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

Modibo Camara† Jason Hartline‡ Aleck Johnsen§

September 14, 2020

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 – centered around *counterfactual internal regret* – that capture the spirit of rationality without relying on beliefs.

*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.

†Department of Economics, Northwestern University. Email: modibokhane@gmail.com.

‡Department of Computer Science, Northwestern University. Email: hartline@northwestern.edu.

§Department of Computer Science, Northwestern University. Email: aleckjohnsen@u.northwestern.edu.

1 Introduction

Mechanism design is a branch of economic theory concerned with the design of social institutions. It encompasses a wide range of phenomena that have historically been of interest to economists, including, but not limited to, auctions (Myerson 1981; Vickrey 1961), matching markets (Gale and Shapley 1962; Roth 1982), taxation (Mirrlees 1971), contracts (Ross 1973; Spence and Zeckhauser 1971), and persuasion (Kamenica and Gentzkow 2011).

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.[1]

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 behavioral assumptions. Specifically, 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 take the form of 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) it has the potential to be widely applicable.[2]

[1] This assumption is limiting in two ways. First, mechanisms based on a common prior may not be practicable, because they rely on knowledge that a real-world designer is unlikely to possess. Second, even if the designer knows the distribution (resp. has beliefs), the participants may not arrive with the same knowledge (resp. share those beliefs).

[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), financial stress-

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 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 – e.g. 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 – e.g. 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 simple 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. Therefore, the company faces a reinforcement learning problem and its benchmark corresponds to the notion of *policy regret* in the literature on bandit learning with adap-

testing (Goldstein and Leitner 2018), and worker motivation (Ely and Szydlowski 2020).

tive adversaries (e.g. Arora, Dekel, et al. 2012). In that setting, Arora, Dekel, et al. (2012) show that guaranteeing sublinear (policy) regret is generally impossible.[3] This fact precludes a simple solution to the company's problem; we must constrain the regulator's behavior.[4]

The standard way to constrain the regulator/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. This will be our approach as well. We build on existing no-regret assumptions, in ways that are intended to refine and better motivate those assumptions.

No-Regret Agents. 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.[5] 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 (see e.g. Littlestone and Warmuth 1994 for ER), 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 in our model 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

[3] Arora, Dekel, et al. (2012) obtain positive results when the adversary satisfies a bounded memory assumption. Ryabko and Hutter (2008) obtain positive results under a different kind of assumption, that the environment is sufficiently "forgiving" of mistakes. These papers reflect two prominent approaches in reinforcement learning: (a) restricting attention to Markov decision processes, and (b) assuming an ability to "reset" the problem (Kearns et al. 1999).

[4] 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. This is reminiscent of a literature on multi-agent reinforcement learning when the state is Markovian (Buşoniu et al. 2010; Hu and Wellman 1998; Littman 1994; Uther and Veloso 2003). Unlike these papers, we do not have the ability in our model to advise all participants simultaneously.

[5] 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. In particular, the agent fails to anticipate the mechanism that the principal is using. As they point out, this leads to odd behavior: the agent may purchase goods at a price exceeding her valuation. In our setting, the agent does not face uncertainty with respect to the mechanism; instead, she faces uncertainty with respect to the state sequence.

mind, there is no need to resort to responsiveness if nonresponsive mechanisms tightly bound the principal's regret.[6]

Counterfactual Internal Regret. Without constraints on the agent's behavior, an early mistake by the principal can result in a permanent, undesirable shift in the agent's behavior. As we will see, this can occur when the agent behaves as if she has additional information about the state of the world that is not accounted for in our description of the model. The agent can make the principal's problem infeasible if she is willing to exploit her information selectively, i.e. based on the principal's choice of policies. Unfortunately, neither no-ER nor no-IR assumptions can rule out selective use of information.

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

[6]This approach seems spiritually similar to that of Immorlica, Mao, 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.

additional context for the forecast. 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

$$\underbrace{O(\epsilon)}_{\text{cost of } \epsilon\text{-robustness}} + \frac{1}{\epsilon}\left(\underbrace{O(\text{CIR})}_{\text{agent's regret}} + \underbrace{\tilde{O}\left(\frac{\delta^{1-n_{\mathcal{Y}}} n_{\mathcal{Y}} n_{\mathcal{R}}^{2n_{\mathcal{P}}}}{T^{1/4}}\right)}_{\text{forecast miscalibration}} + \underbrace{O\left(\delta^{1/2}\right)}_{\text{approximation error}}\right) \tag{1}$$

If the agent satisfies no-CIR, i.e. 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. This mechanism is identical to the first, except its forecast does not use information revealed by the learner. 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

$$\underbrace{O(\epsilon)}_{\text{cost of } \epsilon\text{-robustness}} + \frac{1}{\epsilon}\left(\underbrace{O(\text{CIR})}_{\text{agent's regret}} + \underbrace{\tilde{O}\left(\frac{\delta^{1-n_{\mathcal{Y}}} n_{\mathcal{Y}}}{T^{1/4}}\right)}_{\text{forecast miscalibration}} + \underbrace{O\left(\delta^{1/2}\right)}_{\text{approximation error}}\right) \tag{2}$$

Compared to (1), this drops the exponential dependence on the number $n_{\mathcal{P}}$ of policies. This is because the principal's forecast does not need to take into account the agent's information, which significantly reduces the forecast miscalibration in finite samples.

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 $\hat{\pi}_T$ 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

$$\underbrace{\nabla(\hat{\pi}_T)}_{\text{cost of informational robustness}} + \underbrace{O(\epsilon)}_{\text{cost of } \epsilon\text{-robustness}} + \frac{1}{\epsilon}\left(\underbrace{O(\text{CIR})}_{\text{agent's regret}} + \underbrace{\tilde{O}\left(\frac{\delta^{1-n_y} n_y}{T^{1/4}}\right)}_{\text{forecast miscalibration}} + \underbrace{O\left(\delta^{1/2}\right)}_{\text{approximation error}} \right) \tag{3}$$

Unlike (1) and (2), the principal's regret does not vanish as $T \to \infty$. However, it is vanishing up to the cost of informational robustness $\nabla(\hat{\pi}_T)$ that would also be present under a common prior, if the agent were more informed than the principal.

Finally, although our focus is not on computational complexity, the reader should 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. The literature on sample complexity in mechanism design allows the principal to learn the state distribution from i.i.d. samples (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. However, there has also been work that applies online learning to auction design (e.g. Daskalakis and Syrgkanis 2016; Dudík et al. 2017) 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 potentially forward-looking.

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, Mao, 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 model generalizes to adversarial data.

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 (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 2 introduces the stage game and ϵ-robustness. Section 3 introduces the repeated game. Section 4 defines external, internal, and counterfactual internal regret. Section 4 presents our mechanism and regret bounds when the agent's learner is known. As preparation for the remaining results, section 5 introduces the stage game with private signals. Section 4 presents our mechanism and regret bounds when the agent is uninformed. Section 4 presents our mechanism

and regret bounds when the agent may be more informed than the principal. Section 9 concludes with a discussion of open problems.

Appendix A applies these results to two special cases: our running example, and a principal-agent problem. Appendix B considers the complexity of the agent's learning problem. Appendix C describes our forecasting algorithms in more detail. Appendix D relaxes some of the restrictions on the stage game and generalizes our results. Appendix E collects proofs.

2 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 $V : \mathcal{R} \times \mathcal{P} \times \mathcal{Y} \to [0, 1]$.

Assumption 1 (Regularity)**.** *We impose the following regularity conditions.*

1. *The state space $\mathcal{Y}$ is finite.*
2. *The response space $\mathcal{R}$ is a compact space with metric $d_{\mathcal{R}}$.*
3. *The policy space $\mathcal{P}$ is a compact space with metric $d_{\mathcal{P}}$.*
4. *The utility U is equi-Lipschitz continuous in (r, p) for Lipschitz constants $K_{\mathcal{R}}^U$ and $K_{\mathcal{P}}^U$, i.e.*

$$\forall y \in \mathcal{Y} : \quad |U(r, p, y) - U(\tilde{r}, \tilde{p}, y)| \leq K_{\mathcal{R}}^U d_{\mathcal{R}}(r, \tilde{r}) + K_{\mathcal{P}}^U d_{\mathcal{P}}(p, \tilde{p})$$

5. *The utility V is equi-Lipschitz continuous in (r, p) for Lipschitz constants $K_{\mathcal{R}}^V$ and $K_{\mathcal{P}}^V$, i.e.*

$$\forall y \in \mathcal{Y} : \quad |V(r, p, y) - V(\tilde{r}, \tilde{p}, y)| \leq K_{\mathcal{R}}^V d_{\mathcal{R}}(r, \tilde{r}) + K_{\mathcal{P}}^V d_{\mathcal{P}}(p, \tilde{p})$$

Later on, we will use covers to convert infinite action spaces into discrete approximations. For example, our running example involved an infinite policy space.

Definition 1 (Covers)**.** *Let $\mathcal{X}$ be a metric space with metric $d_{\mathcal{X}}$. Generally, lower case letters x denote elements of $\mathcal{X}$ while upper case letters X denote subsets.*

1. *Fix $\delta_{\mathcal{X}} > 0$. Let the partition $\mathcal{C}_{\mathcal{X}}$ be a $\delta_{\mathcal{X}}$* cover *of $\mathcal{X}$. That is, for every set $X \in \mathcal{C}_X$, any two elements $x, \tilde{x} \in X$ must be within distance $\delta_{\mathcal{X}}$ of one another, i.e. $d_{\mathcal{X}}(x, \tilde{x}) < \delta_{\mathcal{X}}$.*
2. *To reduce notation, we also let $\mathcal{C}_{\mathcal{X}}$ denote a discretized subset of $\mathcal{X}$. That is, for each set $X \in \mathcal{C}_{\mathcal{X}}$, choose a unique $x \in X$ to represent X. In that case, we say $x \in \mathcal{C}_{\mathcal{X}}$.*

3. *Let $x \in \mathcal{X}$ and $\tilde{x} \in C_{\mathcal{X}}$. We say that $\tilde{x}$ is the* discretization *of x if $x, \tilde{x}$ belong to the same subset $X \in C_{\mathcal{X}}$.*

We will refer to covers $C_{\mathcal{P}}$ of the policy space (with metric $d_{\mathcal{P}}$), $C_{\mathcal{R}}$ of the response space (with metric $d_{\mathcal{R}}$), and $C_{\Delta(\mathcal{Y})}$ of the state distributions $\Delta(\mathcal{Y})$ (with the l_1 metric).[7] Of course, if the underlying set $\mathcal{X}$ is finite to begin with, we can simply set $\delta_{\mathcal{X}} = 0$ and let $C_{\mathcal{X}} = \mathcal{X}$.

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 (or a distribution μ over responses) that guarantees her an expected utility that is within an additive constant ϵ of the optimum. 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

$$\alpha_p(\pi, \epsilon) = \min_{\mu \in \Delta(\mathcal{R})} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{r\sim\mu}[V(r, p, y)]\right] \quad \text{s.t.} \quad \max_{\tilde{r}\in\mathcal{R}} \mathrm{E}_{y\sim\pi}[U(\tilde{r}, p, y)] - \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{r\sim\mu}[U(r, p, y)]\right] \leq \epsilon$$

and his best-case utility is described by

$$\beta_p(\pi, \epsilon) = \max_{\mu \in \Delta(\mathcal{R})} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{r\sim\mu}[V(r, p, y)]\right] \quad \text{s.t.} \quad \max_{\tilde{r}\in\mathcal{R}} \mathrm{E}_{y\sim\pi}[U(\tilde{r}, p, y)] - \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{r\sim\mu}[U(r, p, y)]\right] \leq \epsilon$$

The worst-case optimal (or ϵ-robust) policy, defined below, is one of two main ingredients in our proposed mechanisms (the other is a calibrated forecasting algorithm).

Definition 2 (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, policy is*

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

Definition 3 (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 to assume that this cost is growing at most linearly in ϵ, although this assumption is not really necessary (see appendix D).

Assumption 2. *For any distribution π, $\Delta(\pi, \epsilon) = O(\epsilon)$.*

[7] Note that $\mathcal{P}$, $\mathcal{R}$, $\Delta(\mathcal{Y})$ are all compact. Therefore, we can always construct a finite cover.

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. *Assume regularity (assumption 1). For any distribution π, policy p, and constants $\epsilon, \tilde{\epsilon} > 0$, the principal's worst-case and best-case utilities satisfy*

$$\alpha_p(\pi, \epsilon + \tilde{\epsilon}) \geq \alpha_p(\pi, \epsilon) - \frac{\tilde{\epsilon}}{\epsilon} \quad \text{and} \quad \beta_p(\pi, \epsilon + \tilde{\epsilon}) \leq \beta_p(\pi, \epsilon) + \frac{\tilde{\epsilon}}{\epsilon}$$

Appendix A describes two well-known special cases of our model: Bayesian persuasion and the principal-agent problem. For each case, we provide a simple example, check that the example satisfies all relevant assumptions, and evaluate our results. In that sense, these examples serve as sanity checks for the rest of the paper, which involves assumptions and solutions that are sometimes rather abstract.

3 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 in the t^{th} period is given by[8]

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

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 in the t^{th} period 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 some finite set of fixed policies $\mathcal{P}_0 \subseteq \mathcal{P}$ that the principal wishes to consider.[9]

[8] The fact that the response distribution μ_t may depend on realized response history $r_{1:t-1}$ allows the learner to introduce correlation between responses across time, if desired.

[9] Many of our results and definitions can be adapted to any finite set of nonresponsive mechanisms.

Formally, by a fixed policy p, we mean a *constant mechanism* σ^p that selects the same policy

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

in all periods t and for all histories.

To define the principal's regret, we need notation for the agent's behavior under the proposed mechanism σ^*, as well as under the counterfactual mechanisms 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^*. Let μ_t^p describes the agent's behavior under σ^p, i.e.

$$\mu_t^p = L_t(y_{1:t-1}, r_{1:t-1}^p, (\underbrace{p, \ldots, p}_{\text{t times}}))$$

given the realized history of responses $r_{1:t-1}^p$ under σ^p.

Definition 4 (Principal's Regret)**.** *The* principal's regret *relative to the best-in-hindsight fixed policy* $p \in \mathcal{P}_0$ *is*

$$\mathrm{PR}(L, y_{1:T}) = \sup_{p \in \mathcal{P}_0} \frac{1}{T} \sum_{t=1}^{T} \left(\mathrm{E}_{r \sim \mu_t^p}\left[V(r, p, y_t)\right] - \mathrm{E}_{r \sim \mu_t^*}\left[V(r, \sigma^*(y_{1:t-1}, r_{1:t-1}^*, p_{1:t-1}), y_t)\right] \right)$$

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. The following proposition formalizes this simple observation.

Proposition 1 (Impossibility Result for Unrestricted Learners)**.** *In our running example, for every mechanism* σ^**, there exists a learner* L *along with a state sequence* $y_{1:\infty}$ *such that the principal's regret does not vanish, i.e.*

$$\lim_{T \to \infty} \mathrm{PR}(L, y_{1:T}) > 0$$

4 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 5. *Given a sequence $z_{1:T}$ of covariates, the agent's* contextual regret *relative to a best-in-hindsight modification rule $h : \mathcal{Z} \to \mathcal{R}$ is*

$$\mathrm{CR}(p_{1:T}, y_{1:T}) = \max_h \frac{1}{T} \sum_{t=1}^{T} \Big(U(h(z_t), p_t, y_t) - U(r_t, p_t, y_t) \Big)$$

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 6 (Responsiveness)**.** *A 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 period t, 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. In restricting attention to nonresponsive mechanisms, we simplify the agent's problem and make our behavioral assumptions more credible. If our mechanisms were responsive, non-myopic agents would not necessarily satisfy no-regret as defined above. For example, an agent might decide to forgo an otherwise-optimal response if she believes said response would trigger an undesirable policy by the principal going forward.[10] This behavior would be perfectly reasonable but could cause the agent to accumulate regret. Finally, as it turns out, even nonresponsive mechanisms can guarantee vanishing principal's regret in two of the scenarios we study (sections 5 and 7). In these scenarios, there is limited room for responsive mechanisms to improve our guarantees.

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

4.1 External Regret

In our model, external regret is contextual regret where the policy p_t is the context in period t. That is, no-ER requires the agent to perform as well as the best-in-hindsight mapping from policies p_t to responses r_t. Now, why should external regret include the policy as context? Because our stage game is an extensive form. A strategy in the stage game is not a response; it is a function from the observed policy to a response. Our definition of external regret compares the agent's performance to the best-in-hindsight strategy in the stage game.[11]

[10] For instance, in models of repeated sales, a buyer may refuse to purchase a good at a reasonable price if she believes that holding out will cause the seller to reduce prices in the future (Devanur et al. 2019; Immorlica, Lucier, et al. 2017).

[11] Suppose that, instead, we compared the agent's performance to the best-in-hindsight response $r \in \mathcal{R}$. Defining external regret in this way would confound variation in policies with variation in the state, and could lead to odd

An immediate difficulty with defining ER is that the set $\mathcal{P}$ may be continuous.[12] For instance, this is true in our running example. To ensure that the agent's learning problem is feasible in that case, we allow the agent to group together nearby policies according to the cover $\mathcal{C}_\mathcal{P}$ (defined in section 2), and consider regret with respect to this coarser context. Of course, when the policy space $\mathcal{P}$ is finite, there is no need for this, and we can set $\mathcal{C}_\mathcal{P} = \mathcal{P}$.

Definition 7 (External Regret)**.** *The agent's* external regret (ER) *relative to the best-in-hindsight modification rule $h : \mathcal{C}_\mathcal{P} \to \mathcal{R}$ is*

$$\mathrm{ER}(p_{1:T}, y_{1:T}) = \max_h \frac{1}{T} \sum_{t=1}^{T} \Big(U(h(p_t), p_t, y_t) - U(r_t, p_t, y_t) \Big)$$

Note the slight abuse of notation. By $h(p_t)$, we mean $h(P_t)$ where P_t is the unique set in the partition $\mathcal{C}_\mathcal{P}$ that contains p_t.

Although common in the literature (e.g. Nekipelov et al. 2015, Braverman et al. 2018), no-ER assumptions are insufficient for our problem. They do not circumvent the impossibility result (proposition 1) 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.

Proposition 2 (Impossibility Result for No-ER Learners)**.** *In our running example, for every mechanism σ^*, there exists a learner L that guarantees no-ER on all state/policy sequences, i.e.*

$$\lim_{T\to\infty} \sup_{\tilde{p}_{1:T}, \tilde{y}_{1:T}} \mathrm{E}_L \big[\mathrm{ER}(\tilde{p}_{1:T}, \tilde{y}_{1:T}) \big] = 0$$

along with a state sequence $y_{1:\infty}$ such that the principal's regret does not vanish, i.e.

$$\lim_{T\to\infty} \mathrm{PR}(L, y_{1:T}) > 0$$

4.2 Internal and 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).

behavior. For example, consider the"mean-based" learner in Braverman et al. (2018), which never deviates far from the response that maximizes the agent's empirical utility. In that paper, the learner engages in odd behavior, like spending more than the agent's valuation.

Our definition is more similar to that of Hartline, Johnsen, et al. (2019), where agents following a dashboard provided by the mechanism will best respond to an allocation rule given the empirical value distribution, rather than best respond to the empirical bid distribution. This way, the agent adapts sensibly to changes in the principal's policy.

[12] If $\mathcal{P}$ is continuous, the policy p_t may be unique in every period $t = 1, \ldots, \infty$. In that case, requiring no-ER would be equivalent to requiring ex post optimality. That is unreasonably strong.

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 the mechanism σ^* to underperform. In particular, even though the agent may not observe y_t when choosing a response r_t, this cannot prevent the adversary from "correlating" r_t and y_t.[13]

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 the agent's beliefs or to a distribution 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 $r_{1:T}$. To ensure that the agent's learning problem is feasible when the response space $\mathcal{R}$ is infinite, we allow the agent to group together nearby responses according to the cover $\mathcal{C}_{\mathcal{R}}$, and consider regret with respect to this coarser context. Of course, when the response space $\mathcal{R}$ is finite, as in our running example, there is no need for this, and we can set $\mathcal{C}_{\mathcal{R}} = \mathcal{R}$.

Definition 8 (Internal Regret)**.** *The agent's* internal regret (IR) *relative to the best-in-hindsight modification rule* $h : S_{\mathcal{P}} \times S_{\mathcal{R}} \to \mathcal{R}$ *is*

$$\mathrm{IR}(p_{1:T}, y_{1:T}) = \max_h \frac{1}{T} \sum_{t=1}^{T} \Big(U(h(p_t, r_t), p_t, y_t) - U(r_t, p_t, y_t) \Big)$$

Like earlier, note the slight abuse of notation. By $h(p_t, r_t)$*, we mean* $h(P_t, R_t)$ *where* (P_t, R_t) *is the unique set in the collection* $\mathcal{C}_{\mathcal{P}} \times \mathcal{C}_{\mathcal{R}}$ *that contains* (p_t, r_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 fixed policies $p \in \mathcal{P}_0$. The following definitions formalize this.

[13] To be clear, this "correlation" is non-causal. For example, the adversary might choose a state sequence such that $y_t = 1$ on even periods and $y_t = 0$ on odd periods, and a learner L such that $r_t = 1$ on even periods and $r_t = 0$ on odd periods. Empirically-speaking, there would be a correlation between the states and the responses. However, if we subsequently changed the value of state y_t in some period t, this would not affect the response r_t, because the state is not observed and cannot affect the output of the learner L. That is, there is no causal relationship between r_t and y_t.

Definition 9 (Information)**.** *Let the* information partition *be*

$$\mathcal{I} = \underbrace{S_{\mathcal{P}}}_{\text{policy } p_t^*} \times \underbrace{S_{\mathcal{R}}}_{\text{response } r_t^*} \times \underbrace{\left(S_{\mathcal{R}}\right)^{|\mathcal{P}_0|}}_{\text{responses } r_t^p \text{ for } p \in \mathcal{P}_0}$$

and let the information I_t *in period* t *be the unique set in* $\mathcal{I}$ *that satisfies*

$$I_t \ni \left(p_t^*, r_t^*, (r_t^p)_{p \in \mathcal{P}_0}\right)$$

Note that, by definition, the same information I_t is available to the agent regardless of whether the principal follows our mechanism σ^* or deviates to some fixed policy $p \in \mathcal{P}_0$. Intuitively, the principal's choice of mechanism should not affect what information the agent has available.

Definition 10 (Counterfactual Internal Regret)**.** *The agent's* counterfactual internal regret (CIR) *relative to the best-in-hindsight modification rule* $h : S_{\mathcal{P}} \times S_{\mathcal{R}}^{|\mathcal{P}_0|+1} \to \mathcal{R}$ *is*

$$\text{CIR}(p_{1:T}, y_{1:T}) = \max_h \frac{1}{T} \sum_{t=1}^{T} \left(U(h(I_t), p_t, y_t) - U(r_t, p_t, y_t)\right)$$

The discussion in the proof of proposition 2 clarifies how no-CIR rules out the kinds of pathological or irrational behavior that no-ER fails to rule out. In the next section, we will see the crucial role that no-CIR plays in our proving our bounds on the principal's regret. The essential property is that, conditional on information I_t, the agent chooses a roughly constant response that is approximately best-in-hindsight for whichever mechanism the principal is considering.

5 Mechanism for an Informed Principal

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 5 and 6. 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 11 (Information Oracle)**.** *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* $p_t \in \mathcal{P}$ *and the realized history.*

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. To be clear, under a common prior, the fact that the principal knows the agent's prior means that he also 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.

Mechanism 1. *Let the distribution π_t be a forecast of the state y_t generated by a calibrated forecasting algorithm that uses the agent's information as context.*

- *Our forecasting algorithm 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 $\pi \in C_{\Delta(\mathcal{Y})}$ and the loss function is the negated quadratic scoring rule S. In each period, the algorithm makes a prediction π_t and incurs loss $-S(\pi_t, y_t)$. Further details as well as rates of convergence are in appendix C.*

- *The context is the vector of outputs $\Omega(p)$ of the information oracle under discretized policies $p \in C_{\mathcal{P}}$. The forecasting algorithm is run separately for each context.*

Fix a parameter $\bar{\epsilon} > 0$. In period t, the informed-principal mechanism σ^* *chooses the discretization of 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. First, we require some additional notation. Let "$t \in I$" indicate that information I is present in period t, i.e. $I_t = I$. Let $n_I = \sum_{t=1}^{T} \mathbf{1}(t \in I)$ indicate the number of periods with information I. Let $\hat{\pi}_I$ be the empirical distribution conditioned on the agent having 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 agent's response r_t^* is roughly constant, as are her counterfactual responses r_t^p under fixed policies $p \in \mathcal{P}_0$. By regularity (1), slight variations in responses have correspondingly slight impacts on the agent's and principal's utility. Suppose that these responses are exactly constant, i.e. $r_t = r_I$. Note that $p_t = p_I$ is exactly constant as well, across these time periods, for all constant mechanisms σ^p as well as the proposed mechanism σ^*, which uses discretized policies. With everything constant, the principal's average utility across context I takes on a familiar form:

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

Similarly, the agent's average utility is

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

Essentially, within each context I, we have recreated the stage game with common prior $\hat{\pi}_I$. The

agent accumulates regret

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

Under mechanism 1, the principal chooses (roughly) the $\bar{\epsilon}$-robust policy for the forecast π_t. Suppose for the moment that the forecasts are also roughly constant for all periods $t \in I$, i.e. $\pi_t = \pi_I$. Since the forecast is calibrated and uses information I_t as context, π_I cannot be too far in the l_1 distance from $\hat{\pi}_I$ (this is essentially the definition of calibration, and follows from results in appendix C). It follows from regularity that 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. At the same time, it cannot be much better than the best-case optimal utility. More precisely,

$$\max_{\tilde{p}} \beta_{\tilde{p}}(\hat{\pi}_I, \bar{\epsilon}) + \frac{\epsilon_I}{\bar{\epsilon}} \geq \mathrm{E}_{y\sim\hat{\pi}_I}\left[V(r_I, p_I, y)\right] \geq \max_{\tilde{p}} \alpha_{\tilde{p}}(\hat{\pi}_I, \bar{\epsilon}) - \frac{\epsilon_I}{\bar{\epsilon}} \tag{4}$$

By assumption 2, the difference between the upper bound and the lower bound is

$$O(\bar{\epsilon}) + O\left(\frac{\epsilon_I}{\bar{\epsilon}}\right) \tag{5}$$

This pins down the principal's utility under mechanism 1. Moreover, the upper bound in (4) also applies to any constant mechanism σ^p for $p \in \mathcal{P}_0$. Therefore, (5) also bounds the regret accumulated by the principal in context I.

This brings us to our key assumption: the agent's CIR is at most some constant ϵ.

Assumption 3 (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*

$$\epsilon \geq \mathrm{CIR}(y_{1:T}, p^*_{1:T}) \quad \text{and} \quad \epsilon \geq \mathrm{CIR}(y_{1:T}, \underbrace{p, \ldots, p}_{t \text{ times}}),\ \forall p \in \mathcal{P}_0$$

Remark 1. *It is worth emphasizing that this bound applies only to the realized state sequence $y_{1:T}$. That is, the agent does not need to perform well over all state sequences, and her objective need not be worst-case regret minimization. If the agent is Bayesian, for example, she will obtain low CIR as long as her beliefs are well-calibrated.*

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 (5) on the agent's regret ϵ_I in the context of information I, and it follows that the principal's regret is bounded above by

$$O(\bar{\epsilon}) + O\left(\frac{\epsilon}{\bar{\epsilon}}\right)$$

To transform this intuition into a result, we need to address an assumption made along the way: that the forecast π_t is roughly 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 – useless information from the principal's perspective, 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. In appendix D, we avoid this restriction by instead assuming that the principal – using our publicly-announced mechanism – is not more informed than the agent.

Assumption 4. *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 close to one another, then they are also close to the ϵ-robust policy under any convex combination of these distributions. Formally, for any $\lambda \in [0, 1]$,*

$$d_{\mathcal{P}}\left(p^*(\pi, \epsilon), p^*\left(\lambda\pi + (1-\lambda)\tilde{\pi}, \epsilon\right)\right) = O\left(d_{\mathcal{P}}\left(p^*(\pi, \epsilon), p^*(\tilde{\pi}, \epsilon)\right)\right)$$

The following theorem formalizes the preceding discussion and bounds the principal's regret under mechanism 1.

Theorem 1. *Assume regularity (assumption 1), restrictions on the stage game (assumptions 2, 4), and ϵ-bounded CIR (assumption 3). Let σ^* be the mechanism 1. Given access to the information oracle, for any constant $\bar{\epsilon} > 0$, the principal's expected regret $\mathrm{E}_{\sigma^*}\left[\mathrm{PR}(L, y_{1:T})\right]$ is at most*

$$\underbrace{O(\bar{\epsilon})}_{\text{cost of } \bar{\epsilon}\text{-robustness}} + \frac{1}{\bar{\epsilon}} \cdot \left(\underbrace{O(\epsilon)}_{\text{agent's regret}} + \underbrace{\tilde{O}\left(T^{-1/4}\sqrt{|\mathcal{Y}||\mathcal{C}_{\Delta(\mathcal{Y})}||\mathcal{C}_R|^{(|P_0|+|\mathcal{C}_P|)/2}}\right)}_{\text{forecast miscalibration}} + \underbrace{O\left(\delta_{\Delta(\mathcal{Y})}^{1/2}\right) + O(\delta_R) + O(\delta_P)}_{\text{discretization error}} \right)$$

Remark 2. *Here are a few comments on this result.*

1. *The bound depends on the size of the partitions $\mathcal{C}_P$, $\mathcal{C}_R$, and $\mathcal{C}_{\Delta\mathcal{Y}}$. However, if we define these partitions to be as small as possible, we can replace these terms with the covering numbers of $\mathcal{P}$, $\mathcal{R}$, and $\Delta(\mathcal{Y})$, respectively. In that sense, our finite sample bounds will deteriorate as one increases the complexity of the action and state spaces.*

2. *Furthermore, if we define these partitions to be the smallest possible, then theorem 1 implies that the principal's regret vanishes if* $T \to \infty$ *and* $\epsilon, \bar{\epsilon}, \delta_{\Delta(\mathcal{Y})}, \delta_{\mathcal{P}}, \delta_{\mathcal{R}} \to 0$ *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* $\hat{\pi}_{I_t}$*. Formally,*

$$\frac{1}{T}\sum_{t=1}^{T} V(r_t, p_t, y_t) - \frac{1}{T}\sum_{I \in \mathcal{I}} n_I \max_{p \in \mathcal{P}} \beta_p(\hat{\pi}_I, 0) \to 0$$

3. *Finally, note the exponential dependence on the number of alternative mechanisms* $|\mathcal{P}_0|$ *and the size of the policy space cover* $|\mathcal{C}_{\mathcal{P}}|$*. This dependence, which is not present in theorems 2 and 3, reflects the fact that the mechanism 1 uses the agent's information* I_t *as context for its forecast* π_t*. Since our bound is uniform across all learners that satisfy* ϵ*-bounded CIR on the realized state sequence* $y_{1:T}$*, it must accommodate learners that generate a lot of information, regardless of whether that information is useful. As mentioned at the beginning of this section, this is another reason why the "informed principal" setting seems less compelling than the settings studied in sections 7 and 8.*

6 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 7 and 8) and describe the mechanism (section 8), 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 the *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. For instance, if the agent maximizes her expected utility, her responses after signals I would be

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

Finally, the state y is revealed and payoffs are determined.

As in section 2, suppose the agent does not necessarily maximize her expected utility. Instead, she chooses responses r_I (or distributions μ_I over responses) that guarantees her an expected utility

that is within an additive constant ϵ of the optimum. For a given information structure γ, the principal's worst-case utility from following policy p is described by

$$\alpha_p(\pi, \gamma, \epsilon) = \min_{\mu_I \in \Delta(\mathcal{R})} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}\left[\mathrm{E}_{r\sim\mu_I}[V(r,p,y)]\right]\right]$$
$$\text{subject to} \quad \max_{\tilde{r}_I \in \mathcal{R}} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}\left[U(\tilde{r}_I,p,y)\right]\right] - \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}\left[\mathrm{E}_{r\sim\mu_I}[U(r,p,y)]\right]\right] \leq \epsilon$$

and his best-case utility is described by

$$\beta_p(\pi, \gamma, \epsilon) = \max_{\mu_I \in \Delta(\mathcal{R})} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}\left[\mathrm{E}_{r\sim\mu_I}[V(r,p,y)]\right]\right]$$
$$\text{subject to} \quad \max_{\tilde{r}_I \in \mathcal{R}} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}\left[U(\tilde{r}_I,p,y)\right]\right] - \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{I\sim\gamma(\cdot,y)}\left[\mathrm{E}_{r\sim\mu_I}[U(r,p,y)]\right]\right] \leq \epsilon$$

Note that $\alpha(\pi, \epsilon)$, the worst-case utility in the stage game without a private signal, is equivalent to $\alpha(\pi, \gamma, \epsilon)$ when γ is uninformative. The same applies to β.

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 (ϵ-Informational-Robustness)**.** *The worst-case optimal (or* 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 (Cost of ϵ-Informational-Robustness)**.** *Fix a distribution* π *and parameter* $\epsilon > 0$*. The cost of* ϵ-informational-robustness *is the distance between the principal's best-case utility (under the best-case optimal policy (for the best-case information structure) and worst-case utility (under the worst-case optimal policy for the worst-case information structure). Formally,*[14]

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

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, although this assumption is not really necessary (see appendix D).

Assumption 5. *For any distribution* π*,* $\nabla(\pi, \epsilon) = \nabla(\pi) + O(\epsilon)$.

[14] Why do we evaluate the cost of informational robustness under the worst-case information structure? Because the regret guarantee that we obtain in theorem 3 applies uniformly across all learners L. As we will see, different learners will induce different empirical information structures γ. Our cost of informational robustness must accommodate the worst-case information structure, which loosely corresponds to the worst-case learner.

Finally, we verify that lemma 1 still applies in the presence of private signals.

Lemma 2. *Assume regularity (assumption 1). For any distribution π, information structure γ, policy p, and constants $\epsilon, \tilde{\epsilon} > 0$, the principal's worst-case and best-case utilities satisfy*

$$\alpha_p(\pi, \gamma, \epsilon + \tilde{\epsilon}) \geq \alpha_p(\pi, \gamma, \epsilon) - \frac{\tilde{\epsilon}}{\epsilon} \quad \text{and} \quad \beta_p(\pi, \gamma, \epsilon + \tilde{\epsilon}) \leq \beta_p(\pi, \gamma, \epsilon) + \frac{\tilde{\epsilon}}{\epsilon}$$

7 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. To begin, the mechanism is as follows.

Mechanism 2. *Let the distribution π_t be a forecast of the state y_t.*

- *Our forecasting algorithm 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 $\pi \in C_{\Delta(\mathcal{Y})}$ and the loss function is the negated quadratic scoring rule.*

Fix a parameter $\bar{\epsilon} > 0$. In period t, the uninformed-agent mechanism *σ^* chooses the discretization of the $\bar{\epsilon}$-robust policy $p^*(\pi_t, \bar{\epsilon})$ that treats the forecast π_t as a common prior.*

What does it mean for an agent to be uninformed? Following the intuition developed in section 4, 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 adding a lower bound on the agent's ER to our upper bound on the agent's (counterfactual) IR.[15]

Assumption 6 (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*

$$\mathrm{ER}(y_{1:T}, p^*_{1:T}) \geq -\tilde{\epsilon} \quad \text{and} \quad \mathrm{ER}(y_{1:T}, \underbrace{p, \ldots, p}_{t \text{ times}}) \geq -\tilde{\epsilon}, \ \forall p \in \mathcal{P}_0$$

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

To see this, suppose the policy p is fixed and that the learner 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

[15] Although they study a different problem, Blum, Gunasekar, et al. (2018) also use lower bounds on ER to prove results, exploiting the fact that exponential weights guarantees non-negative expected ER (Gofer and Mansour 2016).

that the learner'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 learner's response) are the same in every context. That is, the context is useless. To achieve zero IR, the learner'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 learner'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, for every policy context $P \in \Sigma_{\mathcal{P}}$, it is calibrated to the empirical distribution $\hat{\pi}_P$ conditioned on $p_t \in P$. Formally,

$$\hat{\pi}_P(y) = \frac{1}{n_P} \sum_{t \in P} \mathbf{1}(y_t = y)$$

where $t \in P$ indicates $p_t \in P$ and n_P is the number of periods $t \in P$. The policy context P is coarser than information I, by definition of the latter. So, the principal 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. If there is a unique best-in-hindsight response within policy context P, then the agent will choose roughly the same response $r_t = r_P$ in every period $t \in P$. In other words, the policy context P coincides with the agent's information I, and the principal is correct in assuming that the agent (roughly) optimizes against the empirical distribution $\hat{\pi}_P$. Our previous argument goes through.

Again, we just assumed that there is a unique best-in-hindsight response within policy context P. What if this is not the case, i.e. the best-in-hindsight response is not unique? In general, the argument breaks down. The agent can condition her action on her private information I, which no longer necessarily coincides with 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, if the agent's best response is correlated with the state, this can undermine the principal's utility even if it does not affect the agent's.[16]

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 value of information structure γ to the agent in the stage game with common prior π and policy p is

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

[16]For example, consider a stage game with a binary response $r \in \{0, 1\}$, a binary state $y \in \{0, 1\}$, and a binary policy $p \in \{\text{Risky}, \text{Safe}\}$. The agent's utility is always zero. The principal's utility under the risky policy is 1 if $r = y$ and -1 otherwise. It is slightly negative under the safe policy. If y is drawn from the uniform distribution, and the agent optimizes without a signal, then the principal prefers the risky policy. If the agent receives a signal that is perfectly correlated with the state, and sets $r = 1 - y$, then the principal prefers the safe policy.

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 7. *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, i.e. $\alpha_{p^*(\pi,\epsilon)}(\pi, \epsilon)$, is not much larger than his maxmin payoff with γ, i.e. $\alpha_{p^*(\pi,\epsilon)}(\pi, \gamma, \epsilon)$. 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 $p \in \mathcal{P}$, i.e. $\beta_p(\pi, \gamma, \epsilon)$, is not much larger than his maxmax payoff without γ under the best-case policy, i.e. $\max_{\tilde{p}\in\mathcal{P}} \beta_{\tilde{p}}(\pi, \epsilon)$. That is,*

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

Both parts of assumption 7 would be immediate if the information structure γ were uninformative, because the left-hand sides would be non-positive. Basically, we require useless (to the agent) private signals to be similar to uninformative private signals in these two respects.

Finally, we are ready to bound the principal's regret under mechanism 2.

Theorem 2. *Assume regularity (assumption 1), restrictions on the stage game (assumptions 2, 4, 7), ϵ-bounded CIR (assumption 3), and $\tilde{\epsilon}$-lower-bounded ER (assumption 6). Let σ^* be the uninformed-agent mechanism 2. For any constant $\bar{\epsilon} > 0$, the principal's expected regret $\mathrm{E}_{\sigma^*}\left[\mathrm{PR}(L, y_{1:T})\right]$ is at most*

$$\underbrace{O(\bar{\epsilon})}_{\text{cost of } \bar{\epsilon}\text{-robustness}} + \underbrace{O(\tilde{\epsilon})}_{\text{agent's information}} + \frac{1}{\bar{\epsilon}} \cdot \left(\underbrace{O(\epsilon)}_{\text{agent's regret}} + \underbrace{\tilde{O}\left(T^{-1/4}\sqrt{|\mathcal{Y}||\mathcal{C}_{\Delta(\mathcal{Y})}|}\right)}_{\text{forecast miscalibration}} + \underbrace{O\left(\delta_{\Delta(\mathcal{Y})}^{1/2}\right) + O(\delta_{\mathcal{R}}) + O(\delta_{\mathcal{P}})}_{\text{discretization error}} \right)$$

Remark 3. *If we define the partition $\mathcal{C}_{\Delta\mathcal{Y}}$ to be the smallest possible, then theorem 2 implies that the principal's regret vanishes if $T \to \infty$ and $\epsilon, \bar{\epsilon}, \tilde{\epsilon}, \delta_{\Delta(\mathcal{Y})}, \delta_{\mathcal{P}}, \delta_{\mathcal{R}} \to 0$ 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 $\hat{\pi}_{P_t}$. Formally,*

$$\frac{1}{T}\sum_{t=1}^{T} V(r_t, p_t, y_t) - \frac{1}{T}\sum_{P\in\mathcal{C}_{\mathcal{P}}} n_P \max_{p\in\mathcal{P}} \beta_p(\hat{\pi}_P, 0) \to 0$$

8 Mechanism for an Informed Agent

In section 4, 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 5, 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 cost of informational robustness.

Mechanism 3. *Let the distribution π_t be a forecast of the state y_t.*

- *Our forecasting algorithm applies a generic no-internal-regret algorithm due to Blum and Mansour (2007) in an auxilliary learning problem where the action space consists of the discretized forecasts $\pi \in C_{\Delta(\mathcal{Y})}$ and the loss function is the negated quadratic scoring rule.*

Fix a parameter $\bar{\epsilon} > 0$. In period t, the informed-agent mechanism σ^* *chooses the discretization of 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 4 to the case with private signals.

Assumption 8. *Let $\epsilon > 0$. Let π and $\tilde{\pi}$ be distributions in the stage game. If the ϵ-informationally-robust policies under π and under $\tilde{\pi}$ are close to one another, then they are also close to the ϵ-informationally-robust policy under any convex combination of these distributions. Formally, for any $\lambda \in [0, 1]$,*

$$d_{\mathcal{P}}\left(p^\dagger(\pi, \epsilon), p^\dagger\left(\lambda\pi + (1-\lambda)\tilde{\pi}, \epsilon\right)\right) = O\left(d_{\mathcal{P}}\left(p^\dagger(\pi, \epsilon), p^\dagger(\tilde{\pi}, \epsilon)\right)\right)$$

Next, 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 $P \in C_{\mathcal{P}}$ (where $t \in P$) 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.

$$\hat{\gamma}_P(I, y) = \frac{n_I \hat{\pi}_I(y)}{n_P \hat{\pi}_P(y)} \cdot \mathbf{1}(I \subseteq P)$$

where $I \subseteq P$ is shorthand for $t \in I \implies t \in P$. Before, we could roughly approximate 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 $\hat{\gamma}_P$.

Next, we formalize this discussion and bound the principal's regret under mechanism 3.

Theorem 3. *Assume regularity (assumption 1), restrictions on the stage game (assumptions 5, 8), and ϵ-bounded CIR (assumption 3). Let σ^* be the informed-agent mechanism 3. For any constant $\bar{\epsilon} > 0$, the principal's expected regret $\mathrm{E}_{\sigma^*}\left[\mathrm{PR}(L, y_{1:T})\right]$ is at most*

$$\underbrace{\frac{1}{T}\sum_{P \in C_{\mathcal{P}}} n_P \nabla(\hat{\pi}_P) + O(\bar{\epsilon})}_{\text{cost of } \bar{\epsilon}\text{-informational-robustness}} + \frac{1}{\bar{\epsilon}} \cdot \left(\underbrace{O(\epsilon)}_{\text{agent's regret}} + \underbrace{\tilde{O}\left(T^{-1/4}\sqrt{|\mathcal{Y}||C_{\Delta(\mathcal{Y})}|}\right)}_{\text{forecast miscalibration}} + \underbrace{O\left(\delta_{\Delta(\mathcal{Y})}^{1/2}\right) + O(\delta_{\mathcal{R}}) + O(\delta_{\mathcal{P}})}_{\text{discretization error}} \right)$$

Remark 4. *In contrast to our previous results, this regret bound does not vanish. However, if we define the partition $C_{\Delta\mathcal{Y}}$ to be the smallest possible, the bound does converge to*

$$\frac{1}{T}\sum_{P \in C_{\mathcal{P}}} n_P \nabla(\hat{\pi}_P)$$

as $T \to \infty$ and $\epsilon, \bar{\epsilon}, \delta_{\Delta(\mathcal{Y})}, \delta_{\mathcal{P}}, \delta_{\mathcal{R}} \to 0$ at the appropriate rates. This is the best possible guarantee in a stationary equilibrium of the repeated game where (a) it is common knowledge that y_t is drawn independently from the empirical distribution $\hat{\pi}_{P_t}$ and (b) the agent has access to an unknown information structure γ.

9 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. However, we also want to emphasize that this work can 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 making any assumptions about the data-generating process. This lends credibility to researchers who invoke the common prior for tractability, but do not expect it to be taken literally. However, there are two caveats.

1. 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 (e.g. Artemov et al. 2013; Meyer-ter-Vehn and Morris 2011; Oury and Tercieux 2012). These notions coincide sometimes but not always. In addition, our policies sometimes require informational robustness (Bergemann and Morris 2013).

2. 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 size of the policy and response spaces, and the number of states. These features may also affect the agent's learning rate, which in turn affects our bounds. In that sense, the common prior assumption may be less appealing in more games that are more complex.

Further Work. There are several directions in which to generalize and improve this work. To begin, it is not clear whether our finite sample bounds have a tight dependence on the number of periods T and various other parameters. For example, is it possible to remove the exponential dependence in theorem 1 on the size of the policy space?[17] In addition, there may be opportunities for tightening our results in less abstract settings where the stage game has more structure.

Our analysis was restricted to single-agent problems. Suppose there are multiple agents. From the perspective of any one agent, her opponents correspond to adaptive adversaries (c.f. Arora, Dinitz, et al. 2018) whose future behavior is influenced by the agent's present response. However, if the number of participants is large and the mechanism's outcome preserves the differential privacy of each agent's response history (c.f. McSherry and Talwar 2007), the behavioral assumptions developed here may also be suitable for the multi-agent setting.

We assumed that the principal and agent observe the state after every interaction, but this may be unrealistic in many applications. For instance, in contract theory the state is a function from the agent's actions to outcomes. Let us briefly refer to the principal-agent problem in appendix A.2. There, if the agent chooses to work, we do observe whether the project succeeds or not. However, we may not learn whether the project would have succeeded had the agent shirked. To mitigate this issue, we could consider the case with bandit feedback, where participants observe their own payoffs but not the state. The challenge with bandit feeback is that it requires responsive mechanisms, as the mechanism must depend on the principal's payoffs, which in turn depend on the agent's response.[18]

In section 8, where the agent may be more informed than the principal, the principal's regret did not vanish but rather converged to the cost of informational robustness under a common prior. There is reason to believe that this result is not tight. 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. However, in the repeated game, this approach would require responsive mechanisms.

[17] One approach the principal might take is to attempt to discern the agent's beliefs from the description of her learner L, and substitute those beliefs for his own forecast. If successful, this would tie the principal's forecast miscalibration to the agent's counterfactual internal regret.

[18] Relatedly, Balcan, Blum, Haghtalab, et al. (2015) consider a repeated Stackelberg game where the state is the agent's private type. The principal receives bandit feedback: he never observes the type directly but can infer it from the agent's behavior. The issues associated with responsiveness do not arise in this model as the agent is myopic (or more precisely, there is a sequence of short-lived agents).

As the last two paragraphs illustrate, we need a theory of behavior for responsive mechanisms. The no-regret conditions used here and elsewhere are not as well-motivated when the mechanism (or adversary) is responsive, insofar as they do not generalize traditional rationality assumptions. Extending the logic of no-regret conditions to a larger set of mechanisms – but not necessarily all mechanisms – is a clear priority for further work.

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., Gunasekar, S., Lykouris, T., & Srebro, N. (2018). On preserving non-discrimination when combining expert advice. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 8386–8397). NIPS'18. Montréal, Canada: Curran Associates Inc.

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

Boutilier, C. (2012). Eliciting forecasts from self-interested experts: scoring rules for decision makers. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems - volume 2* (pp. 737–744). AAMAS '12. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems.

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.

Buşoniu, L., Babuška, R., & De Schutter, B. (2010). Multi-agent reinforcement learning: an overview. In D. Srinivasan & L. C. Jain (Eds.), *Innovations in multi-agent systems and applications - 1* (pp. 183–221). Berlin, Heidelberg: Springer Berlin Heidelberg.

Carroll, G. (2015). Robustness and linear contracts. *American Economic Review*, *105*(2), 536–63.

Cesa-Bianchi, N. & Lugosi, G. (2006). *Prediction, learning, and games*. New York, NY, USA: Cambridge University Press.

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). 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.

Devanur, N. R., Peres, Y., & Sivan, B. (2019). Perfect bayesian equilibria in repeated sales. *Games and Economic Behavior*, *118*, 570–588.

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.

Dütting, P., Roughgarden, T., & Talgam-Cohen, I. (2019). Simple versus optimal contracts. In *Proceedings of the 2019 acm conference on economics and computation* (pp. 369–387). EC '19. Phoenix, AZ, USA: ACM.

Ely, J. C. & Szydlowski, M. (2020). Moving the goalposts. *Journal of Political Economy*, *128*(2), 468–506.

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

Gale, D. & Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, *69*(1), 9–15.

Gofer, E. & Mansour, Y. (2016). Lower bounds on individual sequence regret. *Machine Learning*, *103*(1), 1–26.

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.

Hu, J. & Wellman, M. P. (1998). Multiagent reinforcement learning: theoretical framework and an algorithm. In *Proceedings of the fifteenth international conference on machine learning* (pp. 242–250). ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Immorlica, N., Lucier, B., Pountourakis, E., & Taggart, S. (2017). Repeated sales with multiple strategic buyers. (pp. 167–168). EC '17. Cambridge, Massachusetts, USA: Association for Computing Machinery.

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.

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2008). Scoring rules, generalized entropy, and utility maximization. *Operations Research*, *56*(5), 1146–1157.

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

Kearns, M., Mansour, Y., & Ng, A. Y. (1999). Approximate planning in large pomdps via reusable trajectories. In *Proceedings of the 12th international conference on neural information processing systems* (pp. 1001–1007). NIPS'99. Denver, CO: MIT Press.

Littlestone, N. & Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation*, *108*(2), 212–261.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the eleventh international conference on international conference on machine*

learning (pp. 157–163). ICML'94. New Brunswick, NJ, USA: Morgan Kaufmann Publishers Inc.

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.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences*, *42*(9), 654–655.

McSherry, F. & Talwar, K. (2007). Mechanism design via differential privacy. In *Proceedings of the 48th annual ieee symposium on foundations of computer science* (pp. 94–103). FOCS '07. USA: IEEE Computer Society.

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

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *The Review of Economic Studies*, *38*(2), 175–208.

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.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*, *6*(1), 58–73.

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). Full implementation and belief restrictions. *American Economic Review*, *107*(8), 2243–77.

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

Ross, S. A. (1973). The economic theory of agency: the principal's problem. *The American Economic Review*, *63*(2), 134–139.

Roth, A. E. (1982). The economics of matching: stability and incentives. *Mathematics of Operations Research*, *7*(4), 617–628.

Ryabko, D. & Hutter, M. (2008). On the possibility of learning in reactive environments with arbitrary dependence. *Theoretical Computer Science*, *405*(3), 274–284.

Sappington, D. (1983). Limited liability contracts between principal and agent. *Journal of Economic Theory*, *29*(1), 1–21.

Spence, M. & Zeckhauser, R. (1971). Insurance, information, and individual action. *The American Economic Review*, *61*(2), 380–387.

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.

Uther, W. & Veloso, M. (2003). *Adversarial reinforcement learning*.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, *16*(1), 8–37.

A Special Cases

A.1 Bayesian Persuasion

There is an informed sender (i.e. principal) and an uninformed receiver (i.e. agent). The principal designs the process by which information is revealed to the agent. Let $\mathcal{M}$ be a finite set of messages that he can send. Knowing that the agent will react to an informative message, the principal attempts to persuade the agent towards actions that he prefers. Let $\mathcal{A}$ be a finite set of actions that the agent can take. The agent chooses a response $r : \mathcal{M} \to \mathcal{A}$ that maps messages to actions.

A policy is an information structure $p : \mathcal{Y} \to \Delta(\mathcal{M})$. That is, an information structure $p_t(y_t)$ describes the probability of a message m_t being sent, conditional on the state being y_t. The agent receives the message m_t and takes action $a_t = r_t(m_t)$. While the agent may not know the process that generated the state y_t, she understands the process p_t that generates the message m_t conditional on the state. Armed with this understanding, she can infer something about the state y_t based on the message m_t.

All that remains is to specify payoffs. Let $u : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ be the agent's utility function from a given action in a given state. Similarly, let $v : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ be the principal's utility. In the previous subsection, the utility functions U, V depended on the triple (r, p, y) rather than the pair (a, y). To reconcile the two models, we let participants evaluate (r, p, y) by their expected utility conditional on the state. Formally,

$$U(r, p, y) = \sum_{m \in \mathcal{M}} p(m, y) \cdot u(r(m), y) \quad \text{and} \quad V(r, p, y) = \sum_{m \in \mathcal{M}} p(m, y) \cdot v(r(m), y)$$

When the state is fixed, the residual variation in utility is due to the fact that messages are drawn randomly from the distribution $p(y)$. These distributions are common knowledge because the agent observes the principal's policy p before taking an action. Indeed, the fact that the principal commits to an information structure is the defining feature of Bayesian persuasion.

Example 1 (Judge-Prosecutor Game)**.** *The state space is* $\mathcal{Y} = \{\text{Innocent}, \text{Guilty}\}$ *and the action space is* $\mathcal{A} = \{\text{Convict}, \text{Acquit}\}$*. The judge has 0-1 utility u and prefers to convict if the defendant is guilty and acquit if the defendant is innocent. Regardless of the state, the prosecutor's utility v is 1 following a conviction and 0 following an acquittal.*

This example satisfies regularity (1) with the discrete metric on $\mathcal{R}$, the l_1-metric on $\mathcal{P}$, and $K_{\mathcal{R}}^U = K_{\mathcal{R}}^V = K_{\mathcal{P}}^U = K_{\mathcal{P}}^V = 1$.

The worst-case policy $p^*(\pi, \epsilon)$ sends the message "convict" whenever the defendant is guilty. If the defendant is innocent, it sends the message "convict" with probability

$$q = \max\left\{1, \min\left\{0, \frac{p - \epsilon}{1 - p}\right\}\right\}$$

The cost of ϵ-robustness $\Delta(\pi, \epsilon) = O(\epsilon)$ decreases smoothly with ϵ. This game satisfies assumption **??** since q is increasing in p (and hence convex combinations of distributions p will yield q

that is bounded between the ϵ-robust policies for the extremal distributions, which are close by assumption).

The worst-case policy $p^\dagger(\pi, \epsilon)$ for an unknown private signal is full transparency. The cost of informational robustness, i.e. $\nabla(\pi, 0)$, is the difference between the principal's value under the common prior π and his payoff under full transparency. This game satisfies assumption **??** with $M_1 = 1$ and $M_2 = O(\epsilon)$. It trivially satisfies assumptions **??** and **??** since $p^\dagger$ is constant.

A.2 Contract Design

In classic models of moral hazard, the principal incentivizes an agent to put effort into a task the principal cares about. The timing of the game is as follows: (1) the principal commits to a contract, (2) the agent takes a hidden action, (3) nature randomly chooses an outcome, (4) the agent is paid based on the outcome, (5) the game concludes. For concreteness, we consider the limited liability model due to Sappington (1983) where both participants are risk-neutral but the principal is not allowed to charge the agent. This model has been popularized by recent work in robust contract design (see e.g. Carroll 2015, Dütting et al. 2019).

Formally, let $\mathcal{R}$ be a finite set of actions that the agent can take. Let $\mathcal{O}$ be a finite set of outcomes o. The principal observes the outcome but not the action. The state $y : \mathcal{R} \to \mathcal{O}$ describes how actions map to outcomes. The employer commits to a contract $p : \mathcal{O} \to [0, \bar{p}]$ that specifies a non-negative payment for each outcome. The cost function $c : \mathcal{R} \to \mathbb{R}$ describes how costly it is for the agent to take a particular action. The agent's utility function is

$$U(r, p, y) = p(y(r)) - c(r)$$

The benefit function $b : \mathcal{O} \to \mathbb{R}$ describes how beneficial a given outcome is to the principal. The principal's utility function is

$$V(r, p, y) = b(y(r)) - p(y(r))$$

Through the contract p, the principal can incentivize the agent to take actions that, depending on the state, will lead to a more beneficial outcome.

Example 2. *The agent is given a task of unknown difficulty. There are two actions* $\mathcal{A} = \{\text{work}, \text{shirk}\}$, *two outcomes* $\mathcal{O} = \{\text{success}, \text{failure}\}$, *and three states* $\mathcal{Y} = \{\text{trivial}, \text{moderate}, \text{impossible}\}$. *In the trivial state, both actions lead to success. In the impossible state, both actions lead to failure. In the moderate state, work leads to success and shirk leads to failure.*

The principal's benefits are $b(\text{success}) = 2$ *and* $b(\text{failure}) = 0$. *The agent's costs are* $c(\text{work}) = 1$ *and* $c(\text{shirk}) = 0$. *In the impossible and trivial states, the optimal contract pays nothing after both outcpmes and the agent will shirk. In the moderate state, the optimal contract pays* $p(\text{success}) = 5$ *to cover the agent's costs if she works, otherwise* $p(\text{failure}) = 0$. *Generally, if the principal pays the agent after success, the agent will have to take into account the risk that the task turns out to be impossible (where work induces costs without any payment) or trivial (where work is not required for payment). To incentivize work, the contract must compensate the agent accordingly.*

This example satisfies regularity (1) with U, V normalized, the discrete metric on $\mathcal{R}$, the sup-norm-metric on $\mathcal{P}$, and $K_{\mathcal{R}}^U = K_{\mathcal{R}}^V = K_{\mathcal{P}}^U = K_{\mathcal{P}}^V = 1$.

The worst-case policy $p^*(\pi, \epsilon)$ sets $p(\text{failure}) = 0$ and

$$p(\text{success}) = \frac{c(\text{work}) - c(\text{shirk}) + \epsilon}{\pi(\text{moderate})}$$

so long as $p(\text{success}) \leq \bar{p}$ and the principal's π-expected payoff is greater than zero when the agent works. Otherwise, the worst-case policy sets all transfers to zero. The cost of ϵ-robustness $\Delta(\pi, \epsilon) = O(\epsilon)$ decreases smoothly with ϵ.

This game satisfies assumption **??**. To see this, note that as long as working is strictly more costly than shirking, the optimal policies that induce effort are bounded away from the optimal policies that do not. Among the policies that do not induce effort, convex combinations of the distribution will not make inducing effort desirable. Among policies that do induce effort, the fact that the payments following success are decreasing in $\pi(\text{moderate})$ means (as in the last section) that convex combinations of distributions lead to optimal policies that are between the extremal policies.

The worst-case policy $p^\dagger(\pi, \epsilon)$ for an unknown private signal is the same as the optimal policy under a common prior without a private signal. The cost of informational robustness, i.e. $\nabla(\pi, 0)$, is the difference between the principal's value when the agent only works in the "moderate" state and the principal pays her cost of effort conditional on success (assuming the principal prefers this to shirking with zero transfers) and his value in the common prior game without a private signal. This game satisfies assumption **??** with $M_1 = O(\epsilon)$ and $M_2 = 0$.

B Agent's Learning Problem

Upper bounds on external regret are often viewed as compelling assumptions (e.g. Nekipelov et al. 2015, Braverman et al. 2018) because there exist relatively simple algorithms that guarantee vanishing ER as $T \to \infty$. For example, the exponential weights algorithm (a.k.a. hedge algorithm, exponentiated gradient algorithm) satisfies no-ER. In contrast, our behavioral assumptions – e.g. no-FCIR – may appear daunting, insofar as the agent must solve a learning problem with a context space that is exponential in the number of alternative mechanisms, $|\Sigma_0|$. When both the sequence of states $y_{1:T}$ and the learner L are particularly pathological, no-FCIR may indeed be too strong an assumption. When the the learner satisfies additional properties, or sequence of states has some stochastic structure (e.g. is i.i.d. or Markov), no-FCIR may be more reasonable.

In this section, we make one simple observation. There exists a learner that guarantees no-FCIR (and hence no-CIR) for the agent under our mechanism from theorem 2, with the best rate of convergence we can hope for.

Definition 15 (CFL)**.** *Suppose the principal publicizes the forecast π_t in every period t.*[19] *The*

[19] In our view, part of the principal's objective is to make the agent's problem as simple as possible. From a worst-

common forecast learner *(CFL) sets*

$$r_t \in \arg\max_{r \in \mathcal{R}} \mathrm{E}_{y \sim \pi_t}\left[U(r_t, p_t, y_t)\right]$$

Proposition 3 verifies that the CFL satisfies the behavioral assumptions of theorem 2.

Proposition 3. *Let σ^* be the mechanism from theorem 2. Then the CFL satisfies ϵ-bounded FCIR* (4) *in expectation, i.e.*

$$\mathrm{E}_{L,\sigma^*}[\mathrm{FCIR}] \leq \epsilon = \tilde{O}\left(\frac{1}{T^{1/4}}\sqrt{|\mathcal{Y}||\mathcal{F}|}\right) + O\left(\sqrt{|\mathcal{Y}|\delta_{\mathcal{F}}}\right)$$

and $\tilde{\epsilon}$-lower-bounded FER (5)*, where $\tilde{\epsilon} = 0$. Moreover, if the agent uses CFL, the principal's regret bound in theorem 2 applies regardless of whether alignment* (9) *holds.*

Note that these rates preserve the $T^{1/4}$ convergence rate (up to $\delta_{\mathcal{F}}$ error) that is present in all of our mechanisms and reflects miscalibration of the principal. In that sense, the fact that the agent is also learning does not deteriorate the principal's performance at all. Although this has not been our emphasis so far, it would be interesting to see whether (or identify conditions under which) other simple learning algorithms satisfy our behavioral assumptions with decent rates of convergence.

C Calibrated Forecasting

In this appendix, we describe our forecasting algorithm and bound its miscalibration.

A linearly homogeneous, differentiable function H is *strongly convex* with parameter ξ if

$$H(\pi) \geq H(\tilde{\pi}) + \nabla H(\tilde{\pi}) \cdot (\pi - \tilde{\pi}) + \frac{\xi}{2}\|\pi - \tilde{\pi}\|_2^2$$

The gradient of H describes a *proper scoring rule* $S(\pi) = \nabla H(\pi)$ where $H(\pi) = \pi \cdot S(\pi)$ (McCarthy 1956). A scoring rule $S : \Delta(\mathcal{Y}) \to \mathbb{R}^{\mathcal{Y}}$ is proper if the report $\tilde{\pi}$ that maximizes the π-expected score is the distribution π. Strong convexity of H can be thought of as sharpening the incentives for truth-telling (Boutilier 2012).

Specifically, consider the quadratic scoring rule (see e.g. Jose et al. 2008)

$$S_y(\pi) = 2\pi(y) - \sum_{\tilde{y} \in \mathcal{Y}} \pi(\tilde{y})^2$$

where $H(\pi) = \|\pi\|_2^2$ is strongly convex with $\xi = 2$.

Recall that the mechanism σ^* is supposed to be nonresponsive. As a consequence, we cannot determine the principal's beliefs π_t in a given period based on his historical payoffs. To ensure that

case perspective, there is no benefit to hiding this information. With that said, we see no reason why this result should not apply under the weaker assumption that the principal's forecasting algorithm is public knowledge.

the beliefs π_t are well-calibrated, we consider an auxilliary online learning problem based on a scoring rule S. In period t, the principal makes a prediction π_t with loss function $S_{y_t}(\pi_t)$. Specifically, the predictions come from the discretized set of priors $\mathcal{F}_1$, formed by choosing a representative element π from each set in the partition $\mathcal{S}_{\mathcal{F}}$. In terms of the score, this approximation has limited cost. Let $\pi = \left[\hat{\pi}_F\right]_{\mathcal{F}_1}$ be the belief $\pi \in \mathcal{F}_1$ that is closest to the empirical distribution $\hat{\pi}_F$. Then

$$\begin{aligned} S_y(\hat{\pi}_F) - S_y(\pi) &\leq S_y(\hat{\pi}_F) - S_y(\hat{\pi}_F - \delta_{\mathcal{F}}) \\ &= 2\hat{\pi}_F(y) - \sum_{\tilde{y}\in\mathcal{Y}} \hat{\pi}_F(\tilde{y})^2 - 2(\hat{\pi}_F(y) - \delta_{\mathcal{F}}) + \sum_{\tilde{y}\in\mathcal{Y}} (\hat{\pi}_F(\tilde{y}) - \delta_{\mathcal{F}})^2 \\ &= 2\delta_{\mathcal{F}} - \sum_{\tilde{y}\in\mathcal{Y}} \hat{\pi}_F(\tilde{y})^2 + \sum_{\tilde{y}\in\mathcal{Y}} (\hat{\pi}_F(\tilde{y}) - \delta_{\mathcal{F}})^2 \\ &\leq 2\delta_{\mathcal{F}} \end{aligned} \tag{6}$$

where $\hat{\pi}_F - \delta_{\mathcal{F}}$ is shorthand notation for the vector $\left(\hat{\pi}_F(y) - \delta_{\mathcal{F}}\right)_{y\in\mathcal{Y}}$.

In this auxilliary problem, the exponential weights algorithm (see e.g. Cesa-Bianchi and Lugosi 2006) obtains expected external regret at most

$$\sqrt{2T \log |S_{\mathcal{F}}|}$$

relative to the best-in-hindsight $\pi_F^* \in \mathcal{F}_1$. A reduction due to Blum and Mansour (2007) (theorem 5) translates this into a bound on expected internal regret of

$$|S_{\mathcal{F}}| \sqrt{2T \log |S_{\mathcal{F}}|}$$

relative to the best-in-hindsight $\pi_F^* \in \mathcal{F}_1$. Combine this with the maximum approximation error (6) to bound the expected internal regret relative to the best-in-hindsight contextual belief $\pi \in \Delta(\mathcal{Y})$, which must be the empirical distribution $\hat{\pi}_F$ since S is proper. Specifically,

$$|S_{\mathcal{F}}| \sqrt{2T \log |S_{\mathcal{F}}|} + 2T\delta_{\mathcal{F}} \geq \mathrm{E}\left[\sum_{\pi\in\mathcal{F}_1} n_F \hat{\pi}_F \cdot \left(S(\hat{\pi}_F) - S(\pi)\right)\right] \tag{7}$$

where n_F is the number of periods t where $\left[\pi_t\right]_{\mathcal{F}_1} = F_t$. This is a statement about the expected scoring loss, where the expectation reflects randomization in the algorithm. Our next result, lemma 3, translates this into a statement about the l_1 distance between the principal's belief π and the empirical distribution $\hat{\pi}_F$.

Lemma 3. *Let S be a proper scoring rule where the optimal expected score H is ξ-strongly convex. Then*

$$\sqrt{\frac{2|\mathcal{Y}|\kappa}{\xi}} \geq \frac{1}{T} \sum_{\pi\in\mathcal{F}_1} n_F d_1(\pi, \hat{\pi}_F)$$

where

$$\kappa = \frac{1}{T}\sum_{\pi\in\mathcal{F}_1} n_F \hat{\pi}_F \cdot \left(S(\hat{\pi}_F) - S(\pi)\right)$$

Proof. Consider the principal's π-expected regret from predicting $\tilde{\pi}$:

$$\begin{aligned}
\pi \cdot (S(\pi) - S(\tilde{\pi})) &= H(\pi) - \pi \cdot \nabla H(\tilde{\pi}) \\
&\geq H(\tilde{\pi}) - \nabla H(\tilde{\pi}) \cdot \tilde{\pi} + \frac{\xi}{2}\|\pi - \tilde{\pi}\|_2^2 \\
&= \frac{\xi}{2}\|\pi - \tilde{\pi}\|_2^2 \\
&\geq \frac{\xi}{2}\left(\frac{1}{\sqrt{|\mathcal{Y}|}}\|\pi - \tilde{\pi}\|_1\right)^2 \\
&= \frac{\xi}{2|\mathcal{Y}|}\|\pi - \tilde{\pi}\|_1^2
\end{aligned}$$

where the second-to-last line follows from $\|\cdot\|_1 \leq |\mathcal{Y}|^{1/2}\|\cdot\|_2$. It follows that his regret in the auxilliary problem satisfies

$$\kappa \geq \frac{1}{T}\sum_{\pi\in\mathcal{F}_1} n_F \frac{\xi}{2|\mathcal{Y}|} d_1\left(\pi, \hat{\pi}_F\right)^2$$

where, implicitly, F is the forecast context such that $\pi \in F$. Take the square root of both sides of this inequality:

$$\begin{aligned}
\sqrt{\kappa} &\geq \sqrt{\frac{1}{T}\sum_{\pi\in\mathcal{F}_1} n_F \frac{\xi}{2|\mathcal{Y}|} d_1\left(\pi, \hat{\pi}_F\right)^2} \\
&\geq \frac{1}{\sqrt{T}} \cdot \frac{1}{\sqrt{T}} \sum_{\pi\in\mathcal{F}_1} n_F \sqrt{\frac{\xi}{2|\mathcal{Y}|}} d_1\left(\pi, \hat{\pi}_F\right) \\
&\geq \sqrt{\frac{\xi}{2|\mathcal{Y}|}} \cdot \frac{1}{T} \sum_{\pi\in\mathcal{F}_1} n_F d_1\left(\pi, \hat{\pi}_F\right)
\end{aligned}$$

where the first line is the l^2 norm of a vector with T entries, the second line is the l^1 norm of that same vector, and the inequality follows from $\|\cdot\|_1 \leq T^{1/2}\|\cdot\|_2$. Collapse these inequalities and rearrange terms to obtain the desired result. □

If we use the quadratic scoring rule, this lemma implies

$$\mathrm{E}\left[\frac{1}{T}\sum_{\pi\in\mathcal{F}_1} n_F d_1(\pi, \hat{\pi}_F)\right] \leq \sqrt{|\mathcal{Y}||S_{\mathcal{F}}|\sqrt{\frac{2\log|S_{\mathcal{F}}|}{T}} + 2|\mathcal{Y}|\delta_{\mathcal{F}}} \tag{8}$$

To optimize this bound, up to log factors, set $\delta_{\mathcal{F}} = \left(\frac{1}{T}\right)^{\frac{1}{2|\mathcal{Y}|+2}}$, assuming $|S_{\mathcal{F}}| = \left(\left(\frac{1}{\delta_{\mathcal{F}}}\right)^{|\mathcal{Y}|}\right)$.

D Generalized Results

Upper bounds on CIR constitute our rationality assumptions for the agent. However, our results also rely on informational assumptions. Sections 4, 5, and 6 consider environments that differ primarily by how "informed" the agent appears, relative to the principal. In all three cases, however, we require the agent to be at least as informed as the principal. What is the principal's information? Recall that our mechanisms σ^* will be forecast mechanisms (**??**). A calibrated learning algorithm – which we specify later on – will produce a sequence of forecasts $\pi_1, \dots, \pi_T$. It is possible that these forecasts will become correlated with the state, e.g. if there is a trend in the data. We do not rule this out; however, if our forecasts inadvertently pick up useful information, this information should be available to the agent as well (either implicitly or because we publish π_t along with p_t).

The notion of forecastwise regret (and forecastwise CIR) formalizes what we mean by the principal's "information" being available to the agent. The agent's benchmark includes the principal's forecast as additional context. Formally, define the forecast space $\mathcal{F} = \Delta(\mathcal{Y})$. Fix a small constant $\delta_{\mathcal{F}} > 0$ and consider a finite partition $S_{\mathcal{F}}$ of $\mathcal{F}$ where $\pi, \tilde{\pi} \in F \in S_{\mathcal{F}}$ implies $d_\infty(\pi, \tilde{\pi}) \leq \delta_{\mathcal{F}}$. Let $\mathcal{F}_1 \subseteq \mathcal{F}$ contain a single distribution $\pi \in F$ for every $F \in S_{\mathcal{F}}$.

Definition 16. *Let the information partition combine the forecast and CIR context, i.e.*

$$\mathcal{I} = S_{\mathcal{F}} \times \left(S_{\mathcal{R}}\right)^{\Sigma}$$

and let the information $I_t \in \mathcal{I}$ in period t be the unique set that satisfies

$$\left(\pi_t, r_t^*, (r_t^p)_{p \in \mathcal{P}_0}\right) \in I_t$$

Definition 17 (FCIR)**.** *The agent's* forecastwise CIR *relative to a modification rule $h : \mathcal{I} \to \mathcal{R}$ is*

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

The FCIR relative to the best-in-hindsight modification rule is $\mathrm{FCIR} = \max_{h:\mathcal{I}\to\mathcal{R}} \mathrm{FCIR}(h)$.

To state our assumption, we need to define a forecastwise version of ER, just as we defined a forecastwise version of CIR at the end of section 3. Let the forecast context $F_t \in S_{\mathcal{F}}$ in period t be the unique set that satisfies $\pi_t \in F_t$.

Definition 18 (FER)**.** *The agent's* forecastwise external regret *relative to a strategy $h : S_{\mathcal{F}} \to \mathcal{R}$ is*

$$\mathrm{FER}(h) = \frac{1}{T} \sum_{t=1}^{T} \left(U(h(F_t), p_t, y_t) - U(r_t, p_t, y_t)\right)$$

The FER relative to the best-in-hindsight strategy is $\mathrm{FER} = \max_{h:\mathcal{F}\to\mathcal{R}} \mathrm{FER}(h)$.

Theorem 4. *Assume regularity* (1) *and ϵ-bounded FCIR* (4)*. There exists a nonresponsive mechanism σ^* parameterized by the agent's learner L and a constant $\bar{\epsilon} > 0$ such that*

1. *The principal's regret is bounded, i.e.*

$$\mathrm{E}_{\sigma^*}[\mathrm{PR}] \leq \frac{1}{T}\sum_{I\in\mathcal{I}} n_I \Delta(\hat{\pi}_I, \bar{\epsilon}) + \frac{1}{\bar{\epsilon}}\left(O(\epsilon) + \tilde{O}\left(T^{-1/4}\sqrt{|\mathcal{Y}||S_F||S_R|^{(|\Sigma_0|+|S_P|)/2}}\right) + O\left(\delta_F^{1/2}\right) + O(\delta_P)\right)$$

Assumption 9 (Alignment)**.** *The stage game is* (ϵ, M_1, M_2)*-aligned if, for all signals* γ*,*

$$\underbrace{\left(\phi_{p^*(\pi,\epsilon)}(\pi,\epsilon) - \alpha_{p^*(\pi,\epsilon)}(\pi,\gamma,\epsilon)\right)}_{\text{maximum downside of } \gamma \text{ for the principal}} \leq M_1 \underbrace{\max_{r,r_J\in\mathcal{R}} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{J\sim\tilde{\gamma}(\cdot,y)}\left[U(r_J, p^*(\pi,\epsilon), y)\right] - U(r, p^*(\pi,\epsilon), y)\right]}_{\text{usefulness of } \gamma \text{ to the agent}} + M_2$$

and, for all policies $p \in \mathcal{P}_0$*,*

$$\underbrace{\left(\beta_p(\pi,\gamma,\epsilon) - \phi_{p^*(\pi,\epsilon)}(\pi,\epsilon)\right)}_{\text{maximum upside of } \gamma \text{ for the principal}} \leq M_1 \underbrace{\max_{r,r_J\in\mathcal{R}} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{J\sim\tilde{\gamma}(\cdot,y)}\left[U(r_J, p, y)\right] - U(r, p, y)\right]}_{\text{usefulness of } \gamma \text{ to the agent}} + M_2$$

Theorem 5. *Assume regularity* (1)*,* ϵ*-bounded FCIR* (4)*,* $\tilde{\epsilon}$*-lower-bounded FER* (5)*, and* $(\bar{\epsilon}, M_1, M_2)$*-alignment* (9)*. There exists a nonresponsive mechanism* σ^* *parameterized by* $\bar{\epsilon}$ *such that*

1. *The principal's regret is bounded, i.e.*

$$\begin{aligned}\mathrm{E}_{\sigma^*}[\mathrm{PR}] \leq &\frac{1}{T}\sum_{F\in S_F} n_F \Delta(\hat{\pi}_F, \bar{\epsilon}) + \frac{1}{\bar{\epsilon}}\left(O(\epsilon) + \tilde{O}\left(T^{-1/4}\sqrt{|\mathcal{Y}||S_F|}\right) + O\left(\sqrt{\delta_F}\right) + O(\delta_P)\right)\\ &+ O(\tilde{\epsilon}) + M_1\left(O(\tilde{\epsilon}) + O(\epsilon) + \tilde{O}\left(T^{-1/4}\sqrt{|\mathcal{Y}||S_F|}\right) + O\left(\sqrt{\delta_F}\right) + O(\delta_P)\right) + O(M_2)\end{aligned}$$

Theorem 6. *Assume regularity* (1) *and* ϵ*-bounded FCIR* (4)*. There exists a nonresponsive mechanism* σ^* *parameterized by a constant* $\bar{\epsilon} > 0$ *such that*

1. *The principal's regret is bounded, i.e.*

$$\mathrm{E}_{\sigma^*}[\mathrm{PR}] \leq \frac{1}{T}\sum_{F\in S_F} n_F \Delta(\hat{\pi}_F, \bar{\epsilon}) + \frac{1}{\bar{\epsilon}}\left(O(\epsilon) + \tilde{O}\left(T^{-1/4}\sqrt{|\mathcal{Y}||S_F|}\right) + O\left(\delta_F^{1/2}\right) + O(\delta_P)\right)$$

E Omitted Proofs

E.1 Proof of Propositions ?? and ??

Recall that a policy p_t in period t can affect the agent's behavior μ_τ in period $\tau > t$. This raises the prospect that a mistake today can cause irreversible damage to the principal's average utility. By definition, the principal will regret that mistake. This would make the principal's problem infeasible, in that he cannot guarantee low regret for himself.

Generally-speaking, regret bounds can bypass this problem if they restrict how much the agent's response r_t depends on the policy history $p_{1:t-1}$. This is reasonable a priori, since the policy history

$p_{1:t-1}$ appears irrelevant to the agent's problem. It neither affects nor predicts the state y_t, except through its dependence on $y_{1:t-1}$. It is not needed as a predictor of the policy p_t because the agent observes p_t before choosing a response.[20] For these reasons, it seems that the agent can only make herself worse off by allowing irrelevant variables like $p_{1:t-1}$ to affect her response r_t. This would be true if our notion of good performance overall had clear implications for behavior in each period, so that unecessary variation in behavior implies a departure from optimality. Unfortunately, there are various kinds of behavior that obtain low ER. The agent can easily switch between these behaviors while still satisfying no-ER. In the process, she can cause substantial benefit or harm to the principal.

To clarify the problem, we present several examples of learners that we regard as pathological. These are implicit counterexamples to the proposition that no-ER constraints are sufficient for no-regret mechanism design.

Our counterexamples are closely related to the pathological phenomenon of "superefficiency" in statistics. Suppose we are trying to estimate the mean θ of the normal random variable $X \sim N(\theta, 1)$, given an i.i.d. random sample $X_1, \ldots, X_n$. Our objective is to minimize the mean square error, but this depends on the parameter θ. A typical solution is the maximum likelihood estimator (MLE), which in this case outputs the sample mean $n^{-1} \sum_{i=1}^{n} X_i$. For reasons that are unimportant to our discussion, MLE is considered "efficient". However, it is easy to find an estimator that outperforms MLE. For example, a wild-ass guess (WAG) ignores the data and outputs θ^*. If it happens to be the case that $\theta = \theta^*$ then this estimator is optimal.

In the following example, we construct a learner that alternates between a WAG-like predictor and a MLE-like predictor depending on a seemingly irrelevant choice by the principal.

Example 3 (Selective Superefficiency)**.** *Consider a learner L that is capable – either by ingenuity or dumb luck – of predicting the state sequence $y_{1:T}$ perfectly. However, the learner uses this ability only selectively, depending on the state y_1 and policy p_1 in the first period. Despite this seemingly irrational behavior, the learner satisfies vanishing external regret.*

Let $P \subsetneq \mathcal{P}$ be a nonempty subset of policies. Let $Y \subsetneq \mathcal{Y}$ be a nonempty subset of states. Let $r*$ be the best-in-hindsight response by time T. That is, consider some $r^* \in \mathcal{F}$ that happens to be best-in-hindsight given the realized state sequence $y_{1:T}$ but will not be best-in-hindsight uniformly over all state sequences. Given $y_{1:T}$, define the learner L as follows:

1. *If $y_1 \in Y$ and $p_1 \in P$ then use response r^**
2. *If $y_1 \in Y$ and $p_1 \notin P$ then use the response that happens to be optimal given y_t.*
3. *If $y_1 \notin Y$ and $p_1 \in P$ then use the response that happens to be optimal given y_t.*
4. *If $y_1 \notin Y$ and $p_1 \notin P$ then use response r^**

[20] For example, if the agent were Bayesian then there would be no dependence on $p_{1:t}$ at all. If the agent used the exponential weights algorithm then there would only be an indirect dependence, since that algorithm depends on the agent's historical payoffs and these, in turn, depend on the policy history.

In cases 1 and 4, the learner follows the best-in-hindsight response and therefore achieves zero regret. In cases 2 and 3, the learner acts optimally ex post and therefore achieves non-positive regret.

Nonetheless, no mechanism can guarantee no-regret for the principal. Suppose the mechanism chooses $p_1 \in P$. If it turns out that $y_1 \in Y$ then the agent will follow π^*. Otherwise, the agent will be superefficient. Were the mechanism to deviate to $p_1 \notin P$, the situation would be reversed. These constitute permanent changes in the agent's behavior. Suppose one type of behavior is "better" for the principal than another. It is always possible in hindsight that the mechanism's first-period policy was the one that led to the "worse" type of behavior.

Can further assumptions rule out this kind of behavior? Again, consider the analogy with statistics. The WAG estimator – always predict θ^* – will perform very poorly in the counterfactual world where $\theta^* \neq \theta$. Formally, this estimator is not "consistent". Similarly, the learner from example 3 does not guarantee vanishing average external regret under counterfactual state sequences. This reflects a peculiar unresponsiveness to the data.

Unfortunately, imposing consistency or no-regret on all sequences does not rule out these kinds of pathologies. Consider Hodge's (superefficient) estimator, which outputs θ^* unless there is sufficient evidence that $\theta \neq \theta^*$. In that case, it outputs the sample mean. If "sufficient evidence" is defined carefully, this estimator will outperform MLE when $\theta = \theta^*$ and will asympotically match MLE otherwise. We can patch up example 3 in a similar way.

Example 4 (Selective Superefficiency, revised)**.** *We want to modify the learner in example 3 to ensure no-regret on all counterfactual state sequences $\tilde{y}_{1:T}$. This is straightforward. In period $t+1$, if the history $\tilde{y}_{1:t}$ matches the presumed sequence $y_{1:t}$ exactly, then proceed as before. Otherwise, follow any no-ER algorithm. This guarantees vanishing average regret as long as the regret from the first period t where $\tilde{y}_{1:t} \neq y_{1:t}$ is bounded – which it is, since our utility functions are bounded.*

Therefore, for any sequence $y_{1:t}$ there is a learner that satisfies no-regret on all sequences, but exhibits the pathological behavior from example 3 on the realized sequence.

Statisticians deal with superefficiency by arguing that it generically fails to occur. That is, any alternative estimator will weakly underperform MLE on Lebesgue-almost all values θ. For example, we can view Hodge's estimator as asymptotically equivalent to MLE whenever $\theta \neq \theta^*$. In our setting, attempting this argument would necessitate a definition of genericity for sequences of states. While we can provide various definitions, none seem especially compelling.[21]

Another natural restriction to impose is that the learner should not outperform the experts. That is, given the sequence $y_{1:T}$, we only consider learners whose regret at period T is non-negative. Clearly, this rules out the learners in examples 3 and 4, which may obtain negative regret in the sequences where it predicts the state perfectly and acts on that information. Unfortunately, this does not rule out the broader phenomenon, as the following example illustrates.

[21]For example, one could assign equal measure to each sequence $y_{1:T}$ in the set $\mathcal{Y}^T$ of all possible sequences. By this measure, the measure of any constant sequence $(y, \dots, y)$ would converge to zero as $T \to \infty$. Yet it does not seem unreasonable a priori that the world should persist in a fixed state. Alternatively, one could assign equal measure to all permutations of a given sequence $y_{1:T}$. This would effectively return us to an i.i.d. setting.

Example 5 (Selective Superinefficiency)**.** *As in example 3, define a learner L that appears capable of predicting the state sequence $y_{1:T}$ perfectly. This learner will continue to use this ability selectively. Moreover, when the learner uses this ability, she does not always use it to her advantage. Instead, with probability $1-q$ she uses it to her own disadvantage. When q is chosen correctly, the learner satisfies zero regret for all mechanisms.*

Let $P \subsetneq \mathcal{P}$ be a nonempty subset of policies. Let $Y \subsetneq \mathcal{Y}$ be a nonempty subset of states. Let r^* be the best-in-hindsight response by time T. Let $r_t^\dagger$ be the response that happens to be optimal for y_t. Let $\tilde{r}_t$ be the response that minimizes the agent's utility when the state is y_t. Given $y_{1:T}$, define the learner L as follows:

1. *If $y_1 \in Y$ and $p_1 \in P$ then use response r^**
2. *If $y_1 \in Y$ and $p_1 \notin P$ then use $r_t^\dagger$ with probability q and $\tilde{r}_t$ with probability $1-q$*
3. *If $y_1 \notin Y$ and $p_1 \in P$ then use $r_t^\dagger$ with probability q and $\tilde{r}_t$ with probability $1-q$*
4. *If $y_1 \notin Y$ and $p_1 \notin P$ then use prior π^**

In cases 1 and 4, the learner follows the best-in-hindsight prior and therefore achieves zero regret. In cases 2 and 3, as long as $\tilde{r}_t$ underperforms r^* in every period t, by continuity there exists a probability q such that the agent achieves zero regret.

This learner now satisfies both an upper bound and a lower bound on regret. Nonetheless, our difficulties remain. Just as in example 3, the first-period policy can cause permanent changes in the agent's behavior. It is always possible in hindsight that the mechanism's first-period policy was the one that led to the "worse" type of behavior.

We can use this example to prove proposition **??** (proposition **??** is a corollary). In the Bayesian persuasion example, let $y_{2:T}$ be drawn i.i.d. where the defendant is guilty with probability $q = 0.5 - \epsilon$ for a very small $\epsilon > 0$. If the principal chooses p_1 correctly, he can persuade the agent to convict with probability near one. Otherwise, the agent convicts with probability near 0.5.

In the contract theory example, let $y_{2:T}$ be be drawn i.i.d. from some distribution where the principal would find it optimal to pay the agent in the stage game, but both states occur with positive probability. If the principal chooses p_1 correctly, he can pay the agent her cost of effort and achieve the first-best outcome (agent works iff working is effective). Otherwise, the principal has to compensate the agent for her cost of effort in states where working is ineffective.

The fundamental problem with the learners in examples 3, 4, 5 is not that they are well-informed. After all, in some settings we might reasonably expect the agent to be better informed than the analyst. The problem is that they fail to consistently and fully exploit the private information that they clearly possess. Bounds on counterfactual internal regret capture this failure to exploit information and rule out these kinds of pathological behaviors.

Example 6. *Returning to example 3, consider two constant mechanisms σ^p and $\sigma^{\tilde{p}}$ where $p \in P$ and $\tilde{p} \notin P$. Regardless of the state sequence $y_{1:T}$, exactly one of these mechanisms (say σ^p) will*

cause the agent to predict the state perfectly while the other will cause the agent to follow the best-in-hindsight prior. The agent's behavior r^p following σ^p will differ across periods y_t, y_τ if and only if $y_t \neq y_\tau$, while the behavior $r^{\tilde{p}}$ following $\sigma^{\tilde{p}}$ remains constant throughout. The vector $(r^p, r^{\tilde{p}})$ will therefore differ across periods y_t, y_τ if and only if $y_t \neq y_\tau$.

If we require the agent to have no-contextual regret where the context is $(r^p, r^{\tilde{p}})$, it is equivalent to requiring her to predict the state perfectly even if the principal uses $\sigma^{\tilde{p}}$. This is essentially the context used to define CIR. The learner guarantees no-CIR under mechanism σ^p, because it predicts the state perfectly. However, it does not predict the state perfectly under mechanism $\sigma^{\tilde{p}}$, so in this case the agent accumulates CIR.

E.2 Proof of Lemmas ?? and ?? and Additional Results

The lemmas in this section will be used repeatedly in the proofs of theorems 1, 2, and 3.

E.2.1 Proof of Lemmas ?? and ??

Lemma **??** states that for any policy p, information structure γ, constants $\epsilon, \tilde{\epsilon} > 0$, and distribution π, we have

$$\alpha_p(\pi, \gamma, \epsilon + \tilde{\epsilon}) \geq \alpha_p(\pi, \gamma, \epsilon) - \frac{\tilde{\epsilon}}{\epsilon} \quad \text{and} \quad \beta_p(\pi, \gamma, \epsilon + \tilde{\epsilon}) \leq \beta_p(\pi, \gamma, \epsilon) + \frac{\tilde{\epsilon}}{\epsilon}$$

Note that lemma **??** is just a special case where the information structure γ is uninformative. To prove this, define

$$B(\pi, \gamma, \epsilon) = \left\{ \mu_J \in \Delta(\mathcal{R}) \mid \epsilon \geq \max_{\tilde{r}_J \in \mathcal{R}} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{J\sim\gamma(\cdot,y)}\left[U(\tilde{r}_J, p, y) - \mathrm{E}_{r\sim\mu_J}[U(r, p, y)]\right]\right] \right\}$$

and recall that

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

Note that $\alpha_p(\pi, \gamma, \epsilon)$ is decreasing and convex in ϵ. Convexity follows from the fact that $\mu_J \in B(\pi, \gamma, \epsilon)$ and $\tilde{\mu}_J \in B(\pi, \gamma, \tilde{\epsilon})$ implies $\lambda\mu_J + (1-\lambda)\tilde{\mu}_J \in B(\pi, \gamma, \lambda\epsilon + (1-\lambda)\tilde{\epsilon})$. Therefore,

$$\alpha_p(\pi, \gamma, \lambda\epsilon + (1-\lambda)\tilde{\epsilon}) \leq \lambda\alpha_p(\pi, \gamma, \epsilon) + (1-\lambda)\alpha_p(\pi, \gamma, \tilde{\epsilon})$$

Consider any supporting line of α_p at ϵ. It is bounded above by α_p, by definition. Therefore, its slope is at most

$$\frac{\alpha_p(\pi, \gamma, 0) - \alpha_p(\pi, \gamma, \epsilon)}{\epsilon} \leq \frac{1}{\epsilon}$$

since α_p is bounded in the unit interval by our regularity assumption. Therefore, the supporting line will underestimate $\alpha_p(\pi, \gamma, \epsilon + \tilde{\epsilon})$ by at most $\tilde{\epsilon}/\epsilon$ and at least zero. This implies our bound. The argument for β_p is analogous after we observe that it is increasing and concave in ϵ.

E.2.2 Bounds for Misspecified Distributions

The following lemma states that the principal's worst-case utility α_p is not too sensitive to changes in the distribution, for any fixed policy p.

Lemma 4. *For any policy p, information structure γ, constant $\epsilon > 0$, and distributions $\pi, \tilde{\pi}$, we have*

$$\alpha_p(\pi, \gamma, \epsilon) \geq \alpha_p(\tilde{\pi}, \gamma, \epsilon) - \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} - d_1(\pi, \tilde{\pi})$$

Proof. Note that $B(\pi, \gamma, \epsilon) \subseteq B(\tilde{\pi}, \gamma, \epsilon + 2d_1(\pi, \tilde{\pi}))$ since (1) for any $\tilde{r}_J$,

$$\mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{J\sim\gamma(\cdot,y)}\left[U(\tilde{r}_J, p, y)\right]\right] \geq \mathrm{E}_{y\sim\tilde{\pi}}\left[\mathrm{E}_{J\sim\gamma(\cdot,y)}\left[U(\tilde{r}_J, p, y)\right]\right] - d_1(\pi, \tilde{\pi})$$

and (2), for any μ_J,

$$\mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{J\sim\gamma(\cdot,y)}\left[\mathrm{E}_{r\sim\mu_J^*}[U(r, p, y)]\right]\right] \leq \mathrm{E}_{y\sim\tilde{\pi}}\left[\mathrm{E}_{J\sim\gamma(\cdot,y)}\left[\mathrm{E}_{r\sim\mu_J^*}[U(r, p, y)]\right]\right] + d_1(\pi, \tilde{\pi})$$

Therefore,

$$\begin{aligned}
\alpha_p(\pi, \gamma, \epsilon) &\geq \min_{\mu_J \in B(\tilde{\pi}, \gamma, \epsilon + 2d_1(\pi, \tilde{\pi}))} \mathrm{E}_{y\sim\pi}\left[\mathrm{E}_{J\sim\gamma(\cdot,y)}\left[\mathrm{E}_{r\sim\mu_J}[V(r, p, y)]\right]\right] \\
&\geq \min_{\mu_J \in B(\tilde{\pi}, \gamma, \epsilon + 2d_1(\pi, \tilde{\pi}))} \mathrm{E}_{y\sim\tilde{\pi}}\left[\mathrm{E}_{J\sim\gamma(\cdot,y)}\left[\mathrm{E}_{r\sim\mu_J}[V(r, p, y)]\right]\right] - d_1(\pi, \tilde{\pi}) \\
&= \alpha_p(\tilde{\pi}, \gamma, \epsilon + 2d_1(\pi, \tilde{\pi})) - d_1(\pi, \tilde{\pi}) \\
&\geq \alpha_p(\tilde{\pi}, \gamma, \epsilon) - \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} - d_1(\pi, \tilde{\pi})
\end{aligned}$$

□

The following lemma states that the ϵ-robust policy for a distribution $\tilde{\pi}$ that is near the true distribution π will perform almost as well as the ϵ-robust policy for the true distribution π.

Lemma 5. *For any $\epsilon > 0$ and distributions $\pi, \tilde{\pi}$, we have*

$$\alpha_{p^*(\tilde{\pi}, \epsilon)}(\pi, \epsilon) \geq \alpha_{p^*(\pi, \epsilon)}(\pi, \epsilon) - \frac{4d_1(\pi, \tilde{\pi})}{\epsilon} - 2d_1(\pi, \tilde{\pi})$$

Proof. First, observe that

$$\begin{aligned}
\alpha_{p^*(\pi, \epsilon)}(\pi, \epsilon) &\leq \alpha_{p^*(\pi, \epsilon)}(\tilde{\pi}, \epsilon) + \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} + d_1(\pi, \tilde{\pi}) \\
&\leq \alpha_{p^*(\tilde{\pi}, \epsilon)}(\tilde{\pi}, \epsilon) + \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} + d_1(\pi, \tilde{\pi})
\end{aligned}$$

Next, observe that

$$\alpha_{p^*(\tilde{\pi}, \epsilon)}(\tilde{\pi}, \epsilon) \leq \alpha_{p^*(\tilde{\pi}, \epsilon)}(\pi, \epsilon) + \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} + d_1(\pi, \tilde{\pi})$$

Collapse these inequalities to obtain the desired result. □

The following lemma states that the ϵ-informationally-robust policy for a distribution $\tilde{\pi}$ that is near the true distribution π will provide a similar guarantee against the worst-case information structure γ as the ϵ-informationally-robust policy for the true distribution π.

Lemma 6. *For any $\epsilon > 0$ and distributions $\pi, \tilde{\pi}$, we have*

$$\inf_{\gamma} \alpha_{p^\dagger(\pi,\epsilon)}(\pi, \gamma, \epsilon) \leq \inf_{\gamma} \alpha_{p^\dagger(\tilde{\pi},\epsilon)}(\pi, \gamma, \epsilon) + \frac{4d_1(\pi, \tilde{\pi})}{\epsilon} + 2d_1(\pi, \tilde{\pi})$$

Proof. First, observe that

$$\alpha_{p^\dagger(\pi,\epsilon)}(\pi, \gamma, \epsilon) \leq \alpha_{p^\dagger(\pi,\epsilon)}(\tilde{\pi}, \gamma, \epsilon) + \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} + d_1(\pi, \tilde{\pi})$$

which implies

$$\begin{aligned}\inf_{\gamma} \alpha_{p^\dagger(\pi,\epsilon)}(\pi, \gamma, \epsilon) &\leq \inf_{\gamma} \alpha_{p^\dagger(\pi,\epsilon)}(\tilde{\pi}, \gamma, \epsilon) + \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} + d_1(\pi, \tilde{\pi}) \\ &\leq \inf_{\gamma} \alpha_{p^\dagger(\tilde{\pi},\epsilon)}(\tilde{\pi}, \gamma, \epsilon) + \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} + d_1(\pi, \tilde{\pi})\end{aligned}$$

Next, observe that

$$\alpha_{p^\dagger(\tilde{\pi},\epsilon)}(\tilde{\pi}, \gamma, \epsilon) \leq \alpha_{p^\dagger(\tilde{\pi},\epsilon)}(\pi, \gamma, \epsilon) + \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} + d_1(\pi, \tilde{\pi})$$

which implies

$$\inf_{\gamma} \alpha_{p^\dagger(\tilde{\pi},\epsilon)}(\tilde{\pi}, \gamma, \epsilon) \leq \inf_{\gamma} \alpha_{p^\dagger(\tilde{\pi},\epsilon)}(\pi, \gamma, \epsilon) + \frac{2d_1(\pi, \tilde{\pi})}{\epsilon} + d_1(\pi, \tilde{\pi})$$

Collapse these inequalities to obtain the desired result. □

E.3 Proof of Theorem 4

Assume access to a forecast $\pi_t \in \Delta(\mathcal{Y})$ for every period t. We will define this later. In period t, the mechanism computes the policy $p^*(\pi_t, \bar{\epsilon})$ that maximizes the worst-case payoff in the $\bar{\epsilon}$-robust stage game, treating the forecast π_t as the common prior. That is,

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

The mechanism chooses p_t as follows. Let P be the unique policy context that includes the policy $p^*(\pi_t, \bar{\epsilon}) \in P$. Let $p_t = p_P$, where $p_P \in \mathcal{P}_1$ is the representative element of P.

We will refer to the average regret accumulated in each forecast context F, i.e.

$$\epsilon_F = \max_{r \in \mathcal{R}} \frac{1}{n_F} \sum_{t \in F} \big(U(r, p_F, y_t) - U(r_t, p_F, y_t) \big)$$

where $p_F = p_P$ for the unique policy context P associated with forecast context F. We will also refer to the average regret accumulated in each information context $I \in \mathcal{I}$, i.e.

$$\epsilon_I = \max_{r \in \mathcal{R}} \frac{1}{n_I} \sum_{t \in I} \big(U(r, p_I, y_t) - U(r_t, p_I, y_t) \big)$$

where $p_I = p_F$ for the unique forecast context F associated with information I. Note that

$$\epsilon_F = \frac{1}{n_F} \sum_{I \in \mathcal{I}_F} n_I \epsilon_I = \frac{1}{n_F} \sum_{I \in \mathcal{I}_F} \max_{r \in \mathcal{R}} \frac{1}{n_I} \sum_{t \in I} \big(U(r, p_F, y_t) - U(r_t, p_F, y_t) \big)$$

The next two lemmas imply an upper bound on the principal's regret in terms of the quantity

$$\iota \geq \frac{1}{T} \sum_{t=1}^{T} d_1(\pi_t, \hat{\pi}_I)$$

that measures the discrepancy between the forecast π_t and the empirical distribution $\hat{\pi}_I$ conditioned on the agent's information I. Lemma 7 is a lower bound on the principal's payoff under σ^*. Lemma 8 is an upper bound on the his payoff under any constant $\sigma^p \in \Sigma_0$.

Lemma 7. *Suppose the principal runs σ^*. Then*

$$\frac{1}{T} \sum_{I \in \mathcal{I}} \sum_{t \in I} V(r_t, p_t, y_t) \geq \max_{p \in \mathcal{P}} \alpha_p(\hat{\pi}_I, \bar{\epsilon}) - \left(\frac{\epsilon + 4\iota + K^U_{\mathcal{R}} \delta_{\mathcal{R}} + 2K^U_{\mathcal{P}} \delta_{\mathcal{P}}}{\bar{\epsilon}} \right) - \big(2\iota + K^V_{\mathcal{R}} \delta_{\mathcal{R}} + K^V_{\mathcal{P}} \delta_{\mathcal{P}} \big)$$

Proof. Let r_I be a representative element in the response context R associated with information I under mechanism σ^*. By regularity,

$$\begin{aligned} \epsilon_I &\geq \max_{r \in \mathcal{R}} \frac{1}{n_I} \sum_{t \in I} \big(U(r, p_I, y_t) - U(r_I, p_I, y_t) \big) - K^U_{\mathcal{R}} \delta_{\mathcal{R}} \\ &= \max_{r \in \mathcal{R}} \mathrm{E}_{y \sim \hat{\pi}_I} \big[U(r, p_I, y) - U(r_I, p_I, y) \big] - K^U_{\mathcal{R}} \delta_{\mathcal{R}} \end{aligned}$$

It follows, by regularity and definition of α, that

$$\begin{aligned} \frac{1}{n_I} \sum_{t \in I} V(r_t, p_I, y_t) &\geq \mathrm{E}_{y \sim \hat{\pi}_I} \big[V(r_I, p_I, y) \big] - K^V_{\mathcal{R}} \delta_{\mathcal{R}} \\ &\geq \alpha_{p_I}(\hat{\pi}_I, \epsilon_I + K^U_{\mathcal{R}} \delta_{\mathcal{R}}) - K^V_{\mathcal{R}} \delta_{\mathcal{R}} \end{aligned}$$

Summing over information $I \in \mathcal{I}$ and using lemma 1, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{t\in I} V(r_t, p_I, y_t) &\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{t\in I} \alpha_{p_I}(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}}) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} \\
&\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{t\in I} \alpha_{p^*(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} - K_{\mathcal{P}}^V\delta_{\mathcal{P}} \\
&\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{t\in I} \left(\alpha_{p^*(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) - \frac{\epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} - K_{\mathcal{P}}^V\delta_{\mathcal{P}} \\
&\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{t\in I} \alpha_{p^*(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) - \left(\frac{\epsilon + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} - K_{\mathcal{P}}^V\delta_{\mathcal{P}}
\end{aligned}$$

Focus on the first term, i.e.

$$\begin{aligned}
\frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{t\in I} \alpha_{p^*(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) &\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{t\in I} \left(\alpha_{p^*(\hat{\pi}_I,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) - \frac{4d_1(\pi_t, \hat{\pi}_I)}{\bar{\epsilon}} - 2d_1(\pi_t, \hat{\pi}_I)\right) \\
&\geq \frac{1}{T}\sum_{I\in\mathcal{I}} n_I \alpha_{p^*(\hat{\pi}_I,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) - \frac{4\imath}{\bar{\epsilon}} - 2\imath
\end{aligned}$$

Collapsing these inequalities gives us the desired bound. □

Lemma 8. *Suppose the principal runs some constant mechanism $\sigma^p \in \Sigma_0$. Then*

$$\frac{1}{T}\sum_{t=1}^{T} V(r_t, p, y_t) \leq \frac{1}{T}\sum_{I\in\mathcal{I}} n_I \max_{\tilde{p}\in\mathcal{P}} \beta_{\tilde{p}}(\hat{\pi}_I, \bar{\epsilon}) + \left(\frac{\epsilon + K_{\mathcal{R}}^U\delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K_{\mathcal{R}}^V\delta_{\mathcal{R}}$$

Proof. Let r_I be a representative element in the response context R associated with information I under σ^p. By regularity,

$$\begin{aligned}
\epsilon_I &\geq \max_{r\in\mathcal{R}} \frac{1}{n_I}\sum_{t\in I}\Big(U(r, p_I, y_t) - U(r_I, p_I, y_t)\Big) - K_{\mathcal{R}}^U\delta_{\mathcal{R}} \\
&= \max_{r\in\mathcal{R}} \mathrm{E}_{y\sim\hat{\pi}_I}\Big[U(r, p_I, y) - U(r_I, p_I, y)\Big] - K_{\mathcal{R}}^U\delta_{\mathcal{R}}
\end{aligned}$$

It follows, by regularity and definition of β, that

$$\begin{aligned}
\frac{1}{n_I}\sum_{t\in I} V(r_t, p, y_t) &\leq \mathrm{E}_{y\sim\hat{\pi}_I}\Big[V(r_I, p, y)\Big] + K_{\mathcal{R}}^V\delta_{\mathcal{R}} \\
&\leq \alpha_p(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}}) + K_{\mathcal{R}}^V\delta_{\mathcal{R}}
\end{aligned}$$

Summing over information $I \in \mathcal{I}$ and using lemma 1, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{t\in I} V(r_t, p, y_t) &\le \frac{1}{T}\sum_{I\in\mathcal{I}} n_I \beta_p(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U \delta_{\mathcal{R}}) + K_{\mathcal{R}}^V \delta_{\mathcal{R}} \\
&\le \frac{1}{T}\sum_{I\in\mathcal{I}} n_I \left(\beta_p(\hat{\pi}_I, \bar{\epsilon}) + \frac{\epsilon_I + K_{\mathcal{R}}^U \delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K_{\mathcal{R}}^V \delta_{\mathcal{R}} \\
&\le \frac{1}{T}\sum_{I\in\mathcal{I}} n_I \beta_p(\hat{\pi}_I, \bar{\epsilon}) + \left(\frac{\epsilon + K_{\mathcal{R}}^U \delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K_{\mathcal{R}}^V \delta_{\mathcal{R}} \\
&\le \frac{1}{T}\sum_{I\in\mathcal{I}} n_I \max_{\tilde{p}\in\mathcal{P}} \beta_{\tilde{p}}(\hat{\pi}_I, \bar{\epsilon}) + \left(\frac{\epsilon + K_{\mathcal{R}}^U \delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K_{\mathcal{R}}^V \delta_{\mathcal{R}}
\end{aligned}$$

This is the desired bound. □

From these two lemmas, it immediately follows that

$$\mathrm{PR} \le \frac{1}{T}\sum_{I\in\mathcal{I}} n_I \Delta(\hat{\pi}_I, \bar{\epsilon}) + 2\left(\frac{\epsilon + 2\iota + K_{\mathcal{R}}^U \delta_{\mathcal{R}} + K_{\mathcal{P}}^U \delta_{\mathcal{P}}}{\bar{\epsilon}}\right) + \left(2\iota + 2K_{\mathcal{R}}^V \delta_{\mathcal{R}} + K_{\mathcal{P}}^V \delta_{\mathcal{P}}\right)$$

Therefore, to bound the principal's regret, all that remains is to bound ι.

Set $C = S_{\mathcal{R}}^{|\mathcal{P}_0|+|\mathcal{P}_1|}$ where $C_t(\mathcal{P}_1)$ is the vector describing the agent's response contexts R_t under policy history $p^*_{1:t-1}$ and policy choices $p_t \in \mathcal{P}_1$. Note that this is different from the behavior context that we used to define the agent's information, which refers to the response context R_t under policy history $p^*_{1:t}$. Because we are currently designing the mechanism, we cannot refer to p^*_t without attempting to solve a fixed point problem that may not have a solution.

We use the algorithm from appendix **??** to generate π_t, with a modification: run it separately for each context C_t. Adapting equation (8), we obtain

$$\begin{aligned}
\mathrm{E}_{\sigma^*}\left[\frac{1}{T}\sum_{C\in\mathcal{C}}\sum_{F\in\mathcal{F}} n_{F,C} d_1(\pi_t, \hat{\pi}_C)\right] &\le \frac{1}{T}\sum_{C\in\mathcal{C}} n_C \sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{\frac{2\log|\mathcal{F}|}{n_C}} + 2|\mathcal{Y}|\delta_{\mathcal{F}}} \\
&\le \frac{1}{T}\sum_{C\in\mathcal{C}} n_C^{3/4}\sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{2\log|\mathcal{F}|}} + \sqrt{2|\mathcal{Y}|\delta_{\mathcal{F}}} \\
&\le \frac{1}{T}\sum_{C\in\mathcal{C}} \left(\frac{T}{|\mathcal{C}|}\right)^{3/4}\sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{2\log|\mathcal{F}|}} + \sqrt{2|\mathcal{Y}|\delta_{\mathcal{F}}} \\
&= \left(\frac{|\mathcal{C}|}{T}\right)^{1/4}\sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{2\log|\mathcal{F}|}} + \sqrt{2|\mathcal{Y}|\delta_{\mathcal{F}}}
\end{aligned}$$

where $n_{F,C}$ is the number of periods t where $C_t = C$ and $\pi_t \in F$.

Consider any two periods t, τ where $I_t = I_\tau$ but $C_t \neq C_\tau$. Since $I_t = I_\tau$ and information

includes the forecast as context, we know that $\pi_t = \pi_\tau$. Now, consider

$$n_{F_t,C_t} d_1(\pi_t, \hat{\pi}_{C_t}) + n_{F_\tau,C_\tau} d_1(\pi_\tau, \hat{\pi}_{C_\tau})$$

$$= n_{F_t,C_t} d_1(\pi_t, \hat{\pi}_{C_t}) + n_{F_t,C_\tau} d_1(\pi_t, \hat{\pi}_{C_\tau})$$

$$\geq \left(n_{F_t,C_t} + n_{F_t,C_\tau}\right) d_1\left(\pi_t, \frac{1}{n_{F_t,C_t} + n_{F_t,C_\tau}} \left(n_{F_t,C_t} \hat{\pi}_{C_t} + n_{F_t,C_\tau} d_1(\pi_t, \hat{\pi}_{C_\tau})\right)\right)$$

by subadditivity and homogeneity of norms. By continuing this process of combining contexts, we find

$$\frac{1}{T} \sum_{C \in \mathcal{C}} \sum_{F \in \mathcal{F}} n_{F,C} d_1(\pi_t, \hat{\pi}_C) \geq \frac{1}{T} \sum_{I \in \mathcal{I}} n_I d_1(\pi_t, \hat{\pi}_I) = \iota$$

Therefore, the earlier miscalibration bound applies to $\mathrm{E}_{\sigma^*}[\iota]$ as well. Finally, we obtain our bound on the expected principal's regret.

$$\mathrm{E}_{\sigma^*}[\mathrm{PR}] \leq \frac{1}{T} \sum_{I \in \mathcal{I}} n_I \Delta(\hat{\pi}_I, \bar{\epsilon}) + 2\left(\frac{\epsilon + 2\left(\frac{|\mathcal{C}|}{T}\right)^{1/4} \sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{2\log|\mathcal{F}|}} + 2\sqrt{2|\mathcal{Y}|\delta_{\mathcal{F}}} + K^U_{\mathcal{R}}\delta_{\mathcal{R}} + K^U_{\mathcal{P}}\delta_{\mathcal{P}}}{\bar{\epsilon}}\right)$$

$$+ \left(2\left(\frac{|\mathcal{C}|}{T}\right)^{1/4} \sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{2\log|\mathcal{F}|}} + 2\sqrt{2|\mathcal{Y}|\delta_{\mathcal{F}}} + 2K^V_{\mathcal{R}}\delta_{\mathcal{R}} + K^V_{\mathcal{P}}\delta_{\mathcal{P}}\right)$$

E.4 Proof of Theorem 5

Define

$$\hat{\gamma}_P(I, y) = \frac{n_I \hat{\pi}_I(y)}{n_F \hat{\pi}_F(y)} \cdot \mathbf{1}(I \in \mathcal{I}_F)$$

as the empirical information structure conditional on forecast context F. This definition follows from Bayes' rule.

Assume access to a forecast $\pi_t \in \Delta(\mathcal{Y})$ for every period t. We will define this later. In period t, the mechanism computes the policy $p^*(\pi_t, \bar{\epsilon})$ that maximizes the worst-case payoff in the $\bar{\epsilon}$-robust stage game, treating the forecast π_t as the common prior. That is,

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

The mechanism chooses p_t as follows. Let P be the unique policy context that includes the policy $p^*(\pi_t, \bar{\epsilon}) \in P$. Let $p_t = p_P$, where $p_P \in \mathcal{P}_1$ is the representative element of P.

The next two lemmas imply an upper bound on the principal's regret in terms of the quantity

$$\iota \geq \frac{1}{T} \sum_{t=1}^{T} d_1(\pi_t, \hat{\pi}_F)$$

that measures the discrepancy between the forecast π_t and the empirical distribution $\hat{\pi}_F$ conditioned on the forecast context F. Lemma 9 is a lower bound on the principal's payoff under σ^*. Lemma 10 is an upper bound on the his payoff under any constant $\sigma^p \in \Sigma_0$.

Lemma 9. *Suppose the principal runs the mechanism σ^*. Then*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} V(r_t, p_t, y_t) \geq & \frac{1}{T}\sum_{F\in\mathcal{F}} n_F \max_{p\in\mathcal{P}} \alpha_p(\hat{\pi}_F, \bar{\epsilon}) - \left(\frac{\epsilon + 6\iota + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) \\ & - \left(M_1(\epsilon + \tilde{\epsilon} + 2\iota + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}) + M_2 + 3\iota + K_{\mathcal{R}}^V\delta_{\mathcal{R}} + K_{\mathcal{P}}^V\delta_{\mathcal{P}}\right)\end{aligned}$$

Proof. Let r_I be a representative element in the response context R associated with information I under mechanism σ^*. By regularity,

$$\begin{aligned}\epsilon_I &\geq \max_{r\in\mathcal{R}} \sum_{t\in I}\left(U(r, p_I, y_t) - U(r_I, p_I, y_t)\right) - K_{\mathcal{R}}^U\delta_{\mathcal{R}} \\ &= \max_{r\in\mathcal{R}} \mathrm{E}_{y\sim\hat{\pi}_I}\left[U(r, p_I, y) - U(r_I, p_I, y)\right] - K_{\mathcal{R}}^U\delta_{\mathcal{R}}\end{aligned}$$

It follows, by regularity and definition of α, that

$$\begin{aligned}\frac{1}{n_I}\sum_{t\in I} V(r_t, p_I, y_t) &\geq \mathrm{E}_{y\sim\hat{\pi}_I}\left[V(r_I, p_I, y)\right] - K_{\mathcal{R}}^V\delta_{\mathcal{R}} \\ &\geq \alpha_{p_I}(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}}) - K_{\mathcal{R}}^V\delta_{\mathcal{R}}\end{aligned}$$

Summing over information $I \in \mathcal{I}_F$ and using lemma 1, we obtain

$$\begin{aligned}\frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I} V(r_t, p_F, y_t) &\geq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I} \alpha_{p_F}(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}}) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} \\ &\geq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I} \alpha_{p^*(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} - K_{\mathcal{P}}^V\delta_{\mathcal{P}} \\ &\geq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I}\left(\alpha_{p^*(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) - \frac{\epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} - K_{\mathcal{P}}^V\delta_{\mathcal{P}} \\ &= \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I} \alpha_{p^*(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) - \left(\frac{\epsilon_F + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} - K_{\mathcal{P}}^V\delta_{\mathcal{P}}\end{aligned}$$

So far, we have a lower bound for the principal's payoff that nearly matches the principal's worst-case payoff in the stage game if the agent had information structure $\hat{\gamma}_F$ in each forecast context. Furthermore, we know that this information structure cannot be particularly useful to the agent. Define

$$-\tilde{\epsilon}_F = \max_{r\in\mathcal{R}} \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I}\left(U(r, p_F, y_t) - U(r_t, p_F, y_t)\right)$$

as the (possibly negative) regret accumulated in forecast context F relative to the best-in-hindsight

response, rather than the best-in-hindsight function from information to responses. Let π_F be the forecast associated with forecast context F. Note that

$$\begin{aligned}
\epsilon_F + \tilde{\epsilon}_F &= \min_{\tilde{r}\in\mathcal{R}} \frac{1}{n_F} \sum_{I\in\mathcal{I}_F} \max_{r\in\mathcal{R}} \sum_{t\in I} \left(U(r, p_F, y_t) - U(\tilde{r}, p_F, y_t)\right) \\
&\geq \min_{\tilde{r}\in\mathcal{R}} \frac{1}{n_F} \sum_{I\in\mathcal{I}_F} \max_{r\in\mathcal{R}} \sum_{t\in I} \left(U(r, p^*(\pi_F, \bar{\epsilon}), y_t) - U(\tilde{r}, p^*(\pi_F, \bar{\epsilon}), y_t)\right) - 2K_{\mathcal{P}}^U \delta_{\mathcal{P}} \\
&= \min_{r} \max_{r_J} \mathrm{E}_{y\sim\hat{\pi}_F}\left[\mathrm{E}_{J\sim\hat{\gamma}_F(\cdot,y)}\left[U(r_J, p^*(\pi_F, \bar{\epsilon}), y) - U(r, p^*(\pi_F, \bar{\epsilon}), y)\right]\right] - 2K_{\mathcal{P}}^U \delta_{\mathcal{P}} \\
&\geq \min_{r} \max_{r_J} \mathrm{E}_{y\sim\pi_F}\left[\mathrm{E}_{J\sim\hat{\gamma}_F(\cdot,y)}\left[U(r_J, p^*(\pi_F, \bar{\epsilon}), y) - U(r, p^*(\pi_F, \bar{\epsilon}), y)\right]\right] - 2d_1(\pi_F, \hat{\pi}_F) - 2K_{\mathcal{P}}^U \delta_{\mathcal{P}}
\end{aligned}$$

It follows from assumption 9 that

$$M_1(\epsilon_F + \tilde{\epsilon}_F + 2d_1(\pi_F, \hat{\pi}_F) + 2K_{\mathcal{P}}^U \delta_{\mathcal{P}}) + M_2 \geq \alpha_{p^*(\pi_F,\bar{\epsilon})}(\pi_F, \bar{\epsilon}) - \alpha_{p^*(\pi_F,\bar{\epsilon})}(\pi_F, \hat{\gamma}_F, \bar{\epsilon})$$

which can be rewritten as

$$\alpha_{p^*(\pi_F,\bar{\epsilon})}(\pi_F, \hat{\gamma}_F, \bar{\epsilon}) \geq \alpha_{p^*(\pi_F,\bar{\epsilon})}(\pi_F, \bar{\epsilon}) - \left(M_1(\epsilon_F + \tilde{\epsilon}_F + 2d_1(\pi_F, \hat{\pi}_F) + 2K_{\mathcal{P}}^U \delta_{\mathcal{P}}) + M_2\right) \qquad (9)$$

Next, we relate our lower bound on the principal's payoff to the term $\alpha_{p^*(\pi_F,\bar{\epsilon})}(\pi_F, \hat{\gamma}_F, \bar{\epsilon})$. Note that

$$\begin{aligned}
\frac{1}{n_F} \sum_{I\in\mathcal{I}_F} n_I \alpha_{p^*(\pi_F,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) &= \mathrm{E}_{y\sim\hat{\pi}_F}\left[\mathrm{E}_{J\sim\hat{\gamma}(\cdot,y)}\left[\alpha_{p^*(\pi_F,\bar{\epsilon})}(\hat{\pi}_{F|J}, \bar{\epsilon})\right]\right] \\
&\geq \min_{e_J} \mathrm{E}_{y\sim\hat{\pi}_F}\left[\mathrm{E}_{J\sim\hat{\gamma}(\cdot,y)}\left[\alpha_{p^*(\pi_F,\bar{\epsilon})}(\pi_{F|J}, e_J)\right]\right] \quad \text{s.t.} \quad \bar{\epsilon} = \mathrm{E}_{y\sim\pi_F}\left[\mathrm{E}_{J\sim\hat{\gamma}\cdot,y}\left[e_J\right]\right] \\
&= \alpha_{p^*(\pi_F,\bar{\epsilon})}(\hat{\pi}_F, \hat{\gamma}_F, \bar{\epsilon}) \\
&\geq \alpha_{p^*(\pi_F,\bar{\epsilon})}(\pi_F, \hat{\gamma}_F, \bar{\epsilon}) - \frac{2d_1(\pi_F, \hat{\pi}_F)}{\bar{\epsilon}} - d_1(\pi_F, \hat{\pi}_F)
\end{aligned}$$

So combining this with inequality (9) gives

$$\begin{aligned}
&\frac{1}{n_F} \sum_{I\in\mathcal{I}_F} n_I \alpha_{p^*(\pi_F,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) \\
&\geq \alpha_{p^*(\pi_F,\bar{\epsilon})}(\pi_F, \bar{\epsilon}) - \left(M_1(\epsilon_F + \tilde{\epsilon}_F + 2d_1(\pi_F, \hat{\pi}_F) + 2K_{\mathcal{P}}^U \delta_{\mathcal{P}}) + M_2\right) - \left(\frac{2d_1(\pi_F, \hat{\pi}_F)}{\bar{\epsilon}}\right) - d_1(\pi_F, \hat{\pi}_F) \\
&\geq \alpha_{p^*(\hat{\pi}_F,\bar{\epsilon})}(\hat{\pi}_F, \bar{\epsilon}) - \left(M_1(\epsilon_F + \tilde{\epsilon}_F + 2d_1(\pi_F, \hat{\pi}_F) + 2K_{\mathcal{P}}^U \delta_{\mathcal{P}}) + M_2\right) - \left(\frac{6d_1(\pi_F, \hat{\pi}_F)}{\bar{\epsilon}}\right) - 3d_1(\pi_F, \hat{\pi}_F)
\end{aligned}$$

Collapsing these inequalities gives us

$$\frac{1}{n_F}\sum_{t\in F} V(r_t, p_F, y_t) \geq \alpha_{p^*(\hat{\pi}_F,\bar{\epsilon})}(\hat{\pi}_F, \bar{\epsilon}) - \left(\frac{\epsilon_F + 6d_1(\pi_F, \hat{\pi}_F) + K^U_{\mathcal{R}}\delta_{\mathcal{R}} + 2K^U_{\mathcal{P}}\delta_{\mathcal{P}}}{\bar{\epsilon}}\right)$$
$$- \left(M_1(\epsilon_F + \tilde{\epsilon}_F + 2d_1(\pi_F, \hat{\pi}_F) + 2K^U_{\mathcal{P}}\delta_{\mathcal{P}}) + M_2 + 3d_1(\pi_F, \hat{\pi}_F) + K^V_{\mathcal{R}}\delta_{\mathcal{R}} + K^V_{\mathcal{P}}\delta_{\mathcal{P}}\right)$$

Summing over forecast contexts $F \in \mathcal{F}$ gives us the desired result. □

Lemma 10. *Suppose the principal runs some constant mechanism $\sigma^p \in \Sigma_0$. Then*

$$\frac{1}{T}\sum_{t=1}^{T} V(r_t, p, y_t) \leq \frac{1}{T}\sum_{F\in\mathcal{F}} n_F \max_{\tilde{p}\in\mathcal{P}} \beta_{\tilde{p}}(\hat{\pi}_F, \bar{\epsilon}) + \left(\frac{\epsilon + K^U_{\mathcal{R}}\delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + \left(M_1(\epsilon + \tilde{\epsilon}) + M_2 + K^V_{\mathcal{R}}\delta_{\mathcal{R}}\right)$$

Proof. Let r_I be a representative element in the response context R associated with information I under mechanism σ^p. By regularity,

$$\begin{aligned}\epsilon_I &\geq \max_{r\in\mathcal{R}} \sum_{t\in I}\left(U(r, p, y_t) - U(r_I, p, y_t)\right) - K^U_{\mathcal{R}}\delta_{\mathcal{R}} \\ &= \max_{r\in\mathcal{R}} \mathrm{E}_{y\sim\hat{\pi}_I}\left[U(r, p, y) - U(r_I, p, y)\right] - K^U_{\mathcal{R}}\delta_{\mathcal{R}}\end{aligned}$$

It follows, by regularity and definition of β, that

$$\begin{aligned}\frac{1}{n_I}\sum_{t\in I} V(r_t, p, y_t) &\leq \mathrm{E}_{y\sim\hat{\pi}_I}\left[V(r_I, p, y)\right] + K^V_{\mathcal{R}}\delta_{\mathcal{R}} \\ &\leq \beta_p(\hat{\pi}_I, \epsilon_I + K^U_{\mathcal{R}}\delta_{\mathcal{R}}) + K^V_{\mathcal{R}}\delta_{\mathcal{R}}\end{aligned}$$

Summing over information $I \in \mathcal{I}_F$ and using lemma 1, we obtain

$$\begin{aligned}\frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I} V(r_t, p, y_t) &\leq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F} n_I\beta_p(\hat{\pi}_I, \epsilon_I + K^U_{\mathcal{R}}\delta_{\mathcal{R}}) + K^V_{\mathcal{R}}\delta_{\mathcal{R}} \\ &\leq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F} n_I\left(\beta_p(\hat{\pi}_I, \bar{\epsilon}) + \frac{\epsilon_I + K^U_{\mathcal{R}}\delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K^V_{\mathcal{R}}\delta_{\mathcal{R}} \\ &= \frac{1}{n_F}\sum_{I\in\mathcal{I}_F} n_I\beta_p(\hat{\pi}_I, \bar{\epsilon}) + \left(\frac{\epsilon_F + K^U_{\mathcal{R}}\delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K^V_{\mathcal{R}}\delta_{\mathcal{R}}\end{aligned}$$

So far, we have an upper bound for the principal's payoff that nearly matches the principal's worst-case payoff in the stage game if the agent had information structure $\hat{\gamma}_F$ in each forecast context. Furthermore, we know that this information structure cannot be particularly useful to the agent. Define

$$-\tilde{\epsilon}_F = \max_{r\in\mathcal{R}} \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I}\left(U(r, p, y_t) - U(r_t, p, y_t)\right)$$

as the (possibly negative) regret accumulated in forecast context F relative to the best-in-hindsight response, rather than the best-in-hindsight function from information to responses. Let π_F be the forecast associated with forecast context F. Note that

$$\epsilon_F + \tilde{\epsilon}_F = \min_{\tilde{r} \in \mathcal{R}} \frac{1}{n_F} \sum_{I \in \mathcal{I}_F} \max_{r \in \mathcal{R}} \sum_{t \in I} \left(U(r, p, y_t) - U(\tilde{r}, p, y_t) \right)$$

It follows from assumption 9 that

$$M_1(\epsilon_F + \tilde{\epsilon}_F) + M_2 \geq \beta_p(\hat{\pi}_F, \hat{\gamma}_F, \bar{\epsilon}) - \max_{\tilde{p} \in \mathcal{P}} \beta_{\tilde{p}}(\hat{\pi}_F, \bar{\epsilon})$$

which can be rewritten as

$$\beta_p(\hat{\pi}_F, \hat{\gamma}_F, \bar{\epsilon}) \leq \max_{\tilde{p} \in \mathcal{P}} \beta_{\tilde{p}} + M_1(\epsilon_F + \tilde{\epsilon}_F) + M_2$$

Next, we relate our upper bound on the principal's payoff to the term $\beta_p(\hat{\pi}_F, \hat{\gamma}_F, \bar{\epsilon})$. Note that

$$\begin{aligned}
\frac{1}{n_F} \sum_{I \in \mathcal{I}_F} n_I \beta_p(\hat{\pi}_I, \bar{\epsilon}) &= \mathrm{E}_{y \sim \hat{\pi}_F} \left[\mathrm{E}_{J \sim \hat{\gamma}(\cdot, y)} \left[\beta_p(\hat{\pi}_{F|J}, \bar{\epsilon}) \right] \right] \\
&\leq \max_{e_J} \mathrm{E}_{y \sim \hat{\pi}_F} \left[\mathrm{E}_{J \sim \hat{\gamma}(\cdot, y)} \left[\beta_p(\pi_{F|J}, e_J) \right] \right] \quad \text{s.t.} \quad \bar{\epsilon} = \mathrm{E}_{y \sim \pi_F} \left[\mathrm{E}_{J \sim \cdot, \hat{\gamma}} \left[e_J \right] \right] \\
&= \beta_p(\hat{\pi}_F, \hat{\gamma}_F, \bar{\epsilon})
\end{aligned}$$

Collapsing these inequalities gives us

$$\frac{1}{n_F} \sum_{t \in F} V(r_t, p, y_t) \leq \max_{\tilde{p} \in \mathcal{P}} \beta_{\tilde{p}}(\hat{\pi}_F, \bar{\epsilon}) + \left(\frac{\epsilon_F + K_{\mathcal{R}}^U \delta_{\mathcal{R}}}{\bar{\epsilon}} \right) + \left(M_1(\epsilon_F + \tilde{\epsilon}_F) + M_2 + K_{\mathcal{R}}^V \delta_{\mathcal{R}} \right)$$

Summing over forecast contexts $F \in \mathcal{F}$ gives us the desired result. □

From these two lemmas, it immediately follows that

$$\begin{aligned}
\mathrm{PR} \leq & \frac{1}{T} \sum_{F \in \mathcal{F}} n_F \Delta(\hat{\pi}_F, \bar{\epsilon}) + \left(\frac{2\epsilon + 6\iota + 2K_{\mathcal{R}}^U \delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U \delta_{\mathcal{P}}}{\bar{\epsilon}} \right) \\
& + \left(M_1(2\epsilon + 2\tilde{\epsilon} + 2\iota + 2K_{\mathcal{P}}^U \delta_{\mathcal{P}}) + 2M_2 + 3\iota + 2K_{\mathcal{R}}^V \delta_{\mathcal{R}} + K_{\mathcal{P}}^V \delta_{\mathcal{P}} \right)
\end{aligned}$$

Therefore, to bound the principal's regret, all that remains is to bound ι. If we use the algorithm from appendix **??** to generate π_t, this follows directly from equation (8), which states

$$\mathrm{E}_{\sigma^*}[\iota] \leq \sqrt{|\mathcal{Y}||\mathcal{F}| \sqrt{\frac{2 \log |\mathcal{F}|}{T}} + 2|\mathcal{Y}|\delta_{\mathcal{F}}}$$

Finally, we obtain our bound on the expected principal's regret.

$$\begin{aligned}\mathrm{E}_{\sigma^*}[\mathrm{PR}] \leq & \frac{1}{T}\sum_{F\in\mathcal{F}} n_F \Delta(\hat{\pi}_F, \bar{\epsilon}) + \left(\frac{2\epsilon + 6\sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{\frac{2\log|\mathcal{F}|}{T}} + 2|\mathcal{Y}|\delta_{\mathcal{F}}} + 2K_{\mathcal{R}}^{U}\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^{U}\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) \\ & + M_1\left(2\epsilon + 2\tilde{\epsilon} + 2\sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{\frac{2\log|\mathcal{F}|}{T}} + 2|\mathcal{Y}|\delta_{\mathcal{F}}} + 2K_{\mathcal{P}}^{U}\delta_{\mathcal{P}}\right) \\ & + 2M_2 + 3\sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{\frac{2\log|\mathcal{F}|}{T}} + 2|\mathcal{Y}|\delta_{\mathcal{F}}} + 2K_{\mathcal{R}}^{V}\delta_{\mathcal{R}} + K_{\mathcal{P}}^{V}\delta_{\mathcal{P}}\end{aligned}$$

E.5 Proof of Theorem 6

Assume access to a forecast π_t for every period t. We will define this later. The mechanism chooses p_t as follows. Let P be the unique policy context that includes the policy $p^{\dagger}(\pi_t, \bar{\epsilon}) \in P$. Let $p_t = p_P$, where $p_P \in \mathcal{P}_1$ is the representative element of P.

The next two lemmas imply an upper bound on the principal's regret in terms of the quantity

$$\iota \geq \frac{1}{T}\sum_{t=1}^{T} d_1(\pi_t, \hat{\pi}_F)$$

that measures the discrepancy between the forecast π_t and the empirical distribution $\hat{\pi}_F$ conditioned on the forecast context F. Lemma 11 is a lower bound on the principal's payoff under σ^*. Lemma 12 is an upper bound on the his payoff under any constant $\sigma^p \in \Sigma_0$.

Lemma 11. *Suppose the principal runs the mechanism σ^*. Then*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} V(r_t, p_t, y_t) \geq & \frac{1}{T}\sum_{F\in\mathcal{F}} n_F \inf_{\gamma}\max_{p\in\mathcal{P}} \alpha_p(\hat{\pi}_F, \gamma, \bar{\epsilon}) - \left(\frac{\epsilon + 4\iota + K_{\mathcal{R}}^{U}\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^{U}\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) \\ & - \left(2\iota + K_{\mathcal{R}}^{V}\delta_{\mathcal{R}} + K_{\mathcal{P}}^{V}\delta_{\mathcal{P}}\right)\end{aligned}$$

Proof. Let r_I be a representative element in the response context R associated with information I under mechanism σ^*. By regularity,

$$\begin{aligned}\epsilon_I &\geq \max_{r\in\mathcal{R}} \sum_{t\in I}\left(U(r, p_I, y_t) - U(r_I, p_I, y_t)\right) - K_{\mathcal{R}}^{U}\delta_{\mathcal{R}} \\ &= \max_{r\in\mathcal{R}} \mathrm{E}_{y\sim\hat{\pi}_I}\left[U(r, p_I, y) - U(r_I, p_I, y)\right] - K_{\mathcal{R}}^{U}\delta_{\mathcal{R}}\end{aligned}$$

It follows, by regularity and definition of α, that

$$\begin{aligned}\frac{1}{n_I}\sum_{t\in I} V(r_t, p_I, y_t) &\geq \mathrm{E}_{y\sim\hat{\pi}_I}\left[V(r_I, p_I, y)\right] - K_{\mathcal{R}}^V\delta_{\mathcal{R}} \\ &\geq \alpha_{p_I}(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}}) - K_{\mathcal{R}}^V\delta_{\mathcal{R}}\end{aligned}$$

Summing over information $I \in \mathcal{I}_F$, we obtain

$$\begin{aligned}\frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I} V(r_t, p_F, y_t) &\geq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I}\alpha_{p_F}(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}}) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} \\ &\geq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I}\alpha_{p^\dagger(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} - K_{\mathcal{P}}^V\delta_{\mathcal{P}} \\ &\geq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I}\left(\alpha_{p^\dagger(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) - \frac{\epsilon_I + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} - K_{\mathcal{P}}^V\delta_{\mathcal{P}} \\ &= \frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I}\alpha_{p^\dagger(\pi_t,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) - \left(\frac{\epsilon_F + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) - K_{\mathcal{R}}^V\delta_{\mathcal{R}} - K_{\mathcal{P}}^V\delta_{\mathcal{P}}\end{aligned}$$

Focus on the first term, i.e.

$$\begin{aligned}\frac{1}{n_F}\sum_{I\in\mathcal{I}_F} n_I\alpha_{p^\dagger(\pi_F,\bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon}) &= \mathrm{E}_{y\sim\hat{\pi}_F}\left[\mathrm{E}_{J\sim\hat{\gamma}(\cdot,y)}\left[\alpha_{p^\dagger(\pi_F,\bar{\epsilon})}(\hat{\pi}_{F|J}, \bar{\epsilon})\right]\right] \\ &\geq \min_{e_J}\mathrm{E}_{y\sim\hat{\pi}_F}\left[\mathrm{E}_{J\sim\hat{\gamma}(\cdot,y)}\left[\alpha_{p^\dagger(\pi_F,\bar{\epsilon})}(\hat{\pi}_{F|J}, e_J)\right]\right] \quad \text{s.t.} \quad \bar{\epsilon} = \mathrm{E}_{y\sim\hat{\pi}_F}\left[\mathrm{E}_{J\sim\hat{\gamma}\cdot,y}\left[e_J\right]\right] \\ &= \alpha_{p^\dagger(\pi_F,\bar{\epsilon})}(\hat{\pi}_F, \hat{\gamma}_F, \bar{\epsilon}) \\ &\geq \inf_{\gamma}\alpha_{p^\dagger(\pi_F,\bar{\epsilon})}(\hat{\pi}_F, \gamma, \bar{\epsilon}) \\ &\geq \inf_{\gamma}\alpha_{p^\dagger(\hat{\pi}_F,\bar{\epsilon})}(\hat{\pi}_F, \gamma, \bar{\epsilon}) - \frac{4d_1(\pi_F, \hat{\pi}_F)}{\bar{\epsilon}} - 2d_1(\pi_F, \hat{\pi}_F)\end{aligned}$$

Collapsing these inequalities gives us

$$\begin{aligned}\frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I} V(r_t, p_F, y_t) \geq \inf_{\gamma}\alpha_{p^\dagger(\hat{\pi}_F,\bar{\epsilon})}(\hat{\pi}_F, \gamma, \bar{\epsilon}) &- \left(\frac{\epsilon_F + 4d_1(\pi_F, \hat{\pi}_F) + K_{\mathcal{R}}^U\delta_{\mathcal{R}} + 2K_{\mathcal{P}}^U\delta_{\mathcal{P}}}{\bar{\epsilon}}\right) \\ &- \left(2d_1(\pi_F, \hat{\pi}_F) + K_{\mathcal{R}}^V\delta_{\mathcal{R}} + K_{\mathcal{P}}^V\delta_{\mathcal{P}}\right)\end{aligned}$$

Summing over forecast contexts $F \in \mathcal{F}$ gives us the desired result. □

Lemma 12. *Suppose the principal runs some constant mechanism $\sigma^p \in \Sigma_0$. Then*

$$\frac{1}{T}\sum_{t=1}^{T} V(r_t, p, y_t) \leq \sum_{F\in\mathcal{F}} n_F \max_{\tilde{p}\in\mathcal{P}}\beta_{\tilde{p}}(\hat{\pi}_F, \hat{\gamma}_F, \bar{\epsilon}) + \left(\frac{\epsilon + K_{\mathcal{R}}^U\delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K_{\mathcal{R}}^V\delta_{\mathcal{R}}$$

Proof. Let r_I be a representative element in the response context R associated with information I under mechanism σ^p. By regularity,

$$\begin{aligned}\epsilon_I &\geq \max_{r\in\mathcal{R}} \sum_{t\in I} \big(U(r,p,y_t) - U(r_I,p,y_t)\big) - K^U_{\mathcal{R}}\delta_{\mathcal{R}} \\ &= \max_{r\in\mathcal{R}} \mathrm{E}_{y\sim\hat{\pi}_I}\big[U(r,p,y) - U(r_I,p,y)\big] - K^U_{\mathcal{R}}\delta_{\mathcal{R}}\end{aligned}$$

It follows, by regularity and definition of β, that

$$\begin{aligned}\frac{1}{n_I}\sum_{t\in I} V(r_t,p,y_t) &\leq \mathrm{E}_{y\sim\hat{\pi}_I}\big[V(r_I,p,y)\big] + K^V_{\mathcal{R}}\delta_{\mathcal{R}} \\ &\leq \beta_p(\hat{\pi}_I, \epsilon_I + K^U_{\mathcal{R}}\delta_{\mathcal{R}}) + K^V_{\mathcal{R}}\delta_{\mathcal{R}}\end{aligned}$$

Summing over information $I \in \mathcal{I}_F$, we obtain

$$\begin{aligned}\frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I} V(r_t,p,y_t) &\leq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F} n_I\beta_p(\hat{\pi}_I, \epsilon_I + K^U_{\mathcal{R}}\delta_{\mathcal{R}}) + K^V_{\mathcal{R}}\delta_{\mathcal{R}} \\ &\leq \frac{1}{n_F}\sum_{I\in\mathcal{I}_F} n_I\left(\beta_p(\hat{\pi}_I,\bar{\epsilon}) + \frac{\epsilon_I + K^U_{\mathcal{R}}\delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K^V_{\mathcal{R}}\delta_{\mathcal{R}} \\ &= \frac{1}{n_F}\sum_{I\in\mathcal{I}_F} n_I\beta_p(\hat{\pi}_I,\bar{\epsilon}) + \left(\frac{\epsilon_F + K^U_{\mathcal{R}}\delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K^V_{\mathcal{R}}\delta_{\mathcal{R}}\end{aligned}$$

Focus on the first term, i.e.

$$\begin{aligned}\frac{1}{n_F}\sum_{I\in\mathcal{I}_F} n_I\beta_p(\hat{\pi}_I,\bar{\epsilon}) &= \mathrm{E}_{y\sim\hat{\pi}_F}\big[\mathrm{E}_{J\sim\hat{\gamma}(\cdot,y)}\big[\beta_p(\hat{\pi}_{F|J},\bar{\epsilon})\big]\big] \\ &\leq \max_{e_J} \mathrm{E}_{y\sim\hat{\pi}_F}\big[\mathrm{E}_{J\sim\hat{\gamma}(\cdot,y)}\big[\beta_p(\hat{\pi}_{F|J},e_J)\big]\big] \quad \text{s.t.} \quad \bar{\epsilon} = \mathrm{E}_{y\sim\hat{\pi}_F}\big[\mathrm{E}_{J\sim\hat{\gamma}\cdot,y}\big[e_J\big]\big] \\ &= \beta_p(\hat{\pi}_F,\hat{\gamma}_F,\bar{\epsilon}) \\ &\leq \max_{\tilde{p}\in\mathcal{P}} \beta_{\tilde{p}}(\hat{\pi}_F,\hat{\gamma}_F,\bar{\epsilon})\end{aligned}$$

Collapsing these inequalities gives us

$$\frac{1}{n_F}\sum_{I\in\mathcal{I}_F}\sum_{t\in I} V(r_t,p,y_t) \leq \max_{\tilde{p}\in\mathcal{P}} \beta_{\tilde{p}}(\hat{\pi}_F,\hat{\gamma}_F,\bar{\epsilon}) + \left(\frac{\epsilon_F + K^U_{\mathcal{R}}\delta_{\mathcal{R}}}{\bar{\epsilon}}\right) + K^V_{\mathcal{R}}\delta_{\mathcal{R}}$$

Summing over forecast contexts $F \in \mathcal{F}$ gives us the desired result. □

From these two lemmas, it immediately follows that

$$\text{PR} \leq \sum_{F \in \mathcal{F}} n_F \nabla(\hat{\pi}_F, \bar{\epsilon}) + 2\left(\frac{\epsilon + 2\iota + K_{\mathcal{R}}^U \delta_{\mathcal{R}} + K_{\mathcal{P}}^U \delta_{\mathcal{P}}}{\bar{\epsilon}}\right) + \left(2\iota + 2K_{\mathcal{R}}^V \delta_{\mathcal{R}} + K_{\mathcal{P}}^V \delta_{\mathcal{P}}\right)$$

Therefore, to bound the principal's regret, all that remains is to bound ι. If we use the algorithm from appendix **??** to generate π_t, this follows directly from equation (8), which states

$$\mathrm{E}_{\sigma^*}[\iota] \leq \sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{\frac{2\log|\mathcal{F}|}{T}} + 2|\mathcal{Y}|\delta_{\mathcal{F}}}$$

Finally, we obtain our bound on the expected principal's regret.

$$\begin{aligned}\mathrm{E}_{\sigma^*}[\text{PR}] \leq & \frac{1}{T}\sum_{F \in \mathcal{F}} n_F \nabla(\hat{\pi}_F, \bar{\epsilon}) + 2\left(\frac{\epsilon + 2\sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{\frac{2\log|\mathcal{F}|}{T}} + 2|\mathcal{Y}|\delta_{\mathcal{F}}} + K_{\mathcal{R}}^U \delta_{\mathcal{R}} + K_{\mathcal{P}}^U \delta_{\mathcal{P}}}{\bar{\epsilon}}\right) \\ & + \left(2\sqrt{|\mathcal{Y}||\mathcal{F}|\sqrt{\frac{2\log|\mathcal{F}|}{T}} + 2|\mathcal{Y}|\delta_{\mathcal{F}}} + 2K_{\mathcal{R}}^V \delta_{\mathcal{R}} + K_{\mathcal{P}}^V \delta_{\mathcal{P}}\right)\end{aligned}$$

E.6 Proof of Theorem 1

We adapt the proof of theorem 4 to prove theorem 1. This will require only relatively minor changes. Let $\hat{\pi}_{I,F}$ denote the empirical distribution among periods $t \in I \cap F$. Let $n_{I,F}$ indicate the number of such periods. Let π_F denote the (unique) forecast associated with forecast context F. Previously, we defined

$$\iota \geq \frac{1}{T}\sum_{t=1}^{T} d_1(\pi_t, \hat{\pi}_I)$$

Now, we define

$$\iota \geq \frac{1}{T}\sum_{t=1}^{T} d_1(\pi_t, \hat{\pi}_{I,F})$$

Begin at the last line of lemma 7, where it says "focus on the first term". Rewrite that first term as

$$\frac{1}{T}\sum_{I \in \mathcal{I}}\sum_{F \in \mathcal{F}}\sum_{t \in I \cap F} \alpha_{p^*(\pi_F, \bar{\epsilon})}(\hat{\pi}_I, \bar{\epsilon})$$

Now we switch $\hat{\pi}_I$ with π_I, i.e. the convex combination of forecasts,

$$\pi_I = \frac{1}{n_I}\sum_{F \in \mathcal{F}} n_{I,F}\pi_F$$

By lemma 4, this gives us

$$\frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{F\in\mathcal{F}}\sum_{t\in I\cap F}\alpha_{p^*(\pi_F,\bar{\epsilon})}(\hat{\pi}_I,\bar{\epsilon})$$

$$\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{F\in\mathcal{F}}\sum_{t\in I\cap F}\left(\alpha_{p^*(\pi_F,\bar{\epsilon})}(\pi_I,\bar{\epsilon})-\frac{2d_1(\pi_I,\hat{\pi}_I)}{\bar{\epsilon}}-d_1(\pi_I,\hat{\pi}_I)\right)$$

Note that every forecast π_F leads to a policy $p^*(\pi_F,\bar{\epsilon})$ that is in the policy context P associated with information I. By assumption **??**,

$$\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{F\in\mathcal{F}}\sum_{t\in I\cap F}\left(\alpha_{p^*(\pi_I,\bar{\epsilon})}(\pi_I,\bar{\epsilon})-\frac{2d_1(\pi_I,\hat{\pi}_I)}{\bar{\epsilon}}-d_1(\pi_I,\hat{\pi}_I)-O(\delta_{\mathcal{P}})\right)$$

Now we apply lemma 4 again,

$$\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{F\in\mathcal{F}}\sum_{t\in I\cap F}\left(\alpha_{p^*(\pi_I,\bar{\epsilon})}(\hat{\pi}_I,\bar{\epsilon})-\frac{4d_1(\pi_I,\hat{\pi}_I)}{\bar{\epsilon}}-2d_1(\pi_I,\hat{\pi}_I)-O(\delta_{\mathcal{P}})\right)$$

and then lemma 5

$$\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{F\in\mathcal{F}}\sum_{t\in I\cap F}\left(\alpha_{p^*(\hat{\pi}_I,\bar{\epsilon})}(\hat{\pi}_I,\bar{\epsilon})-\frac{8d_1(\pi_I,\hat{\pi}_I)}{\bar{\epsilon}}-4d_1(\pi_I,\hat{\pi}_I)-O(\delta_{\mathcal{P}})\right)$$

By the homogeneity and subadditivity of the l_1 norm,

$$d_1(\pi_I,\hat{\pi}_I)\leq\frac{1}{n_I}\sum_{i=1}^{n}n_{I,F}d_1(\pi_{I,F},\hat{\pi}_{I,F})$$

which gives us

$$\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{F\in\mathcal{F}}\sum_{t\in I\cap F}\left(\alpha_{p^*(\hat{\pi}_I,\bar{\epsilon})}(\hat{\pi}_I,\bar{\epsilon})-\frac{8d_1(\pi_{I,F},\hat{\pi}_{I,F})}{\bar{\epsilon}}-4d_1(\pi_{I,F},\hat{\pi}_{I,F})-O(\delta_{\mathcal{P}})\right)$$

$$\geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{F\in\mathcal{F}}\sum_{t\in I\cap F}\left(\alpha_{p^*(\hat{\pi}_I,\bar{\epsilon})}(\hat{\pi}_I,\bar{\epsilon})\right)-\frac{8\imath}{\bar{\epsilon}}-4\imath-O(\delta_{\mathcal{P}})$$

This is essentially where we were by the end of lemma 7, with the addition of an $O(\delta_{\mathcal{P}})$ term and slightly different constants.

Lemma 8 requires no change. The discussion following lemma 8 requires very little change. Find the line that begins with "Consider any two periods". We rewrite as follows. Consider any two periods t,τ where $I_t=I_\tau$ and $F_t=F_\tau$ but $C_t\neq C_\tau$. Since $F_t=F_\tau$ we know that $\pi_t=\pi_\tau$. Now, consider

$$n_{F_t,C_t}d_1(\pi_t,\hat{\pi}_{C_t})+n_{F_\tau,C_\tau}d_1(\pi_\tau,\hat{\pi}_{C_\tau})$$

$$=n_{F_t,C_t}d_1(\pi_t,\hat{\pi}_{C_t})+n_{F_t,C_\tau}d_1(\pi_t,\hat{\pi}_{C_\tau})$$

$$\geq \left(n_{F_t,C_t} + n_{F_t,C_\tau}\right) d_1\left(\pi_t, \frac{1}{n_{F_t,C_t} + n_{F_t,C_\tau}}\left(n_{F_t,C_t}\hat{\pi}_{C_t} + n_{F_t,C_\tau} d_1(\pi_t, \hat{\pi}_{C_\tau})\right)\right)$$

by subadditivity and homogeneity of norms. By continuing this process of combining contexts, we find

$$\frac{1}{T}\sum_{C\in\mathcal{C}}\sum_{F\in\mathcal{F}} n_{F,C} d_1(\pi_t, \hat{\pi}_C) \geq \frac{1}{T}\sum_{I\in\mathcal{I}}\sum_{F\in\mathcal{F}} n_{I,F} d_1(\pi_t, \hat{\pi}_{I,F}) = \iota$$

Therefore, our bound holds except with the addition of an $O(\delta_{\mathcal{P}})$ term and slightly different constants.

E.7 Proof of Theorem 2

Define

$$\pi_P = \frac{1}{n_P}\sum_{F\in\mathcal{F}} n_{P,F}\pi_F$$

It is straightforward to adapt the proof of theorem 5. Replace all reference to $p^*(\pi_F, \bar{\epsilon})$ with $p^*(\pi_P, \bar{\epsilon})$. This changes U and V (and all derived terms, like α) by at most $O(\delta_{\mathcal{P}})$, by assumption **??**. Replace all remaining references of forecast contexts F to policy contexts P. It remains to verify that

$$\iota \geq \frac{1}{T}\sum_{F\in\mathcal{F}}\sum_{t\in F} d_1(\pi_F, \hat{\pi}_F) \geq \frac{1}{T}\sum_{P\in\mathcal{P}}\sum_{t\in F} d_1(\pi_P, \hat{\pi}_P)$$

which follows from the homogeneity and subadditivity of the l_1 norm, and the fact that $\pi_P, \hat{\pi}_P$ are convex combinations of $\pi_F, \hat{\pi}_F$ for $F \subseteq P$.

E.8 Proof of Theorem 3

Define

$$\pi_P = \frac{1}{n_P}\sum_{F\in\mathcal{F}} n_{P,F}\pi_F$$

It is straightforward to adapt the proof of theorem 6. Replace all reference to $p^\dagger(\pi_F, \bar{\epsilon})$ with $p^\dagger(\pi_P, \bar{\epsilon})$. This changes U and V (and all derived terms, like α) by at most $O(\delta_{\mathcal{P}})$, by assumption **??**. Replace all remaining references of forecast contexts F to policy contexts P. It remains to verify that

$$\iota \geq \frac{1}{T}\sum_{F\in\mathcal{F}}\sum_{t\in F} d_1(\pi_F, \hat{\pi}_F) \geq \frac{1}{T}\sum_{P\in\mathcal{P}}\sum_{t\in F} d_1(\pi_P, \hat{\pi}_P)$$

which follows from the homogeneity and subadditivity of the l_1 norm, and the fact that $\pi_P, \hat{\pi}_P$ are convex combinations of $\pi_F, \hat{\pi}_F$ for $F \subseteq P$.