Available online at www.sciencedirect.com
ScienceDirect
Journal of Economic Theory 146 (2011) 2093–2104
JOURNAL OF Economic Theory
www.elsevier.com/locate/jet

Notes

The robustness of robust implementation ☆

Moritz Meyer-ter-Vehn [a,*], Stephen Morris [b]

[a] *UCLA, Department of Economics, Bunche Hall 8283, Los Angeles, CA 90095, USA*
[b] *Princeton University, Department of Economics, USA*

Received 19 January 2011; accepted 21 January 2011

Available online 3 March 2011

Abstract

We show that a mechanism that robustly implements optimal outcomes in a one-dimensional supermodular environment continues to robustly implement ε-optimal outcomes in all close-by environments. Robust implementation of ε-optimal outcomes is thus robust to small perturbations of the environment. This is in contrast to ex-post implementation which is not robust in this sense as only trivial social choice functions are ex-post implementable in generic environments.

JEL classification: D82

Keywords: Robust implementation; Ex-post implementation; Social choice correspondence; Belief-dependent outcomes

1. Introduction

Fix a mechanism design environment where each agent has a payoff type and agents' payoffs from different allocations depend on the profile of payoff types. If we come up with a mechanism where, whatever the agents' beliefs and higher order beliefs, there is an equilibrium where outcomes are consistent with a social choice correspondence, we say that the mechanism partially robustly implements the social choice correspondence.

☆ This research is supported by NSF grant SES-0850718. We would like to thank the associate editor at JET and three anonymous referees for their thoughtful advice that improved the exposition of this note, and seminar audiences at ASSA 2010, the Princeton Conference on Mechanism Design, Toronto, and UCLA for helpful comments.
* Corresponding author. Fax: +1 310 825 9528.
E-mail addresses: mtv@econ.ucla.edu (M. Meyer-ter-Vehn), smorris@princeton.edu (S. Morris).
URLs: http://www.econ.ucla.edu/people/Faculty/Meyer-ter-vehn.html (M. Meyer-ter-Vehn), http://www.princeton.edu/~smorris (S. Morris).

0022-0531/$ – see front matter
doi:10.1016/j.jet.2011.03.011

If we restrict attention to 'direct mechanisms' where agents truthfully report their payoff types but not their beliefs, a social choice correspondence can be partially robustly implemented only if it can be implemented in ex-post equilibrium.[1] Bergemann and Morris [2] show that this equivalence between robust mechanism design and ex-post mechanism design continues to hold for some environments and social choice correspondences, even if mechanisms eliciting and using beliefs are allowed. In particular, this is true in a quasi-linear environment when there is a unique acceptable social allocation as a function of the payoff types, but transfers can depend on beliefs and are not restricted by budget balance.

In many contexts this observation is negative since ex-post incentive compatibility can be very restrictive. Jehiel, Meyer-ter-Vehn, Moldovanu and Zame [13] show that in a generic class of multi-dimensional quasi-linear environments, only trivial allocation functions can be ex-post implemented. This implies a disturbing discontinuity. In a one-dimensional single-crossing environment, efficient allocations are implementable in ex-post equilibrium and thus robustly. But if the single-crossing environment is embedded in a larger, multi-dimensional payoff environment where single-crossing is perturbed generically, only trivial allocation functions are ex-post implementable even when the mechanism can be tailored to the perturbation.

Here, we show that this negative result and the discontinuity around single-crossing are properties of ex-post implementation, but not properties of robust implementation. When we allow the entire social choice to depend on beliefs and strengthen the solution concept to full implementation in rationalizable strategies, continuity is restored. More precisely, Theorem 1 shows that if a mechanism $\mathcal{M}$ fully implements an optimal social choice function in a one-dimensional, supermodular, baseline environment, then $\mathcal{M}$ also fully implements an almost optimal social choice correspondence whenever agents have approximate common knowledge that the baseline environment is close in a sense we make precise. This observation highlights that robust implementation is more flexible than ex-post implementation in economically important environments. The intuition for the result is that rationalizable strategies vary continuously in underlying payoffs, so if agents have strict incentives to report truthfully in the baseline environment, then reports can be only slightly off in a perturbed environment.

The following example from Jehiel, Meyer-ter-Vehn, Moldovanu and Zame [13] illustrates the main idea. Consider a quasi-linear environment where an object must be allocated to one of two agents. Agent i has a payoff type $\theta_i = (\theta_i^p, \theta_i^c) \in [0,1]^2$, with the interpretation that θ_i^p is a private value component for agent i, while θ_i^c is a component that enters both agents' valuations of the object. The value of the object to agent i is $\theta_i^p + \varepsilon\theta_i^c\theta_{-i}^c$. In this environment, only trivial allocation functions are ex-post implementable. However, if the designer ignores the interdependence term $\varepsilon\theta_i^c\theta_{-i}^c$ and just conducts a second-price auction, it is weakly dominated for type (θ_i^p, θ_i^c) to bid outside $[\theta_i^p, \theta_i^p + \varepsilon]$. The resulting allocation depends on beliefs and is generally not efficient but the welfare loss is less than ε. Thus the social choice correspondence consisting of ε-efficient allocations is robustly implementable.[2,3] One attractive robustness feature that carries over from this example to our general result is that our mechanism $\mathcal{M}$ is defined only with

[1] This observation goes back to Ledyard [14] and Dasgupta, Hammond and Maskin [7] in the case of private values; the extension to interdependent values is straightforward.

[2] While similar in spirit, our notion of implementing ε-efficient allocations is strictly weaker than virtual efficient implementation because it allows for the possibility that the efficient allocation is chosen with probability zero.

[3] Bikhchandani [4] addresses the impossibility theorem of Jehiel et al. [13] in an alternative way. He shows that non-trivial allocation functions are ex-post implementable when the object can be allocated to neither agent, and there are no allocative externalities.

reference to the baseline environment, rather than being tailored to the true type space. Consequently, if $\mathcal{M}$ is efficient and detail-free, in the spirit of Dasgupta and Maskin [8], and Perry and Reny [17], then we can achieve almost efficient, detail-free implementation on all close-by type spaces.

In this paper, we generalize the logic of this example by considering as baseline a class of payoff environments from Bergemann and Morris [3], where payoffs are interdependent and not necessarily quasi-linear, but satisfy a supermodularity condition and a contraction property that limits the interdependence. Theorem 1 shows that if a direct mechanism implements an optimal social choice function in a baseline environment, then the same mechanism implements the social choice correspondence of almost optimal outcomes in any close-by environment.

1.1. Literature

Examples in Bergemann and Morris [2] illustrate the point that robust implementation can occur when ex-post implementation is impossible, using mechanisms that elicit and respond to beliefs. By showing here that close to optimal SCCs are robustly implementable, we highlight the economic significance of this point. Börgers [5], Smith [19], and Yamashita [20] show how mechanisms that elicit and respond to beliefs can robustly improve on ex-post incentive-compatible mechanisms, respectively in a general social choice setting, a public good setting with private values, and a bilateral trade setting.[4] A subtle difference to our note is that Börgers [5] and Smith [19] show that any ex-post mechanism is outperformed by some robust mechanism for some beliefs, while we show that there is a social choice correspondence that is robustly implementable but not ex-post implementable.[5]

Our result that belief-dependent robust mechanisms may improve on ex-post mechanisms contrasts with work by Chung and Ely [6] who show that an ex-post mechanism is the optimal robust mechanism for revenue maximization in a class of single-object allocation settings with quasi-linear payoffs.

2. Setup

Basics: There is a finite set of agents, $i = 1, \ldots, I$, and a compact outcome space X with associated lottery space $Y = \Delta(X)$. A mechanism $\mathcal{M} = (M_1, \ldots, M_I, g)$ is given by measurable message sets M_i and a measurable outcome function $g : M \to Y$.

Types: We represents agents' beliefs and preferences, and higher order beliefs and preferences, with a type space $\mathcal{T} = (T_i, \pi_i, \tilde{u}_i)_{i=1}^I$, where each agent i has a measurable set of types T_i, a measurable function describing the agent's beliefs about others' types $\pi_i : T_i \to \Delta(T_{-i})$, and a measurable, bounded Bernoulli utility function $\tilde{u}_i : X \times T \to [\underline{u}, \overline{u}]$. This utility function extends to lotteries in the standard expected utility way. We put no restrictions on the agents' beliefs and preferences, so this type space can be understood as the universal type space of Harsanyi, and Mertens and Zamir [15] over all possible bounded expected utility functions.

[4] Thus, these papers relate to the negative results about ex-post implementation by Gibbard [11], Satterthwaite [18] and Hagerty and Rogerson [12] in a similar way that this note relates to Jehiel, Meyer-ter-Vehn, Moldovanu and Zame [13].

[5] While Yamashita [20] compares mechanisms by their expected welfare, his two-price mechanism can be shown to also robustly implement a social choice correspondence that is not ex-post implementable.

Interim rationalizability: We ask when the planner can design a mechanism such that rational strategies of the agents lead to acceptable outcomes for the planner. We use the solution concept of interim (correlated) rationalizability (Dekel, Fudenberg and Morris [9]). Agent i is assumed to send some message that survives iterated deletion of messages that are not best responses to any belief over message-type profiles that (a) is consistent with π_i, and (b) puts probability one on profiles that have not yet been deleted. A formal description of the set of *interim rationalizable messages* $\widetilde{R}_i^{\mathcal{M}}(t_i)$ for type t_i of agent i playing mechanism $\mathcal{M}$ is as follows:

$$\widetilde{R}_{i,0}^{\mathcal{M}}(t_i) = M_i,$$

$$\widetilde{R}_{i,k+1}^{\mathcal{M}}(t_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \tilde{\mu}_i \in \Delta(M_{-i} \times T_{-i}) \text{ such that} \\ (1)\ \tilde{\mu}_i(\{(m_{-i}, t_{-i}) \mid m_j \in \widetilde{R}_{j,k}^{\mathcal{M}}(t_j) \text{ for all } j \neq i\}) = 1 \\ (2)\ \operatorname{marg}_{T_{-i}} \tilde{\mu}_i(\cdot) = \pi_i(\cdot|t_i) \\ (3)\ \int \tilde{u}_i(g(m_i, m_{-i}), (t_i, t_{-i}))\, d\tilde{\mu}_i(m_{-i}, t_{-i}) \\ \qquad \geqslant \sup_{m_i'} \int \tilde{u}_i(g(m_i', m_{-i}), (t_i, t_{-i}))\, d\tilde{\mu}_i(m_{-i}, t_{-i}) \end{array} \right\},$$

$$\widetilde{R}_i^{\mathcal{M}}(t_i) = \bigcap_{k \geqslant 1} \widetilde{R}_{i,k}^{\mathcal{M}}(t_i).$$

Dekel, Fudenberg and Morris [9] argue that interim rationalizable messages are those that are consistent with common knowledge of rationality among the agents and that are played in some subjective correlated equilibrium. We use rationalizability rather than equilibrium as solution concept as it is more natural and strengthens our results.

Implementation: The outcomes acceptable to the planner are represented by a social choice correspondence (SCC) $F : T \to Y$.

Definition 1. Mechanism $\mathcal{M}$ robustly implements SCC F at type profile $t \in T$ if

$$m \in \widetilde{R}^{\mathcal{M}}(t) \quad \Rightarrow \quad g(m) \in F(t).$$

This definition does not require existence of rationalizable messages. We argue in Section 4 that this does not matter for the interpretation of Theorem 1.[6]

Approximate optimality: We will be particularly interested in approximately optimal SCCs. Label a planner as agent 0 with utility $\tilde{u}_0 : Y \times T \to \mathbb{R}$. A SCC $F : T \to Y$ is λ-optimal at type profile $t \in T$ if

$$\tilde{u}_0\big(y', t\big) \geqslant \sup_y \big\{\tilde{u}_0(y, t)\big\} - \lambda \quad \forall y' \in F(t).$$

If the planner maximizes the sum of the agents' utilities, so $\tilde{u}_0(y, t) = \sum \tilde{u}_i(y, t)$, and in addition the environment is quasi-linear, then SCC F is λ-*optimal* if and only if allocations are λ-efficient.

The purpose of the paper is to give new sufficient conditions for the implementation of approximately optimal SCCs.

[6] To ensure existence of rationalizable messages we would need to assume compactness of type spaces T_i and message spaces M_i, as well as continuity of utility functions $\tilde{u}_i$, belief functions π_i, and the outcome function g of the mechanism.

3. Baseline payoff environments

The payoff (type) environment $(\Theta, u) = ((\Theta_i)_{i=1}^I, (u_i)_{i=0}^I)$ is given by compact one-dimensional payoff type spaces $\Theta_i \subseteq \mathbb{R}$ for agents $i = 1, \ldots, I$ and continuous Bernoulli utility functions $u_i : Y \times \Theta \to \mathbb{R}$ for $i = 0, 1, \ldots, I$.

Our sufficient condition relies on approximate common knowledge that payoffs are close to those generated by such a payoff environment that satisfies additional properties. A payoff environment does not specify agents' beliefs and higher order beliefs: we will be allowing them to have any beliefs and higher order beliefs.

We first describe the notion of closeness capturing approximate common knowledge of a payoff environment. Then we describe an alternative solution concept defined directly on the payoff environment that constrains behavior of types close to the payoff environment. Then we report the additional properties of the payoff environment that we use in our positive results.

3.1. Approximate common knowledge

We say that the payoff environment (Θ, u) is *γ-approximate common knowledge at type profile t*, if there is an event $E = (E_i)_{i=1}^I \subseteq T$ with $t \in E$ and measurable mappings to payoff types $\tilde{\theta}_i : E_i \to \Theta_i$ such that:

1. Baseline payoffs and true payoffs are γ-close on E:
$$\left|u_i\left(y, \tilde{\theta}(t)\right) - \tilde{u}_i(y,t)\right| \leqslant \gamma \quad \forall i, y \in Y,\ t \in E.$$
2. The event E is common $(1-\gamma)$-belief among agents:
$$\pi_i(E_{-i}|t_i) \geqslant 1-\gamma \quad \forall i, t_i \in E_i.$$

This definition allows for two perturbations of the payoff environment. Condition 1 allows payoffs to be misspecified by up to γ at every type profile. Outside the set E payoffs can be completely misspecified, but condition 2 requires the set E to be $(1-\gamma)$-belief closed. For an illustration, if T is any type space over Θ with perturbed utilities $u_i(\cdot)$ satisfying $|u_i(y,\theta) - \tilde{u}_i(y,\theta)| \leqslant \gamma$, then the payoff environment (Θ, u) is γ-approximate common knowledge on all of T.

3.2. Payoff environment solution concept

Instead of applying the solution concept of Definition 1 directly, we connect it to an approximate version of rationalizability defined directly on the payoff environment (Θ, u). For a fixed mechanism $\mathcal{M}$, $\delta > 0$ and each payoff type θ_i of each agent i, we iteratively delete messages that are not within δ of a best response for any belief over other agents' remaining payoff types and messages. Formally, we set

$$R_{i,0}^{\mathcal{M},\delta}(\theta_i) = M_i,$$

$$R_{i,k+1}^{\mathcal{M},\delta}(\theta_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \mu_i \in \Delta(M_{-i} \times \Theta_{-i}) \text{ such that} \\ (1)\ \mu_i(\{(m_{-i}, \theta_{-i}) \mid m_j \in \widetilde{R}_{j,k}^{\mathcal{M},\delta}(\theta_j) \text{ for all } j \neq i\}) = 1 \\ (2)\ \int u_i(g(m_i, m_{-i}), (\theta_i, \theta_{-i}))\, d\mu_i(m_{-i}, \theta_{-i}) \\ \qquad \geqslant \sup_{m_i'} \int u_i(g(m_i', m_{-i}), (\theta_i, \theta_{-i}))\, d\mu_i(m_{-i}, \theta_{-i}) - \delta \end{array} \right\},$$

$$R_i^{\mathcal{M},\delta}(\theta_i) = \bigcap_{k \geqslant 1} R_{i,k}^{\mathcal{M},\delta}(\theta_i)$$

and say that *message m_i is δ-rationalizable for payoff type θ_i* in mechanism $\mathcal{M}$, if $m_i \in R_i^{\mathcal{M},\delta}(\theta_i)$. If message m_i is interim rationalizable for a type t_i and the payoff environment (Θ, u) is approximate common knowledge at t_i, then m_i is δ-rationalizable for the corresponding payoff type θ_i in mechanism $\mathcal{M}$:

Lemma 1. *Fix a payoff environment (Θ, u) and a mechanism $\mathcal{M} = (M_1, \ldots, M_I, g(\cdot))$. For any $\delta > 0$ there exists $\gamma > 0$, such that whenever (Θ, u) is γ-approximate common knowledge at $t = (t_i, t_{-i})$, then any interim rationalizable message m_i of type t_i is also δ-rationalizable for the payoff type $\tilde{\theta}_i(t_i)$. Formally,*

$$\widetilde{R}_i^{\mathcal{M}}(t_i) \subseteq R_i^{\mathcal{M},\delta}\big(\tilde{\theta}_i(t_i)\big).$$

Proof. In Appendix A. □

This result is a consequence of the upper hemicontinuity of interim rationalizable outcomes (Dekel, Fudenberg and Morris [9]). It is an incomplete information analogue of a result of Ely [10] on complete information rationalizability. Battigalli and Siniscalchi [1] define a class of rationalizability solution concepts (called 'Δ-rationalizability') for incomplete information environments where common knowledge of certain first-order beliefs is built; here we are analogously building approximate common knowledge of the payoff environment.

As a corollary of this lemma we have:

Corollary 1. *For any $\delta > 0$ and mechanism $\mathcal{M}$ there exists $\gamma > 0$, such that $\mathcal{M}$ implements SCC F at type profile $t \in T$ whenever*

1. *payoff environment (Θ, u) is γ-approximate common knowledge at t, and*
2. *$m \in R^{\mathcal{M},\delta}(\tilde{\theta}(t)) \Rightarrow g(m) \in F(t)$.*

3.3. One-dimensional, contractive, supermodular payoff type environments

We now introduce the restrictions on the payoff environment that our main result appeals to. Recall that each Θ_i is a compact subset of $\mathbb{R}$ and that payoff functions $u_i(y, \theta), u_0(y, \theta)$ are continuous in θ. We consider slightly stronger assumptions than those in Bergemann and Morris [3].

A1 **Assumption – Monotone aggregator.**
The type profile θ affects payoffs u_i via (across-agents) aggregators $h_i : \Theta \to \mathbb{R}$, so that

$$u_i(y, \theta) = v_i\big(y, h_i(\theta_i, \theta_{-i})\big).$$

The aggregator h_i is continuous, and strictly increasing in θ_i, and payoffs $v_i : Y \times \mathbb{R} \to \mathbb{R}$ are continuous in aggregated types.

A2 **Assumption – Supermodularity.**
There is an order $\prec_i$ on Y such that the payoff function $v_i(y, \phi)$ is weakly supermodular.

A3 **Assumption – Contraction property.**
There exists a strictly increasing function $\alpha_{CP} : \mathbb{R}^+ \to \mathbb{R}^+$, such that for all $\xi > 0$

$$h_i(\theta_i, \theta_{-i}) - h_i\left(\theta_i', \theta_{-i}'\right) > \alpha_{CP}(\xi)$$

if $\theta_i - \theta_i' \geqslant \xi$ and $|\theta_j - \theta_j'| < \xi$ for all $j \neq i$.[7]

We will also consider a uniform version of ex-post incentive compatibility. Social choice function $f : \Theta \to Y$ is *uniformly ex-post incentive compatible*, if there exists a strictly increasing function $\alpha_{IC} : \mathbb{R}^+ \to \mathbb{R}^+$, such that for all $i, \theta_i \neq \theta_i'$ and θ_{-i} the payoff-loss when type θ_i reports θ_i' in the direct revelation mechanism $f : \Theta \to Y$ is bounded below by α_{IC}:

$$u_i\left(f(\theta_i, \theta_{-i}), (\theta_i, \theta_{-i})\right) - u_i\left(f\left(\theta_i', \theta_{-i}\right), (\theta_i, \theta_{-i})\right) \geqslant \alpha_{IC}\left(\left|\theta_i - \theta_i'\right|\right).$$

Instead of justifying these assumptions directly, we show that they are satisfied in the two following examples from Bergemann and Morris [3], and assert that they are also satisfied in the further examples 'Binary allocation' and 'Information aggregation' in that paper.

Single-object allocation: Quasi-linear utilities with type spaces $\Theta_i = [0, 1]$, valuations $v_i(\theta) = \theta_i + \gamma \sum_{j \neq i} \theta_j$, and $\gamma < 1/(I-1)$ satisfy assumptions A1–A3. Writing $\hat{\theta}_i$ for the report of agent i in the direct mechanism, the generalized Vickrey mechanism allocates to $i \in \arg\max_j \{\hat{\theta}_j\}$ and charges the winner $p_i = \max_{j \neq i}\{\hat{\theta}_j\} + \gamma \sum_{j \neq i} \hat{\theta}_j$. Truthtelling is ex-post incentive compatible and gives rise to an efficient allocation. To render this mechanism uniformly ex-post incentive compatible, it is played with probability $1 - \varepsilon$, while with probability $\varepsilon\hat{\theta}_i / I$ the object is allocated to i for a payment of $\hat{\theta}_i/2 + \gamma\hat{\theta}_i \sum_{j \neq i} \hat{\theta}_j$, and with probability $\varepsilon(I - \sum_j \hat{\theta}_j)$ the object is not allocated.

Public good provision: Quasi-linear utilities with valuations $v_i(x, \theta) = (\theta_i + \gamma \sum_{j \neq i} \theta_j)x$, cost functions $\frac{1}{2}x^2$, and $\gamma < 1/(I-1)$ satisfy assumptions A1–A3. The generalized Vickrey mechanism is ex-post efficient and uniformly ex-post incentive compatible.

4. Main result

Theorem 1. *Assume that* (Θ, u) *satisfies* A1–A3, *mechanism* $\mathcal{M} = (\Theta_1, \ldots, \Theta_N, f)$ *is uniformly ex-post incentive compatible, and the social choice function* f *is* λ*-optimal at every* θ. *Fix* $\varepsilon > 0$.

Then there exists $\gamma > 0$, *such that the same mechanism* $\mathcal{M}$ *robustly implements a social choice correspondence* F *that is* $(\lambda + \varepsilon)$*-optimal at any type profile* t *with* γ*-approximate common knowledge of the payoff environment* (Θ, u).

In the introductory auction example the mechanism that performs well in the perturbed environment is just the second-price auction. Similarly here, the $(\lambda + \varepsilon)$-optimal mechanism $\mathcal{M}$ does not depend on the perturbed true environment. This independence is important, because it implies that Theorem 1 does not exploit Definition 1 by constructing a mechanism $\mathcal{M}$ in which rationalizable messages do not exist.

We prove Theorem 1 in two steps. Lemma 2 shows that rationalizable reports of t_i must be close to $\tilde{\theta}_i(t_i)$. Then it essentially suffices to show that the objective function is continuous in the reports.

[7] We are grateful to an anonymous referee for pointing out this simplified definition.

Lemma 2. *Assume that the payoff environment* (Θ, u) *satisfies* A1–A3*, and that* $\mathcal{M} = (\Theta_1, \ldots, \Theta_N, f)$ *is uniformly ex-post incentive compatible. Then for every* $\xi > 0$*, there exists* $\delta > 0$*, such that for all types* θ_i *only reports* $\theta_i' \in [\theta_i - \xi, \theta_i + \xi]$ *are* δ*-rationalizable.*

Proof. If this is not the case, then there is $\xi > 0$ such that for any $\delta > 0$

$$\sup\{|\theta_i' - \theta_i| \colon i, \theta_i, \theta_i' \in R_i^{\mathcal{M},\delta}(\theta_i)\} \geqslant \xi.$$

Let the LHS be maximal for agent i. Applying the contraction property, there exist θ_i and $\theta_i' \in R_i^{\mathcal{M},\delta}(\theta_i)$ with $\theta_i \geqslant \theta_i' + \xi$ such that

$$h_i(\theta_i, \theta_{-i}) - h_i(\theta_i', \theta_{-i}') > \alpha_{CP}(\xi) \quad \forall \theta_{-i} \in \Theta_{-i},\ \theta_{-i}' \in R_{-i}^{\mathcal{M},\delta}(\theta_{-i}).$$

The aggregator h_i is uniformly continuous, so we can choose $\zeta \in (0, \xi)$ with

$$h_i(\theta_i, \theta_{-i}) \geqslant h_i(\theta_i' + \zeta, \theta_{-i}') \quad \forall \theta_{-i} \in \Theta_{-i},\ \theta_{-i}' \in R_{-i}^{\mathcal{M},\delta}(\theta_{-i}).$$

Fix any $\theta_{-i}' \in R_{-i}^{\mathcal{M},\delta}(\theta_{-i})$. The payoff-loss of type $\theta_i' + \zeta$, when he reports θ_i' instead of $\theta_i' + \zeta$ and others truthfully report θ_{-i}', is at least $\alpha_{IC}(\zeta)$ by uniform ex-post incentive compatibility:

$$u_i\big(f(\theta_i' + \zeta, \theta_{-i}'), (\theta_i' + \zeta, \theta_{-i}')\big) - u_i\big(f(\theta_i', \theta_{-i}'), (\theta_i' + \zeta, \theta_{-i}')\big) \geqslant \alpha_{IC}(\zeta).$$

This payoff-loss is weakly greater for type θ_i and any true type of others θ_{-i}

$$\begin{aligned} &u_i\big(f(\theta_i' + \zeta, \theta_{-i}'), (\theta_i, \theta_{-i})\big) - u_i\big(f(\theta_i', \theta_{-i}'), (\theta_i, \theta_{-i})\big) \\ &\quad \geqslant u_i\big(f(\theta_i' + \zeta, \theta_{-i}'), (\theta_i' + \zeta, \theta_{-i}')\big) - u_i\big(f(\theta_i', \theta_{-i}'), (\theta_i' + \zeta, \theta_{-i}')\big) \geqslant \alpha_{IC}(\zeta) \end{aligned}$$

because (1) the aggregated type is weakly greater $h_i(\theta_i, \theta_{-i}) \geqslant h_i(\theta_i' + \zeta, \theta_{-i}')$ by the above, (2) the aggregated value function v_i is supermodular in outcome and aggregator, and (3) outcomes are increasing $f(\theta_i' + \zeta, \theta_{-i}') \succcurlyeq_i f(\theta_i', \theta_{-i}')$ by supermodularity and uniform ex-post incentive compatibility.

Since this argument holds for any $\theta_{-i}' \in R_{-i}^{\mathcal{M},\delta}(\theta_{-i})$, we obtain the desired contradiction $\theta_i' \notin R_i^{\mathcal{M},\delta}(\theta_i)$ for any $\delta < \alpha_{IC}(\zeta)$. □

Proof of Theorem 1. We first show that there exists $\delta > 0$ such that every δ-rationalizable reports θ' of types θ lead to $(\lambda + \varepsilon/2)$-optimal outcomes in the payoff environment. For any $\varepsilon > 0$, fix $\xi > 0$ such that $|u_0(y, \theta') - u_0(y, \theta)| \leqslant \varepsilon/4$ for all $\theta' \in \prod[\theta_i - \xi, \theta_i + \xi]$. By Lemma 2 there exists $\delta > 0$, such that only reports $\theta_i' \in [\theta_i - \xi, \theta_i + \xi]$ are δ-rationalizable. So, for type profile θ, δ-rationalizable reports $\theta' \in R^{\mathcal{M},\delta}(\theta)$, and any outcome $y \in Y$ we obtain

$$u_0(y, \theta) - u_0\big(f(\theta'), \theta\big) \leqslant u_0\big(y, \theta'\big) - u_0\big(f(\theta'), \theta'\big) + \varepsilon/2 \leqslant \lambda + \varepsilon/2,$$

proving that outcome $f(\theta')$ is $(\lambda + \varepsilon/2)$-optimal for types θ in the payoff environment.

To finish the proof, choose $\gamma < \varepsilon/4$ such that all rationalizable messages θ' of types $t \in E$ are δ-rationalizable for the γ-close payoff types $\theta = \tilde{\theta}(t)$. Then

$$\tilde{u}_0(y, t) - \tilde{u}_0\big(f(\theta'), t\big) \leqslant u_0(y, \theta) - u_0\big(f(\theta'), \theta\big) + \varepsilon/2 \leqslant \lambda + \varepsilon$$

finishing the proof. □

5. Discussion

We have reported sufficient conditions for implementing almost optimal social choice correspondences. Theorem 1 implies that for a fixed type space Θ and an open set of quasi-linear payoff functions, ε-efficient social choice correspondences are robustly implementable.

To what extent can these sufficient conditions be weakened? Bergemann and Morris [3] show that in many payoff environments without the contraction property A3, it is impossible to fully implement efficient outcomes. Then, a fortiori, efficient outcomes cannot be approximated in close-by environments, and arguments by Oury and Tercieux [16] suggest that it may be impossible to even partially robustly implement an almost efficient correspondence. However, there may be baseline payoff environments that violate our sufficient conditions but that still allow for interesting mechanisms with unique rationalizable messages. In such environments – that assume the result of Bergemann and Morris [3] rather than its conditions – some regularity assumptions ensure that only close-by messages are δ-rationalizable as in Lemma 2. The proof of Theorem 1 then bounds the payoff-loss when true types have almost common knowledge of the baseline environment. These regularity assumptions are satisfied if either (1) the baseline environment is finite, or (2) it is compact and the mechanism is continuous.[8]

Abandoning the assumptions on the payoff environment more broadly can lead to a more dramatic failure of robust implementation. We illustrate this with a concluding example of a quasi-linear environment where values are not close to supermodular, and a mechanism designer with no information about agents' first-order beliefs weakly prefers a default choice over designing a mechanism that tries to take into account agents' information.

Example.[9] There are two agents, Rowena and Colin, with binary payoff types $\Theta_R = \{u, d\}$, $\Theta_C = \{\ell, r\}$ and binary allocations $x \in X = \{0, 1\}$ with lotteries $\Delta(X) = [0, 1]$. Utilities are quasi-linear $u_i(x, \theta, p) = x v_i(\theta) + p_i$ and values $v_i(\theta)$ are given by

Values	ℓ	r
u	$2, -1$	$-2, 1$
d	$-2, 1$	$2, -1$

Consider type spaces $T_i = \Theta_i \times \Delta(\Theta_{-i})$ with typical element $t_i = (\theta_i, \pi_i)$; we extend beliefs over payoff types $\pi_i \in \Delta(\Theta_{-i})$ to beliefs over types $\Delta(T_{-i})$ by assuming beliefs over first-order beliefs to be uniform.

In Appendix A we prove that for any incentive-compatible mechanism $f : T \to [0, 1]$ and any planner preferences over allocations $u_0(x, \theta)$, there is a default allocation $y_0 \in [0, 1]$ that beats the worst-case performance of $\mathcal{M}$: For all payoff types θ there exist belief-types π such that[10]

$$u_0\big(f(\theta, \pi), \theta\big) \leqslant u_0(y_0, \theta).$$

This is in stark contrast to the mechanisms that robustly lead to ε-optimal outcomes in close to supermodular environments. The reason for this negative result is that, say, Rowena does

[8] We are grateful to an anonymous referee for pointing this out to us. The argument that the rationalizability correspondence is upper hemicontinuous is analogue to Theorem 3 in Oury and Tercieux [16].

[9] This environment is taken from a working paper version of Jehiel, Meyer-ter-Vehn, Moldovanu, Zame [13].

[10] The argument extends immediately to implementation in rationalizable strategies: In any mechanism $\mathcal{M} = (M_R, M_C, g)$ where every type has a rationalizable message, for every payoff types θ there exist belief types π and rationalizable messages $m \in R^{\mathcal{M}}(\theta, \pi)$ such that $u_0(g(m), \theta) \leqslant u_0(y_0, \theta)$.

not know whether her type u values allocation $x = 1$ higher than her type d; that depends on her belief about Colin's type. The circularity of the monotonicity constraints in the payoff matrix implies that no mechanism can robustly separate the payoff types, i.e. ensure allocations $f(\theta, \pi) < y_0 < f(\theta', \pi)$ for some θ, θ', y_0 and all π.

The contrast with the introductory auction example is instructive: There, the order of types depends on beliefs only for types with close-by private value components. However, type $\theta_i = (1, 1)$ is unambiguously higher than type $\theta_i = (0, 0)$. This partial order ensures that ε-efficient, allocation correspondences are robustly implementable.

Appendix A

Proof of Lemma 1. For $\delta > 0$, let $\gamma = \min\{\delta/(2 + 2(\bar{u} - \underline{u})), 1/2\}$. We argue by induction and suppose that $\widetilde{R}^{\mathcal{M}}_{i,k-1}(t_i) \subseteq R^{\mathcal{M},\delta}_{i,k-1}(\tilde{\theta}_i(t_i))$ for $(k-1)$th level rationalizable actions of all types $t_i \in E_i$.

We fix $m_i \in \widetilde{R}^{\mathcal{M}}_{i,k}(t_i)$ and will show that $m_i \in R^{\mathcal{M},\delta}_{i,k}(\tilde{\theta}_i(t_i))$. By the definition, there exists a belief $\tilde{\mu}_i \in \Delta(M_{-i} \times T_{-i})$ that rationalizes message m_i for type t_i:

$$(1)\quad \tilde{\mu}_i\big(\big\{(m_{-i}, t_{-i}) \,\big|\, m_j \in \widetilde{R}^{\mathcal{M}}_{j,k-1}(t_j) \text{ for all } j \neq i\big\}\big) = 1,$$

$$(2)\quad \operatorname{marg}_{T_{-i}} \tilde{\mu}_i = \pi_i(t_i),$$

$$(3)\quad \int \tilde{u}_i\big(g(m_i, m_{-i}), (t_i, t_{-i})\big)\, d\tilde{\mu}_i(m_{-i}, t_{-i}) \geqslant \sup_{m_i'} \int \tilde{u}_i\big(g(m_i', m_{-i}), (t_i, t_{-i})\big)\, d\tilde{\mu}_i(m_{-i}, t_{-i}).$$

We now define a belief $\mu_i \in \Delta(M_{-i} \times \Theta_{-i})$ that rationalizes message m_i for type $\theta_i = \tilde{\theta}_i(t_i)$. Let μ_i be the image of the measure $\tilde{\mu}_i \in \Delta(M_{-i} \times E_{-i})$ (restricted to E_{-i}) under the mapping $\tilde{\theta}_{-i} : E_{-i} \to \Theta_{-i}$

$$\mu_i(X_{-i}) = \frac{\tilde{\mu}_i((m_{-i}, t_{-i}) \colon (m_{-i}, \tilde{\theta}_{-i}(t_{-i})) \in X_{-i})}{\tilde{\mu}_i(M_{-i}, E_{-i})}.$$

Now

$$\mu_i\big(\big\{(m_{-i}, \theta_{-i}) \,\big|\, m_j \in R^{\mathcal{M},\delta}_{j,k-1}(\theta_j) \text{ for all } j \neq i\big\}\big) = 1$$

because $\tilde{\mu}_i(\{(m_{-i}, t_{-i}) \mid m_j \in \widetilde{R}^{\mathcal{M}}_{j,k-1}(t_j) \text{ for all } j \neq i\}) = 1$ by assumption, and $\widetilde{R}^{\mathcal{M}}_{j,k-1}(t_j) \subseteq \widetilde{R}^{\mathcal{M},\delta}_{j,k-1}(\theta_j)$ by the induction hypothesis.

Secondly, for all m_i'

$$\begin{aligned}
&\tilde{\mu}_i(M_{-i}, E_{-i}) \int\limits_{M_{-i}\times\Theta_{-i}} u_i\big(g(m_i, m_{-i}), \big(\tilde{\theta}_i(t_i), \theta_{-i}\big)\big)\, d\mu_i(m_{-i}, \theta_{-i}) \\
&\quad\geqslant \int\limits_{M_{-i}\times E_{-i}} \big(\tilde{u}_i\big(g(m_i, m_{-i}), (t_i, t_{-i})\big) - \gamma\big)\, d\tilde{\mu}_i(m_{-i}, t_{-i}) \\
&\quad\geqslant \int\limits_{M_{-i}\times T_{-i}} \big(\tilde{u}_i\big(g(m_i, m_{-i}), (t_i, t_{-i})\big) - \gamma\big)\, d\tilde{\mu}_i(m_{-i}, t_{-i}) - \gamma\bar{u}
\end{aligned}$$

$$\geqslant \int_{M_{-i}\times T_{-i}} \big(\tilde{u}_i\big(g\big(m_i', m_{-i}\big), (t_i, t_{-i})\big) - \gamma\big)\, d\tilde{\mu}_i(m_{-i}, t_{-i}) - \gamma\bar{u}$$

$$\geqslant \int_{M_{-i}\times E_{-i}} \big(\tilde{u}_i\big(g\big(m_i', m_{-i}\big), (t_i, t_{-i})\big) - \gamma\big)\, d\tilde{\mu}_i(m_{-i}, t_{-i}) - \gamma(\bar{u} - \underline{u})$$

$$\geqslant \tilde{\mu}_i(M_{-i}, E_{-i}) \int_{M_{-i}\times\Theta_{-i}} \big(u_i\big(g\big(m_i', m_{-i}\big), \big(\tilde{\theta}_i(t_i), \theta_{-i}\big)\big) - 2\gamma\big)\, d\mu_i(m_{-i}, \theta_{-i}) - \gamma(\bar{u} - \underline{u}).$$

As $t_i \in E_i$ it must be that $\tilde{\mu}_i(M_{-i}, E_{-i}) = \pi_i(E_i) \geqslant 1-\gamma$, and so m_i is a $\gamma(2+(\bar{u}-\underline{u})/(1-\gamma))$-best reply to the assessment μ_i. This establishes $m_i \in R_{i,k}^{\mathcal{M},\delta}(\tilde{\theta}_i(t_i))$. □

Proof of concluding example. Fix planner's preferences $u_0(x,\theta)$ and an IC direct mechanism $f: T \to [0,1]$; depending on the context we write type profiles as $t = (\theta_R, \theta_C, \pi)$ or $t = (\theta, \pi)$. For fixed beliefs $\pi = (\pi_R, \pi_C)$ the standard IC constraints not to misreport one's payoff type imply the following monotonicity constraints:

$$\sum_{\theta_C} \pi_R(\theta_C)\big(v_R\big(\theta_R', \theta_C\big) - v_R(\theta_R, \theta_C)\big)\big(f\big(\theta_R', \theta_C, \pi\big) - f(\theta_R, \theta_C, \pi)\big) \geqslant 0, \qquad \text{(Mon-R)}$$

$$\sum_{\theta_R} \pi_C(\theta_R)\big(v_C\big(\theta_R, \theta_C'\big) - v_C(\theta_R, \theta_C)\big)\big(f\big(\theta_R, \theta_C', \pi\big) - f(\theta_R, \theta_C, \pi)\big) \geqslant 0. \qquad \text{(Mon-C)}$$

We now show that for all payoff types $\theta, \theta' \in \{0,1\}^2$ there exist beliefs π such that the monotonicity constraints imply $f(\theta,\pi) \leqslant f(\theta',\pi)$. This implies that there exists $y_0 \in [0,1]$ such that for every θ there exist beliefs $\underline{\pi}_\theta$ and $\overline{\pi}_\theta$ with $f(\theta, \underline{\pi}_\theta) \leqslant y_0 \leqslant f(\theta, \overline{\pi}_\theta)$ and thus finishes the proof.

To do so fix $\theta = (u,\ell)$, say. To show $f(u,\ell,\pi) \leqslant f(u,r,\pi)$ and $f(u,\ell,\pi) \leqslant f(d,r,\pi)$, we consider beliefs π with $\pi_C(u) = 1$ so that ℓ is the 'low type' for Colin who values allocation $x = 1$ at -1, and $\pi_R(r) = 1$ so that u is the 'low type' for Rowena. Then, first (Mon-C) implies that

$$\big(v_C(u,\ell) - v_C(u,r)\big)\big(f(u,\ell,\pi) - f(u,r,\pi)\big) = (-1-1)\big(f(u,\ell,\pi) - f(u,r,\pi)\big) \geqslant 0$$

so $f(u,\ell,\pi) \leqslant f(u,r,\pi)$; second (Mon-R) implies that

$$\big(v_R(d,r) - v_R(u,r)\big)\big(f(d,r,\pi) - f(u,r,\pi)\big) = (1+1)\big(f(d,r,\pi) - f(u,r,\pi)\big) \geqslant 0$$

so $f(u,r,\pi) \leqslant f(d,r,\pi)$ and thus $f(u,\ell,\pi) \leqslant f(d,r,\pi)$.

The last case, i.e. showing that $f(u,\ell,\pi') \leqslant f(d,\ell,\pi')$ for some π', needs a different argument: When Rowena knows $\pi_R(\ell) = 1$ her incentive constraints goes into the wrong direction and (Mon-R) implies $f(u,\ell,\pi) \geqslant f(d,\ell,\pi)$. Rather we fix π' with $\pi_R'(r) = 1$ and $\pi_C'(u) = \pi_C'(d) = 1/2$. By the above, we know $f(u,r,\pi) \leqslant f(d,r,\pi)$. By (Mon-C)

$$\begin{aligned} 0 &\leqslant \big(v_C(u,r) - v_C(u,\ell)\big)\big(f(u,r,\pi) - f(u,\ell,\pi)\big)/2 \\ &\quad + \big(v_C(d,r) - v_C(d,\ell)\big)\big(f(d,r,\pi) - f(d,\ell,\pi)\big)/2 \\ &= \big(f(u,r,\pi) - f(u,\ell,\pi)\big) - \big(f(d,r,\pi) - f(d,\ell,\pi)\big) \\ &\leqslant -f(u,\ell,\pi) + f(d,\ell,\pi) \end{aligned}$$

and hence $f(u,\ell,\pi) \leqslant f(d,\ell,\pi)$. One way to understand this argument is that the utility (net of transfers) of Colin's type r is weakly negative when he reports truthfully, because $f(u,r,\pi) \leqslant f(d,r,\pi)$. If $f(u,\ell,\pi) > f(d,\ell,\pi)$ he could get a strictly positive net utility from misreporting his type as ℓ. □

References

[1] P. Battigalli, M. Siniscalchi, Rationalization and incomplete information, Adv. Theor. Econ. 3 (2003), Article 3.
[2] D. Bergemann, S. Morris, Robust mechanism design, Econometrica 73 (2005) 1771–1813.
[3] D. Bergemann, S. Morris, Robust implementation in direct mechanisms, Rev. Econ. Stud. 76 (2009) 1175–1206.
[4] S. Bikhchandani, Ex post implementation in environments with private goods, Theoretical Econ. 1 (2006) 369–393.
[5] T. Börgers, Undominated strategies and coordination in normalform games, Soc. Choice Welfare 8 (1991) 65–78.
[6] K. Chung, J. Ely, Foundations of dominant-strategy mechanisms, Rev. Econ. Stud. 74 (2007) 447–476.
[7] P. Dasgupta, P. Hammond, E. Maskin, The implementation of social choice rules: Some results on incentive compatibility, Rev. Econ. Stud. 46 (1979) 185–216.
[8] P. Dasgupta, E. Maskin, Efficient auctions, Quart. J. Econ. 115 (2000) 341–388.
[9] E. Dekel, D. Fudenberg, S. Morris, Interim correlated rationalizability, Theoretical Econ. 2 (2007) 15–40.
[10] J. Ely, Rationalizability and approximate common knowledge, http://www.kellogg.northwestern.edu/research/math/papers/1324.pdf, 2001.
[11] A. Gibbard, Manipulation of voting schemes, Econometrica 41 (1973) 587–601.
[12] K. Hagerty, W. Rogerson, Robust trading mechanisms, J. Econ. Theory 42 (1987) 94–107.
[13] P. Jehiel, M. Meyer-ter-Vehn, B. Moldovanu, W. Zame, The limits of ex post implementation, Econometrica 74 (2006) 585–610.
[14] J. Ledyard, Incentive compatibility and incomplete information, J. Econ. Theory 18 (1978) 171–189.
[15] J.-F. Mertens, S. Zamir, Formulation of Bayesian analysis for games of incomplete information, Int. J. Game Theory 14 (1985) 1–29.
[16] M. Oury, O. Tercieux, Continuous implementation, IAS Working Paper 90, 2009.
[17] M. Perry, P. Reny, An efficient auction, Econometrica 70 (2002) 1199–1213.
[18] M. Satterthwaite, Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions, J. Econ. Theory 10 (1975) 187–217.
[19] D. Smith, A prior free efficiency comparison of mechanisms for the public good problem, http://www-personal.umich.edu/~dougecon/Doug_Smith_JMP.pdf, 2010.
[20] T. Yamashita, Robust welfare guarantees in bilateral trading mechanisms, http://www.stanford.edu/~takuroy/files/wi.pdf, 2011.