Online Appendix
Robust Implementation with Costly Information

Harry Pei[*] Bruno Strulovici[†]

October 23, 2022

A Extension: Robust Implementation with a Continuum of States

This appendix extends Theorems 1 and 2 to environments in which (i) there is a continuum of states, (ii) the state space Θ is compact, and (iii) agents' payoff functions in the unperturbed environment and the social choice function f are all continuous with respect to θ.

Formally, let Θ be a compact set in some normed vector space with norm $||\cdot||$. Let $q \in \Delta(\Theta)$ denote the objective distribution of θ, which we assume to have full support and no atom. A social choice function $f : \Theta \rightarrow \Delta(Y)$ is *continuous* if for every $\varepsilon > 0$, there exists $\delta > 0$ such that $||f(\theta) - f(\theta')||_{TV} \leq \varepsilon$ for every $||\theta - \theta'|| \leq \delta$. This definition corresponds to uniform continuity, which is equivalent to continuity since Θ is compact. The same comment applies to the continuity of agents' payoff functions introduced below.

Agent $i \in \{1,2\}$ can observe the realization of θ at cost $c_i \in [0,+\infty)$. Agent i's payoff function in the unperturbed environment is $u_i(\theta,y) + t_i - c_i d_i$.

We say that $u_i(\theta,y)$ is continuous with respect to θ if for every $y \in Y$ and $\varepsilon > 0$, there exists $\delta > 0$ such that $|u_i(\theta,y) - u_i(\theta',y)| \leq \varepsilon$ for every $||\theta - \theta'|| \leq \delta$. The notion of η-perturbation remains the same as in the baseline model, that is, agents' payoff functions coincide with those in the unperturbed environment with probability at least $1-\eta$, and with complementary probability, they can have arbitrary preferences over state-contingent outcomes $\widetilde{u}_i(\omega,\theta,y)$, arbitrary costs of learning $\widetilde{c}_i(\omega)$, and arbitrary beliefs and higher-order beliefs about each other's preferences over outcomes and costs of learning as long as these beliefs can be derived from a common prior.

[*] Department of Economics, Northwestern University. Email: harrydp@northwestern.edu

[†] Department of Economics, Northwestern University. Email: b-strulovici@northwestern.edu

We emphasize that we do not require agent i's payoff in the perturbed environment $\widetilde{u}_i(\omega, \theta, y)$ to be continuous with respect to θ when type $Q_i(\omega)$ of agent i is not a normal type.

Corollary 1. *Suppose Θ is compact, q has full support and has no atom, and both f and (u_1, u_2) are continuous with respect to θ. For every $\varepsilon > 0$, there exist $\eta > 0$ and a finite mechanism $\mathcal{M}$ such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium σ under $(\mathcal{M}, \mathcal{G})$ such that $\max_{\theta \in \Theta} ||g_\sigma(\theta) - f(\theta)|| \leq \varepsilon$.*

When there is a continuum of states and the objective state distribution has no atom, we can dispense the generic assumption on q in the case where Θ is a finite set. Intuitively, this is because we can always partition Θ into several connected subsets such that the probability of one of these subsets is strictly greater than the probability of every other subset.

We explain how to modify the proof of Theorem 2 to show this corollary. For simplicity, we focus on the case in which $u_1(\theta, y) = u_2(\theta, y) = 0$. The generalization to general utility functions $u_1(\theta, y)$ and $u_2(\theta, y)$ follows the same steps as those of Appendix A. Since the state space Θ is compact and the desired social choice function f is continuous, for every $\varepsilon > 0$ one can construct a finite partition of Θ using the finite cover theorem that satisfies the following three conditions:

1. Every partition element occurs with positive probability under q.

2. There exists a partition element that occurs with strictly higher probability compared to every other partition element.

3. For every pair θ, θ' that belong to the same partition element, we have $||f(\theta) - f(\theta')||_{TV} \leq \frac{\varepsilon}{2}$.[1]

Fix any partition that satisfies the above requirements. Denote the partition elements by $\{\Theta^1, ..., \Theta^n\}$. For every $j \in \{1, 2, ..., n\}$, let θ^j be an arbitrary element in Θ^j. We introduce a new social choice function $\widetilde{f} : \Theta \to \Delta(Y)$ such that $\widetilde{f}(\theta) = f(\theta^j)$ for every $\theta \in \Theta^j$ and $j \in \{1, 2, ..., n\}$.

Consider the mechanism constructed in the proof of Theorem 2, in which every agent has $2n-1$ messages. With a continuum of states, each agent is asked to report which element of the partition the realized θ belongs to. The proof of Theorem 2 implies that there exists a mechanism $\mathcal{M}$ such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma^*(\mathcal{G})$ such that $\max_{\theta \in \Theta} ||g_{\sigma^*(\mathcal{G})}(\theta) - \widetilde{f}(\theta)||_{\mathrm{TV}} < \varepsilon/2$. Since $||\widetilde{f}(\theta) - f(\theta)||_{TV} = ||f(\theta^j) - f(\theta)||_{TV} \leq \varepsilon/2$, the triangular inequality implies that $\max_{\theta \in \Theta} ||g_{\sigma^*(\mathcal{G})}(\theta) - f(\theta)||_{\mathrm{TV}} < \varepsilon$. Hence, the said mechanism robustly implements f.

[1] For general $u_1(\theta, y)$ and $u_2(\theta, y)$ that are continuous with respect to θ, we can find a partition that satisfies the above requirements while also making sure that $|u_i(\theta, y) - u_i(\theta', y)| < \varepsilon/2$ for every $y \in Y$, $i \in \{1, 2\}$, and θ, θ' belonging to the same partition element.

Remark: When there is a continuum of states, in order to implement a social choice function that is ε-close to f, our mechanism requires each agent to have more messages as ε goes to zero. This is because there are two sources of approximation errors: one of them is caused by the perturbation on agents' preferences and beliefs, and the other one is caused by approximating f via $\widetilde{f}$. The second source of approximation error vanishes to zero when the partition on Θ becomes finer. In another word, the number of messages in the mechanism depends on our tolerance of approximation errors ε. This stands in contrast to environments with a finite number of states, in which the planner can robustly implement the desired social choice function using a mechanism where each agent has $2|\Theta| - 1$ messages, regardless of the required approximation error.

B General Information Acquisition Technologies

We extend our main result to environments in which the agents can choose any partition of the state space Θ as their information structures, and different partitions of the state space may have different costs. We start from describing the general environment.

Let Θ be a finite set of states and $q \in \Delta(\Theta)$ denote the prior distribution of θ. Let Y denote the set of outcomes. The planner commits to a mechanism $\mathcal{M} \equiv \{M_1, M_2, t_1, t_2, g\}$, where M_i is a finite set of messages for agent $i \in \{1, 2\}$, $t_i : M_1 \times M_2 \to \mathbb{R}$ is the transfer to agent i, and $g : M_1 \times M_2 \to \Delta(Y)$ is the implemented outcome.

After observing $\mathcal{M}$, agents simultaneously and independently decide what information to acquire. Each agent can choose any partition of Θ as his information structure. Let $\mathcal{P}$ be the set of partitions of Θ. Let $P_i \in \mathcal{P}$ denote the partition chosen by agent i. Let P^* denote the finest partition. Agent $i \in \{1, 2\}$ observes the element of P_i the realized θ belongs to and sends a message $m_i \in M_i$. The planner makes transfers and implements an outcome according to $\mathcal{M}$. Agent i's payoff is:

$$u_i(\theta, y) + t_i - c_i(P_i), \tag{B.1}$$

where $c_i : \mathcal{P} \to [0, +\infty)$ is agent i's information acquisition cost function. A *perturbation* is characterized by

$$\mathcal{G} \equiv \Big\{\Omega, \Pi, (Q_i)_{i\in\{1,2\}}, (\widetilde{u}_i)_{i\in\{1,2\}}, (\widetilde{c}_i)_{i\in\{1,2\}}\Big\},$$

where Ω is a countable set of *circumstances*, whose typical element is denoted by $\omega \in \Omega$, $\Pi \in \Delta(\Omega)$ is the distribution of ω, and is assumed to be independent of θ, and Q_i is agent i's information partition on Ω. For every $\omega \in \Omega$, let $Q_i(\omega)$ denote the partition element of Q_i that contains ω. Agent i's payoff function is given by

$$\widetilde{u}_i(\omega, \theta, y) + t_i - \widetilde{c}_i(\omega, P_i). \tag{B.2}$$

Type $Q_i(\omega)$ is a *normal type* if $\widetilde{u}_i(\omega', \theta, y) = u_i(\theta, y)$ and $\widetilde{c}_i(\omega, P_i) = c_i(P_i)$ for every $\omega' \in Q_i(\omega)$. For every $\eta > 0$, we say that $\mathcal{G}$ is an η-perturbation if the probability of the event that both agents are normal is at least $1 - \eta$. For every $\eta > 0$ and $\overline{c} > 0$, we say that $\mathcal{G}$ is a $\overline{c}$-bounded η-perturbation if it is an η-perturbation where $\widetilde{c}_i(\omega, P^*) \leq \overline{c}$ for every $i \in \{1, 2\}$ and $\omega \in \Omega$.

For any $f : \Theta \to \Delta(Y)$, we describe a mechanism where each agent has n messages that can robustly implement f for all $\overline{c}$-bounded perturbations. We focus on the case where $u_1 = u_2 = 0$ since generalizing the proof to arbitrary u_1 and u_2 resembles the arguments in Appendix A.

Let $M_1 = M_2 = \{1, 2, ..., n\}$. The outcome function $g : M_1 \times M_2 \to \Delta(Y)$ is given by:

- $g(j, j) = f(\theta^j)$ for every $j \in \{1, 2, ..., n\}$.
- $g(i, j) = g(j, i)$ for every i, j.
- For every $j > i$, $g(j, i) = \sum_{k=1}^{j-1} \frac{1}{j-1} g(k, i)$, i.e., when agent 2 reports i, agent 1 reporting j and reporting 1 to $j - 1$ uniformly at random lead to the same distribution over outcomes.

The last step of the construction also implies that for every $j > i$, when agent 2 reports i, agent 1 reporting j and reporting 1 to i uniformly at random lead to the same distribution over outcomes. The transfer function to agent $i \in \{1, 2\}$ is given by

$$t_i(m_1, m_2) = \begin{cases} 0 & \text{if } m_i \neq m_{-i} \\ R_i^j & \text{if } m_i = m_{-i} = j, \end{cases}$$

where $R_i^1, ..., R_i^n$ satisfy $R_i^j > R_i^{j-1} + \frac{2c_i(P^*)}{q(\theta^j)}$ for every $j \geq 2$, and $R_i^1 \geq \frac{(n-1)\overline{c}}{\min_{\theta \in \Theta} q(\theta)}$.

Step 1: Let $\Sigma \equiv M^n$ be the set of strategies, with $(1, 2, ..., n) \in \Sigma$ the *truthful strategy*. Let

$$\Sigma^* \equiv \Big\{ (m^1, ..., m^n) \in \Sigma \text{ such that } m^j \leq j \text{ for every } j \in \{1, 2, ..., n\} \Big\}. \tag{B.3}$$

Intuitively, Σ^* is the set of strategies where agent's report does not exceed the index of the state. We show that there exists $\gamma < 1/2$ such that in the auxiliary game where agents' payoffs are

$\{t_1 - c_1(P_1), t_2 - c_2(P_2)\}$ and both agents are only allowed to choose strategies supported in Σ^*, then both agents being truthful is a γ-dominant equilibrium. To see this, for every $i \in \{1, 2\}$, suppose agent i believes that

1. agent $-i$'s strategy is supported in Σ^*,

2. agent $-i$ plays his truthful strategy $(1, 2, ..., n)$ with probability at least $1/2$.

Since $R_i^j > R_i^{j-1} > ... > R_i^1$ and $R_i^j > R_i^{j-1} + \frac{2c_i(P^*)}{q(\theta^j)}$, we know that conditional on the state being θ^j, agent i's expected transfer from reporting message j is strictly greater than his expected transfer from reporting any message strictly lower than j, and this difference in expected transfer is strictly greater than $c_i(P^*)$. When agent i is only allowed to report truthfully or to report a lower state, he strictly prefers his truthful strategy $(1, 2, ..., n)$ to any other strategy in Σ^*. Since Θ is finite, there exists $\gamma < 1/2$ such that both agents being truthful is a γ-dominant equilibrium in the auxiliary game.

Step 2: For any perturbation $\mathcal{G}$, consider a *perturbed auxiliary game* where agent i's payoff is $\widetilde{u}_i(\omega, \theta, y) + t_i - \widetilde{c}_i(\omega, P_i)$ and both agents are only allowed to use strategies supported in Σ^*. The critical path lemma in Kajii and Morris (1997) implies that for very $\varepsilon > 0$, there exists $\eta > 0$, such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ in the perturbed auxiliary game where the probability with which both agents using the truthful strategy is at least $1 - \varepsilon$. Since $g(j, j) = f(\theta^j)$ for every $j \in \{1, 2, ..., n\}$, social choice function f is implemented with probability more than $1 - \varepsilon$ when agents behave according to $\sigma(\mathcal{G})$.

Step 3: We show that $\sigma(\mathcal{G})$ remains an equilibrium when both agents are allowed to choose any strategy supported in Σ, not only strategies that are supported in Σ^*. Suppose by way of contradiction that type $Q_1(\omega)$ strictly prefers some strategy $(m^1, ..., m^n) \notin \Sigma^*$ to all strategies supported in Σ^*. Assuming that agent 2 behaves according to $\sigma(\mathcal{G})$, which means that his strategy is supported in Σ^*, we compare type $Q_1(\omega)$'s expected payoff from $(m^1, ..., m^n)$ to his expected payoff from the following mixed strategy $(m_\dagger^1, ..., m_\dagger^n)$, where

- if $m^j \leq j$, then $m_\dagger^j = m^j$;

- if $m^j > j$, then $m_\dagger$ is the mixed strategy of reporting $\{1, 2, ..., j\}$ each with probability $\frac{1}{j}$.

One can verify that $(m_\dagger^1, ..., m_\dagger^n)$ is supported in Σ^*, and furthermore, as long as player 2's strategy is supported in Σ^*, the implemented outcome is the same no matter whether agent 1 uses strategy

$(m^1, ..., m^n)$ or strategy $(m^1_\dagger, ..., m^n_\dagger)$. In addition, for every j such that $m^j > j$, $m^j_\dagger$ attaches strictly positive probability to every element in the set $\{1, 2, ..., j\}$. Hence, by reporting $m^j_\dagger$ instead of m^j in state θ^j, agent 1's expected transfer increases by at least $\frac{q(\theta_j)R^1_1}{n-1}$. Type $Q_1(\omega)$ prefers $(m^1_\dagger, ..., m^n_\dagger)$ to $(m^1, ..., m^n)$ if $\frac{q(\theta_j)R^1_1}{n-1} > \overline{c}$. Hence, for every perturbation $\mathcal{G}$, the equilibrium in the auxiliary perturbed game $\sigma(\mathcal{G})$ remains an equilibrium when both agents are allowed to choose any strategy supported in Σ, which implies that our mechanism robustly implements f.

C Proof of Proposition 1: General Utility Functions and Costs

We extend the proof of Proposition 1 in Appendix F of the main text to general utility functions u_1 and u_2 and general learning costs c_1 and c_2. Consider the mechanism whose outcome function is given by:

$$g(m_1, m_2) = \begin{cases} f(\theta^{|m_1|}) & \text{if } |m_1| = |m_2| \\ f(\theta^1) & \text{otherwise,} \end{cases} \tag{C.1}$$

and whose transfer functions are given by:

$$t_1(m_1, m_2) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_1 \leq 1 \text{ but } (m_1, m_2) \neq (1,1) \\ R^0 - x & \text{if } m_1 \geq 2 \text{ and } m_2 \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.2}$$

$$t_2(m_1, m_2) = \begin{cases} R^j & \text{if } m_1 = m_2 = j \geq 1 \\ R^0 & \text{if } m_2 \leq 1 \text{ but } (m_1, m_2) \neq (1,1) \\ R^0 - x & \text{if } m_2 \geq 2 \text{ and } m_1 \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.3}$$

where the parameters $\{R^n, ..., R^1, R^0, x\}$ satisfy $R^n, ..., R^0 > x > \frac{\max\{c_1, c_2\}}{q(\theta^n)}$

$$R^1 - R^0 > \frac{2\max\{c_1, c_2\}}{q(\theta^1)} + 2 \max_{i \in \{1,2\}} \Big\{ \max_{y \in Y} u_i(\theta^1, y) - \min_{y \in Y} u_i(\theta^1, y) \Big\}, \tag{C.4}$$

$$R^j - R^1 - x > \frac{2\max\{c_1, c_2\}}{q(\theta^j)} + 2 \max_{i \in \{1,2\}} \Big\{ \max_{y \in Y} u_i(\theta^j, y) - \min_{y \in Y} u_i(\theta^j, y) \Big\} \text{ for every } j \in \{2, 3, ..., n\}, \tag{C.5}$$

and

$$\frac{x}{R^j - R^0} \geq \frac{q(\theta^j)}{1-q(\theta^j)} \text{ for every } j \in \{2,3,...,n\}. \tag{C.6}$$

We modify the first step of our proof in which we show that both agents being truthful is a γ-dominant equilibrium for some $\gamma < \frac{1}{2}$.

Recall that each agent has $2n-1$ messages and that we are considering a *restricted game without perturbation* where for every $i \in \{1,2\}$, agent i is only allowed to use strategies that belong to $\Delta(\Sigma_i^*)$ where

$$\Sigma_i^* \equiv \Big\{(m^1, ..., m^{|S_i|}) \in \Sigma \text{ such that for every } k \in \{1,...,|S_i|\},\, m^k \in \{-n,...,-2,1\} \cup \{h_i(s_i^k)\}\Big\}.$$

Suppose agent 1 believes that agent 2 intends to be truthful with probability at least $\frac{1}{2}$,

- For every $j \geq 2$, conditional on agent 1 receiving a signal $s_1 \in S_1$ that satisfies $h_1(s_1) = j$, if agent 1's realized message is j, then he receives an expected transfer of

  $$\Pr(m_2 = j|s_1)R^j + \Pr(m_2 \leq 1|s_1)(R^0 - x),$$

  and if agent 1's realized message is no more than 1, then he receives an expected transfer of

  $$\Pr(m_2 = 1|s_1)R^1 + \Pr(m_2 \neq 1|s_1)R^0.$$

  Since $\pi(h_2(s_2) = h_1(s_1)|s_1) \geq 1-\overline{\tau}$ when π is of size $\overline{\tau}$, we have $\Pr(m_2 = j|s_1) \geq \frac{1-\tau}{2}(1-\overline{\tau})$ and $\Pr(m_2 = 1|s_1) \leq 1 - \frac{1-\tau}{2}(1-\overline{\tau})$. When inequality (C.5) is satisfied, $\overline{\tau}$ is close to 0, and $\tau \leq \overline{\tau}$, we have $q(\theta^j)\Big(\Pr(m_2 = j|s_1)R^j + \Pr(m_2 \leq 1|s_1)(R^0 - x)\Big) - q(\theta^j)\Big(\Pr(m_2 = 1|s_1)R^1 + \Pr(m_2 \neq 1|s_1)R^0\Big) > \max\{c_1, c_2\} + q(\theta^j)\Big\{\max_{y,y' \in Y} u_1(\theta^j, y) - u_1(\theta^j, y')\Big\}$. Therefore, agent 1 strictly prefers to send message j when he receives any signal $s_1 \in S_1$ that satisfies $h_1(s_1) = j$ when he believes that agent 2 intends to be truthful with probability at least $\frac{1}{2}$.

- Conditional on agent 1 receiving a message s_1 that satisfies $h_1(s_1) = 1$, his expected transfer when his realized message is 1 is $\Pr(m_2 = 1|s_1)R^1 + \Pr(m_2 < 0|s_1)R^0$ and his expected transfer when his realized message is negative is R^0. When inequality (C.4) is satisfied and $\overline{\tau}$ is close enough to 0, agent 1 prefers to message 1 as his intended message to any negative message as his intended message when he believes that agent 2's strategy belongs to $\Delta(\Sigma_2^*)$ and agent 2 intends to be truthful with probability at least $\frac{1}{2}$, even taking into account his cost of learning c.

Since agent 1 has a strict incentive to be truthful when he believes that agent 2 intends to be truthful with probability at least $\frac{1}{2}$, there exists $\gamma < \frac{1}{2}$ such that he also has a strict incentive to do so when he believes that agent 2 intends to be truthful with probability at least γ. Therefore, both agents intending to be truthful is a γ-dominant equilibrium.

D Proof of Proposition 2

We adapt the proof of Theorem 1 to show Statement 1. We focus on the case where $u_1 = u_2 = 0$ and $c_1 = c_2 = c$. Extending the proof to general (u_1, u_2) and heterogeneous learning costs is analogous to the generalization in Appendix A of the main text, and modifying the proof of Theorem 2 to show Statement 2 follows a similar argument to the one given here. The details are available upon request.

Let $\mathbf{q}_i$ denote the set of interim beliefs of player $i \in \{1, 2\}$. Player i's *pure strategy* is

$$\{m_i^1(q_i), ..., m_i^n(q_i)\}_{q_i \in \mathbf{q}_i}, \tag{D.1}$$

where $m_i^j(q_i)$ is the message he sends when his interim belief is q_i and the state is θ^j. Let Σ_i denote the set of pure strategies for agent i. Let $\Sigma_i^* \subset \Sigma_i$ be such that a strategy belongs to Σ_i^* if and only if $m_i^j(q_i) \in \{1, j\}$ for every $j \in \{1, 2, ..., n\}$ and $q_i \in \mathbf{q}_i$. Agent i's strategy is *truthful* if $m_i^j(q_i) = j$ for every $j \in \{1, 2, ..., n\}$ and $q_i \in \mathbf{q}_i$. Consider the status quo rule with ascending transfers constructed in Section 4.1 of the main text where the parameters $R^n, ..., R^1$ satisfy

$$R^j > R^1 \text{ for every } j \geq 2, \tag{D.2}$$

$$\sum_{j=2}^{n} (R^j - R^1) q(\theta^j) > 2c \text{ and } R^1 q(\theta^1) \geq \overline{c} \text{ for every } q \in \mathbf{q}. \tag{D.3}$$

Such $R^n, ..., R^1$ exist when $\mathbf{q}$ is interior. The rest of the proof is similar to that of Theorem 1.

First, let us examine the restricted game without any perturbation where for every $i \in \{1, 2\}$, agent i is only allowed to choose strategies in $\Delta(\Sigma_i^*)$. If agent i believes that agent j is truthful with probability at least $\frac{1}{2}$, then conditional on each q_i, the expected transfer he receives is strictly greater when he uses his truthful strategy. Hence, there exists $\gamma < \frac{1}{2}$ such that both agents being truthful is a γ-dominant equilibrium. Next, let us consider the restricted game with perturbation

$\mathcal{G}$. The critical path lemma implies that for every $\varepsilon > 0$, there exists $\eta > 0$ such that for every η-perturbation $\mathcal{G}$, there exists an equilibrium $\sigma(\mathcal{G})$ induced by $(\mathcal{M}, \mathcal{G})$ in which both agents use their truthful strategies with probability more than $1 - \varepsilon$. In this equilibrium, f is implemented with probability more than $1 - \varepsilon$. In the last step, let us consider the unrestricted game with perturbation. Similar to the proof of Theorem 1, the second part of (D.3) implies that $\sigma(\mathcal{G})$ remains an equilibrium when agents can use any strategies in Σ_i, not just those in Σ_i^*. This verifies that our mechanism can robustly implement f for every $(q_1, q_2) \in \mathbf{q} \times \mathbf{q}$.