## Abstract

This paper considers the disclosure problem of a sender who wants to use hard evidence to persuade a receiver towards higher actions. When the receiver hopes to make inferences based on the distribution of the data, the sender has an incentive to drop observations to mimic the distributions observed under better states. Selecting equilibria under a criterion that strategies remain optimal after allowing for announcements about messaging strategies, we find that, in the limit when datasets are large, it is optimal for senders to play an *imitation strategy*, under which they submit evidence to prove they have enough data corresponding to a desirable state. The receiver makes inferences by checking if senders meet a sufficient burden of proof to take a high action. The outcome exhibits partial pooling: senders are honest when either they have little data or the state is good, but they try to deceive the receiver when they have access to a lot of data and the state is bad.

# Inference under Selectively Disclosed Data

February 22, 2022

## 1 Introduction

In order to take appropriate actions, decision-makers often rely on data and evidence supplied by self-interested informants, such as companies or individuals. Their informants, however, are often motivated by private concerns about the conclusions the decision-maker draws from the data, and have strong preferences about what action the decision-maker should take. For example, researchers carrying out experiments might aim to support a particular hypothesis, either out of personal bias or because their research is sponsored by an interested party (e.g. soda manufacturers and drug companies). Public companies that release accounting and performance data aim to benefit their shareholders, and therefore generally prefer to disclose data that increases the value of their stock. The amount and specificity of data available to report in both of these cases is increasing as data becomes easier to generate and store – both the models used to analyze experiments, and those used to predict financial outcomes, often take as inputs many individual datapoints over a variety of possible outcomes.

We model this scenario as a communication game between an uncommitted sender with known preferences and a sophisticated receiver. There is a finite number of states, each of which is associated with a distribution over a finite set of outcomes. The sender and the receiver share a common model about the state of the world and state-induced outcome distributions, but the sender observes the true dataset of outcomes, while the receiver does not, and has uncertainty about how many draws there are.

Because disclosure is voluntary, senders will selectively withhold information from receivers if it "looks bad". A sophisticated receiver anticipates this, and accounts for the sender's omission strategy when updating from the data they are shown. It is known in a single-datum case that when senders act strategically and receivers are unsure if senders hold evidence, then sophisticated receivers cannot fully separate their uncertainty about the whether the sender is informed from their uncertainty about the implications of the data about the state of the world – there is partial pooling Dye (1985). The same is true of a larger dataset of uncertain size, when senders choose a way to disclose part of their dataset rather than either disclose or not. If datasets are large to begin with, the only value of additional data to a sender is in benefiting their ability to game the receiver's beliefs – they have more flexibility over datasets to send, and are able to send the same datasets as senders in better states that are endowed with less data. There is a large set of possible strategies, and generally many equilibria, but it is intuitive that coalitions of senders will pool in ways that are optimal for the entire coalition, and this selects a particular equilibrium outcome.

In this paper, we show that in the limit as $N$ approaches infinity, partial pooling arises in equilibrium because senders with "excess" data target the same strategies as senders under more favorable states that they can mimic. The payoff to senders in the limit can be thought of as the outcome of an equilibrium of a continuous-dataset approximation of big data, in which datasets no longer consist of individual draws, but of a mass of data corresponding to a particular state's distribution. Within the continuum model, senders can implement an optimal targeting strategy either by *imitating* the distribution of data under the target state exactly, or by providing a great-enough mass of data of all the observations that maximize the likelihood ratio under the target state and less-desirable states. Under either approach, the receiver's inference problem comes down to demanding that senders meet a burden of proof in order to elicit a higher action, which is an amount of data that may differ depending on which state of the world senders target. Senders choose which state to imitate by weighing a combination of the inherent desirability of the state, and their relative advantage at targeting it, which depends on the similarity of the state's outcome distribution to the true distribution. Relative to full revelation, the partial pooling equilibrium advantages senders with good access to data under bad states at the expense of senders in good states with little data.

The paper is laid out as follows. Section 2 begins by outlining a model of communication with a finite dataset. In Section 3, we solve an example game and introduce a notion of robustness to an "inclusive" credible announcement, related to those in Matthews et al. (1991), that is satisfied by reasonable equilibria. In Section 4, we show that robustness to these coalitional deviations is equivalent to optimality under a lexicographic order over equilibrium outcomes, and give an algorithm that constructs the unique equilibrium that survives them. We then turn to the continuous-dataset approximation to the communication model in Section 5, and in it we propose an *imitation equilibrium outcome* with the property that types with large data endowments target the strategies of better-state, lower-data types in order to deceive the receiver into taking higher-payoff actions. In Section 6, we show that the imitation equilibrium outcome in the continuous-data approximation exactly describes the large-data limit of lexicographically optimal equilibrium outcomes. Section 7 discusses how to substitute the assumption of exogenous data generation for costly, endogenous data acquisition, and concludes.

## 1.1 Review of literature

A number of papers show how the unraveling results of Grossman (1981) and Milgrom (1981) can fail when the sender is endowed with a random amount of evidence. Dye (1985) models a single-datum case, under which a nonzero probability of senders failing to receive evidence results in pooling between those senders and senders with unfavorable evidence. Subsequent papers by Shin ('94, '03) show that sender-optimality of the "sanitization strategy" that reveals only sufficiently favorable evidence extends to games with multiple pieces of evidence, so long as the payoff-relevant state is binary (success vs. failure).

Like us, Dzuida (1985) investigates the question of how senders pool with one another when evidence is disposable by approximating large datasets with a continuous measure of evidence. Her model assumes a binary state and binary signal realizations, as well as a positive measure of honest types who must disclose their entire set of evidence. The results resemble our findings in the binary-state case, and she focuses on an outcome with payoffs continuous in data endowment that coincides with our outcome selected by lexicographic optimality and immunity to coalitional announcements. Another observation common to our analysis and Dzuida is that, relative to the case with a symmetrically informed receiver, outcomes are worse for high-state, low-evidence senders, who cannot distinguish themselves from low-state senders; and they are better for low-state, high-evidence senders, who can pretend to be better types. Following Dzuida, Felgenhauer and Schulte (2014) model the discretionary disclosure of binary evidence with an endogenous and sequential process of data acquisition. We consider endogenous information acquisition preceding disclosure under our model, with a focus on comparative outcomes rather than incentives to invest.

Our results speak to a discussion of persuasion using hard evidence in fields like scientific research and corporate asset management. Shin (2003) applies the sanitation strategy to the disclosure of independent successes in maximizing the market value of corporate stocks; in comparison, we analyze incomplete disclosure strategies when the market uses large datasets to inform more complex models in which inference of the state depends on signals' joint distributions. Relatedly, there is a large body of work examining the effects of publication bias that arises due to the systematic omission of negative or inconclusive results. Simonsohn et al. (2014) and Andrews and Kasy (2019) propose methods to identify and correct for the bias induced by selective reporting of scientific findings, using observable distortions in the distribution of reported data (e.g. the "p-curve"). Although these studies do not consider *strategic* data omission, their inference problem is similar to that faced by our receiver against the strategy of the sender.

# 2 Model

There is a sender ($s$), who wishes to communicate to a receiver ($r$) about an unknown state of the world. The receiver is uninformed, and relies on the sender to provide them with evidence in order to make a choice that affects both themselves and the sender. However, the sender's and receiver's incentives are misaligned: the sender's preferred action for the receiver does not depend on the true state, and instead, the sender always wants the receiver to take a higher action (i.e. one that is more beneficial to the sender). Furthermore, the sender is able to drop data as they please: the dataset they submit to the receiver may be incomplete, and the receiver must make inferences assuming that the sender will omit data when it is in their strategic interest.

**States and payoffs.** The sender and receiver share a common prior $\beta_0(\cdot)$ on the state of the world $\theta \in \Theta \subseteq \mathbb{R}$. The support of the prior — the set of states they consider possible — is finite, $\Theta = \{\theta_1, \ldots, \theta_J\}$, with $\theta_j$ increasing in $j$. I assume that the receiver takes the action $a_r = \mathbb{E}[\theta]$ that matches the expectation of their belief over $\Theta$.[1] In short, the receiver's optimal action is increasing in their expectation of the state of the world.[2]

The sender is wishes the receiver to take as high an action as possible, and thus aims to persuade the receiver that the expected state is high. Their payoff from persuading the receiver to adopt a given belief $\beta$ is

$$u_s(\beta) = \tilde{u}_s(a_r) = \tilde{u}_s(\mathbb{E}_\beta[\theta]).$$

We assume $\tilde{u}_s(a_r)$ is increasing in $a_r$, and so $u_s(\beta)$ is increasing in $\mathbb{E}_\beta[\theta]$.

**Evidence.** The private information of the sender is communicable: it comes in the form of hard evidence about the state of the world. In particular, the sender has access to a dataset. Each *datapoint* in the dataset is an observation within a space of outcomes $\mathcal{D} = \{1, \ldots, D\}$, and each state of the world induces a different distribution of observations – when the state is $\theta_j$, the distribution of outcomes of a single experiment is $f_j$. I assume that, while all $f_j$ share full support over $\mathcal{D}$, they are distinct, so that any two states are distinguishable by the distribution of outcomes they generate.

The entire dataset consists of a finite collection of i.i.d. draws of $f_j$. Different senders differ in how much data they can acquire, and ex-ante, the *mass distribution* of data, $g(n)$, is known to both parties, but the true number of observations $n$ is not. Nevertheless, the number of observations possible is assumed to be bounded, and when the support of $g(n)$ is in $\{1, \ldots, N\}$, the sender's dataset, or *type*, is given by

$$t = \frac{1}{N}(n_1, \ldots, n_D),$$

where $t(d) := \frac{n_d}{N}$ is the normalized total mass of experiments in which the outcome is $d$. The total normalized mass of the dataset is $\frac{n}{N} = \frac{1}{N}\sum_{d=1}^{D} n_d$, and alternately denoted as $|t|$.

[1] Note that elements of $\Theta$ and actions $a_r$ are assumed to already be appropriately normalized: if the receiver's optimal action is intead $a'_r = h(\mathbb{E}[v(\theta')])$ where $v$ and $h$ are increasing functions, the mappings $\theta = v(\theta')$ and $a_r = h(a'_r)$ renormalize the state and action space to the correct form.

[2] A canonical example of a payoff function that justifies this choice is $u_r = -(a_r - \theta)^2$.

We denote the ex-ante probability that the sender will be of type $t$ by $q(t)$, and the posterior over the state conditional on the sender receiving $t$ as $\pi(\cdot|t)$.[3]

**Messaging and inference.** The receiver does not directly observe the sender's type. Instead, after receiving a dataset, the sender voluntarily submits a message to the receiver, consisting of observations from the dataset. I assume that the sender's access to data determines whether it is feasible to submit a particular body of evidence to the receiver:

**Assumption 2.1** *The sender can send any message $m = \frac{1}{N}(\tilde{n}_1, \ldots, \tilde{n}_D)$ that is a* subset *of their dataset ($m \subseteq t$), where*

$$m \subseteq t \;\Leftrightarrow\; m(d) \leq t(d) \; \forall d \in \mathcal{D}.$$

The disclosure game with these parameters is $\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g, u_s)$, with type space $\mathcal{T}_N$ and message space $\mathcal{M}_N$ that are isomorphic to each other, containing all vectors $\frac{1}{N}(n_1, \ldots, n_D)$ with the sum of nonnegative integers $n_1 + \ldots + n_D \leq N$. Irrespective of $N$, the spaces $\mathcal{T}_N$ and $\mathcal{M}_N$ can be embedded into a *global data space* $\mathcal{F} = [0,1] \times \Delta\mathcal{D}$, consisting of all vectors $(w_1, \ldots, w_D)$ of nonnegative real weights with $\sum_{d=1}^D w_d \leq 1$.

In this game, senders can choose which feasible message to send given their type according to a possibly mixed messaging strategy, $\sigma(\cdot|t) : \mathcal{T}_N \to \Delta\mathcal{M}_N$ – their choice of a message is the only means by which they can influence the receiver's action and their own payoffs.

The receiver's belief over states is $\beta \in \Delta\Theta$. After receiving message $m$, the receiver updates their beliefs according to $\beta(\cdot|m) : \mathcal{M}_N \to \Delta\Theta$. More primitively, though only of indirect consequence to the sender, the receiver holds beliefs, denoted $\beta[\cdot|m]$ with square brackets, about the sender's type, which imply their beliefs about the state:

$$\beta(\theta_j|m) = \frac{\sum_{t \in \mathcal{T}_N} \beta[t|m]\pi(\theta_j|t)}{\sum_{t \in \mathcal{T}_N} \beta[t|m]}.$$

**PBE and outcomes.** Following the convention in signaling games, our base solution concept is PBE (alternately referred to as PBE or "equilibrium").

Importantly, we assume the sender is unable to commit ex-ante to a messaging policy, in which they give up playing optimally when endowed with certain datasets in exchange for more lenient inferences in other scenarios. While doing so successfully may indeed benefit the sender in expectation, there is little incentive for the sender to keep the commitment in the interim stage, both in a one-shot setting, and when the sender is anonymous in a

[3] The exact expressions are

$$q(t) = \frac{n!}{\Pi_{d=1}^D n_d!} g(n) \sum_{j'} \beta_0(\theta_{j'}) \Pi_{d=1}^D f_{j'}(d)^{n_d}, \quad \text{and} \quad \pi(\theta_j|t) = \frac{\beta_0(\theta_j)\Pi_{d=1}^D f_j(d)^{n_d}}{\sum_{j'} \beta_0(\theta_{j'})\Pi_{d=1}^D f_{j'}(d)^{n_d}}.$$

large population. Thus, we expect the sender under each type to optimize $\sigma(\cdot|t)$ given their anticipation of the receiver's response.

**Definition** An equilibrium is $(\sigma^*, \beta^*)$ where

1. $\sigma^*$ prescribes the highest-payoff feasible message to a sender of each type:

$$\sigma^*(\cdot|t) \in \arg\max_{m \subseteq t} u_s(\beta^*(\cdot|m)).$$

2. $\beta^*$ is consistent with Bayesian updating given knowledge that the sender plays to $\sigma^*$:

$$\beta^*[t|m] = \frac{q(t)\sigma^*(m|t)}{\sum_{t \in \mathcal{T}_N} q(t)\sigma^*(m|t)} \qquad \text{for all on-path } m,$$

and $\beta^*[t|m] = 0$ if $m \not\subseteq t$.

There is off-path indeterminacy in the receiver's beliefs, so there may be multiple $\beta^*$, differing on off-path messages, that jointly form an equilibrium along with a given $\sigma^*$. However, we can define

$$\beta_{\sigma^*}[t|m] := \begin{cases} \frac{q(t)\sigma^*(m|t)}{\sum_{t \in \mathcal{T}_N} q(t)\sigma^*(m|t)} & \text{for all on-path } m, \\ \mathbb{1}(\arg\min_{t' \subseteq m} \mathbb{E}_{\pi(\cdot|t')}[\theta]) & \text{for all off-path } m \in \mathcal{F} \end{cases}$$

that, firstly, extends the receiver's inference function to all messages in $\mathcal{F}$, and therefore all of $\mathcal{M}_N$; and secondly, makes all off-path messages minimally attractive for the sender. Given the following lemma, if $(\sigma^*, \beta_{\sigma^*})$ is a PBE, we will often suppress $\beta$ and call $\sigma^*$ an equilibrium:

**Lemma 2.2** *A strategy $\sigma^*$ constitutes a PBE with along with some $\beta$ if and only if $(\sigma^*, \beta_{\sigma^*})$ is a PBE.*

In general, when $N$ is large, the game has many PBE due to self-reinforcing expectations about both off- and on-path play. Rather than using equilibrium as a final solution concept, in the following two sections we will propose a refinement, *lexicographic optimality*, that selects the equilibria we consider most reasonable, due to their robustness to deviations by a coalition of types of senders. We postpone the discussion of details of equilibrium selection until then.

Finally, an *outcome* of an equilibrium is the mapping from a dataset in $\mathcal{F}$ to the payoff[4] that a sender endowed with the dataset receives by best-responding to $\beta_{\sigma^*}$,

$$u_{\sigma^*}(t) = \max_{m \in \mathcal{F}: m \subseteq t} u_s(\beta_{\sigma^*}(\cdot|m)).$$

[4] Since payoffs are monotone in actions, this is equivalent to a mapping from types to the actions induced by their messages in equilibrium.

While it is straightforward to define outcomes for types in $\mathcal{T}_N$ as they payoff they obtain in equilibrium, and $u_{\sigma^*}(t)$ coincides with this definition for positive-probability types, the extension to all of $\mathcal{F}$ allows comparison of outcomes across games with different type spaces, under the thought experiment: "what payoff would a sender with dataset $t$ obtain if they know they are playing against a receiver who believes they are in equilibrium $\sigma^*$ of game $\mathcal{G}_N$, even if $t$ is not a possible type in $\mathcal{G}_N$"?

# 3 Example: 2 states

When the state of the world is binary, $\Theta = \{\theta_1, \theta_2\}$, the receiver's belief is a single number $\beta(\theta_2) \in [0, 1]$, and the sender's problem boils down to convincing the receiver that the state is $\theta_2$ with as high a probability as possible.

## 3.1 Binary outcomes, binary states

To further simplify the problem, suppose that the domain of $f_j$ is also binary, $\mathcal{D} = \{1, 2\}$. Let $f_2(2) = p_2$ and $f_1(2) = p_1$, with $p_2 > p_1$, so that outcome 2 is more likely under state 2 than state 1. For a given $N$, assuming $g(n)$ has full support on $\{0, \dots, N\}$, the set of possible types of the sender, $\mathcal{T}_N$, is illustrated below, with the notation $t = (n_1, n_2)$.

| | | | | | | |
|---|---|---|---|---|---|---|
| | $(0, N)$ | | | | | |
| | $(0, N-1)$ | $(1, N-1)$ | | | | |
| $n_2$ | $(0, N-2)$ | $(1, N-2)$ | $(2, N-2)$ | | | |
| | $(0, 1N-3)$ | $(1, N-3)$ | $(2, N-3)$ | $(3, N-3)$ | | |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | |
| | $(0, 0)$ | $(1, 0)$ | $(2, 0)$ | $(3, 0)$ | $\dots$ | $(N, 0)$ |
| | | | $n_1$ | | | |

Table 1: $\mathcal{T}_N$ when $\Theta = \{1, 2\}$ and $\mathcal{D} = \{1, 2\}$.

The set of possible messages, $\mathcal{M}_N$, is identical to the type space. Table 1 above illustrates the set of messages available to type $t = (1, N-2)$ in blue, and the set of types capable of sending message $m = (1, N-2)$ in red.

Whenever $N \geq 2$, there are multiple equilibria. For instance, when $N = 2$, the data mass distribution is $g(0) = g(1) = g(2) = \frac{1}{3}$, the prior is $\beta(2) = \frac{1}{2}$, and the distribution of outcomes is $p_2 = 0.9$, $p_1 = 0.8$, the game has the 3 equilibria in Table 2. The first equilibrium separates senders into 3 pools, which all obtain different payoffs, while the outcomes of the remaining 2 equilibria are identical, and involve 2 different payoffs, depending on the sender's type. While types $(0, 1)$ and $(1, 1)$ could obtain a higher payoff than they do in $\sigma_2^*$ and $\sigma_3^*$ by separating from the other 3 types, they do not do so, because of adverse beliefs about the receiver's response to message $(0,1)$. In $\sigma_2^*$, because $(0,1)$ is off-path, the receiver may believe that the sender's type is $(1,1)$ with high probability if $(0,1)$ is observed, which makes the message unattractive. In $\sigma_3^*$, $(0,0)$ is on-path, but played by $(1,1)$ with greater probability than it is played by $(0,1)$, despite the fact that both types are indifferent between playing it and $(0,0)$: this worsens message $(0,1)$ and improves message $(0,0)$, which in turn supports the indifference between the two messages that gives rise to these counterintuitive mixing probabilities. If types $(0,1)$ and $(1,1)$ could together announce to the receiver that they plan to play as in equilibrium $\sigma_1^*$ and be believed, they would, and would then keep their word, even without commitment.

| | | (0,2) | (0,1) | (1,1) | (0,0) | (1,0) | (2,0) |
|---|---|---|---|---|---|---|---|
| $\sigma_1^*$ | Messages | (0,2) | (0,1) | (0,1) | (0,0) | (0,0) | (0,0) |
| | Payoffs | 1.56 | 1.49 | 1.49 | 1.47 | 1.47 | 1.47 |
| $\sigma_2^*$ | Messages | (0,2) | (0,0) | (0,0) | (0,0) | (0,0) | (0,0) |
| | Payoffs | 1.56 | 1.48 | 1.48 | 1.48 | 1.48 | 1.48 |
| $\sigma_3^*$ | Messages | (0,2) | (0,1) and (0,0) | (0,1) and (0,0) | (0,0) | (0,0) | (0,0) |
| | Payoffs | 1.56 | 1.48 | 1.48 | 1.48 | 1.48 | 1.48 |

Table 2: 3 equilibria of $\mathcal{G}_2$ with $p = 0.9$, $q = 0.8$

The equilibrium $\sigma_1^*$ is not vulnerable to such announcements, and has a simple form: senders send as many observations of outcome 2 as they can, and none of outcome 1. Indeed, for all $N$ there exists an immune equilibrium, and in cases where $|\Theta| = |\mathcal{D}| = 2$, it entails disclosing only observations of outcome 2.

**Definition** Given an outcome $u_{\sigma^*}$, a set of types $T$ has a *credible inclusive announcement* that they will play a partial strategy $\hat{\sigma}_M$ over message set $M$ for payoff $v$ if

- $\hat{\sigma}_M : M \times T \to \mathbb{R}$ is such that $\sum_{t \in T} \hat{\sigma}_M(m|t) = 1$ for all $m \in M$, $\sum_{m \in M} \hat{\sigma}_M(m|t) = 1$ for all $t \in T$, and $u_s(\beta_{\hat{\sigma}_M}(\cdot|m)) = v$ for all $m \in M$.
- $T = \{t : u_{\sigma^*}(t) \leq v \text{ and } \exists m \in M \text{ s.t. } m \subseteq t\}$, and there is some $t \in T$ with $u_\sigma(t) < v$.

Credible inclusive announcements are related to the concept of a credible announcement **?**, which does not impose that all types that weakly prefer to obtain $v$ to their equilibrium payoff participate in the announcement if possible, but rather only types that strictly prefer $v$. In our context, robustness to credible announcements is too strong, and often rules out all equilibria: types that are indifferent between a base equilibrium and an announcement may no longer able to obtain their payoff from the base equilibrium once the announcement is made and believed, and such announcements may not correspond to any equilibrium at all. In contrast, a sequence of improvements from credible inclusive announcements can always be used to construct an equilibrium, as I show in the following section.

**Claim 3.1** *When $|\Theta| = |\mathcal{D}| = 2$, the unique equilibrium with an outcome immune to credible* *inclusive announcements takes the following form:*

- *On-path messages are* $\{(0, n_2[k])\}_{k=1}^{K}$*, with* $n_2[1] = N$ *and*

$$n_2[k] = \arg \max_{n < n_2[k-1]} u_s \left( \frac{\sum_{n_2=n}^{n_2[k-1]} \sum_{n_1=0}^{N-n_2} \pi(\theta_2 | (n_1, n_2)) q((n_1, n_2))}{\sum_{n_2=n}^{n_2[m-1]} \sum_{n_1=0}^{N-n_2} q((n_1, n_2))} \right)$$

  *for all* $k > 1$*.*

- *A sender plays the most demanding on-path message they can send:*

$$\sigma^*(t) = (0, \max(n_2[k] : n_2[k] \leq t(2))).$$

# 4 Lexicographic optimality

When are outcomes not improvable by announcing that some messages will be used, and ought to be interpreted, differently than in equilibrium? Intuitively, if senders are already using all messages optimally – that is, conditional on the behavior of types they cannot imitate, senders use messages to form pools that give them the highest potential payoffs – then they can do no better.

It turns out that equilibria that are immune to credible inclusive announcements are the same as those with outcomes that satisfy a *lexicographic optimality* condition: across all equilibria, they give the senders who have the highest potential equilibrium payoffs their best possible payoffs, and conditional on this, they also maximize the payoffs to the next-highest-potential-payoff group of senders, and so on.

To state the definition, let $t_\sigma^+(u)$ be the set of possible types that obtain a payoff of at least $u$ under outcome $u_\sigma$.

**Definition** We say $u_\sigma(\cdot)$ **weakly lexicographically dominates** $u_{\sigma'}(\cdot)$ (i.e., $u_\sigma(\cdot) \succeq_l u_{\sigma'}(\cdot)$) if either there exists an element $u$ of

$$U := \{u : t_\sigma^+ \setminus t_{\sigma'}^+(u) \text{ is nonempty}\}$$

that is greater than or equal to every element $u'$ of

$$U' := \{u' : t_{\sigma'}^+ \setminus t_\sigma^+(u) \text{ is nonempty}\},$$

or $U'$ is empty.

**Definition** $u_\sigma(\cdot)$ **strictly lexicographically dominates** $u_{\sigma'}(\cdot)$ (i.e., $u_\sigma(\cdot) \succ_l u_{\sigma'}(\cdot)$) if $u_\sigma(\cdot) \neq u_{\sigma'}(\cdot)$ and $u_\sigma(\cdot) \succeq_l u_{\sigma'}(\cdot)$.

Lexicographic dominance defines a partial order on outcomes. When the poset of outcomes has a maximal element, we call it lexicographically optimal:

**Definition** $u_\sigma(\cdot)$ is **lexicographically optimal** if it strictly lexicographically dominates all other equilibrium outcomes.

In general, there is no equilibrium of $\mathcal{G}_N$ that is Pareto optimal for the entire set of sender types. Additionally, unlike the example of Section 3, when $N$ is large there is generally no Blackwell dominant equilibrium that is most informative for the receiver, and best maximizes their payoff over arbitrary payoff functions. However, in cases where either exists, it must coincide with the lexicographically optimal equilibrium.

## 4.1 Construction, existence, and uniqueness

In order to show that the lexicographically optimal equilibrium outcome exists and is unique, we construct it.

We begin with some useful notation. First, fix abstractly a set of types $T \subset \mathcal{T}_N$. Given $\mathcal{T}$, define for every message $m \in \mathcal{M}_N$ and set of messages $M \subseteq \mathcal{M}_N$

$$T^+(m) = \{t \in T : m \subseteq t\} \quad \text{and} \quad T^+(M) = \bigcup_{m \in M} T^+(m),$$

the set of types in $T$ capable of sending $\tilde{f}$ or any $\tilde{f} \in M$, respectively.

Denote the receiver's belief over states after updating their prior based on knowledge that the sender's type is in set $T$ by

$$\beta(\theta|T) = \frac{\sum_{t \in T} \pi(\theta|t) q(t)}{\sum_{t \in T} q(t)}.$$

We say a set of messages $M = \{m_1, \ldots, \tilde{m}_I\}$ implements a pool of sender types $T_{\hat{\sigma}_M}$ if there is an associated partial strategy $\hat{\sigma}_M : M \times \mathcal{T} \to \mathbb{R}$ with $\hat{\sigma}(\cdot|t) \in \Delta M$, satisfying:

A. $t \in T_{\hat{\sigma}_M}(m_i) > 0$ only if $m_i \subseteq t$.

B. $\sum_i \hat{\sigma}_M(m_i|t) = 1$ for all $t \in T_{\hat{\sigma}_M}$.

C. $\sum_{t \in T_{\hat{\sigma}_M}} \hat{\sigma}_M(m_i|t) = 1$ for all $m_i \in M$.

D. $u_s(\beta_{\hat{\sigma}_M}(\cdot|\tilde{f}_i)) = u_s(\beta_{\hat{\sigma}_M}(\cdot|\tilde{f}_j))$ for all $i, j$.

The payoff to a pool is $u(T_{\hat{\sigma}_M}) := u(\beta(\cdot|T_{\hat{\sigma}_M}))$. Note that types in $T_{\hat{\sigma}_M}$ do not pool in the traditional sense of sending the exact same message (and thus being indistinguishable to the receiver). Instead, they may indeed send different messages that induce different beliefs over the mixture of types; however, these beliefs will result in the receiver taking the same action, and are therefore outcome-equivalent.

Finally, with reference to type set $T$, define the set of *upper pools* to be the collection of message sets that implement the pooling of the set of all types in $T$ capable of sending them.

$$\mathcal{P}_T = \{M \subseteq \mathcal{M} : M \text{ implements the pooling of } T^+(M)\}$$

Fixing the strategy of the receiver, if we let $M$ be the set of messages such that the receiver's response yields payoff $u^*$ to the sender, and let $T$ be the set of senders incapable of sending any message that yields payoff greater than $u^*$, then the best response of all senders in $T^+(M)$ to the receiver's strategy is to play some message in $M$. If, in addition, $M \in \mathcal{P}_T$ and $u(T^+(M)) = u^*$, then there exists a best response by senders in $T^+(M)$ that preserves the payoff to $M$ when the receiver best-responds in turn to the updated strategy.

**Lemma 4.1** *For every message set $M$, there is an upper pool $M'$ consisting of types $T_{\hat{\sigma}_{M'}} \subseteq T^+(M)$ such that $u(T_{\hat{\sigma}_{M'}}) \geq u(T^+(M))$; the inequality is strict if $M$ is not itself an upper pool.*

**Lemma 4.2** *For any $T \subseteq \mathcal{T}_N$, the set of utility-maximizing upper pools in $\mathcal{M}_N$, i.e. $\arg\max_{M \in \mathcal{P}_T \bigcup \mathcal{M}_N} u(T_{\hat{\sigma}_M})$, is an upper semilattice in the inclusion order on the set of participating types.*

Observe that it is possible to construct a strategy profile in the following way.

**Algorithm.**

1. Let $T_1 = \mathcal{T}_N$, and define $\mathcal{P}_{T_1}$ to be the set of upper pools over $T_1$. Find the upper pool in $\mathcal{P}_{T_1} \bigcap \mathcal{M}_N$ that yields the highest payoff to participating senders:
$$M_1 \in \arg \max_{M \in \mathcal{P}_{T_1} \bigcap \mathcal{M}_N} u(T_1^+(M)).$$
If there are multiple such pools, then we take their union, which is also in $\mathcal{P}_{T_1} \bigcup \mathcal{M}_N$ by Lemma 4.2.
2. For $s = 2$ onwards, restrict the set of types to $T_s = T_{s-1} \setminus T_{s-1}^+(M_{s-1})$, and find (the union of)
$$M_s \in \arg \max_{M \in \mathcal{P}_{T_s} \bigcap \mathcal{M}_N} u(T_s^+(M)).$$
3. Continue until $T_s \setminus T_s^+(M_s) = \emptyset$, and define $\sigma^*$ by $\sigma^*(m|t) = \hat{\sigma}_{M_s}(m)$ where $M_s$ is the pool containing $m$.

**Theorem 4.3** *$\sigma^*$ is an equilibrium.*

The theorem is immediate from the following lemma, which states that payoffs to the iteratively-constructed pools are strictly decreasing.

**Lemma 4.4** *$u(T_m^+(M_m)) > u(T_{m+1}^+(M_{m+1}))$ for all $m$.*

Indeed, a necessary and sufficient condition for $\sigma^*$ to be an equilibrium is that $u(T_m^+(M_m)) \geq u(T_{m+1}^+(M_{m+1}))$ for all $m$; it is additionally true in this case that the inequality is strict, so each successive pool obtains a different payoff.

By construction, $u_{\sigma^*}$ is unique[5] and lexicographically optimal among equilibrium outcomes. Indeed, all equilibria can be constructed via a version of the algorithm in which the pool chosen in each step need not be the maximal-payoff upper pool. By imposing that we take the largest maximal-payoff pool, we ensure that if $u_{\sigma^*_{alt}}$ coincides with $u_{\sigma^*}$ for all payoffs greater than $v$, then the set of types obtaining payoff $v$ is at least as large under $u_{\sigma^*}$ as it is under $u_{\sigma^*_{alt}}$.

**Theorem 4.5** *$u_{\sigma^*}$ is the unique outcome of equilibria immune to credible inclusive announcements.*

**Proof** By construction, if $u_{\sigma^*_{alt}} \neq u_{\sigma^*}$, then there exists a $v$ such that the set of pools achieving a payoff greater than $v$ is identical in $u_{\sigma^*_{alt}}$ and $u_{\sigma^*}$, but the pool of types $T$ achieving payoff $v$ under $u_{\sigma^*}$ is a strict superset of that under $u_{\sigma^*_{alt}}$. Then types in $T$ can make a credible inclusive announcement that they will play as they do in $\sigma^*$.

If $u_{\sigma^*_{alt}}$ is not lexicographically optimal, then a sequence of improvements by credible inclusive announcements, starting with the senders that achieve the highest payoffs under $\sigma^*$, will terminate in an equilibrium with the lexicographically optimal outcome.

# 5 Large-dataset approximations

We would like to characterize the actions taken by the receiver in the lexicographically optimal equilibrium outcome. However, it is challenging to exactly construct lexicographically optimal outcomes for large datasets, as the runtime for the algorithm is exponential in $N$. An infinite-data approximation to the limit delivers a key simplification: the sender's dataset becomes deterministic given the state and the mass of data they receive, and randomness in individual draws ceases to matter.

## 5.1 Modeling infinite data

When data are finite, the amount of data a sender possesses determines two things: how informative their dataset is about the state of the world, and how much leeway they have to manipulate their message to the receiver by removing datapoints. Having more data improves both of these things, but the first concern becomes less and less important as the amount of data grows: as the number of datapoints increases to infinity, a dataset becomes close to

[5] A slight caveat here is that, while the equilibrium *outcome* that is lexicographically optimal and immune to credible inclusive announcements is unique, in corner cases there can be multiple equilibria that implement it, differing only in the mixing probabilities of different types in the same pool; so we refrain from saying the equilibrium itself is unique.

perfectly informative about the state, so the additional informational value of *any* additional number of datapoints vanishes. Because the impact of additional data on a sender's feasible message set is nonvanishing, having additional data relative to an already-large dataset impacts the sender's outcome almost entirely through the manipulability channel, rather than through its informativeness.

To approximate the big-data scenario, we suppose that a sender is endowed with a *continuous mass* of datapoints. Although any continuum of data is perfectly informative about the state, the mass of data received by senders of different types may differ, and affect their ability to imitate each other. For instance, if the state is $\theta_j$ and they receive a total mass $\mu$ of data, then they receive a measure $\mu f_j(1)$ of observations of outcome 1, $\mu f_j(2)$ of observations of outcome 2, and so on. We assume that the density $g(\mu)$ describing the probability of obtaining a measure $\mu$ of data is continuous on its support $[0, 1]$, and vanishing to 0 at 1. A sender's type is $t = \mu f_j$ when they receive a mass $\mu$ of data and the state is $\theta_j$.

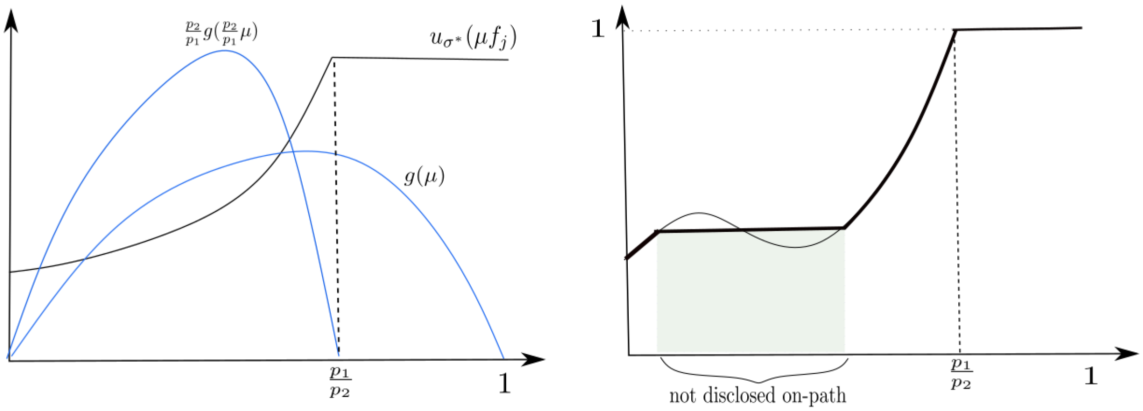
The set of possible types is $\mathcal{T}_\infty = [0, 1] \times \Theta \subset \mathcal{F}$, and we let the set of potential messages be $\mathcal{M}_\infty = \mathcal{F}$ – that is, we place no restrictions on what distributions of data the sender may show to the receiver, except that a type $t$ can only send a message $m$ if $m \subseteq t$. Call this infinite-data game $\mathcal{G}_\infty$. Observe that the lexicographic dominance ordering over equilibrium outcomes applies as well to outcomes of equilibria of $\mathcal{G}_\infty$. So, as a first guess to approximating lexicographically optimal outcomes in big-data settings, we may look to equilibria with lexicographically optimal outcomes in $\mathcal{G}_\infty$.

## 5.2 A binary-state example

To examine outcomes in the infinite-data approximation, let us return to the example with $|\Theta| = |\mathcal{D}| = 2$. Because $f_1 = (1-p_1, p_1)$ and $f_2 = (1-p_2, p_2)$, types take the form $\mu(1-p_1, p_1)$ or $\mu(1-p_2, p_2)$. Since outcome 2 is better proof that the state is 2 than outcome 1 is, let us focus on equilibria in which, like in the finite-$N$ case, senders disclose only observations of outcome 2.

Figure 1a shows that when $g(\frac{p_2}{p_1}\mu)/g(\mu)$ is monotone in $\mu$, disclosing more observations of 2 is always better: conditional on observing a mass $\mu p_2$ of observations of 2, the receiver believes the sender either has a mass $\mu$ of data and the state is 2, or the sender has a mass $\frac{p_2}{p_1}\mu$ of data and the state is 1, so the sender's payoff is $1 + \dfrac{g(\mu)}{\frac{p_2}{p_1}g(\frac{p_2}{p_1}\mu) + g(\mu)}$.

Equilibrium outcomes must always be monotone: when $t \subseteq t'$, then $u_{\sigma^*}(t) \leq u_{\sigma^*}t'$, since all messages available to $t$ are also available to $t'$. However, it is possible for $g(\frac{p_2}{p_1}\mu)/g(\mu)$ to be nonmonotone in $\mu$. In this case the strategy "disclose as many observations of 2 as possible" does not respect payoff monotonicity. Instead, the lexicographically optimal outcome involves *ironing* the putative payoff function $1 + \dfrac{g(\mu)}{\frac{p_2}{p_1}g(\frac{p_2}{p_1}\mu) + g(\mu)}$. Figure 1b gives



(a) Payoffs as a function of $\mu$ and $f_j$ when $g(\frac{p_2}{p_1}\mu)/g(\mu)$ is monotone.

(b) Payoffs when $g(\frac{p_2}{p_1}\mu)/g(\mu)$ is non-monotone.

an illustration.

Note how the construction coincides with the equilibrium of Claim 3.1: when $g(\frac{p_2}{p_1}\mu)/g(\mu)$ is monotone, payoffs to disclosing increasing amounts of outcome 2 are increasing as well. On the other hand, when $\frac{g(\frac{p_2}{p_1}\mu)}{g(\mu)}$ is not, they are not, and in the finite-data equilibrium, the set of on-path messages $\{(0, n_2[k])\}_{k=1}^{K}$ is a strict subset of $\{(0, n)\}_{n=0}^{N}$, with some types pooling with types that have fewer observations of 2.

Finally, note that there is indeterminacy in the equilibrium strategies that would implement $u_{\sigma^*}$. Unlike in the finite-data setting, senders could just as well have imitated the entire distribution $f_2$ by sending $\mu f_2$ instead of sending only $(0, \mu p_2)$, since they prove the same thing: the same set of types under both state 1 and state 2 are capable of sending either. We call equilibria in which all on-path messages take the form $\mu f_j$ *imitation equilibria*.

It is straightforward to extend the construction of the imitation equilibrium outcome with $|\Theta| = |\mathcal{D}| = 2$ to construct the imitation equilibrium outcome when the state is binary, but the space of experimental outcomes is an arbitrary finite set. Where $\frac{p_2}{p_1}$ gives ratio of the maximum measure of data distributed $f_2$ that a sender has under state 2 to the measure that a sender endowed with the same total amount of data under state 1 has, the same ratio can be constructed for arbitrary outcome spaces: we define for any particular observation the relative likelihood under distributions $f$ and $f'$ to be

$$LR(f, f'|d) = \frac{f(d)}{f'(d)}$$

and the maximum of $LR(f_j, f_{j'}|d)$ over all $d$ to be

$$r_{j'}(j) = \max_{d} \left( \frac{f_j(d)}{f_{j'}(d)} \right).$$

Then the equilibrium outcome constructed, replacing $\frac{p_2}{p_1}$ by $r_1(2)$, for the general case is the analogous imitation equilibrium, and is also lexicographically optimal.

## 5.3 Imitation with $J > 2$ states

We now extend a characterization of an imitation equilibrium outcome to the case with $J > 2$ states. In particular, we look for an equilibrium in which payoffs under every state are a continuous function of $\mu$. While imitation equilibria are not unique, the construction of the continuous-payoff equilibrium follows the intuition that we prioritize awarding high payoffs to senders that have high potential payoffs in equilibrium. Here, we focus on characterizing the equilibrium and its outcome, but we will show in Section 6 that when a sequence of finite models converges to an infinite-data game, the limit of the corresponding lexicographically optimal equilibrium outcomes must converge to exactly the outcome of this imitation equilibrium.

The central object defining the imitation equilibrium is a "burden of proof" associated with each payoff and state, which gives the volume of data imitating the given state distribution that is necessary to obtain the desired payoff. Senders endowed with different datasets will best meet the burden of proof in different ways. Indeed, the equilibrium can be summarized by a vector-valued *burden-of-proof function*, $\hat{\mu}(u) = (\hat{\mu}_1(u), \ldots, \hat{\mu}_J(u))$, such that each sender need only consider the maximal level of utility $u$ such that they can meet the burden of proof for some component $j$ of the associated vector. Their optimal strategy is then to imitate state $j$ using a measure $\hat{\mu}_j(u)$ of data. Correspondingly, the payoff obtained by disclosing $(\mu f_j)$ is $u_j(\mu)$, which is the (continuous) inverse of $\hat{\mu}$ in that

$$\hat{\mu}_j(u_j(\mu)) = \min\{\mu' : u_j(\mu') = u_j(\mu)\} \qquad \text{and} \qquad u_j(\hat{\mu}_j(u)) = u.$$

where $\hat{\mu}_j(u)$ may also be empty if there is no $\mu \in [\underline{\mu}, \bar{\mu}]$ such that $u_k(\mu) = u$. Indeed, the domain of $\hat{\mu}_j$ will turn out to be $[u_s(\mathbb{1}_1), u_s(\mathbb{1}_j)]$ – imitating the state-$j$ distribution never yields a greater payoff than having the state thought to be $j$ for sure.

As in the case of a binary state, the set set $\{r_{j'}(j)\}_{j',j \in 1,\ldots,J}$ fully characterizes the *pairwise* comparisons between $f_1, \ldots, f_J$, which are the only relevant features for masquerading across states, as they encode how advantaged the data distribution under each state is in imitating another based on their relative similarity.

**Theorem 5.1** *There exists a unique*[6] *vector-valued function* $\hat{\mu}(u) : [0, \theta_J] \to \mathbb{R}^J$ *such that*

1. $u_j(\mu)$ *is continuous and (weakly) increasing in* $\mu$ *for all* $j$.
2. *There is a strategy* $\sigma^*$ *with* $\sigma^*(\mu f_j)$ *supported on*

$$\tilde{S}_j(\mu) = \{(\hat{\mu}_k(u) f_k) : k \in \arg\max_k u_k(\frac{\mu}{r_j(k)})\}$$

[6] Unique up to (outcome-irrelevant) indeterminacy when no amount of data distributed $f_j$ would convince the sender to award a payoff of $u$.

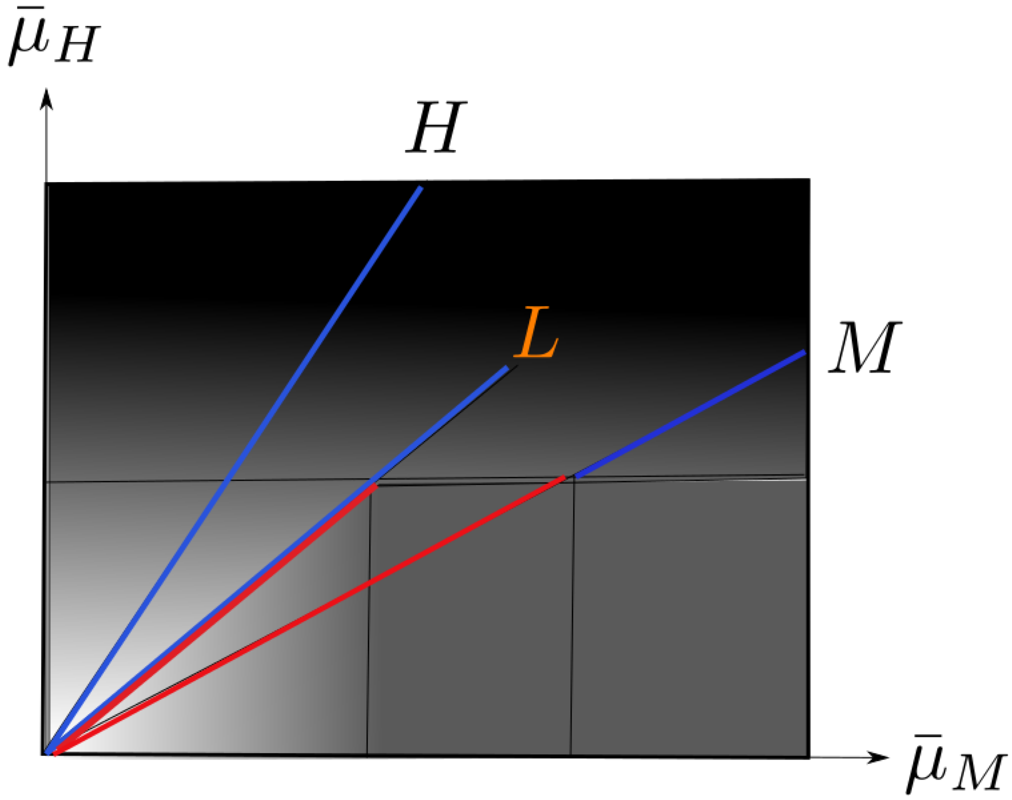


Figure 2: An illustration of sender's disclosure policy in the equilibrium $\sigma^*$ with 3 states: high (H), medium, (M), and low (L). A blue line represents types who masquerade under the high-state distribution; a red line represents types that masquerade under the medium-state distribution; and the coexistence of both denotes mixing.

*with* $\sigma(\hat{\mu}_k(u)f_k|\hat{\mu}_k(u)f_k) = 1$ *for all* $k$ *such that* $\theta_k \geq u$ *and such that for each* $u$ *and* $k$,

$$u_s\left(\beta_{\sigma^*}(\cdot|\hat{\mu}_k(u)f_k)\right) = u.$$

*Then* $\sigma^*$ *is an equilibrium sender strategy profile, and* $\hat{\mu}(\cdot)$ *is the corresponding burden-of-proof function.*

The equilibrium can be constructed step-by-step – see the Appendix. Intuitively, the construction notes that for a target utility level $u \in (\theta_j, \theta_{j+1})$, strategies must consist of senders imitating a state in $j+1, \ldots, J$, so the burden of proof can be projected down to $J - j$ dimensions. Then, using a system of differential equations, we may set the rate of change of each component of $\hat{\mu}(u)$ and $\{\sigma^*(\hat{\mu}_k(u)f_k|\cdot)\}_{k=j+1}^J$ such that $\frac{du(\beta^*(\cdot|\hat{\mu}_k(u)f_k))}{du} = 1$, except in the case of nonmonotonicities, which, as in the binary case, we handle with ironing.

Figure 3 shows the result of this process in a setting with 3 states. For any number of states, the resulting sender strategy profile is always part of an equilibrium of the disclosure game. To see why, observe that when the receiver believes off-path messages are negative signals (i.e. are sent by the worst type they could be sent by), then it is necessary to match some dimension of burden-of-proof $\hat{\mu}(u)$ in order to obtain a payoff $u$. Therefore, the best that any sender in $\tilde{S}_j(u)$ can do is indeed to send a measure $\hat{\mu}_j(u)$ of $f_j$.

As in the case of binary states, there are many equilibria and not all are likely. The one proposed here, however, has the appealing feature that all senders are either truthful, or achieve a higher payoff than they would if their identity was known; this contrasts with equilibria in which senders refrain from sending even positive off-path information, for fear of it being interpreted unfavorably.

I conclude the discussion of the equilibrium by summarizing some descriptive features.

**Theorem 5.2** *Under the equilibrium $\sigma^*$, there are thresholds $z_j^* > z_j^{**} \geq 0$ for each state such that:*

- *Whenever the sender's type is $\mu f_j$ with $\mu > z_j^*$, the sender masquerades as a higher type, and receives a payoff $u_{\sigma^*}(\mu f_j) > \theta_j$.*
- *Whenever $\mu \in (z_j^{**}, z_j^*]$, the sender is honest and the receiver knows it upon receiving the data: $u_{\sigma^*}(\mu f_j) = \theta_j$.*
- *Whenever $\mu \leq z_j^{**}$, the sender is honest, but the receiver believes they are a worse type with positive probability, and $u_{\sigma^*}(\mu f_j) < \theta_j$.*

# 6 Convergence of lexicographically optimal equilibria

A sequence of games of finite data, $(\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g_N, u_s))_{N=1}^\infty$, converges to $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty, u_s)$ if $N g_N(\lfloor N\mu \rfloor)$ converges uniformly to $g(\mu)$.

**Definition** A sequence of equilibria $(\sigma_1, \sigma_2, \ldots)$ of games $\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g_N, u_s))_{N=1}^\infty$ has *outcomes that converge* to the outcome of an equilibrium $\sigma_\infty$ of the limit infinite-data game $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty, u_s)$ if the payoffs $u_{\sigma_N}(t)$ converge uniformly to $u_\sigma(t)$ over $\mathcal{T}_\infty$.

Note payoffs are only required to converge for types that are possible in the limit, which is consistent with the fact that lexicographic optimality does not constrain payoffs for types that occur with probability (density) 0.

Nevertheless, here they do, and they converge to the imitation equilibrium $\sigma^*$ of the limit infinite-data game, as our 2nd main theorem shows.

**Theorem 6.1** *If $\sigma_\infty^*$ is the imitation equilibrium in $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty, u_s)$ and $u_{\sigma_\infty^*}$ is strictly increasing in $\mu$ for each $\theta$, then along any sequence $\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g_N, u_s))_{N=1}^\infty$ that converges to $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty, u_s)$, the LD equilibrium outcomes converge to $u_{\sigma_\infty^*}$.*

The qualifier that $u_{\sigma^*_\infty}$ be strictly increasing in $\mu$ entails a restriction on $g_\infty$ and $\{f_j\}_{j=1}^J$ that, while made for the sake of tractability, is nevertheless satisfied by some broad classes of functions: for example, it is always satisfied when $g_\infty$ is concave on $[0,1]$. We conjecture that it is not, in fact, necessary, and though we lack a proof in general, we confirm the conjecture in the case of a binary state:

**Theorem 6.2** *If $\sigma^*_\infty$ is the imitation equilibrium in $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty, u_s)$ where $|\Theta| = 2$, then the LD equilibrium outcomes along any sequence $\mathcal{G}_N(\Theta, \{f_j\}_{j=1}^J, g_N, u_s))_{N=1}^\infty$ converging to $\mathcal{G}_\infty(\Theta, \{f_j\}_{j=1}^J, g_\infty, u_s)$ converge to $u_{\sigma^*_\infty}$.*

The full proof is in the Appendix, and a sketch is as follows. When $N$ is very large, there is a type in $\mathcal{T}_N$ close to any type $t \in \mathcal{T}_\infty$. Under the algorithm that generates $\sigma^*_N$, that type must obtain the payoff of the maximal upper pool at the step $m$ in which its payoff is assigned. For any $\epsilon$, define

$$T(m) = \{t' \in \mathcal{T}_N : t' \text{ remains at step } m \text{ and } u_{\sigma^*}(t') \geq u_{\sigma^*}(t) - \epsilon\}.$$

The payoff to the maximal upper pool is lower-bounded by the receiver's belief about the state conditional on the sender being in $T(m)$, since there remains a set of messages $M$ such that all types in $T(m)$ can send at least one message in $M$, but no types outside $T(m)$ can do so. When the receiver forms their belief about the state conditional on the sender being in $T(m)$, the sender's payoff is bounded below, with the bound approaching $u_{\sigma^*}(t) - \epsilon$ as $N \to \infty$; intuitively, this comes from the fact that whenever a set of types in the neighborhood of $\mu f_j \in \mathcal{T}_\infty$ is in the possible set, a corresponding measure of types in the neighborhood of the type that $\mu f_j$ imitates under the imitation equilibrium must also be in the set. Since the imitated types correspond to better states, the belief given the set of types must be at least as favorable. In the limit as $N \to \infty$, no equilibrium outcome of $\mathcal{G}_N$ can be unilaterally better for all senders in $\mathcal{T}_\infty$, due to Bayes plausibility, and they cannot be worse for any sender, so the two outcomes must coincide.

# 7 Extension: Endogenous data acquisition

We have assumed so far that the distribution of $\mu$ is exogenous and identical for all senders. This captures some sources of variation in the data volume, such as invalid trials due to human error or dropouts. But the volume of data generated may also vary because of sender-specific differences in data-gathering ability – either different capacities (e.g. time constraints) or costs of obtaining more evidence.

The case of exogenous capacities is simple, and there is a one-to-one mapping between capacity constraints and distributions of attained data. The outcome of the game is unchanged if, instead of assuming that senders are randomly endowed with a measure $\mu$ of data following the outcome of trials, we suppose that each sender knows their capacity $K$ for data collection prior to experiments, which is uncorrelated with the state. Then the sender's optimal strategy in the data-collection stage is to meet their capacity exactly (set $\mu = K$), the distribution of $\mu$ over the population is the same as the distribution of $K$, and the receiver draws identical conclusions.

It is more challenging to map costs of data acquisition to disclosure game outcomes. Nevertheless, it is trivially true that every distribution of data endowments can be founded on *some* cost structure, as cost functions

$$c(\mu) = \begin{cases} 0, & \mu \leq K \\ u_s(\mathbb{1}_{\theta_1}) + 1, & \mu > K \end{cases}$$

mimic capacities in that the (possibly weakly) optimal choice is $\mu = K$, and so any $g$ can be imitated by a corresponding distribution over $K$ among such cost functions. Conversely, for most reasonable distributions of cost functions over the population, there must exist $g$ such that the equilibrium outcome of the augmented game with data acquisition is the lexicographically optimal outcome of the disclosure game in which $g$ is the distribution of endowments.

# References

Isaiah Andrews and Maximillian Kasy. Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–2794, 2019.

Ronald A. Dye. Disclosure of nonproprietary information. *Journal of Accounting Research*, 23(1):123–145, 1985.

Wioletta Dzuida. Strategic argumentation. *Journal of Economic Theory*, 146(4), 1985.

Mike Felgenhauer and Elisabeth Schulte. Strategic private experimentation. *American Economic Journal: Microeconomics*, 6(4):74–105, 2014.

Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3):461, 1981.

Steven Matthews, Masahiro Okuno-Fujiwara, and Andrew Postlewaite. Refining cheap-talk equilibria. *Journal of Economic Theory*, 55(2):247–273, 1991.

Paul R. Milgrom. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12(2):380–391, 1981.

Hyun Song Shin. Disclosures and asset returns. *Econometrica*, 71(1):105–133, 2003.

Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. p-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 2014.

# A Appendix A: Construction of imitation equilibrium

## A.1 Proof of Theorem 5.1

Analogously to the finite case, given any (finite) collection of messages $M$, letting $\mathcal{T} \subset T$ denote any arbitrary type set, let $\mathcal{T}^+(M)$ be the subset of types in $\mathcal{T}$ capable of sending a message in $M$.

For a given equilibrium $\sigma$, denote $\tau_\sigma(m)$ to be the set of types who find it (weakly) optimal to send message $m$ under $\sigma$, and let $\tau_\sigma^{++}(m)$ to be the set of types that send $m$ with positive probability under $\sigma$.

**Proof of Theorem 5.1** To create the burden-of-proof vector, we hypothesize, to begin with, that it is associated with a strategy $\sigma$ in which the support of play by $\mu f_j$ is

$$\text{supp } \sigma(\cdot|\mu f_j) \subseteq B(\mu f_j) = \{\frac{\mu}{r_j(k)} f_k : k \in \arg\max_{k'} u_{k'}(\frac{\mu}{r_j(k')})\}, \tag{1}$$

that is, each type plays the strategy "send a message corresponding to as great a mass as possible of some $f_k$, where $k$ is chosen to maximize the payoff from doing so".[7]

The broad approach to constructing $\sigma$ over all types is iterative and comes at the end of the proof. It requires details of constructing $\sigma$ near a fixed payoff $v$, which I will discuss here. Consider fixing a payoff $v$, and suppose that $v \in [\theta_l, \theta_{l+1})$. Suppose we are given $\hat{\mu}^+(u)$ and $\sigma^+$ that are are a burden-of-proof vector and associated strategy for the game with type space $T^+(\{\hat{\mu}_k^+(v) f_k\}_{k=l+1}^J)$ and the type distribution given by the same relative probabilities between types in $T^+(\{\hat{\mu}_k^+(v) f_k\}_{k=l+1}^J)$ as in the original game.

Define the set $M(v) = \{(\hat{\mu}_k(v) f_k)\}_{k=l+1}^J$ of all messages that yield payoff $v$ under $\sigma^+$. Observe that the payoff to sending a message in $M(v)$ must be the payoff to the receiver knowing that the sender is one of the types that sends a message in $M(v)$ under strategy $\sigma^+$. This payoff is

$$U(M(v)) = \frac{\sum_{k=l+1}^J \sum_j \theta_j g(\frac{\hat{\mu}_k(v)}{r_j(k)}) r_j(k) \sigma^+(\hat{\mu}_k(v) f_k | \frac{\hat{\mu}_k(v)}{r_j(k)} \theta_j)}{\sum_{k=l+1}^J \sum_j g(\frac{\hat{\mu}_k(v)}{r_j(k)}) r_j(k) \sigma^+(\hat{\mu}_k(v) f_k | \frac{\hat{\mu}_k(v)}{r_j(k)} \theta_j)} = v.$$

Analogously, the payoff under $\sigma^+$ to a subset $C$ of $M(v)$ is $U(C) = v$. Alternatively, define the payoff to knowing only that the sender is one of the types in $M(v)$ who can send a message in $C$ (even if they don't do so with probability 1) by

$$W(C) = \frac{\sum_{k:\hat{\mu}_k(v) f_k \in C} \sum_{j:\mu_k(v) f_k \in B(r_j(k)\mu_k(v) f_j)} \theta_j g(\frac{\hat{\mu}_k(v)}{r_j(k)}) r_j(k)}{\sum_{k=l+1}^J \sum_j g(\frac{\hat{\mu}_k(v)}{r_j(k)}) r_j(k)}.$$

[7] This is analogous to the first step in the binary-state construction of revealing as much as possible of $f_2$, prior to ironing.

We may define a directed graph with nodes in $M(v)$ and a link $m \to m'$ if and only if $\tau_\sigma(m') \bigcap \tau_\sigma^{++}(m)$ is nonempty. If there is a type that mixes between $m$ and $m'$ with strictly interior probability, then $m \to m'$ and $m' \to m$. If, on the other hand, there is a type that is able to send both $m$ and $m'$, and obtains their best possible payoff from either, but no such types send $m'$ with positive probability (though some send $m$ with positive probability), then $m \to m'$, but not vice-versa.

A connected component of this graph is $C \subseteq M(v)$ such that for all $m, m' \in C$, either there is a path $m \to \ldots \to m'$, or a path $m' \to \ldots \to m$, and there is no $m'' \in M(v) \setminus C$ with either an ingoing or outgoing link to any $m \in C$. A strongly connected component is a component $C'$ such that there are (directed) paths $m \to \ldots \to m'$ and $m' \to \ldots \to m$ for all $m, m' \in C'$, and not for any $m \in C'$ and $m' \notin C'$. If $C$ is any subset of nodes of $M$ not connected to any other nodes, and there is no subset $C' \subset C$ with $U(C) \geq v$, then $C$ must be a strongly connected component of $M$, since if it is disconnected, then its sub-components contain at least one group of messages for which all types achieving a payoff of $v$ that can send one of the messages do, but this would yield a lower payoff than $v$.

If there is a connected component $C \subset M(v)$ that is also strongly connected, then there exists small enough $\epsilon$ such that for all $v' \in [v - \epsilon, v]$, the burden-of-proof vector satisfies

$$\frac{\hat{\mu}_k(v')}{\hat{\mu}_k(v)} = \frac{\hat{\mu}_{k'}(v')}{\hat{\mu}_{k'}(v)} \qquad \forall\ k : \hat{\mu}_k(v) f_k \in C,\ k' : \hat{\mu}_{k'}(v) f_{k'} \in C \tag{2}$$

because if not, then indifference for mixing types is not preserved, which in turn means that the payoffs to messages in putative frontier $(\hat{\mu}_k(v')f_k)_{k=l+1}^J$ fails to hold.

If $C$ is a connected component consisting of multiple, disjoint strongly connected components, then it must be that there is a directed acyclic graph of such components in $C$, in which, for two strongly connected components $C'$ and $C''$, $C \to C''$ if there are types that send a message in $C'$ that could send a message in $C''$, but not vice-versa. Then, we compare the rate of change of payoffs from playing a message $C'$ relative to a message in $C''$ for a type that can send either, and consider $C'$ and $C''$ as if they were a single component if the rate is weakly greater for $C'$ than $C''$, and not if vice versa.

Using these insights, the construction of $u_k(\mu)$ proceeds iteratively:

1. Start with $l = J$ and $\hat{\mu}_J(\theta_J) = 1$.
2. Construct $u_s(\cdot)$ as follows:
   (a) Initialize $v_1^l = \theta_l$.
   (b) For each $v_i^l$, partition $M(v_i^l) = \{(\hat{\mu}_k(v_i^l)f_k)\}_{k=l+1}^J$ into components as described above, calculate payoffs for each component, and calculate $M(v - \epsilon)$ given constraint 2. Given these messages, set

   $$v_{i+1}^l = \max\big\{u \leq v_i^l : M(u) \text{ not partitionable into strongly connected connected components}\big\},$$

and repeat starting from $v^l_{i+1}$ until $v^l_{i+1}$ does not exist. Observe that $\{u_k\}_{k=1}^J$ thus constructed must be consistent with an equilibrium strategy as long as they are monotone in $\mu$.

Let $\hat{\mu}_k(u) = \min\{\tilde{\mu}' : u_k(\tilde{\mu}') = u\}$ for $k \in \{l, \ldots, J\}$, and let $u_j(\mu) = \max_{k \geq l} u_k(\mu / r_j(k))$ for $j' < l$.

3. If at any point $u_k(\mu)$ is nonmonotone in $\mu$, then iron, and repeat step (2) from the lower bound of the ironing interval.
4. If $l > 1$, decrease $l$ by 1, and let $\hat{\mu}_j(\theta_j) = \max_{j' < j}\{\frac{1}{r_{j'}(j)}\hat{\mu}_{j'}(\theta_j)\}$. Then return to (2).

## A.2 Proofs of convergence (Theorems 6.1 and 6.2)

**Proof** The proof of theorems 6.1 and 6.2 has 3 steps. First, we give a lemma establishing that under the conditions of the theorems, for any set of messages $M$, when the set of all types in $\mathcal{T}_\infty \setminus \mathcal{T}^+_\infty(M)$ that attain a payoff of at least $v$ in $\sigma^*_\infty$ is nonempty, heir payoff when they form a pool is at least $v$. Using this, we show that $u_{\sigma^*}$ is a lower bound on payoffs for types in $\mathcal{T}_\infty$ in the limit. Finally, Bayes plausibility implies that

$$\lim_{N \to \infty} \sum_{j=1}^{j} \beta_0(\theta_j) \int_{\mu=0}^{1} u_{\sigma_N}(\mu f_j) g(\mu) d\mu = \mathbb{E}_{q_\infty}[\mathbb{E}_{\beta(\cdot|\sigma(t))}[\theta]|t] = \mathbb{E}_{\beta_0}[\theta],$$

which in conjunction with the lower bound implies that in the limit outcomes must coincide exactly with $u_{\sigma^*}$ for types in $\mathcal{T}_\infty$.

**Lemma A.1** *Suppose that either $|\Theta| = 2$ or payoffs under $u_{\sigma^*_\infty}$ are strictly increasing in $\mu$ for each $\theta$. If $M$ is a collection of messages and $(\underline{\mu}_1 f_1, \ldots, \underline{\mu}_i f_i; \underline{\mu}_{i+1} f_{i+1}, \ldots, \underline{\mu}_J f_J)$ is the frontier of types achieving a payoff of at least $v$ under $\sigma^*_\infty$, where $\theta_i < v \leq \theta_{i+1}$, then*

$$\mathbb{E}[\theta | t \in \mathcal{T}^+_\infty(\{\underline{\mu}_j f_j\}_{j=1}^J) \setminus \mathcal{T}^+_\infty(M)] \geq v$$

*whenever $\mathcal{T}^+_\infty(\{\underline{\mu}_j f_j\}_{j=1}^J) \setminus \mathcal{T}^+_\infty(M)$ is nonempty.*

**Proof of Lemma** Denote $T(v, M) = \mathcal{T}^+_\infty(\{\underline{\mu}_j f_j\}_{j=1}^J) \setminus \mathcal{T}^+_\infty(M)$. Let $(\bar{\mu}_1, \ldots, \bar{\mu}_i; \bar{\mu}_{i+1}, \ldots, \bar{\mu}_J)$ be the minimum masses of data distributed like $f_1, \ldots, f_i; f_{i+1}, \ldots, f_J$, respectively, necessary to send some message in $M$. Then

$$\mathbb{E}[\theta | t \in T(v, M)] = \frac{\sum_{j=1}^J \beta_0(\theta_j)\theta_j(G(\bar{\mu}_j) - G(\underline{\mu}_j))}{\sum_{j=1}^J \beta_0(\theta_j)(G(\bar{\mu}_j) - G(\underline{\mu}_j))}.$$

If $(\bar{\mu}_{i+1}, \ldots, \bar{\mu}_J) \leq (\underline{\mu}_{i+1}, \ldots, \underline{\mu}_J)$ pointwise, then $T(v, M)$ is empty. Otherwise, let the states $j_1, \ldots, j_A$ be the maximal set such that $(\bar{\mu}_{j_1}, \ldots, \bar{\mu}_{j_A}) > (\underline{\mu}_{j_1}, \ldots, \underline{\mu}_{j_A})$ pointwise. Call the set of types in $T(v,M)$ that send $\mu' f_{j_a}$ with positive probability under $\sigma^*$ by $t_{\sigma^*}(\mu' f_{j_a})$, and let $\theta(t)$ refer to the state corresponding to the distribution of dataset $t$.

$$\begin{aligned}\mathbb{E}_{\hat{\sigma}_v}[v(\theta)|\mu' f_{j_a}] &= \frac{\sum_{t\in t_{\sigma^*}(\mu' f_{j_a})\bigcap T(v,M)} \theta(t) g(\frac{\mu}{r_{\theta(t)}(j_a)})\sigma^*(\mu' f_{j_a}|t)\frac{\beta_0(\theta)}{r_{\theta(t)}(j_a)}}{\sum_{t\in t_{\sigma^*}(\mu' f_{j_a})\bigcap T(v,M)} g(\frac{\mu}{r_{\theta(t)}(j_a)})\sigma^*(\mu' f_{j_a}|t)\frac{\beta_0(\theta)}{r_{\theta(t)}(j_a)}} \\ &\geq \frac{\sum_{t\in t_{\sigma^*}(\mu' f_{j_a})} \theta(t) g(\frac{\mu}{r_{\theta(t)}(j_a)})\sigma^*(\mu' f_{j_a}|t)\frac{\beta_0(\theta)}{r_{\theta(t)}(j_a)}}{\sum_{t\in t_{\sigma^*}(\mu' f_{j_a}))} g(\frac{\mu}{r_{\theta(t)}(j_a)})\sigma^*(\mu' f_{j_a}|t)\frac{\beta_0(\theta)}{r_{\theta(t)}(j_a)}} \\ &\geq v.\end{aligned} \tag{3}$$

In the case when payoffs under $u_{\sigma^*_\infty}$ are strictly increasing, the first inequality comes from the fact that $\theta_{j_a} \geq v > \theta(t)$ whenever $\theta(t) \neq \theta_{j_a}$, and $\mu' f_{j_a} \in T(v,M)$ only if all types that play it under $\sigma^*$ are also in $T(v,M)$.

When $|\Theta| = 2$, the inequality also holds when ironing occurs in the equilibrium construction. If $\mu' f_H$ is a message that supports the pooling of types in an ironing interval, then we may further break the set of types in $t_{\sigma^*(\mu' f_H)}$ by the message they would send under $\bar{\sigma}$, the unironed, “send as much of $f_H$ as possible” strategy. If the ironing interval is from $\mu' f_H$ to $\mu'' f_H$, then

$$\begin{aligned}\mathbb{E}_{\hat{\sigma}_v}[v(\theta)|\mu' f_H] &= \frac{\beta_0(\theta_H)\int_{\mu'}^{\min(\mu'',\bar{\mu}_H)} g(\mu)\theta_H d\mu + \beta_0(L)\int_{\frac{\mu'}{r_L(H)}}^{\frac{\min(\mu'',\bar{\mu}_L)}{r_L(H))}} g(\mu)v(L)d\mu}{\beta_0(\theta_H)[G(\min(\mu'',\bar{\mu}_H)) - G(\mu')] + \beta_0(\theta_L)[G(\frac{\min(\mu'',\bar{\mu}_L)}{r_L(H))}) - G(\frac{\mu'}{r_L(H)})]} \\ &\geq \frac{\beta_0(\theta_H)\int_{\mu'}^{\min(\mu'',\bar{\mu}_H)} g(\mu)\theta_H d\mu + \beta_0(L)\int_{\frac{\mu'}{r_L(H)}}^{\frac{\min(\mu'',\bar{\mu}_H)}{r_L(H))}} g(\mu)v(L)d\mu}{\beta_0(\theta_H)[G(\min(\mu'',\bar{\mu}_H)) - G(\mu')] + \beta_0(\theta_L)[G(\frac{\min(\mu'',\bar{\mu}_H)}{r_L(H))}) - G(\frac{\mu'}{r_L(H)})]}.\end{aligned} \tag{4}$$

The latter half of the inequality is the expectation of the state's value when the set of possible types includes $[\mu' f_H, \min(\mu'',\bar{\mu}_H)f_H]$ and $[\frac{\mu'}{r_L(H)}f_L, \frac{\min(\mu'',\bar{\mu}_H)f_H}{r_L(H)}f_L]$. Since $\mu' < \min(\mu'',\bar{\mu}_H) \leq \mu'$, by the construction of the ironing interval, this value exceeds the expectation of the state's value over a set of types including $[\mu' f_H, \mu'' f_H]$ and $[\frac{\mu'}{r_L(H)}f_L, \frac{\mu'' f_H}{r_L(H)}f_L]$, which is the original expectation of the state's value under message $\mu' f_H$, and is no less than $v$.

In either case, the expectation of $\theta$ given that the sender's type is in $T(v,M)$ is a weighted average of $\mathbb{E}_{\hat{\sigma}_v}[\theta|\mu' f_{j_a}]$ over on-path messages $\mu' f_{j_a}$ in $T(v,M)$. We have shown that each component is no less than $v$, and so the weighted average is also at least $v$. ▮

Before proceeding to construct bounds on payoffs in the finite games, it is helpful to define a neighborhood of $\mathcal{T}_\infty$ as the set of types in each finite game with datasets distributed similarly to the underlying distribution in some state. For $\eta \in (0,1]$ and $k \in [0,1]$, define

$$S_N(\eta,k) = \{t \in \mathcal{T}_N : |t| \geq k \text{ and } \exists\theta \text{ s.t. } \sup_d |t(d) - |t|f_\theta(d)| \leq \eta\}.$$

Fix an integer $n$. Conditional on $|t| = n$ and the true state being $\theta_j$, the Glivenko-Cantelli theorem states that there is a bound on the probability that $\sup_d |\sum_{x=1}^d t(d) - \frac{n}{N} F_j(d)| > \eta$ that decreases to 0 for large $n$, irrespective of $N$. Because data have a discrete distribution, this implies a similar bound on the empirical probability mass function: if $|t| = n$ and $\theta_j$ is the true state, the probability that $\sup_d |t(d) - \frac{n}{N} f_j(d)| > \eta$ is at most $b_=(n, \eta)$, with $\lim_{n\to\infty} b_=(n,\eta) = 0$ for all $\eta > 0$. If the true state is $\theta_{j'} \neq \theta_j$ and $|t| = n$, then the probability that $\sup_d |t(d) - \frac{n}{N} f_j(d)| > \eta$ is at least $b_{\neq}(n,\eta)$, with $\lim_{\eta\to 0} \lim_{n\to\infty} b_{\neq}(n,\eta) = 1$.

When $N$ and $k$ are large, the proportion of types that lie in $S_N(\eta, k)$ is close to 1, for all $\eta$. In particular, $\lim_{k\to 0} \lim_{\eta\to 0} \lim_{N\to\infty} q_N(S_N(\eta,k)) = 1$, since:

- With probability decreasing to 0 as $k \to 0$, $|t| < k$.
- For fixed $k$ and $\eta$, the probability that there does not exist $\theta$ such that $\sup_d |t(d) - |t| f_\theta(d)| \leq \eta$ given that $|t| \geq k$ decreases to 0 as $Nk \to \infty$.

We may further subdivide $S_N(\eta, k)$ into a set of types associated with each state,

$$S_N^j(\eta, k) = \{t \in S_N(\eta, k) : \sup_d |t(d) - |t| f_j(d)| \leq \eta\}.$$

A further consequence of the convergence of empirical distributions is that, when $Nk \to \infty$ and $\eta \to 0$, the sets $(S_N^j(\eta,k))_{j=1}^J$ are disjoint. Additionally, for all $t \in S_N^j(\eta, k)$, there is a uniform lower bound on the probability that the state is $\theta_j$ given that the sender is of type $t$, which we call $w(k, \eta, N)$, with $\lim_{k\to 0} \lim_{\eta\to 0} \lim_{N\to\infty} w(k,\eta,N) = 1$.

In addition, we can lower-bound $q_N(\{t \in S_N^j(\eta,k) : \underline{\mu} f_j \subseteq t \subseteq \bar{\mu} f_j\})$ for all $k < \underline{\mu} < \bar{\mu}$. Let $\Delta(N)$ be a bound on $\sup_d |(\sum_{x=1}^d g_N(x)) - G(d)|$ that goes to 0 as $N \to \infty$. Observe that if $\underline{\mu} + \eta < |t| < \bar{\mu} - \eta$ and $t \in S_N^j(\eta, k)$, then $\underline{\mu} f_j \subseteq t \subseteq \bar{\mu} f_j$, so a lower bound is

$$q_N(\{t \in S_N^j(\eta,k) : \underline{\mu} f_j \subseteq t \subseteq \bar{\mu} f_j\}) \geq \beta_0(\theta_j)(1 - b_=(Nk,\eta))(G(\bar{\mu} - D\eta) - G(\underline{\mu} + D\eta) - \Delta(N)). \tag{5}$$

Similarly, there is an upper bound on $q_N(\{t \in S_N^j(\eta,k) : t \not\subseteq \underline{\mu} f_j \text{ and } \bar{\mu} f_j \not\subseteq t\})$:

$$q_N(\{t \in S_N^j(\eta,k) : t \not\subseteq \underline{\mu} f_j \text{ and } \bar{\mu} f_j \not\subseteq t\}) \leq \beta_0(\theta_j)(G(\bar{\mu} + D\eta) - G(\underline{\mu} - D\eta) + \Delta(N)) + (1 - \beta_0(\theta_j)) b_{\neq}(kN, \eta). \tag{6}$$

Now we proceed to construct a lower bound for $u_{\sigma_N}(\hat{\mu} f_{\hat{j}})$. First, recall that $u_{\sigma_N}(\mu f_j) \geq \max_{\{f \in \mathcal{T}_N : t \subseteq \mu f_j\}} u_{\sigma_N}(t)$. Observe that there exists a dataset $\hat{t} = \frac{1}{N}(\lfloor N \hat{\mu} f_{\hat{\theta}}(1) \rfloor, \ldots, \lfloor N \hat{\mu} f_{\hat{\theta}}(k) \rfloor)$ in $\mathcal{T}_N$ and that $u_{\sigma_N}(\hat{\mu} f_{\hat{j}}) \geq u_{\sigma_N}(\hat{t})$.

For a given $N$, suppose $\hat{t}$ belongs to the $m$th upper pool under the algorithm that constructs $\sigma_N$. Denote by $\hat{M}_N(m-1)$ the set of messages that implement the upper pools in step $1, \ldots, m-1$, and fix $\mathcal{T}_{N,m} = \mathcal{T}_N^+(\hat{M}_N(m-1))$ to be the set of remaining types at the start of the $m$th step of the algorithm that constructs $\sigma_N$; therefore, $\hat{t}$ belongs to $\mathcal{T}_{N,m}$.

Let $\underline{M}_\infty(\epsilon, N)$ be the set of on-path messages that result in a payoff of $u_\sigma((\hat{\mu}-\epsilon)f_{\hat{j}})$ under infinite data. We see that the set of types in $\mathcal{T}^+_{N,m}(\underline{M}_\infty(\epsilon, N))$ includes $\hat{f}$ when $N$ is large enough. From Lemma A.1, there is an upper pool in $\mathcal{T}_{\hat{N},m}$ that achieves a payoff of at least $u(\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon, N)))$, so $u_{\sigma_N}(\hat{\mu}f_{\hat{j}})$ is lower-bounded by $u(\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon, N)))$.

Let $(\underline{\mu}_1(\epsilon, N), \dots, \underline{\mu}_J(\epsilon, N))$ be a vector that gives the minimum mass of data under distributions $f_1, \dots, f_J$, respectively, such that the dataset contains some message in $\underline{M}_\infty(\epsilon, N)$, and let $(\bar{\mu}_1(N), \dots, \bar{\mu}_J(N))$ be the maximum mass of data under each distribution such that there does not exist $t \in \mathcal{T}_{\hat{N},m}$ such that $t \subseteq \bar{\mu}_j f_j$. All $t \in \mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))$ satisfy $t \not\subseteq \underline{\mu}_j(\epsilon, N)f_j$ and $\bar{\mu}_j(N)f_j \not\subseteq t$, and all $t$ satisfying $\underline{\mu}_j(\epsilon, N)f_j \subseteq t \subseteq \bar{\mu}_j(N)f_j$ for some $j$ are in $\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))$.

We may rewrite

$$u(\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))) = \frac{\sum_{j=1}^J \sum_{t\in\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))} q_N(t)\theta_j\pi_N(\theta_j|t)}{\sum_{t\in\mathcal{T}^+_{\hat{N},m}(\underline{M}_\infty(\epsilon))} q_N(t)}. \tag{7}$$

Let the numerator be $Q(N, \hat{\mu}f_{\hat{j}}, \epsilon)$ and the denominator be $R(N, \hat{\mu}f_{\hat{j}}, \epsilon)$. Analogously to eq. 5, a lower bound for $Q(N, \hat{\mu}f_{\hat{j}}, \epsilon)$ is

$$\underline{Q}(N, \hat{\mu}f_{\hat{j}}, \epsilon) = \sum_j \beta_0(\theta_j)\theta_j[G(\bar{\mu}_j(N)-\eta D)-G(\max(\underline{\mu}_j(\epsilon, N)+\eta D, k))-\Delta(N)]w(k,\eta,N)(1-b_=(k,\eta)), \tag{8}$$

and it follows from eq. 6 that an upper bound for $R$ is

$$\begin{aligned}\bar{R}(N, \hat{\mu}f_{\hat{j}}, \epsilon) = &\left(\sum_j \beta_0(\theta_j)[G(\bar{\mu}_j(N)+\eta D) - G(\underline{\mu}_j(\epsilon, N) - \eta D) + \Delta(N)]\right) \\ &+ J(1-b_{\neq}(k,\eta)) + (1 - q_N(S_N(\eta, k))).\end{aligned} \tag{9}$$

We have

$$\lim_{k\to 0}\lim_{\eta\to 0}\lim \inf_{N\to\infty} \underline{Q} \geq \lim \inf_{N\to\infty} \sum_{j=1}^J \beta_0(\theta_j)\theta_j(G(\bar{\mu}_j(N)) - G(\underline{\mu}_j(\epsilon, N)))$$

and

$$\lim_{k\to 0}\lim_{\eta\to 0}\lim \inf_{N\to\infty} \bar{R} \leq \lim \inf_{N\to\infty} \sum_{j=1}^J \beta_0(\theta_j)(G(\bar{\mu}_j(N)) - G(\underline{\mu}_j(\epsilon, N))).$$

Both of the RHS are finite and strictly positive for all $N$ and $\epsilon > 0$; therefore,

$$\begin{aligned}\lim_{k\to 0}\lim_{\eta\to 0}\lim\inf_{N\to\infty}\frac{Q}{\bar{\bar{R}}} &\geq \lim\inf_{N}\frac{\sum_{j=1}^{J}\beta_0(\theta_j)\theta_j(G(\bar{\mu}_j(N))-G(\underline{\mu}_j(\epsilon,N)))}{\sum_{j=1}^{J}\beta_0(\theta_j)(G(\bar{\mu}_j(N))-G(\underline{\mu}_j(\epsilon,N)))}\\ &= \lim\inf_{N}\mathbb{E}[\theta|t\in T(u_{\sigma_\infty}((\hat{\mu}-\epsilon)f_{\hat{j}}),\hat{M}_N(m-1))]\\ &\geq u_{\sigma_\infty}((\hat{\mu}-\epsilon)f_{\hat{j}}),\end{aligned} \tag{10}$$

where the last inequality follows from Lemma A.1.

Because $k$ and $\eta$ are arbitrary variables used to obtain the bound, it follows from this that $\lim_{N\to\infty}u(\mathcal{T}^{+}_{\hat{N},m}(\underline{M}_\infty(\epsilon)))\geq u_{\sigma_\infty}((\hat{\mu}-\epsilon)f_{\hat{j}})$. Finally, because payoffs are continuous, taking a sequence of bounds as $\epsilon\to 0$ implies that $\lim\inf_{N\to\infty}u_{\sigma_N}(\hat{\mu}f_{\hat{j}})\geq\lim_{\epsilon\to 0}\lim\inf_{N\to\infty}u(\mathcal{T}^{+}_{\hat{N},m}(\underline{M}_\infty(\epsilon)))\geq u_{\sigma_\infty}(\hat{\mu}f_{\hat{j}})$.

The last step is to show that

$$\lim_{N\to\infty}\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{1}u_{\sigma_N}(\mu f_j)g(\mu)d\mu=\mathbb{E}_{\beta_0}[\theta].$$

Since we know already that

$$\lim_{N\to\infty}\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{1}u_{\sigma_\infty}(\mu f_j)g(\mu)d\mu=\mathbb{E}_{\beta_0}[\theta]$$

and $\lim\inf_{N\to\infty}u_{\sigma_N}(\mu f_j)\geq u_{\sigma_\infty}(\mu f_j)$ for all $\mu f_j\in\mathcal{T}_\infty$, this additional fact suffices to ensure that $u_{\sigma_N}(\cdot)=u_{\sigma_\infty}(\cdot)$ over $\mathcal{T}_\infty$.

The proof comes from dividing $\mu\in(k,1)$ into $X$ chunks, with the $x$th chunk given by $(\mu_{x-1},\mu_x]$ where $\mu_x=x\frac{1-k}{X}+k$.

Consider types $t\in S^j_N(\eta,k)$ such that $\mu_{x-1}f_j\subseteq t\subseteq\mu_x f_j$: their payoff under $\sigma_N$ has to be in $[u_{\sigma_N}(\mu_{x-1}f_j),u_{\sigma_N}(\mu_x f_j)]$. This implies that

$$\begin{aligned}\underline{V}_N(k,\eta,X)=&\sum_{j=1}^{J}\beta_0(\theta_j)\sum_{x=1}^{X}u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1}-\eta D)-G(\mu_x+\eta D)-\Delta(N)](1-b_{=}(k,\eta))\\ &\leq\mathbb{E}_{\beta_0}[\theta],\end{aligned} \tag{11}$$

since $\underline{V}_N(k,\eta,X)$ is a lower bound for the total probability-weighted sum of payoffs under $\sigma_N$ over $t\in\mathcal{T}_N\bigcup S_N(\eta,k)$, while $\mathbb{E}_{\beta_0}[\theta]$ is equal to the total probability-weighted sum of payoffs under $\sigma_N$ of all types in $\mathcal{T}_N$.

Finally, the difference between $\sum_{j=1}^{J} \beta_0(\theta_j) \int_{\mu=0}^{1} u_{\sigma_N}(\mu f_j)g(\mu)d\mu$ and $\underline{V}_N(k,\eta,X)$ vanishes as $X \to \infty$, $k \to 0$, $\eta \to 0$, and $N \to \infty$. To see this, observe that if $c$ is an upper bound on $g$ (which exists because $g$ is continuous on compact interval $[0,1]$),

$$\begin{aligned}
V_N(k,\eta,X) \geq \sum_{j=1}^{J} \beta_0(\theta_j)\Big(\sum_{x=1}^{X} & u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1}) - G(\mu_x)] \\
&- (\theta_J b_{=}(k,\eta)[G(\mu_{x+1}) - G(\mu_x)] + 2c\eta D + \Delta(N))\Big) \\
\geq \sum_{j=1}^{J} \beta_0(\theta_j) \sum_{x=1}^{X} & u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1}) - G(\mu_x)] \\
& -JX\theta_J(b_{=}(k,\eta) + 2c\eta D + \Delta(N)).
\end{aligned} \tag{12}$$

Then, for any $\epsilon$ and $j$, define $\xi_N^j(\epsilon, X)$ to be the set of values of $x$ such that $u_{\sigma_N}(\mu_{x+1}f_j) - u_{\sigma_N}(\mu_x f_j) > \epsilon$. The size of $\xi_N^j(\epsilon, X)$ is at most $\frac{\theta_J}{\epsilon}$. For all $x \notin \xi_N^j(\epsilon, X)$, we have the bound $\int_{\mu_x}^{\mu_{x+1}} u_{\sigma_N}(\mu f_j)g(\mu)d\mu - u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1}) - G(\mu_x)] < \epsilon[G(\mu_{x+1}) - G(\mu_x)]$. So,

$$\begin{aligned}
&\sum_{j=1}^{J} \beta_0(\theta_j) \int_{\mu=0}^{2} u_{\sigma_N}(\mu f_j)g(\mu)d\mu - \underline{V}_N(k,\eta,X) \\
&\quad \leq \left(\sum_{j=1}^{J} \beta_0(\theta_j) \sum_{x=1}^{X} \left(\int_{\mu_x}^{\mu_{x+1}} u_{\sigma_N}(\mu f_j)g(\mu)d\mu - u_{\sigma_N}(\mu_x f_j)[G(\mu_{x+1}) - G(\mu_x)]\right)\right) \\
&\quad\quad + JX\theta_J(b_{=}(k,\eta) + 2c\eta D + \Delta(N) + (1 - q_N(S_N(\eta,k)))) \\
&\quad \leq \sum_{j=1}^{J} \left(\beta_0(\theta_j) \left(\sum_{x \notin \xi_N^j(\epsilon,X)} \epsilon[G(\mu_{x+1}) - G(\mu_x)]\right) + \left(\sum_{x \in \xi_N^j(\epsilon,X)} \theta_J[G(\mu_{x+1}) - G(\mu_x)]\right)\right) \\
&\quad\quad + JX\theta_J(b_{=}(k,\eta) + 2c\eta D + \Delta(N) + (1 - q_N(S_N(\eta,k)))) \\
&\quad \leq \epsilon + J\frac{c(1-k)}{X}\frac{\theta_J^2}{\epsilon} + JX\theta_J\Big(b_{=}(k,\eta) + 2c\eta D + \Delta(N) + (1 - q_N(S_N(\eta,k)))\Big)
\end{aligned} \tag{13}$$

since $\sum_{x \in \xi_N^j(\epsilon,X)} [G(\mu_{x+1}) - G(\mu_x)] \leq \frac{c(1-k)}{X}\frac{\theta_J}{\epsilon}$. Then

$$\lim_{\epsilon \to 0} \lim_{X \to \infty} \lim_{k \to 0} \lim_{\eta \to 0} \lim_{N \to \infty} \sum_{j=1}^{J} \beta_0(\theta_j) \int_{\mu=0}^{2} u_{\sigma_N}(\mu f_j)g(\mu)d\mu - \underline{V}_N(k,\eta,X) = \lim_{\epsilon \to 0} \lim_{X \to \infty} \epsilon + J\frac{c(1-k)}{X}\frac{\theta_J^2}{\epsilon} = 0.$$

Again, since $\epsilon$, $X$, $k$, and $\eta$ were all constructed variables, this implies that

$$\lim_{N \to \infty} \sum_{j=1}^{J} \beta_0(\theta_j) \int_{\mu=0}^{2} u_{\sigma_N}(\mu f_j)g(\mu)d\mu = \lim_{\epsilon \to 0} \lim_{X \to \infty} \lim_{k \to 0} \lim_{\eta \to 0} \lim_{N \to \infty} \underline{V}_N(k,\eta,X) \leq \mathbb{E}_{\beta_0}[\theta].$$

As it is already clear from the lower bound on $u_{\sigma_N}(\mu f_j)$ that $\lim_{N\to\infty}\sum_{j=1}^{J}\beta_0(\theta_j)\int_{\mu=0}^{2}u_{\sigma_N}(\mu f_j)g(\mu)d\mu \geq \mathbb{E}_{\beta_0}[\theta]$, equality obtains.

## A.3 Proof of Lemmas A.1, 4.2 and 4.4

**Proof of Lemma A.1** Consider a game in which the type set is $\mathcal{T}^+(M)$, each type's action set is the set of messages in $M$ that they are able to send, and the payoff to playing $\hat{\sigma}(\cdot|t)$ against the receiver's putative strategy profile $\hat{\sigma}'$ is $\sum_{\tilde{f}} u(\beta_{\hat{\sigma}'}(\cdot|\tilde{f}))\hat{\sigma}(\tilde{f}|t)$, the utility to the sender of the receiver's updated belief conditional on seeing them play $\tilde{f}$ when the population is expected to play according to $\hat{\sigma}'$.

Payoffs are continuous in $\hat{\sigma}$ and $\hat{\sigma}'$. Let the best response correspondence be given by

$$r_t(\hat{\sigma}') = \arg\max_{\hat{\sigma}(\cdot|t)} \sum_{\tilde{f}\in M} u(\beta_{\hat{\sigma}'}(\cdot|\tilde{f}))\hat{\sigma}(\tilde{f}|t).$$

A fixed point $\hat{\sigma}^*$ of $r$ corresponds to a PBE of the constructed game, and a standard Nash existence argument shows that there must be at least one. Then let $M'$ be the set of messages that achieve the highest payoff under $\hat{\sigma}^*$; along with the restriction of $\hat{\sigma}^*$ to $T^+(M')$, it forms an upper pool.

If $M$ is not itself an upper pool, then $\mathcal{T}^+(M)\setminus\mathcal{T}^+(M')$ is nonempty and contains types that do worse than those in $\mathcal{T}^+(M')$. Then,

$$u(\mathcal{T}^+(M')) > u(\mathcal{T}^+(M)) > u(\mathcal{T}^+(M)\setminus\mathcal{T}^+(M)).$$

$\blacksquare$

**Proof of Lemma 4.2** Consider 2 such pools, $M = \{\tilde{f}_1,\ldots,\tilde{f}_I\}$ and $M' = \{\tilde{f}'_1,\ldots,\tilde{f}'_J\}$, with type sets $\mathcal{T}^+(M)$ and $\mathcal{T}^+(M')$. We aim to show their union is also a utility-maximizing upper pool. Let $A = \mathcal{T}^+(M)\setminus\mathcal{T}^+(M')$, $B = \mathcal{T}^+(M')\setminus\mathcal{T}^+(M)$, and $C = \mathcal{T}^+(M)\bigcap\mathcal{T}^+(M')$. Observe that if we let $M''$ be the message set that includes $\tilde{f}_i \vee \tilde{f}'_j$ for every $i \leq I$, $j \leq J$ (where $\vee$ is the pointwise max operator on datasets), then $C = \mathcal{T}^+(\mathcal{M}'')$.

We have $u(\mathcal{T}^+(M)) = u(\alpha A + (1-\alpha)C) = u(\mathcal{T}^+(M')) = u(\alpha' B + (1-\alpha')C) = u^*$. So $u(\mathcal{T}^+(M)\bigcup\mathcal{T}^+(M')) \geq u^*$ unless $u(A) < u^*$, $u(B) < u^*$, and $u(C) > u^*$; but by the previous lemma, the last of these would imply that $C$ contains a higher-utility upper pool than $M$ and $M'$. Since this is not true, $u(\mathcal{T}^+(M)) = u^*$ and it is an upper pool itself (otherwise it would contain a strictly better upper pool, a contradiction). $\blacksquare$

**Proof of Lemma 4.4** Suppose to the contrary that $u(\mathcal{T}^+_m(M_m)) \leq u(\mathcal{T}^+_{m+1}(M_{m+1}))$. Then,

$$u(\mathcal{T}^+_m(M_m\bigcup M_{m+1})) = u(\alpha'\beta(\cdot|\mathcal{T}^+_m(M_m)) + (1-\alpha')\beta(\cdot|\mathcal{T}^+_{m+1}(M_{m+1}))) \geq u(\mathcal{T}^+_m(M_m)),$$

which implies (by Lemma A.1) that either $M_m\bigcup M_{m+1}$ must itself be an upper pool with respect to $\mathcal{T}_m$, or that there exists $M' \subset M_m\bigcup M_{m+1}$ such that $u(\mathcal{T}^+_m(M')) > u(\mathcal{T}^+_m(M_m))$. Either of these would contradict that $M_m$ is a maximal upper pool in $\mathcal{T}_m$.

## A.4 Proof of Claim 3.2

**Proof** We proceed inductively. Since $\pi(0,N) > \pi(n_1,n_2)$ for all $(n_1,n_2)$, and $(0,N)$ cannot be imitated by any other type, $M_1 = (0,N)$ and $T_{\hat{\sigma}_{M_1}} = \{(0,N)\}$.

Now suppose $M_m = (0,\tilde{n}_2[m])$ for $m = 1,\dots,j$. Then $\mathcal{T}_{j+1} = \{(n_1,n_2)\}_{n_2 \leq \tilde{n}_2[j]-1, n_1 \leq N-n_2}$. Consider any $(\tilde{n}_1,\tilde{n}_2)$. Then by combinatorial identity,

$$\beta(H|\mathcal{F}^+(\tilde{n}_1,\tilde{n}_2)) = \pi_H(\tilde{n}_1,\tilde{n}_2).$$

There may be a set of types who are able to send $(\tilde{n}_1,\tilde{n}_2)$ but are already included in $T_{\hat{\sigma}_{M_m}}$ for some $m \leq j$. This set, $\mathcal{F}^+(\tilde{n}_1,\tilde{n}_2) \setminus \mathcal{T}_{j+1}$, satisfies

$$\beta(H|\mathcal{F}^+(\tilde{n}_1,\tilde{n}_2) \setminus \mathcal{T}_{j+1}) = \beta(H|\mathcal{F}^+(\tilde{n}_1,\tilde{n}_2[j])) = \pi_H(\tilde{n}_1,\tilde{n}_2[j]) > \pi_H(\tilde{n}_1,\tilde{n}_2).$$

Therefore,

$$\beta(H|\mathcal{T}^+_{j+1}(\tilde{n}_1,\tilde{n}_2)) < \pi_H(\tilde{n}_1,\tilde{n}_2).$$

This implies that a single message $(\tilde{n}_1,\tilde{n}_2)$ with $\tilde{n}_1 > 0$ cannot be a highest-payoff pool, since the type set of the upper pool consisting of message $(0,\tilde{n}_2)$ yields strictly higher payoff. To see this, observe that

$$\beta(H|\mathcal{T}^+_{j+1}(0,\tilde{n}_2)) = \alpha\beta(H|\mathcal{T}^+_{j+1}(\tilde{n}_1,\tilde{n}_2)) + (1-\alpha)\beta(H|\mathcal{T}^+_{j+1}(0,\tilde{n}_2) \setminus \mathcal{T}^+_{j+1}(\tilde{n}_1,\tilde{n}_2)),$$

and $\beta(H|\mathcal{T}^+_{j+1}(0,\tilde{n}_2) \setminus \mathcal{T}^+_{j+1}(\tilde{n}_1,\tilde{n}_2)) > \pi_H(\tilde{n}_1,\tilde{n}_2)$: that is, the receiver's belief conditional on the sender being in $\mathcal{T}_{j+1}$ and being able to send $\tilde{n}_2$ high signals but not $\tilde{n}_1$ low signals, is better than their belief when the sender is able to send at least $\tilde{n}_1$ low signals, therefore their belief is better when the burden of proof does not require any low signals be sent.

In addition, the highest-payoff pool cannot correspond to a set of distinct messages $M = \{(\tilde{n}_1^1,\tilde{n}_2^1),\dots,(\tilde{n}_1^L,\tilde{n}_2^L)\}$, such that $n_2^l < n_2^{l-1}$ and $n_1^l > n_1^{l-1}$ for all $l$.[8] To see this, first focus on

$$T_L := \{(n_1,n_2) : \tilde{n}_2^L \leq n_2 < \tilde{n}_2^{L-1}, \tilde{n}_1^L \leq n_1 \leq N - n_2\},$$

the set of types in $\mathcal{T}_{j+1}$ that can send $(\tilde{n}_1^L,\tilde{n}_2^L)$ but no other messages in $M$. Observe as before that $\beta(H|T_L) \leq \pi_H(\tilde{n}_1^L,\tilde{n}_2^L)$. Consider 2 cases:

- If $\beta(H|T_L) \geq \beta(H|\mathcal{T}_{j+1}(M))$, then let $M' = \{(\tilde{n}_1^1,\tilde{n}_2^1),\dots,(\tilde{n}_1^{L-2},\tilde{n}_2^{L-2}),(\tilde{n}_1^{L-1},\tilde{n}_2^L)\}$, i.e. replace messages $(\tilde{n}_1^{L-1},\tilde{n}_2^{L-1})$ and $(\tilde{n}_1^L,\tilde{n}_2^L)$ with a single message that is their (pointwise) minimum.
- If $\beta(H|T_L) < \beta(H|\mathcal{T}_{j+1}(M))$, then letting $M' = \{(\tilde{n}_1^1,\tilde{n}_2^1),\dots,(\tilde{n}_1^{L-1},\tilde{n}_2^{L-1})\}$, i.e. drop $(\tilde{n}_1^L,\tilde{n}_2^L)$ from the message set.

[8] A set of messages that does not satisfy these properties can either be reordered to do so, or is redundant in that there are some $l$, $l'$ such that $(n_1^l,n_2^l) \subset (n_1^{l'},n_2^{l'})$; so sets of messages satisfying these criteria are exhaustive of possible upper pools.

In either case, we have $\beta(H|\mathcal{T}_{j+1}(M')) \geq \beta(H|\mathcal{T}_{j+1}(M))$, and $M'$ is a strictly smaller set of messages than $M$. Repeat on $M'$ and iterate until the message set is a singleton; then it is a commuting upper pool that yields strictly better belief than $M$.

The above argument shows that message set $M_{j+1}$ of the unique upper pool chosen in the $j+1^{st}$ step of the algorithm is of the form $\{(0, \tilde{n}_2[j+1])\}$ where $\tilde{n}_2[j+1] < \tilde{n}_2[j]$. It is immediate the value of $\tilde{n}_2[j+1]$ that maximizes payoff to the pool is as given in the claim. Given the choice of $\tilde{n}_2[j+1]$, the payoff to $M_m$ is decreasing in $m$; therefore, the strategy profile constructed is an equilibrium. ▮