scientific, 2017, vol. I.

[7] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. MIT Press, 2012.

[8] T. J. Sargent, *Dynamic Macroeconomic Theory*. Harvard University Press, 2009.

[9] N. L. Stokey, *Recursive Methods in Economic Dynamics*. Harvard University Press, 1989.

[10] S. M. Ross, *Introduction to Stochastic Dynamic Programming*. Academic Press, 2014.

[11] A. G. Schwing, C. Zach, Y. Zheng, and M. Pollefeys, "Adaptive random forest - How many experts to ask before making a decision?" in *2011 IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 1377–1384.

[12] N. C. Sevüktekin, A. G. Schwing, and A. C. Singer, "Distributed estimation via opinion dynamics," in *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, 2019, pp. 476–480.

[13] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[14] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *Journal of Parallel and Distributed Computing*, vol. 67, no. 1, pp. 33–46, 2007.

[15] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[16] F. S. Cattivelli and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Transactions on Automatic Control*, vol. 55, no. 9, pp. 2069–2084, 2010.

[17] D. Berend and A. Kontorovich, "A finite sample analysis of the naïve Bayes classifier," *Journal of Machine Learning Research*, vol. 16, pp. 1519–1545, 2015.

[18] D. Berend and J. Paroush, "When is Condorcet's jury theorem valid?" *Social Choice and Welfare*, vol. 15, no. 4, pp. 481–488, 1998.

[19] E. Baharad, J. Goldberger, M. Koppel, and S. Nitzan, "Distilling the wisdom of crowds: Weighted aggregation of decisions on multiple issues," *Autonomous Agents and Multi-Agent Systems*, vol. 22, no. 1, pp. 31–42, 2011.

[20] R. Ben-Yashar and J. Paroush, "A nonasymptotic Condorcet jury theorem," *Social Choice and Welfare*, vol. 17, no. 2, pp. 189–199, 2000.

[21] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[22] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in Neural Information Processing Systems*, 2009, pp. 2035–2043.

[23] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet EM: A provably optimal algorithm for crowdsourcing," in *Advances in Neural Information Processing Systems*, 2014, pp. 1260–1268.

[24] J. Goldberger, "Combining soft decisions of several unreliable experts," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2334–2338.

[25] P. Welinder and P. Perona, "Online crowdsourcing: Rating annotators and obtaining cost-effective labels," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 25–32.

[26] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Advances in Neural Information Processing Systems*, 2011, pp. 1953–1961.

[27] Q. Liu, J. Peng, and A. T. Ihler, "Variational inference for crowdsourcing," in *Advances in Neural Information Processing Systems*, 2012, pp. 692–700.

[28] J. Ok, S. Oh, J. Shin, and Y. Yi, "Optimality of belief propagation for crowdsourced classification," in *International Conference on Machine Learning*, 2016, pp. 535–544.

[29] A. Khetan and S. Oh, "Achieving budget-optimality with adaptive schemes in crowdsourcing," in *Advances in Neural Information Processing Systems*, 2016, pp. 4844–4852.

[30] F. Parisi, F. Strino, B. Nadler, and Y. Kluger, "Ranking and combining multiple predictors without labeled data," *Proceedings of the National Academy of Sciences*, vol. 111, no. 4, pp. 1253–1258, 2014.

[31] U. Shaham, X. Cheng, O. Dror, A. Jaffe, B. Nadler, J. Chang, and Y. Kluger, "A deep learning approach to unsupervised ensemble learning," in *International Conference on Machine Learning*, 2016, pp. 30–39.

[32] R. Hegde and N. R. Shanbhag, "Soft digital signal processing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 9, no. 6, pp. 813–823, 2001.

[33] N. R. Shanbhag, R. A. Abdallah, R. Kumar, and D. L. Jones, "Stochastic Computation," in *Design Automation Conference*. IEEE, 2010, pp. 859–864.

[34] S. Zhang and N. R. Shanbhag, "Embedded algorithmic noise tolerance for signal processing and machine learning systems via data path decomposition," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3338–3350, 2016.

[35] R. A. Abdallah and N. R. Shanbhag, "An energy-efficient ECG processor in 45-nm CMOS using statistical error compensation," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 11, pp. 2882–2893, 2013.

[36] E. P. Kim, D. J. Baker, S. Narayanan, D. L. Jones, and N. R. Shanbhag, "Low power and error resilient PN code acquisition filter via statistical error compensation," in *2011 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2011, pp. 1–4.

[37] D. Seo and L. R. Varshney, "Information-theoretic limits of algorithmic noise tolerance," in *2016 IEEE International Conference on Rebooting Computing (ICRC)*. IEEE, 2016, pp. 1–4.

[38] E. P. Kim and N. R. Shanbhag, "Statistical analysis of algorithmic noise tolerance," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 2731–2735.

[39] N. C. Sevüktekin and A. C. Singer, "The good, the bad, algorithmic noise tolerance (ANT), the ugly," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5366–5370.

[40] J. Von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components," *Automata Studies*, vol. 34, pp. 43–98, 1956.

[41] E. F. Moore and C. E. Shannon, "Reliable circuits using less reliable relays," *Journal of the Franklin Institute*, vol. 262, no. 3, pp. 191–208, 1956.

[42] H. Nyquist, "Thermal agitation of electric charge in conductors," *Physical Review*, vol. 32, no. 1, p. 110, 1928.

[43] H. A. Haus, R. B. Adler, and T. Teichmann, "Circuit theory of linear noisy networks," *Physics Today*, vol. 13, p. 61, 1960.

[44] N. C. Sevüktekin, M. Raginsky, and A. C. Singer, "Linear noisy networks with stochastic components," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 5386–5391.

[45] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[46] M. J. A. marquis de Condorcet, *Essai sur L'application de L'analyse a la Probabilite des Decisions: Rendues a la Pluralite de Voix*. De l'Imprimerie Royale, 1785.

[47] J. Zhang, R. S. Blum, and H. V. Poor, "Approaches to secure inference in the internet of things: Performance bounds, algorithms, and effective attacks on IoT sensor networks," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 50–63, 2018.

[48] I. Ochoa-Alvarez, "Genomic data compression and processing: Theory, models, algorithms, and experiments," Ph.D. dissertation, Stanford University, 2016.

[49] S. M. Ross, *Applied Probability Models with Optimization Applications*. Courier Corporation, 2013.

[50] R. Durrett, *Probability: Theory and Examples*. Cambridge University Press, 2019, vol. 49.

[51] S. Bubeck, N. Cesa-Bianchi et al., "Regret analysis of atochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[52] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[53] A. P. Dempster, "A generalization of Bayesian inference," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 2, pp. 205–232, 1968.

[54] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976, vol. 42.

[55] L. A. Zadeh, "Fuzzy logic," *Computer*, vol. 21, no. 4, pp. 83–93, 1988.

[56] C.-C. Lee, "Fuzzy logic in control systems: Fuzzy logic controller," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, no. 2, pp. 404–418, 1990.

[57] R. Dorfman, "The detection of defective members of large populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, December 1943.

[58] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933.

[59] J. Paroush, "Stay away from fair coins: A Condorcet jury theorem," *Social Choice and Welfare*, vol. 15, no. 1, pp. 15–20, 1997.

[60] B. Hajek, *Random Processes for Engineers*. Cambridge University Press, 2015.

[61] D. Berend and A. Kontorovich, "A sharp estimate of the binomial mean absolute deviation with applications," *Statistics & Probability Letters*, vol. 83, no. 4, pp. 1254–1259, 2013.

[62] M. Kearns and L. Saul, "Large deviation methods for approximate probabilistic inference," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998, pp. 311–319.

[63] M. Raginsky, I. Sason et al., "Concentration of measure inequalities in information theory, communications, and coding," *Foundations and Trends in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013.

[64] S. S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2005.

[65] W. Rudin, *Functional Analysis*. McGrawHill, 1991.

[66] I. N. Hajj, "Circuit theory in circuit simulation," *IEEE Circuits and Systems Magazine*, vol. 16, no. 2, pp. 6–10, 2016.

[67] F. Dörfler, J. W. Simpson-Porco, and F. Bullo, "Electrical networks and algebraic graph theory: Models, properties, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 977–1005, 2018.

[68] W. Ellens, F. Spieksma, P. Van Mieghem, A. Jamakovic, and R. Kooij, "Effective graph resistance," *Linear Algebra and Its Applications*, vol. 435, no. 10, pp. 2491–2506, 2011.

[69] A. Ghosh, S. Boyd, and A. Saberi, "Minimizing effective resistance of a graph," *SIAM Review*, vol. 50, no. 1, pp. 37–66, 2008.

[70] P.-M. Samson et al., "Concentration of measure inequalities for Markov chains and Φ-mixing processes," *The Annals of Probability*, vol. 28, no. 1, pp. 416–461, 2000.

[71] A. Kontorovich, "Obtaining measure concentration from Markov contraction," *Markov Processes and Related Fields*, vol. 18, pp. 613–638, 2012.

[72] A. W. Drake, *Fundamentals of Applied Probability Theory*. McGraw-Hill, 1967.

[73] C. D. Meyer, Jr, "Generalized inversion of modified matrices," *SIAM Journal on Applied Mathematics*, vol. 24, no. 3, pp. 315–323, 1973.