Mechanisms for a No-Regret Agent: Beyond the Common Prior

Modibo K. Camara, Jason D. Hartline, Aleck Johnsen
Department of Economics; Department of Computer Science
Northwestern University
Evanston, IL, United States
mcamara@u.northwestern.edu, hartline@northwestern.edu, aleckjohnsen@u.northwestern.edu

2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS) | 978-1-7281-9621-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/FOCS46700.2020.00033

***Abstract*—A rich class of mechanism design problems can be understood as incomplete-information games between a principal who commits to a policy and an agent who responds, with payoffs determined by an unknown state of the world. Traditionally, these models require strong and often-impractical assumptions about beliefs (a common prior over the state). In this paper, we dispense with the common prior. Instead, we consider a repeated interaction where both the principal and the agent may learn over time from the state history. We reformulate mechanism design as a reinforcement learning problem and develop mechanisms that attain natural benchmarks without any assumptions on the state-generating process. Our results make use of novel behavioral assumptions for the agent – based on *counterfactual internal regret* – that capture the spirit of rationality without relying on beliefs.[1]**

I. Introduction

Mechanism design is a branch of economic theory concerned with the design of social institutions. It encompasses a wide range of social phenomena, such as auctions, matching markets, taxation, contracts, and persuasion. Despite this field's potential, it is often unclear whether and how mechanisms derived from economic theory can be implemented in practice. In particular, one modeling practice stands out as a barrier to implementation: the *common prior* assumption. Many mechanism design problems are only interesting in the presence of uncertainty, and this uncertainty is typically modeled as stochasticity. The *state* of the world is drawn according to some distribution and, importantly, the distribution is commonly known by the designer and all participants in the mechanism.

This paper will dispense with the common prior assumption. In its place, we consider a model of adversarial online learning where the principal and a single agent are learning about the state, over time, using data. The static mechanism design problem is a Stackelberg game of incomplete information. The principal chooses a policy p, the agent chooses a response r, nature chooses a state y, and payoffs are realized. In the online problem, this game is repeated T times, where state y_t is revealed at the end of period t. The sequence of states is arbitrary and the principal's mechanism should perform well without prior knowledge of the sequence. The principal's present choices can affect the agent's future behavior; this makes mechanism design a reinforcement learning problem in our model.

In the absence of distributional assumptions, standard restrictions on the agent's behavior, like Bayesian rationality, become toothless. In its place, we define *counterfactual internal regret* (CIR) and assume that the agent obtains low CIR. This is an ex post definition of rationality that includes Bayesian rationality (with a well-calibrated prior) as a special case. We develop data-driven mechanisms that are guaranteed to perform well under our assumptions. That is, we prove upper bounds on the principal's regret from following our mechanism, relative to the single fixed policy that performs best in hindsight. Our results are reductions from the principal's problem to robust versions of static mechanism design with a common prior.

Running Example: Bayesian persuasion is a model of strategic communication, due to Kamenica and Gentzkow (2011). It has received considerable attention from economists and, more recently, algorithmic game theorists (e.g. Dughmi and Xu 2016, Cummings et al. 2020). It is a useful test case for our framework because (a) it is interesting even with only one agent, (b) the optimal solution varies with the agent's beliefs, and (c) researchers have identified a number of applications.[2]

Our running example is adapted from Kamenica and Gentzkow (2011). A drug company (the principal) seeks approval from a regulator (the agent) for a newly-developed drug. The state $y \in \{\text{High}, \text{Low}\}$ describes the drug's quality. Neither the regulator nor the company know the quality in advance. The company needs to design a clinical trial that will generate (possibly noisy) information about the drug's quality. Roughly, a trial p specifies the probability $p(m, y)$ of sending a message m to the regulator, conditional on the drug quality y. Informally, the message describes the outcome of the trial. After hearing the message, the regulator decides whether to approve the drug. The regulator receives a payoff if it approves a high-quality drug or rejects a low-quality drug. The company receives a payoff if the regulator approves, regardless of quality. Its challenge is to design a

[1]For the full version of this paper, see https://arxiv.org/abs/2009.05518.

[2]Bayesian persuasion has been used to study a wide range of topics, including recommendation systems (Mansour et al. 2016), traffic congestion (Das et al. 2017), congested social services (Anunrojwong et al. 2020), and financial stress-testing (Goldstein and Leitner 2018).

2575-8454/20/$31.00 ©2020 IEEE
DOI 10.1109/FOCS46700.2020.00033

clinical trial that convinces the regulator to approve as many drugs as possible.

To predict behavior in incomplete-information games, we need to make assumptions about how the agents deal with uncertainty. The common prior is one such assumption. In our running example, the common prior would specify a probability $q \in [0, 1]$ that the drug is high quality. Consider the case $q = 1/3$. If the company does not run a trial – it recommends "approve" in every state – the regulator would never approve, as the drug is more likely to be low quality than high quality ex ante. If the company runs the most thorough trial possible – it recommends "approve' if and only if the drug is high quality – the regulator would approve with probability $1/3$. Finally, consider the optimal trial. The optimal trial always recommends "approve" if the drug is high quality. If the drug is low quality, it recommends "approve" and "reject" with equal probability. After hearing "approve", the regulator's posterior puts equal weight on both states, and so it might as well approve. Here, the regulator approves with probability $2/3$.

Online Mechanism Design: In our model, both the company and the regulator would be learning about drug quality over time. New drugs arrive sequentially. For each drug, the company designs a clinical trial and generates a message. The regulator hears the message and decides whether to approve. Regardless of whether the drug is approved, both parties eventually learn the drug's true quality, and the next drug arrives. The company's strategy, called a *mechanism*, maps the drug (i.e. state) history and the approval decision (i.e. response) history to a trial for the current drug. The regulator's strategy, called a learning algorithm or *learner*, maps the drug quality history and the trial (i.e. policy) history to an approval decision for the current drug. This model is *online* because the company and regulator must make decisions while the drugs are still arriving. It is *adversarial* in the sense that we impose no assumptions on the sequence of drugs, and so any results (e.g. claiming that a mechanism performs well) must hold for all such sequences.

The company's problem is to develop a mechanism that performs as well as the best-in-hindsight trial. That is, the company should not regret following its mechanism relative to any alternative where it picks the same trial p in every period. To evaluate what would have happened under an alternative sequence of trials, the company must take into account how the regulator's behavior would have changed. So, the company faces a reinforcement learning problem and its benchmark corresponds to the notion of *policy regret* in the literature on bandit learning with adaptive adversaries. In that setting, Arora, Dekel, et al. (2012) show that guaranteeing sublinear (policy) regret is generally impossible. This precludes a simple solution to the company's problem; we must constrain the regulator's behavior.[3]

No-Regret Agents: The standard way to constrain the regulator, or agent's behavior – i.e. to capture "self-interest" in the absence of a meaningful notion of ex ante optimality – is to impose upper bounds on the agent's regret. We build on existing no-regret assumptions, but also highlight their limitations.

Two notions of regret have been used historically: external and internal (or swap) regret (ER and IR). For example, Nekipelov et al. (2015) show how ER bounds combined with bidding data can be used to partially identify bidder valuations in a dynamic auction. Braverman et al. (2018) consider a dynamic pricing problem against no-ER agents.[4] Their analysis is generalized by Deng et al. (2019), who study repeated Stackelberg games of complete information. Furthermore, the literature on no-regret learning in games has established that if agents satisfy a no-ER (resp. no-IR) property in a repeated game, the empirical distribution of their actions will converge to a coarse correlated equilibrium (resp. correlated equilibrium) (Blum, Hajiaghayi, et al. 2008; Foster and Vohra 1997; Hart and Mas-Colell 2001; Hartline, Syrgkanis, et al. 2015).

Both ER and IR can be thought of as "non-policy" regret, because they do not take into account how the agent's behavior affects the behavior of others. The justification for these regret bounds is that (a) they are satisfied by well-known learning algorithms, and (b) they generalize optimality conditions associated with a stationary equilibrium. Nonetheless, these regret bounds can be problematic. Effectively, they assume that agents are (a) sophisticated enough to obtain low non-policy regret, but (b) not aware that their true objective is policy regret. Keep in mind that an agent who minimizes policy regret can easily obtain high non-policy regret, and thereby violate the regret bounds.

To avoid this problem, the principal can commit to a mechanism that is *nonresponsive* to the agent's behavior: the policy p_t depends on the state history but not on the agent's response history. When mechanisms are nonresponsive, non-policy regret and policy regret coincide for the agent. Then, bounds on the agent's regret are permissive assumptions that allow a wide range of sophisticated and self-interested behavior, including Bayesian rationality. Keep in mind, there is no need to resort to responsiveness if nonresponsive

[3] Similarly, Arora, Dinitz, et al. (2018) consider policy regret in a repeated game and use the self-interest of the adaptive adversary to motivate behavioral restrictions. They identify a class of stable no-ER algorithms such that, if all participants use an algorithm in this class, all participants obtain low policy regret. Our approach differs in that we do not make assumptions on the algorithm that the agent uses, other than bounding the agent's regret on the realized state sequence.

[4] In their model, the agent is learning an appropriate response to the principal's pricing strategy. If the agents use naive mean-based learners, Braverman et al. (2018) provide a mechanism that extracts the full surplus. Our setting differs in that the agent faces uncertainty about a state of the world, rather than about the mechanism.

mechanisms tightly bound the principal's regret.[5]

Counterfactual Internal Regret: Our no-regret assumption is motivated by the following observation. Even if the agent satisfies no-ER or no-IR, an early mistake by the principal can result in a permanent, undesirable shift in the agent's behavior. This can occur when the agent behaves as if she has additional information about the state that is not explicitly provided by the model. The agent can make the principal's problem infeasible if she exploits her information selectively, i.e. based on the principal's chosen policies.

Our notion of rationality requires the agent to fully and consistently exploit her information, regardless of the principal's chosen policies. Existing benchmarks like external and internal regret cannot capture this requirement. To see why, it helps to consider the fable of the tortoise and the hare. Both animals have an hour to traverse a one-mile track. For the tortoise, this requirement is feasible and binding: finishing in time means hustling, without substantial breaks or detours. For the hare, however, the requirement is hardly restrictive: it may stop for a break, walk rather than run, or even run around in circles while still finishing the race in time. Benchmarks like external or internal regret imply reasonable behavior for an uninformed agent (i.e. the tortoise). But for an informed agent (i.e. the hare), these benchmarks are easy enough to satisfy that it may engage in all kinds of frivolous behavior – possibly to the detriment of the principal.

The solution to our analogy is to strengthen the hare's benchmark. If the hare has to traverse the track in three minutes, it needs to hustle, like the tortoise. Similarly, if the agent has to obtain no-regret with her information as additional context, this would preclude the kind of frivolous behavior that makes the principal's problem infeasible. Of course, setting this benchmark requires us to know the nature and quality of the agent's information, just as we needed to know the top speed of the hare. The idea behind counterfactual internal regret is that we can identify the agent's information with her past behavior under counterfactual mechanisms. Intuitively, any information that is useful should eventually reveal itself through variation in behavior.

Main Results: This paper considers three variations on our model: one where the principal knows the agent's information, one where the agent has no private information, and one where the agent may have private information. In each case, we propose a mechanism and bound on the principal's regret in terms of the agent's counterfactual internal regret (CIR).

Our first mechanism is intended as a warmup. It requires oracle access to the agent's information and has poor performance in finite samples, but avoids some complications associated with information asymmetry between the principal and agent. First, the mechanism produces a calibrated forecast of the state in the current period using off-the-shelf algorithms, using the oracle as additional context for the forecast. The forecast miscalibration error is

$$F_T = \tilde{O}\left(T^{-1/4}\delta^{1-n_{\mathcal{Y}}} n_{\mathcal{Y}} n_{\mathcal{R}}^{2n_{\mathcal{P}}}\right)$$

Then, it chooses the worst-case optimal policy in a (hypothetical) ϵ-robust version of the common prior game. In that game, the agent's response only needs to be ϵ-approximately optimal, and the mechanism substitutes its forecast for the prior.

Theorem 1 bounds the principal's regret under this mechanism, under some restrictions on the stage game. Suppose there are $n_{\mathcal{Y}}$ states, $n_{\mathcal{P}}$ policies, and $n_{\mathcal{R}}$ responses. Fix a parameter $\epsilon > 0$ (controlling robustness) and $\delta > 0$ (controlling the fineness of a grid). Our bound is

$$O(\epsilon) + \frac{1}{\epsilon} O(\mathrm{CIR}) + O(F_T) + O(\delta^{1/2})$$

If the agent satisfies no-CIR, i.e. $\mathrm{CIR} \to 0$ as $T \to \infty$, then the principal's regret vanishes in T as long as $\epsilon, \delta \to 0$ at the appropriate rates. Moreover, the principal's average payoffs converge to a natural benchmark: what he would have obtained in a stationary equilibrium of the repeated game with a common prior (the empirical distribution conditioned on agent's information).

Our second mechanism applies when the agent is as uninformed as the principal. It is identical to the first, except its forecast does not use information revealed by the learner, and so the forecast miscalibration error is

$$G_T = \tilde{O}\left(T^{-1/4}\delta^{1-n_{\mathcal{Y}}} n_{\mathcal{Y}}\right)$$

We formalize "uninformedness" by assuming that the agent's external regret is non-negative (in conjunction with no-CIR). Theorem 2 bounds the principal's regret under this mechanism, under some additional restrictions on the stage game. Our bound is

$$O(\epsilon) + \frac{1}{\epsilon} O(\mathrm{CIR}) + O(G_T) + O(\delta^{1/2})$$

Our third mechanism applies even when the agent is more informed than the principal. Here, we consider an "informationally robust" version of the stage game, due to Bergemann and Morris (2013), where the agent receives a private signal from an unknown information structure. Like before, we formulate an ϵ-robust version of this game, where the agent's response need only be ϵ-approximately optimal. Our mechanism is identical to the second mechanism, except that it chooses the worst-case optimal policy in the ϵ-informationally-robust game instead of the ϵ-robust game.

Theorem 3 bounds the principal's regret under this mechanism, under some restrictions on the stage game. Let $\hat{\pi}_T$

[5]This approach seems philosophically similar to that of Immorlica et al. (2020), who develop mechanisms that incentivize efficient social learning. By restricting attention to simple disclosures (i.e. unbiased subhistories), they significantly simplify the agents' inferential problem and can motivate a permissive notion of frequentist rationality. Having restricted disclosure in this manner, they nonetheless design mechanisms with optimal rates of convergence.

denote the empirical distribution of states $y_{1:T}$. Given a common prior π, let $\nabla(\pi)$ be the difference between the principal's maxmin payoff and his maxmax payoff across all possible information structures. Roughly, our bound is

$$\nabla(\hat{\pi}_T) + O(\epsilon) + \frac{1}{\epsilon} O(\text{CIR}) + O(G_T) + O(\delta^{1/2})$$

Here, the principal's regret does not vanish as $T \to \infty$. However, it is vanishing up to the cost of informational robustness $\nabla(\hat{\pi}_T)$.

Finally, although our focus is not on computational complexity, note that the computational tractability of our mechanisms will depend critically on our ability to solve robust mechanism design problems under a common prior. So, while our bounds on the principal's regret apply to a large class of games, evaluating tractability may require a case-by-case analysis.

Additional Related Work: Within computer science, many researchers share our goal of replacing prior knowledge in mechanism design with data. For example, a number of papers have applied online learning to auction design (e.g. Blum and Hartline 2005; Blum, Kumar, et al. 2004; Daskalakis and Syrgkanis 2016; Dudík et al. 2017; Kleinberg and Leighton 2003) and Stackelberg security games (e.g. Balcan, Blum, Haghtalab, et al. 2015). Here, agents are either short-lived or myopic, whereas our agent is long-lived and forward-looking. In addition, the literature on sample complexity in mechanism design allows the principal to learn the state distribution from i.i.d. samples (e.g. Balcan, Blum, Hartline, et al. 2008; Cole and Roughgarden 2014; Morgenstern and Roughgarden 2015; Syrgkanis 2017). Here, the data arrives as a batch rather than online, there is no repeated interaction and the question of responsiveness does not arise.

These papers can avoid the agent's learning problem because they emphasize applications where the agent does not face uncertainty, or where truthfulness is a dominant strategy. In contrast, Cummings et al. (2020) and Immorlica et al. (2020) study problems that are closer to our own, insofar as both the principal and the agent must learn from data. They impose behavioral assumptions that are suited for i.i.d. data, whereas our behavioral assumptions apply to arbitrary data-generating processes.

Within economics, research has focused on relaxing prior knowledge, rather than replacing it entirely. Part of the literature on robust mechanism design relaxes the common prior to some kind of approximate agreement on the distribution (e.g. Artemov et al. 2013; Jehiel et al. 2012; Meyer-ter-Vehn and Morris 2011; Ollár and Penta 2017; Oury and Tercieux 2012). Our approach will suggest ϵ-robustness and ϵ-informational-robustness as alternatives to "approximate agreement".

Organization: Section II introduces the stage game and ϵ-robustness. Section III introduces the repeated game. Section IV defines counterfactual internal regret. Section V presents our results when the principal is informed. Section VI introduces the stage game with private signals. Section VII presents our results for an uninformed agent. Section VIII presents our results when the agent may be more informed than the principal. Section IX concludes. In addition, the full version of this paper includes more details and proofs, including examples and a discussion on the complexity of the agent's no-CIR learning problem.

II. Stage Game

Our model features three participants: a male principal, a female agent, and nature. As advertised, we are interested in a repeated interaction between these participants. To begin with, however, we describe the stage game, which will constitute a single-round of the repeated game. In the stage game, the principal moves first and commits to a policy $p \in \mathcal{P}$. Next, the agent observes the policy p and then chooses a response $r \in \mathcal{R}$. Utility functions depend on the response r, the policy p, and an unknown state of the world $y \in \mathcal{Y}$, chosen by nature. Formally, the agent's utility function is $U : \mathcal{R} \times \mathcal{P} \times \mathcal{Y} \to [0, 1]$ while the principal's utility function is $V : \mathcal{R} \times \mathcal{P} \times \mathcal{Y} \to [0, 1]$.

The state space $\mathcal{Y}$ is finite, with $n_{\mathcal{Y}}$ elements. In this paper, we will also treat $\mathcal{P}$ and $\mathcal{R}$ as finite, with $n_{\mathcal{P}}$ and $n_{\mathcal{R}}$ elements respectively. This is done for ease of exposition but not required for our results.

The stage game plays an important role in our analysis. Two of our results (theorems 1 and 2) are best understood as reducing the online mechanism design problem to the simpler task of finding a "locally-robust" policy in the stage game. In the locally-robust problem, we maintain the traditional common prior assumption: that is, the state y is drawn from a commonly known distribution π. However, we relax the assumption that the agent maximizes her expected utility $\mathrm{E}_{y\sim\pi}[U(r, p, y)]$. Instead, she chooses a response that guarantees her an expected utility within an additive constant ϵ of the optimum. Let $B(\pi, \epsilon)$ be the set of response distributions μ consistent with this assumption, i.e. where

$$\epsilon \geq \max_{\tilde{r}\in\mathcal{R}} \mathrm{E}_{y\sim\pi}[U(\tilde{r}, p, y)] - \mathrm{E}_{y\sim\pi}[\mathrm{E}_{r\sim\mu}[U(r, p, y)]]$$

Since this assumption only partially identifies the agent's behavior, the principal's utility can take on a range of values. The principal's worst-case utility from following policy p is described by the function

$$\alpha_p(\pi, \epsilon) = \min_{\mu\in B(\pi,\epsilon)} \mathrm{E}_{y\sim\pi}[\mathrm{E}_{r\sim\mu}[V(r, p, y)]]$$

and his best-case utility is described by

$$\beta_p(\pi, \epsilon) = \max_{\mu\in B(\pi,\epsilon)} \mathrm{E}_{y\sim\pi}[\mathrm{E}_{r\sim\mu}[V(r, p, y)]]$$

Definition 1 (ϵ-Robustness)**.** *The* ϵ-robust *policy is worst-case optimal over all response distributions* μ *that achieve at*

least the agent's optimal expected utility minus ϵ. Formally, it is

$$p^*(\pi, \epsilon) \in \arg\max_{p \in \mathcal{P}} \alpha_p(\pi, \epsilon)$$

Definition 2 (Cost of ϵ-Robustness)**.** *Fix a distribution π and parameter $\epsilon > 0$. The cost of ϵ-robustness is the distance between the principal's best-case utility (under the best-case optimal policy) and worst-case utility (under the worst-case optimal policy). Formally,*

$$\Delta(\pi, \epsilon) = \max_{p \in \mathcal{P}} \beta_p(\pi, \epsilon) - \alpha_{p^*(\pi,\epsilon)}(\pi, \epsilon)$$

The cost of ϵ-robustness will be a key variable in our upper bounds on the principal's regret in the repeated game. It will be convenient (but not necessary) to assume that this cost is growing at most linearly in ϵ.

Assumption 1. $\forall \pi \in \Delta(\mathcal{Y})$, $\Delta(\pi, \epsilon) = O(\epsilon)$.

Finally, the following lemma will be important to our results. Suppose that the principal misjudges the agent. Instead of choosing a response that achieves at least her optimal expected utility minus ϵ, the agent only achieves her optimal expected utility minus $\epsilon + \tilde{\epsilon}$, for $\tilde{\epsilon} > 0$. Nonetheless, if the principal uses the ϵ-robust policy, his utility degrades smoothly in the residual $\tilde{\epsilon}$.

Lemma 1. *Fix distribution $\pi \in \Delta(\mathcal{Y})$, policy $p \in \mathcal{P}$, and constants $\epsilon, \tilde{\epsilon} > 0$. The principal's worst-case utility satisfies*

$$\alpha_p(\pi, \epsilon + \tilde{\epsilon}) \geq \alpha_p(\pi, \epsilon) - \frac{\tilde{\epsilon}}{\epsilon}$$

and his best-case utility satsifies

$$\beta_p(\pi, \epsilon + \tilde{\epsilon}) \leq \beta_p(\pi, \epsilon) + \frac{\tilde{\epsilon}}{\epsilon}$$

In the full version, we consider two special cases of our model: Bayesian persuasion and contract design. For each case, we provide a simple example in which we verify our assumptions and evaluate our results.

III. Repeated Game

In the repeated game, the stage game is repeated T times. In period t, the principal chooses policy p_t, the agent chooses response r_t, and nature chooses the state y_t. At the end of period t, the state y_t is revealed to both the principal and the agent. The agent's repeated game strategy (henceforth, *learner* L) maps the state history $y_{1:t-1}$, the response history $r_{1:t-1}$, the policy history $p_{1:t-1}$, and the current policy p_t to a distribution μ_t over responses. Formally, the response distribution is given by

$$L_t : \mathcal{Y}^{t-1} \times \mathcal{R}^{t-1} \times \mathcal{P}^t \to \Delta(\mathcal{R})$$

The principal's repeated game strategy (henceforth, *mechanism* σ) maps the state history $y_{1:t-1}$, the response history $r_{1:t-1}$, and the policy history $p_{1:t-1}$ to a distribution ν_t over policies. Formally, the policy distribution is given by

$$\sigma_t : \mathcal{Y}^{t-1} \times \mathcal{R}^{t-1} \times \mathcal{P}^{t-1} \to \Delta(\mathcal{P})$$

Our goal is to design a mechanism σ^* that the principal would not regret using, relative to a finite set of alternative mechanisms. Regret – which we define momentarily – measures the gap in performance between σ^* and the alternative mechanism σ that performed best in hindsight, given the realized sequence of states $y_{1:T}$. We consider a simple set of alternative mechanisms, corresponding to the *constant mechanisms* σ^p that select the same policy

$$\sigma^p_t(y_{1:t-1}, r_{1:t-1}, p_{1:t-1}) = p$$

in all periods t and for all histories.

To define the principal's regret, we need notation for the agent's behavior under both the mechanism σ^* and any constant mechanism σ^p. Fix the state sequence $y_{1:T}$. Let μ^*_t describe the agent's behavior under σ^*, i.e.

$$\mu^*_t = L_t\left(y_{1:t-1}, r^*_{1:t-1}, p^*_{1:t}\right)$$

given the realized history of responses $r^*_{1:t-1}$ and policies $p^*_{1:t}$ under σ^*. Similarly, let μ^p_t describe the agent's behavior under σ^p.

Definition 3 (Principal's Regret)**.** *The* principal's regret $\mathrm{PR}(L, y_{1:T})$ *relative to the best-in-hindsight σ^p is*

$$\max_{p \in \mathcal{P}_0} \frac{1}{T} \sum_{t=1}^{T} \left(\mathrm{E}_{r \sim \mu^p_t}[V(r, p, y_t)] - \mathrm{E}_{r \sim \mu^*_t}[V(r, p^*_t, y_t)]\right)$$

The mechanism σ^* satisfies no-regret if the principal's regret is $o(1)$, i.e. it vanishes as $T \to \infty$. Recall that the no-regret mechanism design problem is infeasible without further assumptions on the learner L.

IV. Behavioral Assumptions

In this section, we develop a restriction on the learner L that captures "rational" behavior by the agent, without requiring assumptions on the state sequence $y_{1:T}$. In particular, we build on no-regret assumptions pioneered in the literature on learning in games.

In online learning, regret measures how much better or worse off the agent would have been had she followed the best-in-hindsight "simple" strategy instead of her learner. Different notions of regret correspond to different definitions of simplicity. All of the regret notions used in this paper will be special cases of *contextual regret*, defined as follows. Given a sequence $z_{1:T}$ of variables in some arbitrary set $\mathcal{Z}$, contextual regret considers a strategy "simple" if, for any two periods t and τ, sharing the same context $z_t = z_\tau$ implies taking the same response $r_t \neq r_\tau$.

Definition 4. *Given sequence $z_{1:T}$ of covariates, the agent's* contextual regret $\mathrm{CR}(p_{1:T}, y_{1:T})$ *relative to the best-in-hindsight modification rule $h : \mathcal{Z} \to \mathcal{R}$ is*

$$\max_h \frac{1}{T} \sum_{t=1}^{T} (U(h(z_t), p_t, y_t) - U(r_t, p_t, y_t))$$

Note that, unlike our definition of the principal's regret, the agent's contextual regret does not take into account how changes in her past behavior would have also affected the principal's behavior. This omission is justified when the mechanism is nonresponsive.

Definition 5. *The mechanism σ is* nonresponsive *if*

$$\sigma_t(y_{1:t-1}, r_{1:t-1}, p_{1:t-1}) = \sigma_t(y_{1:t-1}, \tilde{r}_{1:t-1}, p_{1:t-1})$$

for any state history $y_{1:t-1}$, policy history $p_{1:t-1}$, and response histories $r_{1:t-1}, \tilde{r}_{1:t-1}$.

Our mechanisms will be nonresponsive. This is a design choice, not an assumption. As discussed in the introduction, restricting attention to nonresponsive mechanisms simplifies the agent's problem and makes our behavioral assumptions more credible. As it turns out, there exist nonresponsive no-regret mechanisms in two of the scenarios we study, so there is no need for responsive mechanisms in these settings.

In the rest of this section, we define three special cases of contextual regret: *external regret* (ER), *internal regret* (IR), and *counterfactual internal regret* (CIR).

A. External Regret

In our model, external regret is contextual regret where the policy p_t is the context in period t.[6]

Definition 6. *The* external regret $\mathrm{ER}(p_{1:T}, y_{1:T})$ *relative to the best-in-hindsight modification rule $h : \mathcal{P} \to \mathcal{R}$ is*

$$\max_h \frac{1}{T} \sum_{t=1}^{T} (U(h(p_t), p_t, y_t) - U(r_t, p_t, y_t))$$

Although common in the literature, no-ER assumptions are insufficient for our problem. They do not circumvent the infeasibility of no-regret mechanism design that motivated us to restrict the agent's behavior in the first place. In particular, this is because they fail to rule out certain pathological behaviors. Because these pathological behaviors are clearly not in the agent's best interest, we also conclude that no-ER fails to rule out "irrational" behavior and is therefore not a good definition of "rationality". The following proposition (and its proof) clarifies the issue.

[6]In other words, we compare the agent's performance to her best-in-hindsight strategy in the stage game (which is technically a function $\mathcal{P} \to \mathcal{R}$). If we instead compared the agent's performance to the best-in-hindsight response $r \in \mathcal{R}$, this would confound variation in policies with variation in the state. Our approach is similar to that of Hartline, Johnsen, et al. (2019), where agents best respond to an allocation rule given the empirical value distribution, rather than naively best respond to the empirical bid distribution.

Proposition 1. *In our running example, for every mechanism σ^*, there exists a learner L that guarantees no-ER on all histories and a state sequence $y_{1:\infty}$ such that the principal's regret does not vanish.*

B. Counterfactual Internal Regret

Before defining CIR, we provide a brief intuition: what went wrong with external regret? Recall the tortoise and hare analogy in the introduction. For a behavioral assumption to rule out pathological behaviors, it may have to adapt to the information of the agent (or the speed of the animal).

What do we mean by information? Implicit in most stochastic models is the idea that the state is fundamentally unpredictable. But there is no ex ante sense in which the deterministic sequence $y_{1:T}$ is predictable or not. In particular, the agent may behave as if she possesses "private information" about the sequence of states that goes beyond the "public information" inherent in the description of the model. In practice, the agent may have access to data that the principal lacks, notice a pattern that did not occur to the principal, or succeed through dumb luck. Formally, this reflects an adversary who simultaneously chooses the state sequence $y_{1:T}$ and the learner L to cause our mechanism to underperform. In particular, even though the agent cannot observe y_t when choosing response r_t, this does not prevent the adversary from "correlating" r_t and y_t.

No-CIR requires the agent to consistently and fully exploit her private information. In the spirit of revealed preference, private information is identified with her behavior across counterfactual mechanisms. Intuitively, if the agent is able to distinguish between periods t, τ and finds it useful to do so, then her behavior should also differ between those two periods. If her behavior under one mechanism reveals private information, this information should also be accessible to her under a different mechanism. This logic allows us to define a purely ex post notion of rationality that does not refer to beliefs or distributions over state sequences.

No-CIR refines no-IR, a weaker condition that was developed in the literature on calibration (e.g. Foster and Vohra 1997). Internal regret is contextual regret where the context is the agent's own behavior $r_{1:T}$.

Definition 7. *The agent's* internal regret $\mathrm{IR}(p_{1:T}, y_{1:T})$ *relative to the best-in-hindsight modification rule $h : \mathcal{P} \times \mathcal{R} \to \mathcal{R}$ is*

$$\max_h \frac{1}{T} \sum_{t=1}^{T} (U(h(p_t, r_t), p_t, y_t) - U(r_t, p_t, y_t))$$

Counterfactual internal regret is contextual regret where the context is the concatenation of: the policy p_t^* under the proposed mechanism σ^*; the agent's behavior $r_{1:T}^*$ under σ^*; and her counterfactual behavior $r_{1:T}^p$ under the constant mechanisms σ^p.

Definition 8. *Let the* information partition *be*

$$\mathcal{I} = \mathcal{P} \times \mathcal{R} \times (\mathcal{R})^{n_{\mathcal{P}}}$$

and let the information I_t *in period* t *be*

$$I_t = (p_t^*, r_t^*, (r_t^p)_{p \in \mathcal{P}})$$

Crucially, the same information I_t is available to the agent regardless of whether the principal follows our mechanism σ^* or deviates to a constant mechanism σ^p. Intuitively, the principal's choice of mechanism should not affect what information the agent has available.

Definition 9. *The agent's* counterfactual internal regret $\mathrm{CIR}(p_{1:T}, y_{1:T})$ *relative to the best-in-hindsight modification rule* $h : \mathcal{I} \to \mathcal{R}$ *is*

$$\max_h \frac{1}{T} \sum_{t=1}^{T} (U(h(I_t), p_t, y_t) - U(r_t, p_t, y_t))$$

The discussion in the proof of proposition 1 clarifies how no-CIR rules out the kinds of pathological or irrational behavior that no-ER fails to rule out.

V. Mechanism for a Known Learner

Our first result should be viewed as pedagogical. It bounds the principal's regret under a mechanism that requires oracle access to the agent's learner. This requirement is unrealistic and will be removed in sections V and VI. Likewise, the bound itself will feature an exponential dependence on the size of the policy space. This dependence will also be removed in later sections.

Definition 10. *The* information oracle $\Omega_t : \mathcal{P} \to \mathcal{I}$ *specifies the information* I_t *that the learner* L *would generate in period* t *given any policy* $p_t \in \mathcal{P}$.

This case is a convenient starting point because it avoids the bulk of the information asymmetries between the principal and the agent that our later results need to address. That follows from the fact that any private information generated by the learner can be anticipated by the principal with access to the information oracle. This case is also a convenient point of departure from the common prior assumption because it permits a wider range of agent behavior without relaxing the principal's knowledge of said behavior.[7]

Definition 11. *The forecasting algorithm* Forecast *applies a generic no-internal-regret algorithm due to Blum and Mansour (2007) in an auxilliary learning problem where the action space consists of discretized forecasts* $\pi \in \mathcal{C}_{\Delta(\mathcal{Y})}$ *and the loss function is the negated quadratic scoring rule* S. *In each period, the algorithm makes a prediction* π_t *and incurs loss* $-S(\pi_t, y_t)$.

Mechanism 1. *Let the distribution* π_t *be a forecast of the state* y_t *given by the* Forecast *algorithm, using the agent's information as additional context. Formally, the context is the vector of outputs* $\Omega(p)$ *of the information oracle under policies* $p \in \mathcal{P}$. *Now, fix a parameter* $\bar{\epsilon} > 0$. *In period* t, *the* informed-principal mechanism σ^* *chooses the* $\bar{\epsilon}$*-robust policy* $p^*(\pi_t, \bar{\epsilon})$ *that treats the forecast* π_t *as a common prior.*

Before stating the theorem in full, we present the reasoning behind the result and clarify the components of the regret bound, as well as the assumptions required. Let "$t \in I$" indicate that $I_t = I$. Let $n_I = \sum_{t=1}^{T} \mathbf{1}(t \in I)$ be the number of periods with information I. Let $\hat{\pi}_I$ be the empirical distribution conditioned on information I, i.e.

$$\hat{\pi}_I(y) = \frac{1}{n_I} \sum_{t \in I} \mathbf{1}(y_t = y)$$

We begin with a straightforward but important observation: across all periods $t \in I$, the responses $r_t^* = r_I^*$ and policies $p_t^* = p_I^*$ are constant, as are her counterfactual responses $r_t^p = r_I^p$ under the constant mechanisms σ^p. As a result, the principal's average utility across context I takes on a familiar form:

$$\frac{1}{n_I} \sum_{t \in I} V(r_I, p_I, y_t) = \mathrm{E}_{y \sim \hat{\pi}_I}[V(r_I, p_I, y)]$$

Similarly, the agent's average utility is

$$\frac{1}{n_I} \sum_{t \in I} U(r_I, p_I, y_t) = \mathrm{E}_{y \sim \hat{\pi}_I}[U(r_I, p_I, y)]$$

Within each context I, we have recreated the stage game with common prior $\hat{\pi}_I$. The agent accumulates regret

$$\epsilon_I = \max_{\tilde{r}} \mathrm{E}_{y \sim \hat{\pi}_I}[U(\tilde{r}, p, y)] - \mathrm{E}_{y \sim \hat{\pi}_I}[U(r_I, p, y)]$$

Under mechanism 1, the principal chooses the $\bar{\epsilon}$-robust policy for the forecast π_t. Suppose for the moment that the forecast is constant across all periods $t \in I$, i.e. $\pi_t = \pi_I$. Since the forecast is well-calibrated by design and uses information I_t as context, π_I cannot be too far in the l_1 distance from $\hat{\pi}_I$. Therefore, the $\bar{\epsilon}$-robust policy for π_I is nearly $\bar{\epsilon}$-robust for $\hat{\pi}_I$.

At this point, the principal has (roughly) applied the $\bar{\epsilon}$-robust policy for the empirical distribution $\hat{\pi}_I$, to an agent that obtains regret ϵ_I. In that sense, the principal has misjudged the agent's capacity to make mistakes. However, recall lemma 1: this affects the principal's best-case and worst-case utilities by at most $\epsilon_I / \bar{\epsilon}$. It follows that, roughly-speaking, the principal's utility is not much worse than the worst-case optimal utility, i.e.

$$\mathrm{E}_{y \sim \hat{\pi}_I}[V(r_I, p_I, y)] \geq \max_{\tilde{p}} \alpha_{\tilde{p}}(\hat{\pi}_I, \bar{\epsilon}) - \frac{\epsilon_I}{\bar{\epsilon}}$$

[7] Under a common prior, the principal knows the agent's prior and therefore has precise knowledge of the agent's learner. In addition, since the agent is Bayesian, the agent does not find it beneficial to randomize and her learner will typically be deterministic. Essentially, the common prior provides an information oracle for free.

At the same time, it cannot be much better than the best-case optimal utility. More precisely,

$$\mathrm{E}_{y\sim\hat{\pi}_I}[V(r_I, p_I, y)] \leq \max_{\tilde{p}} \beta_{\tilde{p}}(\hat{\pi}_I, \bar{\epsilon}) + \frac{\epsilon_I}{\bar{\epsilon}} \qquad (1)$$

By assumption 1, the difference between the upper bound and the lower bound is $O(\bar{\epsilon}) + O(\epsilon_I/\bar{\epsilon})$. This pins down the principal's utility under mechanism 1. Moreover, the upper bound (1) also applies to the principal's payoffs under any constant mechanism σ^p. So, the regret accumulated by the principal in context I is also at most

$$O(\bar{\epsilon}) + O(\epsilon_I/\bar{\epsilon})$$

This brings us to our key assumption: the agent's CIR is at most some constant ϵ. Note that we do not require CIR to be bounded on all state sequences, only the realized $y_{1:T}$. For instance, a Bayesian agent will obtain low CIR as long as her beliefs are well-calibrated.

Assumption 2 (Bounded CIR)**.** *Let $y_{1:T}$ be the realized state sequence and let $p^*_{1:T}$ be the policy sequence generated by the proposed mechanism σ^*. There exists a constant $\epsilon \geq 0$ such that $\epsilon \geq \mathrm{CIR}(y_{1:T}, p^*_{1:T})$ and $\epsilon \geq \mathrm{CIR}(y_{1:T}, (p, \ldots, p))$, $\forall p \in \mathcal{P}$.*

Since CIR is contextual regret with information I_t as context, bounded CIR ensures that $\epsilon \geq \frac{1}{T}\sum_{I\in\mathcal{I}} n_I \epsilon_I$. Combine this with our bound on the agent's regret ϵ_I in the context of information I, and it follows that the principal's regret is at most

$$O(\bar{\epsilon}) + O(\epsilon/\bar{\epsilon})$$

To transform this intuition into a result, we need to address an assumption made along the way: that the forecast π_t is constant across all periods $t \in I$. This is not necessarily true. The adversary can choose a sequence of states $y_{1:T}$ that makes the principal appear more informed than the agent. Indeed, variation in forecasts can be interpreted as private information of the principal, even if it is spurious. On the other hand, any variation in π_t that affects the policy p_t will also be included in the agent's information I_t. What remains is variation in π_t that does not affect the policy – information that is useless to the principal, but not necessarily useless to the agent. If the principal expects the agent to exploit this information and the agent does not, this can lead to a suboptimal policy choice.

The following assumption restricts attention to stage games where this problem does not arise; that is, the agent's failure to exploit information that is useless to the principal does not affect the principal's utility. Alternatively, we can avoid this restriction by allowing the agent to use the principal's forecast π_t as additional context.

Assumption 3. *Let $\epsilon > 0$. Let π and $\tilde{\pi}$ be distributions in the stage game. If the ϵ-robust policies under π and under $\tilde{\pi}$ are the same, then they are also equal to the ϵ-robust policy under any convex combination $\breve{\pi} = \lambda\pi + (1-\lambda)\tilde{\pi}$ of these distributions. That is,*

$$p^*(\pi, \epsilon) = p^*(\tilde{\pi}, \epsilon) \implies p^*(\pi, \epsilon) = p^*(\breve{\pi}, \epsilon)$$

Theorem 1. *Assume restrictions on the stage game (assumptions 1, 3), and ϵ-bounded CIR (assumption 2). Let σ^* be the mechanism 1. Given access to the information oracle, for any constants $\bar{\epsilon}, \delta > 0$, the principal's expected regret $\mathrm{E}_{\sigma^*}[\mathrm{PR}(L, y_{1:T})]$ is at most*

$$O(\bar{\epsilon}) + \frac{1}{\bar{\epsilon}} \cdot \tilde{O}\left(\epsilon + T^{-1/4}\delta^{1-n_{\mathcal{Y}}} n_{\mathcal{Y}} n_{\mathcal{R}}^{2n_{\mathcal{P}}} + \delta^{1/2}\right)$$

Theorem 1 implies that the principal's regret vanishes if $T \to \infty$ and $\epsilon, \bar{\epsilon}, \delta \to 0$ at the appropriate rates. It also follows from the proof that the principal's payoffs converge to a natural benchmark: what he would have obtained in a stationary equilibrium of the repeated game where it is common knowledge that y_t is drawn independently from the empirical distribution $\hat{\pi}_{I_t}$.

VI. Stage Game with Private Signals

In general, we cannot expect the principal to have access to an information oracle. Fortunately, we can still construct mechanisms σ^* that obtain vanishing or bounded principal's regret without any knowledge of the learner. However, in order to state the relevant assumptions (sections VII and VIII) and describe the mechanism (section VIII), we need to consider scenarios where the agent has private information that the principal lacks. This requires a brief detour. In this section, we revisit the stage game in order to introduce terminology that reflects agent's private information.

Suppose that the state y is drawn from a known distribution π, but the agent has access to a private signal $I \in \mathcal{I}$ generated by *information structure* γ.

Definition 12 (Information Structure)**.** *An* information structure *is a function $\gamma : \mathcal{I} \times \mathcal{Y} \to [0, 1]$ where $\gamma(\cdot, y)$ is a probability distribution over $\mathcal{I}$.*

The game proceeds as follows. First, nature chooses a hidden state $y \sim \pi$. Second, the principal chooses a policy p. Third, the agent observes a signal $I \sim \gamma(\cdot, y)$ and chooses a response r_I. Finally, the state y is revealed and payoffs are determined.

As in section II, suppose the agent does not necessarily maximize her expected utility. Instead, she chooses responses r_I that guarantees her an expected utility that is within an additive constant ϵ of the optimum. Let $B(\pi, \gamma, \epsilon)$ be the set of response distributions μ_I consistent with this assumption, i.e. where

$$\epsilon \geq \max_{\tilde{r}_I \in \mathcal{R}} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}[U(\tilde{r}_I, p, y) - \mathrm{E}_{r\sim\mu_I}[U(r, p, y)]]\right]$$

For a given information structure γ, the principal's worst-case utility $\alpha_p(\pi, \gamma, \epsilon)$ under policy p is

$$\min_{\mu_I \in B(\pi,\gamma,\epsilon)} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}[\mathrm{E}_{r\sim\mu_I}[V(r, p, y)]]\right]$$

and his best-case utility $\beta_p(\pi, \gamma, \epsilon)$ is given by

$$\max_{\mu_I \in B(\pi,\gamma,\epsilon)} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}[\mathrm{E}_{r\sim\mu_I}[V(r,p,y)]]\right]$$

Recall that our theorem 1 could be interpreted as reducing the online mechanism design problem to the simpler task of finding a ϵ-robust policy in the stage game without a private signal. The same is true of our next result, theorem 2. In contrast, theorem 3 reduces the online problem to solving for a robust policy when the agent has a private signal generated by an unknown information structure. This corresponds to notion of informational robustness introduced by Bergemann and Morris (2013) and applied by Bergemann, Brooks, et al. (2017), applied to our single-agent setting.

Definition 13. *The* ϵ-informationally-robust *policy for an unknown information structure* γ *is*

$$p^\dagger(\pi, \epsilon) \in \arg\max_{p\in\mathcal{P}} \inf_{\gamma} \alpha_p(\pi, \gamma, \epsilon)$$

Definition 14. *Fix* $\pi \in \Delta(\mathcal{Y})$ *and* $\epsilon > 0$*. The* cost of ϵ-informational-robustness *the principal's best-case optimal minus his worst-case optimal utility. Formally,*

$$\nabla(\pi, \epsilon) = \max_{p\in\mathcal{P}} \sup_{\gamma} \beta_p(\pi, \gamma, \epsilon) - \max_{p\in\mathcal{P}} \inf_{\gamma} \alpha_p(\pi, \gamma, \epsilon)$$

Let $\nabla(\pi) = \nabla(\pi, 0)$ denote the cost of informational robustness in the traditional setting where the agent is optimizing exactly ($\epsilon = 0$). It will be convenient to assume that the cost is growing at most linearly in ϵ.

Assumption 4. $\forall \pi \in \Delta(\mathcal{Y})$, $\nabla(\pi, \epsilon) = \nabla(\pi) + O(\epsilon)$.

VII. Mechanism for an Uninformed Agent

Our second result bounds the principal's regret under a mechanism that does not require detailed knowledge of the learner L. Instead, this result assumes that the agent is not more informed than the principal.

Mechanism 2. *Let the distribution* π_t *be a forecast of the state* y_t *given by the* FORECAST *algorithm. Fix a parameter* $\bar{\epsilon} > 0$*. In period* t*, the* uninformed-agent mechanism σ^* *chooses* $\bar{\epsilon}$*-robust policy* $p^*(\pi_t, \bar{\epsilon})$ *that treats forecast* π_t *as a common prior.*

What does it mean for an agent to be uninformed? The agent's behavior cannot reveal an understanding of the state sequence that goes far beyond the principal's forecast. This can be formalized by bounding the agent's ER from below and her IR (or CIR) from above.

Assumption 5 (Lower-Bounded ER)**.** *Let* $y_{1:T}$ *be the realized state sequence and let* $p^*_{1:T}$ *be the policy sequence generated by the proposed mechanism* σ^**. There exists a constant* $\tilde{\epsilon} \geq 0$ *such that* $\mathrm{ER}(y_{1:T}, p^*_{1:T}) \geq -\tilde{\epsilon}$ *and* $\mathrm{ER}(y_{1:T}, (p, \ldots, p)) \geq -\tilde{\epsilon}$, $\forall p \in \mathcal{P}$.

We claimed that there is no ex ante sense in which the deterministic sequence $y_{1:T}$ is predictable or not. However, this combination of bounds can be seen as an ex post definition of unpredictability. If an agent fully exploits the information she reveals under mechanism σ^* (no-IR) without outperforming the best use of public information (non-negative ER), her private information cannot be that useful. Fully exploiting useless information generally means ignoring it.

To see this, suppose the policy p is fixed and that the agent obtains non-positive IR and non-negative ER. It is trivial to show that IR is non-negative and bounded below by ER, so it follows that the agent's IR and ER both equal zero. In turn, IR and ER can only be equal when the best-in-hindsight responses conditional on the context (i.e. the agent's response) are the same in every context. That is, the context is useless. To achieve zero IR, the agent's response must equal some best-in-hindsight response conditional on the context. If the best-in-hindsight response is unique, this means that the agent's response is the same in every period.

What this amounts to, essentially, is that our reasoning for theorem 1 largely applies to theorem 2. Let us recall the first steps of that argument. Previously, we considered all periods $t \in I$ with information I as context. It followed immediately from the definition of information that the agent's responses r_t were roughly some constant r_I. Furthermore, since the principal's forecasts used I_t as context, the constant policy p_I was calibrated to the empirical distribution $\hat{\pi}_I$.

Now, our mechanism does not have access to I_t and is not calibrated to $\hat{\pi}_I$. Instead, it is calibrated to the empirical distribution $\hat{\pi}_{p_t}$ conditioned on policy p_t, i.e.

$$\hat{\pi}_P(y) = \frac{1}{n_p} \sum_{t=1}^{T} \mathbf{1}(y_t = y, p_t = p)$$

where n_p is the number of periods where $p_t = p$. The policy context p is coarser than information I, by definition of the latter. So, the principal following mechanism 2 behaves as if the agent shares his prior $\hat{\pi}_p$, while the agent behaves as if she receives I as a private signal.

This is where non-negative ER comes in. The agent's information I is useless to her. Suppose for now that there is a unique best response given policy p and distribution $\hat{\pi}_p$. Then the agent will choose the same response $r_t = r_p$ in every period where $p_t = p$. In other words, the policy context p coincides with the agent's information I, and the principal is correct in assuming that the agent optimizes against the empirical distribution $\hat{\pi}_p$. Our previous argument goes through.

What if our supposition fails, i.e. there are multiiple best responses given policy p and distribution $\hat{\pi}_p$? In general, the argument breaks down. The agent can condition her action on her private information I, which no longer necessarily coincides with the policy context p. To be clear, this private signal I remains useless to the agent. Moreover, the $\bar{\epsilon}$-robust policy is by definition robust to multiplicity of

best responses. However, because the agent's best response may be correlated with the state, this can undermine the principal's utility even if it does not affect the agent's.

The following assumption restricts attention to stage games where this issue does not arise. Informally, it asserts that if a private signal is useless to the agent, then it has limited relevance to the principal, assuming that the principal is following (nearly) optimal policies. Formally, the agent's value $\phi_p(\pi, \gamma)$ from information structure γ in the stage game with common prior π is

$$\max_{r, r_I \in \mathcal{R}} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}[U(r_I, p, y)] - U(r, p, y)\right]$$

This is the expected utility of the agent that optimizes given information structure γ minus the expected utility of the agent if she does not receive a private signal.

Assumption 6. *Let π be a distribution, $\epsilon > 0$ be a constant, and γ be an information structure (intuitively, one that is not useful to the agent).*

1) *If the principal uses ϵ-robust policy $p^*(\pi, \epsilon)$, his maxmin payoff without γ is not much larger than his maxmin payoff with γ. That is,*

$$\alpha_{p^*(\pi,\epsilon)}(\pi, \epsilon) - \alpha_{p^*(\pi,\epsilon)}(\pi, \gamma, \epsilon) = O\left(\phi_{p^*(\pi,\epsilon)}(\pi, \gamma)\right) + O(\epsilon)$$

2) *The principal's maxmax payoff with γ under any policy $p \in \mathcal{P}$ is not much larger than his maxmax-optimal payoff without γ. That is,*

$$\beta_p(\pi, \gamma, \epsilon) - \max_{\tilde{p}\in\mathcal{P}} \beta_{\tilde{p}}(\pi, \epsilon) = O\left(\phi_p(\pi, \gamma)\right) + O(\epsilon)$$

Theorem 2. *Assume restrictions on the stage game (assumptions 1, 3, 6), ϵ-bounded CIR (assumption 2), and $\tilde{\epsilon}$-lower-bounded ER (assumption 5). Let σ^* be the uninformed-agent mechanism 2. For any constant $\bar{\epsilon} > 0$, the principal's regret $\mathrm{E}_{\sigma^*}[\mathrm{PR}(L, y_{1:T})]$ is at most*

$$O(\bar{\epsilon}) + \frac{1}{\bar{\epsilon}} \cdot \tilde{O}\left(\epsilon + T^{-1/4}\delta^{1-n_{\mathcal{Y}}} n_{\mathcal{Y}} + \delta^{1/2}\right)$$

Theorem 2 implies that the principal's regret vanishes if $T \to \infty$ and $\epsilon, \bar{\epsilon}, \tilde{\epsilon}, \delta \to 0$ at the appropriate rates. It also follows from the proof that the principal's payoffs converge to a natural benchmark: what he would have obtained in a stationary equilibrium of the repeated game where it is common knowledge that y_t is drawn independently from the empirical distribution $\hat{\pi}_{p_t}$.

VIII. Mechanism for an Informed Agent

In section IV, we assumed that the principal knows the agent's learner L. The implication of this assumption is that the principal is as informed as the agent. In section V, we assumed that the agent is as uninformed as the principal. In this section, we allow the agent to be more informed than the principal. This generality comes at a cost: we no longer ensure vanishing principal's regret. Instead, we show that, in the limit, the following mechanism guarantees regret that is no greater than the empirical cost of informational robustness.

Mechanism 3. *Let the distribution π_t be a forecast of the state y_t given by the* FORECAST *algorithm. Fix a parameter $\bar{\epsilon} > 0$. In period t, the* informed-agent mechanism σ^* *chooses the $\bar{\epsilon}$-informationally-robust policy $p^\dagger(\pi_t, \bar{\epsilon})$ that treats the forecast π_t as a common prior.*

Theorem 3 builds on the same reasoning as theorems 1 and 2. First, we need to adapt assumption 3.

Assumption 7. *Fix $\pi, \tilde{\pi} \in \Delta(\mathcal{Y})$ and $\epsilon > 0$. If the ϵ-informationally-robust policies under π and under $\tilde{\pi}$ are the same, then they are also equal to the ϵ-informationally-robust policy under any convex combination any convex combination $\breve{\pi} = \lambda\pi + (1-\lambda)\tilde{\pi}$. That is,*

$$p^\dagger(\pi, \epsilon) = p^\dagger(\tilde{\pi}, \epsilon) \implies p^\dagger(\pi, \epsilon) = p^\dagger(\breve{\pi}, \epsilon)$$

Second, recall how, in the previous section, we were concerned that the principal's policy p_t in period t was calibrated to the empirical distribution $\hat{\pi}_p$ given policy context p, rather than the empirical distribution $\hat{\pi}_I$ given information $I = I_t$. There, we resolved that problem by assuming the agent was uninformed (non-negative ER). Here, our solution is even simpler: choose a policy p_t that is robust to the agent's private information I, whatever that may be.

To be more precise, recall that the policy context p is coarser than information I. We can interpret periods $t \in I$ as those periods in which the agent received a private signal I. By looking at the frequency of information I within policy context p, we can define an empirical information structure $\hat{\gamma}_p$ using Bayes' rule, i.e.

$$\hat{\gamma}_p(I, y) = \frac{n_I \hat{\pi}_I(y)}{n_p \hat{\pi}_p(y)}$$

for all information $I \in \mathcal{I}$ that is consistent with policy p. Before, we could approximately treat the principal's and agent's utility as their expected utility in the stage game where the state y was drawn from the empirical distribution $\hat{\pi}_I$. Now, the approximation is the expected utility in the stage game where $y \sim \hat{\pi}_P$ and the agent receives private signal I from the empirical information structure $\hat{\gamma}_P$. Of course, the principal's policy p_t is robust to all information structures γ, including $\hat{\gamma}_P$.

Theorem 3. *Assume restrictions on the stage game (assumptions 4, 7), and ϵ-bounded CIR (assumption 2). Let σ^* be the informed-agent mechanism 3. For any constant $\bar{\epsilon} > 0$, the principal's regret is at most*

$$\mathrm{E}_{\sigma^*}[\mathrm{PR}(L, y_{1:T})] \le \frac{1}{T}\sum_{p\in\mathcal{P}} n_p \nabla(\hat{\pi}_p) + O(\bar{\epsilon}) + \frac{1}{\bar{\epsilon}} \cdot \tilde{O}\left(\epsilon + T^{-1/4}\delta^{1-n_{\mathcal{Y}}} n_{\mathcal{Y}} + \delta^{1/2}\right)$$

In contrast to our previous results, this regret bound does not vanish. However, the bound does converge to the empirical cost of informational robustness as $T \to \infty$ and $\epsilon, \bar{\epsilon}, \delta \to 0$ at the appropriate rates. Nonetheless, there is reason to believe that this bound is not tight – but improving it would require us to use responsive mechanisms. Here is why: although the principal will never have access to the private signal I of the agent, he may attempt to learn (via the agent's past behavior) about the information structure γ that generates it. In turn, the agent may anticipate this and attempt to manipulate the principal's policy by feigning (partial) ignorance of her private signal. This suggests a less conservative definition of informational robustness, where the principal learns the quality of any information that the agent decides to exploit.

IX. Conclusion

We studied single-agent mechanism design where the common prior assumption is replaced with repeated interaction and frequent feedback about the world. Our primary motivation was to remove a barrier (the common prior) that makes it difficult to implement mechanisms in practice, but this work can also be viewed as a learning foundation for (robust) mechanism design. Indeed, our results show that policies similar to those predicted by a common prior can perform well even without distributional assumptions.

However, there are two caveats to this interpretation. First, our policies are robust to agents that behave suboptimally by up to some $\epsilon > 0$. In contrast, most papers on local robustness involve an optimizing agent with misspecified beliefs. These notions coincide sometimes but not always. In addition, our policies sometimes require informational robustness. Second, the number of interactions T required for our mechanisms to approximate the static common prior game may be large. In particular, our bounds depend on features of the stage game, like the number of policies, responses, and states. These features may also affect the agent's learning rate, which in turn affects our bounds. In this sense, the common prior assumption may be less appealing in games that are more complex.

Interesting directions for future work include generalizing to multi-agent problems, relaxing observability of the state to observability of payoffs (i.e. bandit feedback), and tightening the regret bound in theorem 3 by allowing the principal to learn about the information structure of the agent. The latter two changes would require a theory of behavior under responsive mechanisms.

Acknowledgment

This work began as part of the 2018 Special Quarter on Data Science and Online Markets at Northwestern. We are especially grateful to Simina Brânzei and Katya Khmelnitskaya for their early contributions to this project. We are also grateful to Eddie Dekel, Marciano Siniscalchi, and several anonymous referees for helpful comments, in addition to audiences at the 70th Midwest Theory Day and Northwestern. Jason Hartline and Aleck Johnsen were supported in part by NSF grant CCF-1618502.

References

Anunrojwong, J., Iyer, K., & Manshadi, V. (2020). Information design for congested social services: optimal need-based persuasion. In *Proceedings of the 21st acm conference on economics and computation* (pp. 349–350). EC '20. Virtual Event, Hungary: Association for Computing Machinery.

Arora, R., Dekel, O., & Tewari, A. (2012). Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th international coference on international conference on machine learning* (pp. 1747–1754). ICML'12. Edinburgh, Scotland: Omnipress.

Arora, R., Dinitz, M., Marinov, T. V., & Mohri, M. (2018). Policy regret in repeated games. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 6733–6742). NIPS'18. Montréal, Canada: Curran Associates Inc.

Artemov, G., Kunimoto, T., & Serrano, R. (2013). Robust virtual implementation: Toward a reinterpretation of the Wilson doctrine. *Journal of Economic Theory*, *148*(2), 424–447.

Balcan, M.-F., Blum, A., Haghtalab, N., & Procaccia, A. D. (2015). Commitment without regrets: online learning in stackelberg security games. In *Proceedings of the sixteenth acm conference on economics and computation* (pp. 61–78). EC '15. Portland, Oregon, USA: Association for Computing Machinery.

Balcan, M.-F., Blum, A., Hartline, J. D., & Mansour, Y. (2008). Reducing mechanism design to algorithm design via machine learning. *Journal of Computer and System Sciences*, *74*(8), 1245–1270.

Bergemann, D., Brooks, B., & Morris, S. (2017). First-price auctions with general information structures: implications for bidding and revenue. *Econometrica*, *85*(1), 107–143.

Bergemann, D. & Morris, S. (2013). Robust predictions in games with incomplete information. *Econometrica*, *81*(4), 1251–1308.

Blum, A., Hajiaghayi, M., Ligett, K., & Roth, A. (2008). Regret minimization and the price of total anarchy. In *Proceedings of the fortieth annual acm symposium on theory of computing* (pp. 373–382). STOC '08. Victoria, British Columbia, Canada: ACM.

Blum, A. & Hartline, J. D. (2005). Near-optimal online auctions. In *Proceedings of the sixteenth annual acm-siam symposium on discrete algorithms* (pp. 1156–1163). SODA '05. Vancouver, British Columbia: Society for Industrial and Applied Mathematics.

Blum, A., Kumar, V., Rudra, A., & Wu, F. (2004). Online learning in online auctions. *Theoretical Computer Science*, *324*(2), 137–146.

Blum, A. & Mansour, Y. (2007, December). From external to internal regret. *J. Mach. Learn. Res. 8*, 1307–1324.

Braverman, M., Mao, J., Schneider, J., & Weinberg, M. (2018). Selling to a no-regret buyer. In *Proceedings of the 2018 acm conference on economics and computation* (pp. 523–538). EC '18. Ithaca, NY, USA: ACM.

Cole, R. & Roughgarden, T. (2014). The sample complexity of revenue maximization. In *Proceedings of the forty-sixth annual acm symposium on theory of computing* (pp. 243–252). STOC '14. New York, New York: ACM.

Cummings, R., Devanur, N. R., Huang, Z., & Wang, X. (2020). Algorithmic price discrimination. In *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '20. Salt Lake City, Utah, USA.

Das, S., Kamenica, E., & Mirka, R. (2017, October). Reducing congestion through information design. In *2017 55th annual allerton conference on communication, control, and computing (allerton)* (pp. 1279–1284).

Daskalakis, C. & Syrgkanis, V. (2016). Learning in auctions: regret is hard, envy is easy. In *2016 ieee 57th annual symposium on foundations of computer science (focs)* (pp. 219–228).

Deng, Y., Schneider, J., & Sivan, B. (2019). Strategizing against no-regret learners. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 1579–1587). Curran Associates, Inc.

Dudík, M., Haghtalab, N., Luo, H., Schapire, R. E., Syrgkanis, V., & Vaughan, J. W. (2017). Oracle-efficient online learning and auction design. In *2017 ieee 58th annual symposium on foundations of computer science (focs)* (pp. 528–539).

Dughmi, S. & Xu, H. (2016). Algorithmic Bayesian persuasion. In *Proceedings of the forty-eighth annual acm symposium on theory of computing* (pp. 412–425). STOC '16. Cambridge, MA, USA: ACM.

Foster, D. P. & Vohra, R. V. (1997). Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, *21*(1), 40–55.

Goldstein, I. & Leitner, Y. (2018). Stress tests and information disclosure. *Journal of Economic Theory*, *177*, 34–69.

Hart, S. & Mas-Colell, A. (2001). A general class of adaptive strategies. *Journal of Economic Theory*, *98*(1), 26–54.

Hartline, J. D., Johnsen, A., Nekipelov, D., & Zoeter, O. (2019). Dashboard mechanisms for online marketplaces. In *Proceedings of the 2019 acm conference on economics and computation* (pp. 591–592). EC '19. Phoenix, AZ, USA: ACM.

Hartline, J. D., Syrgkanis, V., & Tardos, É. (2015). No-regret learning in bayesian games. In *Proceedings of the 28th international conference on neural information processing systems - volume 2* (pp. 3061–3069). NIPS'15. Montreal, Canada: MIT Press.

Immorlica, N., Mao, J., Slivkins, A., & Wu, Z. S. (2020). Incentivizing exploration with selective data disclosure. In *Proceedings of the 21st acm conference on economics and computation* (pp. 647–648). EC '20. Virtual Event, Hungary: Association for Computing Machinery.

Jehiel, P., Meyer-ter-Vehn, M., & Moldovanu, B. (2012). Locally robust implementation and its limits. *Journal of Economic Theory*, *147*(6), 2439–2452.

Kamenica, E. & Gentzkow, M. (2011, October). Bayesian persuasion. *American Economic Review*, *101*(6), 2590–2615.

Kleinberg, R. & Leighton, T. (2003). The value of knowing a demand curve: bounds on regret for online posted-price auctions. In *Proceedings of the 44th annual ieee symposium on foundations of computer science* (p. 594). FOCS '03. USA: IEEE Computer Society.

Mansour, Y., Slivkins, A., Syrgkanis, V., & Wu, Z. S. (2016). Bayesian exploration: incentivizing exploration in bayesian games. In *Proceedings of the 2016 acm conference on economics and computation* (p. 661). EC '16. Maastricht, The Netherlands.

Meyer-ter-Vehn, M. & Morris, S. (2011). The robustness of robust implementation. *Journal of Economic Theory*, *146*(5), 2093–2104.

Morgenstern, J. & Roughgarden, T. (2015). The pseudo-dimension of near-optimal auctions. In *Proceedings of the 28th international conference on neural information processing systems - volume 1* (pp. 136–144). NIPS'15. Montreal, Canada: MIT Press.

Nekipelov, D., Syrgkanis, V., & Tardos, E. (2015). Econometrics for learning agents. In *Proceedings of the sixteenth acm conference on economics and computation* (pp. 1–18). EC '15. Portland, Oregon, USA: ACM.

Ollár, M. & Penta, A. (2017, August). Full implementation and belief restrictions. *American Economic Review*, *107*(8), 2243–77.

Oury, M. & Tercieux, O. (2012). Continuous implementation. *Econometrica*, *80*(4), 1605–1637.

Syrgkanis, V. (2017). A sample complexity measure with applications to learning optimal auctions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 5358–5365). NIPS'17. Long Beach, California, USA: Curran Associates Inc.