Mechanism Design with a Common Dataset *

Modibo K. Camara†

November 10, 2021

Working Paper

Latest version available here.

Abstract

I propose a new approach to mechanism design: rather than assume a common prior belief, assume access to a common dataset. I restrict attention to incomplete information games where a designer commits to a policy and a single agent responds. I proposed a penalized policy that performs well under weak assumptions on how the agent learns from data. Policies that are too complex, in a precise sense, are penalized because they lead to unpredictable responses by the agent. This approach leads to new insights in models of vaccine distribution, prescription drug approval, performance pay, and product bundling.

Acknowledgements. I am grateful to Eddie Dekel, Jinshuo Dong, Jeff Ely, Jason Hartline, Matt Notowidigdo, Ludvig Sinander, Marciano Siniscalchi, Quitzé Valenzuela-Stookey, Sam Taggart, and audiences at Northwestern, for discussions and comments that improved this paper. All errors are my own.

†Department of Economics, Northwestern University. Email: mcamara@u.northwestern.edu.

Contents

1 Introduction

It is a truism in economics that beliefs are important determinants of behavior. In any number of settings, ranging from vaccine distribution to compensation policy, understanding agent's beliefs can make the difference between a successful policy and one that fails dramatically. Unfortunately, in many instances, the rich behavioral or survey data needed to identify agents' beliefs may not be available. This can limit our ability to provide clear policy recommendations.

While existing modeling techniques can circumvent our ignorance of agents' beliefs, they are widely recognized as imperfect. On one extreme, it is common to assume that the agent shares the policymaker's beliefs or that she knows the data-generating process. These assumptions have been criticized as unrealistic, both in economic theory (e.g. Wilson 1987, Bergemann and Morris 2005) and in econometrics (e.g. Manski 1993, Manski 2004a). On the other extreme lie robust solution concepts that make no assumptions about the agents' beliefs. But these solution concepts lead to unreasonably conservative recommendations in many applications.

This paper proposes a new, less extreme approach: rather than assume a common prior belief, assume access to a common dataset. The high-level idea is straightforward. If the data convincingly demonstrates some fact about the world, the agent should believe that fact. But if there is insufficient data to reach a particular conclusion, the agent's beliefs are undetermined.

I formalize this approach by integrating a model of statistical learning into mechanism design. To develop this approach, I restrict attention to single-agent problems where a policymaker commits to a policy, an agent responds, and payoffs are determined by a state of nature. Both the policymaker and agent have access to an i.i.d. dataset that they can use to learn about the distribution of the state. *Regret bounds* limit how suboptimal the agent's action can be with respect to the true distribution. A regret bound is *feasible* if even an ignorant agent can satisfy it using off-the-shelf learning rules. Typical regret bounds tighten as the sample size grows.

I derive feasible regret bounds and propose *penalized policies* that highlight new tradeoffs for the policymaker. Policy choices can influence the complexity of the agent's learning problem, which in turn affects her regret bound. Policies that are too complex, in a precise sense, can increase the likelihood that the agent makes a mistake, as well as the severity of that mistake. Penalized policies implicitly penalize complexity, by guarding against the worst-case mistake by the agent.

I propose a *data-driven penalized policy* and present the key technical results of the paper. Since the optimal penalized policy depends on the true data-generating process, it is generally not feasible. However, the policymaker can learn from the common dataset, just like the agent. The fact that the policymaker is learning at the same time as the agent poses theoretical challenges that appear to be new to both statistics and economics. Nonetheless, I develop a data-driven penalized policy and characterize its rate of convergence, which is approximately optimal. In the limit as the sample size grows, this policy coincides with the optimal policy under the common prior.

Finally, I show that penalization can lead to new insights through illustrative examples. Specifically, I consider models of vaccine distribution, prescription drug approval, performance pay, and product bundling. This framework captures important dimensions of policy complexity and highlights trade-offs that are obscured by both the common prior and robust solution concepts.

Model. I consider a rich class of incomplete information games where a policymaker commits to a policy and a single agent responds. Payoffs are determined by the policy, the response, and a hidden state of nature. The state is drawn from some unknown distribution. This setup captures a number of classic design problems in economics, like monopoly regulation, Bayesian persuasion, and contract design. Traditionally, these models are solved by assuming the state distribution is common knowledge, or that the policymaker and the agent have common prior beliefs.

Rather than assume a common prior belief, I assume that both the policymaker and the agent have access to a common dataset. This dataset consists of n i.i.d. draws from the state distribution. Both participants have strategies that map the data to actions. If there were only one participant, this would be a standard statistical decision problem (see e.g. Wald 1950, Manski 2004b).

Agent's Behavior. The goal of this paper is to produce a strategy for the policymaker that performs well under reasonable assumptions on the agent's behavior. To formalize these assumptions, I adapt ideas from statistical learning theory.

I impose bounds on the agent's *regret* that vary with sample size and the policymaker's strategy. Regret measures how suboptimal the agent's strategy is, in expectation, according to the true distribution. An agent that knows the true distribution can guarantee zero regret. An agent whose beliefs are only somewhat misspecified will obtain low regret. An agent who fails to learn from the available data, or has deeply misspecified beliefs, will obtain high regret.

I restrict attention to *feasible* regret bounds. These are regret bounds that the agent can satisfy using off-the-shelf heuristics like empirical utility maximization, even if she has no prior knowledge of the true distribution. I do not assume that the agent uses these heuristics. Instead, I assume that she does not underperform these heuristics. She can be Bayesian or non-Bayesian, well-informed or poorly-informed, and still satisfy a feasible regret bound.

I provide sufficient conditions for a regret bound to be feasible by borrowing a central concept from statistical learning theory: *Rademacher complexity*. This measures the complexity of a statistical learning problem. Naturally, an agent facing a more complex learning problem should be more likely to accumulate regret. Since the agent's learning problem is linked to the policy choice, Rademacher complexity captures a form of policy complexity from the agent's perspective.

To complete the specification of the sufficient condition, I introduce a new concept called *sample privacy*. This concept is closely related to differential privacy (Dwork, McSherry, et al. 2006). It measures how aggressively the policymaker makes use of the realized sample. Sample privacy is a bridge between traditional statistical learning problems and economic problems where multiple participants learn from the same data. The agent's learning problem looks more traditional if the policymaker limits how aggressively he uses the available data, and standard tools like Rademacher complexity remain useful.

Penalized Policy. This model formalizes a sense in which complex policies are undesirable. Policies that are more complex or more sensitive to the data lead to looser regret bounds. As a result, the agent's behavior becomes less predictable. For a policymaker that is concerned with worst-case

guarantees, less predictable behavior can only make him worse off.

I propose *penalized policies* as a way to handle this new trade-off. To make the agent's behavior more predictable, a policymaker can set policies that are simpler and less sensitive to the data. But these changes can be costly. A penalized policy balances the advantages of complexity with the disadvantages of unpredictability. More precisely, it evaluates policies according to the policymaker's payoff under the worst-case agent response that respects the regret bound. This implicitly penalizes policies that are too complex for the agent, given the amount of data available.[1]

Data-Driven Penalized Policy The optimal penalized policy can generate useful insights, but it still depends on the true distribution. Fortunately, this paper shows that it is possible to approximate the optimal penalized policy by using the available data.

Specifically, I construct a *data-driven penalized policy*. There are three steps. First, it evaluates the policymaker's expected payoff with respect to the empirical distribution. Second, it adds white noise to the policymaker's payoff, following the exponential mechanism of McSherry and Talwar (2007). Third, it estimates the set of behaviors that satisfy the agent's regret bound, and optimizes against the worst case behavior in this estimated set.

Theorem 1 shows that the data-driven penalized policy satisfies sample privacy. They key challenge here is that sample privacy is a property of the policymaker's strategy, but I cannot define the strategy without specifying sample privacy parameters for the agent's regret bound.

Theorem 2 shows that the rate of convergence of the data-driven penalized policy is approximately optimal. First, I show that the policymaker's expected payoff converges as the sample size grows. More precisely, it converges to his optimal payoff in a hypothetical model where both participants know the true distribution. Next, I evaluate the rate of convergence. The optimal rate is that of the optimal penalized policy, where the policymaker knows the true distribution but the agent is still learning. If the optimal rate is $n^{-\gamma}$, I show that the data-driven penalized policy converges at a $n^{-\frac{\gamma}{1+2\gamma}}$ rate. In typical applications where $\gamma = 1/2$, the rate of convergence is $n^{-1/4}$.

Illustrative Examples. I show that penalized policies can provide new insights in four examples.

First, I consider a model of vaccine distribution. Here, the penalized policy waits for statistically-significant clinical trial results before distributing vaccines. While this is common practice, it has been criticized (Wasserstein and Lazar 2016) and conflicts with the recommendations of the treatment choice literature (Manski 2019). My model offers a *strategic* reason to insist on statistical significance: the population needs to be convinced of vaccine quality. People may not take up the vaccine if it has not been proven effective beyond a reasonable doubt. If there are fixed costs to vaccine distribution, this outcome is worse than not distributing the vaccine at all.

Second, I consider a model of prescription drug approval by a regulator. Here, the penalized policy restricts doctors' ability to prescribe drugs that haven't been proven effective in clinical trials. Limiting the number of drugs approved reduces the risk that doctors fall prey to false positives, in

[1] This is an important point of contrast with the large literature on penalization and regularization in statistics. Here, the goal is not to penalize policies that are too complex for the policymaker to learn effectively. Instead, the policymaker wants to penalize policies that complicated *the agent's* learning problem.

which an ineffective drug appears to be effective by random chance. The penalized policy sets a standard for approval that increases as more drugs are approved, as in stepwise methods for multiple hypothesis testing (e.g. Holm 1979, Romano and Wolf 2005). In contrast, alternative approaches lead to extreme recommendations. If the doctors knew the true distribution of treatment effects, the optimal policy would approve all drugs. Meanwhile, robust policies approve no drugs at all.

Third, I consider a model of performance pay. An employer incentivizes an employee to exert costly effort by paying wages contingent on observed performance. If both participants know the distribution of performance conditional on effort, the optimal contract only compensates the employee at extreme levels of performance. The payment conditional on extreme performance may be extreme as well. In my model, this does not work well. Given limited data, it is not clear that it is worth investing effort for a small chance of receiving a bonus. To address this, the penalized contract caps and flattens the wage schedule. As before, wages are zero until performance reaches a threshold, at which point they jump and remain flat. But the threshold is lower and the payments smaller, allowing the agent to obtain moderate wages for moderate performance.

Finally, I consider a model of product bundling. A firm has several products for sale and wants to sell them in a way that maximizes expected profit. Here, the penalized policy favors selling large bundles of products, or even bundling all products together into a grand bundle. The reason for this is that consumers learn about their value for the product through reviews. If there are many products, but few reviews per product, consumers can be confident in the value of the grand bundle while being uncertain about the value of any given product. In that case, all else equal, it is easier to convince consumers to buy the bundle. I contrast this conclusion with prior work that recommends selling all items separately (Carroll 2017).

Related Literature This work contributes to three research efforts. For robust mechanism design, it is a principled way to interpolate between two extremes: the common prior and prior-freeness. In that respect, I share a goal with Artemov et al. (2013) and Ollár and Penta (2017), although my methods are quite different. For learning in games, I provide a convenient behavioral assumption that relies on tools from statistical learning theory. Liang (2020) studies a similar model but takes a more abstract approach to modeling agent behavior. For data-driven mechanism design, I study scenarios in which the agent is learning from data, not just the policymaker. Related work includes Camara et al. (2020), Cummings et al. (2020), and Immorlica et al. (2020).

I leave a more detailed discussion of the prior literature to Section 7.

Organization. I introduce the model in Section 2. Section 3 formalizes the behavioral assumptions, with Rademacher complexity defined in Subsection 3.2 and sample privacy in Subsection 3.3. Section 4 defines the penalized policy. Section 5 proposes and evaluates the data-driven penalized policy. Section 6 presents four illustrative examples. Section 7 discusses related literature. Section 8 concludes. Omitted proofs can be found in Appendix A.

2 Model

This paper studies Stackelberg games of incomplete information. There are two players: a policymaker and an agent. The policymaker moves first and commits to a policy $p \in \mathcal{P}$. After observing the policy p, the agent chooses a response $r \in \mathcal{R}$, or possibly a mixed response $\pi^r \in \Delta(\mathcal{R})$. Finally, the state of nature $s \in S$ is drawn from a distribution $\pi^s \in \Delta(S)$. The policymaker's utility is given by $u^P(p, r, s)$ and the agent's utility by $u^A(p, r, s)$.

So far, this setup is quite general. It can be used to model a variety of single-agent problems in mechanism design, contract design, information design, and other areas of interest to economists. Features like transfers, messages between the policymaker and agent, and asymmetric information about the state can be incorporated by defining the policy and response spaces appropriately.[2]

I maintain the following assumption throughout the paper.

Assumption 1. *The space $\mathcal{P}$ consists of $n_{\mathcal{P}} < \infty$ policies and utility functions u^A, u^P are bounded.*

The reader is welcome to think of $\mathcal{P}$ as a discrete approximation to an infinite policy space. Assumption 1 also ensures that, for any fixed policy p, the maximum variation in each participant's utility function is finite. Let $i \in \{A, P\}$ indicate the agent or policymaker, and define

$$\Delta^i(p) = \sup_{r,s} u^i(p, r, s) - \inf_{r,s} u^i(p, r, s) < \infty \tag{1}$$

I will impose additional regularity assumptions, as needed, later on in the paper.

Common Knowledge. One way to solve this model is to assume that both the policymaker and the agent know the distribution π^s, or equivalently, that π^s represents their shared prior belief over the state s. Here, the agent is free to choose any response r that maximizes her expected utility, i.e.

$$r \in \arg\max_{r'} \mathrm{E}_{\pi^s}\left[u^A(p, r', s)\right] \tag{2}$$

In turn, the policymaker chooses a policy p that maximizes his expected utility after taking into account how the agent will respond.

If equation (33) has multiple solutions, then the policymaker may not know which response r the agent will choose. One way to deal with this is to specify a tie-breaking rule that determines which response r the agent chooses when she is indifferent between multiple responses.[3] Another

[2] For example, if the policymaker has access to transfers, then the policy p must specify the transfers made (if any). Alternatively, if the agent has private information about some aspect of the state s, then her response r should map that information to some action. Finally, if the agent reports her private information to the policymaker, then p should map that report to some action.

[3] A typical tie-breaking rule assumes that the agent breaks ties in favor of the policymaker. This assumption is convenient in games with an infinite policy space because it ensures that an optimal policy exists. Furthermore, it is often innocuous insofar as small perturbations of the optimal policy can be used to break indifferences in favor of the policymaker with negligible loss of optimality. Here, the policy space is finite so existence of an optimal policy is not an issue. Furthermore, in games where the favorable tie-breaking assumption is innocuous, the common prior benchmark will be roughly the same regardless of whether I assume favorable tie-breaking or not.

approach is to remain agnostic to how the agent breaks ties and optimize against the worst-case best response. In that case, the policymaker can guarantee an expected utility of

$$\begin{aligned} \mathrm{CK}(\pi^s) = \max_{p} \min_{r} \mathrm{E}_{\pi^s}\left[u^P(p, r, s)\right] \\ \text{s.t.} \quad r \in \arg\max_{r'} \mathrm{E}_{\pi^s}\left[u^A(p, r', s)\right] \end{aligned} \tag{3}$$

I refer to $\mathrm{CK}(\pi^s)$ as the *common knowledge benchmark*.

The common knowledge assumption has two obvious drawbacks. First, the optimal policy will generally depend on the true distribution π^s and the policymaker may not know π^s. If our goal is to consult the policymaker, we cannot recommend a policy to him that relies on information that he does not have access to. Second, the agent may not know the true distribution π^s. In that case, she may choose responses r that are suboptimal in the sense that they do not solve equation (2), and the policy that solves (3) may no longer be optimal.

Similar issues arise when the distribution π^s is interpreted as a common prior belief. The policymaker may not be willing to specify a prior belief over the state s, or be concerned that his beliefs are too uninformed. The agent may not agree with the policymaker's prior belief, especially if that belief is not well-informed.

Prior-Free Solution Concepts. Another way to solve this model is to not make any assumptions about the agent's beliefs $\tilde{\pi}$. These approaches are called prior-free, belief-free, or belief-robust.

Suppose the policymaker knows the true distribution π^s, but does not know anything about the agent's beliefs. In that case, the *maximin policy* maximizes the policymaker's worst-case expected utility. That is,

$$\begin{aligned} \mathrm{MM}(\pi^s) = \max_{p} \min_{\tilde{\pi}^s, r} \mathrm{E}_{\pi^s}\left[u^P(p, r, s)\right] \\ \text{s.t.} \quad r \in \arg\max_{r'} \mathrm{E}_{\tilde{\pi}^s}\left[u^A(p, r', s)\right] \end{aligned} \tag{4}$$

The objective is evaluated with respect to the worst-case belief $\tilde{\pi}^s \in \Delta(S)$.

There are alternatives to the maximin policy. For example, the *minimax regret policy* minimizes the policymaker's worst-case regret from following policy p, rather than the policy p' that would have been optimal given the agent's response r. That is,

$$\begin{aligned} \mathrm{MR}(\pi^s) = \min_{p} \max_{\tilde{\pi}^s, r} \left(\max_{p', r'} \mathrm{E}_{\pi^s}\left[u^P(p', r', s)\right] - \mathrm{E}_{\pi^s}\left[u^P(p, r, s)\right] \right) \\ \text{s.t.} \quad r, r' \in \arg\max_{r'} \mathrm{E}_{\tilde{\pi}^s}\left[u^A(p, r', s)\right] \end{aligned} \tag{5}$$

Again, the objective is evaluated with respect to the worst-case belief $\tilde{\pi}^s \in \Delta(S)$.

The advantage of maximin and minimax regret policies are that they do not require the agent to know the true distribution π^s, or to agree with the policymaker's beliefs. The disadvantage is that they can be extremely conservative: $\mathrm{MM}(\pi^s)$ or $\mathrm{MR}(\pi^s)$ may only be a small fraction of $\mathrm{CK}(\pi^s)$.

This is especially true in applications like contract design and Bayesian persuasion where the agent's optimal response varies considerably in her beliefs over the state.

Common Dataset. I propose a third way to solve this model: one that avoids key drawbacks of the two existing approaches. Rather than assume common knowledge of the distribution, I assume access to a common dataset. Formally, a dataset consists of n i.i.d. observations of the state, i.e.

$$S_1, \dots, S_n \sim \pi^s$$

Each participant's strategy is now a statistical decision rule. The *policymaker's strategy* maps the dataset to a distribution over policies:

$$\sigma_n^P : \mathcal{S}^n \to \Delta(\mathcal{P})$$

The *realized policy* is

$$P_n \sim \sigma_n^P(S_1, \dots, S_n)$$

The *agent's strategy* maps the dataset and the policymaker's policy to a distribution over responses:

$$\sigma_n^A : \mathcal{S}^n \times \mathcal{P} \to \Delta(\mathcal{R})$$

The *realized response* is

$$R_n \sim \sigma_n^A(S_1, \dots, S_n, P_n)$$

Both P_n and R_n are random variables, because (i) they depend on the random sample and (ii) the policymaker and agent may use mixed strategies.

Remark 1. It is worth emphasizing that this dataset is ideal in several ways.

1. First, the states of nature S_i are directly observed. This is a good starting point because the data clearly identifies the distribution π^s. In practice, however, the state of nature may not be observed directly, and the distribution π^s may not be point identified by the available data. That raises the additional hurdle of partial identification.

2. Second, the observations S_i are drawn independently from the true distribution π^s. This is a standard assumption, but may not hold in dynamic environments where historical data does not fully reflect the present. It is possible to drop this assumption with an alternative approach developed in Camara et al. (2020). However, that approach requires the policymaker to interact repeatedly with the same agent. Here, I only require a single interaction.

3. Third, the dataset is available to both participants. More precisely, any data that the policymaker uses must also be available to the agent (it is not a problem if the agent has access to additional data). In some cases, the policymaker may be able to guarantee that this assumption holds by sharing his data with the agent. In other cases, the policymaker may prefer to keep the data confidential.

These are important caveats, but at least they are explicit. By being explicit about what data is available, it is possible to have a productive discussion over what the participants are likely to know and what they may not know. In contrast, common knowledge assumptions bypass any discussion of how the policymaker and agent actually arrived at said knowledge.

In the next section, I propose new rationality assumptions that restrict the agent's strategy σ_n^A.

3 Agent's Behavior

I propose new rationality assumptions that restrict the agent's behavior. The basic idea is that, since she has access to a dataset, the agent should not underperform standard heuristics for learning from data. I formalize this assumption by introducing *feasible regret bounds*. Then I provide sufficient conditions for a regret bound to be feasible.

The key object of interest is the agent's *regret*, which captures how suboptimal her strategy is.[4]

Definition 1. *The agent's* regret *is the difference between her optimal expected utility and the expected utility she achieves by following her strategy σ_n^A. Formally,*

$$\text{Regret}_n^A\left(\sigma_n^A, \sigma_n^P, \pi^s\right) = \max_r \mathrm{E}_{\pi^s}\left[u^A(P_n, r, s)\right] - \mathrm{E}_{\pi^s}\left[u^A(P_n, R_n, s)\right]$$

To be clear, expectations are taken with respect to the state s, the realized sample $S_1, \ldots, S_n$, and any internal randomization associated with mixed strategies. The sample $S_1, \ldots, S_n$ comes in through the realized response R_n and policy P_n.

It turns out that a refined notion of regret will be more useful. The policymaker not only cares about whether the agent makes mistakes, but also about how those mistakes are correlated with the policies that he chooses.

Definition 2. *The agent's* conditional regret *is her regret conditional on the realized policy $P_n = p$.*

$$\text{Regret}_n^A\left(\sigma_n^A, \sigma_n^P, \pi^s \mid P_n = p\right) = \max_r \mathrm{E}_{\pi^s}[u(p, r, s)] - \mathrm{E}_{\pi^s}\left[u(p, R_n, s) \mid P_n = p\right]$$

To be clear, this is not related to contextual regret, or regret conditioned on private information (as in Camara et al. 2020). The realized policy does not convey any information about the distribution π^s that the agent does not already have access to. However, the realized response R_n and the policy P_n may be correlated because they depend on the same random sample. So the conditional expectation

$$\mathrm{E}_{\pi^s}\left[u(p, R_n, s) \mid P_n = p\right]$$

may be different from the unconditional expectation used to define regret.

[4] This is not the same as the ex-post notion of regret used in the literature on learning in games.

3.1 Regret Bounds

A *regret bound* $B\left(\sigma_n^P, \pi^s, p\right)$ is a function of the policymaker's strategy σ_n^P, the realized policy $P_n = p$ and the distribution π^s that bounds the agent's conditional regret. The agent is allowed to use any strategy σ_n^A that satisfies

$$\text{Regret}_n^A\left(\sigma_n^A, \sigma_n^P, \pi^s \mid P_n = p\right) \leq B\left(\sigma_n^P, \pi^s, p\right)$$

This is an upper bound: it is always possible that the agent outperforms this bound. In particular, an agent that knows the true distribution π^s will satisfy this bound because her regret is zero.

Remark 2. Allowing the agent to obtain positive regret (or conditional regret) is one way to relax the prior knowledge assumption. I argue that this approach has several important advantages.

1. I do not require the researcher to specify a set of prior beliefs that the agent could reasonably possess (see e.g. Artemov et al. 2013, Ollár and Penta 2017).[5] This is difficult to do without guidance on which prior beliefs are reasonable. In contrast, I can provide guidance for how to choose regret bounds by considering measures of statistical complexity.

2. I do not rule out the possibility that the agent is Bayesian or that she has prior knowledge about the distribution π^s that goes beyond the common dataset. For example, the unbiased inference procedures in Salant and Cherry (2020) rule out this possibility. Regret bounds are intended to relax standard assumptions, not to contradict them.

3. I do not require the agent to be Bayesian: she is welcome to use strategies σ_n^A that are not consistent with Bayesian updating, as long as they satisfy the regret bound. Non-Bayesian (e.g. frequentist) methods for learning from data are commonly used in practice, including by empirical economists. It is reassuring that regret bounds allow for this possibility.

Not all regret bounds are compelling. I restrict attention to *feasible* regret bounds, which the agent can satisfy even if she has no prior knowledge of the distribution π^s.

Definition 3. *A regret bound B is* feasible *given policymaker's strategy σ_n^P if there exists an agent strategy $\tilde{\sigma}_n^A$ such that, for all policies p and distributions $\tilde{\pi}^s$,*

$$\text{Regret}_n^A\left(\tilde{\sigma}_n^A, \sigma_n^P, \tilde{\pi}^s \mid P_n = p\right) \leq B\left(\sigma_n^P, \tilde{\pi}^s, p\right)$$

The strategy $\tilde{\sigma}_n^A$ may be distinct from the agent's actual strategy σ_n^A.

Feasible regret bounds make for reasonable rationality assumptions. After all, an agent whose strategy σ_n^A systematically underperforms a feasible regret bound would benefit from deviating to the strategy $\tilde{\sigma}_n^A$ that satisfies it. This deviation would reduce her regret, or equivalently, increase

[5] This often involves committing to a metric ρ on the space of beliefs, and assuming that the agent's beliefs are within some distance ϵ of the true distribution. That approach is consistent with a regret bound for reasonable choices of ρ where optimizing against approximately correct beliefs leads to approximately optimal responses.

her expected utility according to the true distribution. It would require no prior knowledge of the distribution π^s, since $\tilde{\sigma}_n^A$ does not depend on π^s.

Infeasible regret bounds, however, make for dubious rationality assumptions. In that case, there does not exist a single strategy $\tilde{\sigma}_n^A$ that satisfies the regret bound across all distributions $\tilde{\pi}^s$. It is still possible for the agent to satisfy the regret bound under the true distribution π^s, but she would need prior knowledge about π^s that goes beyond the dataset. Although I do not want to *rule out* the possibility that the agent has prior knowledge, I also do not want to *require* prior knowledge.

3.2 Rademacher Complexity

In this subsection, I provide a sufficient condition for a regret bound to be feasible in the special case where the policymaker ignores the data. I begin by defining Rademacher complexity, a key concept from statistical learning theory, and stating the sufficient condition. Then I provide intuition.

A Rademacher random variable σ is uniformly distributed over $\{-1, 1\}$.

Definition 4 (Bartlett and Mendelson 2003)**.** *The* Rademacher complexity *induced by policy p is*

$$\mathcal{RC}_n^A(p, \pi^s) = \mathrm{E}_{\pi^s}\left[\max_r \frac{1}{n}\sum_{i=1}^n \sigma_i \cdot u^A\left(p, r, S_i\right)\right]$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables.

The policymaker's strategy σ_n^P is *constant* if there exists a distribution $\pi^p \in \Delta(\mathcal{P})$ such that, for all sample realizations $S_1, \ldots, S_n$,

$$\sigma_n^P(S_1, \ldots, S_n) = \pi^p$$

Proposition 1. *Let the policymaker's strategy σ_n^P be constant. Then B is a feasible regret bound if*

$$\forall p, \tilde{\pi}^s, \quad B\left(\sigma_n^P, \tilde{\pi}^s, p\right) \geq 4\mathcal{RC}_n^A(p, \tilde{\pi}^s)$$

The Rademacher complexity bounds the agent's regret if she uses a particular strategy called *empirical utility maximization*. This strategy will play the role of $\tilde{\sigma}_n^A$ in the definition of feasibility.

Definition 5. Empirical utility maximization $\hat{\sigma}_n^A$ *is an agent strategy. It chooses the response r that maximizes the agent's expected utility with respect to the empirical distribution $\hat{\pi}^s$. Formally,*

$$\hat{\sigma}_n\left(S_1, \ldots, S_n\right) \in \arg\max_r \frac{1}{n}\sum_{i=1}^n u^A\left(P_n, r, S_i\right)$$

Let $\hat{R}_n = \hat{\sigma}_n\left(S_1, \ldots, S_n\right)$ denote the empirical utility maximizer.

If the policymaker's strategy σ_n^P is constant, then

$$\forall p, \tilde{\pi}^s, \quad \mathrm{Regret}_n^A\left(\hat{\sigma}_n^A, p, \tilde{\pi}^s\right) \leq 4\mathcal{RC}_n^A(p, \tilde{\pi}^s) \tag{6}$$

This follows immediately from well-known results in Bartlett and Mendelson (2003).

Intuitively, Rademacher complexity measures the potential for empirical utility maximization to overfit to sampling noise. First, it trivializes the agent's learning problem by randomizing the sign of the agent's utility function. That is, it replaces the utility function $u^A(p, r, s)$ with $\sigma \cdot u^A(p, r, s)$, where σ is a Rademacher random variable. This modified learning problem is trivial since all responses are equally good: expected utility is zero for every response r, i.e.

$$\mathrm{E}_{\pi^s}\left[\sigma \cdot u^A(p, r, s)\right] = 0$$

Second, Rademacher complexity asks how much the empirical utility maximizer $\hat{R}_n$ will overfit to this modified problem. Formally, the empirical utility is

$$\frac{1}{n}\sum_{i=1}^{n} \sigma_i \cdot u^A\left(p, r, S_i\right)$$

where σ_i are i.i.d. Rademacher random variables. For any given response r, the empirical utility is zero in expectation. On the other hand, the empirical utility of $\hat{R}_n$ will generally have a positive expected value in finite samples. This expected value is the Rademacher complexity, and reflects how severely the agent can be misled by the sampling noise in $\sigma_1, \ldots, \sigma_n$.

Note that the Rademacher complexity will typically converge to zero as $n \to \infty$. In other words, the agent's learning problem typically becomes easier as she obtains more data. In addition, the Rademacher complexity generally depends on the policy p. The policymaker can increase or decrease the Rademacher complexity through his choice of policies.

3.3 Sample Privacy

In this subsection, I generalize the sufficient condition for a regret bound to be feasible. The condition in proposition 1 only applies when the policymaker ignores the data. I show that a similar condition applies as long as the policymaker does not use the data too aggressively. I formalize this requirement through a new concept called *sample privacy*.

Definition 6. *The policymaker's strategy σ_n^P satisfies* (ϵ, δ)-sample privacy *if there exists an event $E \subseteq S^n$ where (i) E occurs high probability, i.e.*

$$\Pr_{\pi^s}\left[\left(S_1, \ldots, S_n\right) \in E\right] \geq 1 - \delta$$

and (ii) after conditioning on E, the sample is nearly independent of the realized policy, i.e.

$$\left(S_1, \ldots, S_n\right) \in E \implies \Pr_{\pi^s}\left[P_n = p \mid S_1, \ldots, S_n\right] \leq e^{\epsilon} \cdot \Pr_{\pi^s}\left[P_n = p \mid E\right]$$

Note that the realized policy P_n may be random even after conditioning on the sample $S_1, \ldots, S_n$, if the policymaker uses a mixed strategy.

The concept is closely related to *differential privacy* (Dwork, McSherry, et al. 2006). When applied to the policymaker's strategy σ_n^P, differential privacy ensures that the realized policy P_n does not change much when any one observation $S_i = s$ is replaced with another value $S_i = s'$. In contrast, sample privacy ensures that the policy P_n does not change much when the entire sample $S_1, \dots, S_n$ is dropped and replaced with a new sample $S_1', \dots, S_n'$. Critically, the new sample is drawn from the same distribution π^s as the original sample.

If the policymaker's strategy σ_n^P is constant then it satisfies $(0, 0)$-sample privacy. This is because the realized policy P_n is independent of the sample $S_1, \dots, S_n$. Indeed, proposition 1 applies to any strategy σ_n^P that satisfies $(0, 0)$-sample privacy. From that perspective, proposition 2 shows that the lower bound in proposition 1 increases smoothly as the parameters (ϵ, δ) increase.

Proposition 2. *Let the policymaker's strategy σ_n^P satisfy (ϵ, δ)-sample privacy. Then B is a feasible regret bound if*

$$\forall p, \tilde{\pi}^s, \quad B\left(\sigma_n^P, \tilde{\pi}^s, p\right) \geq 4e^{\epsilon} \cdot \mathcal{RC}_n^A(p, \tilde{\pi}^s) + \delta \cdot \Delta^A(p)$$

The proof shows that if the policymaker's strategy σ_n^P satisfies sample privacy then conditioning on $P_n = p$ does not have a meaningful impact on any moment of the sample. In particular, the agent's conditional regret is a moment of the sample. There is a similar result in differential privacy (Dwork and Roth 2014, section 2.3.1).

The fact that sample privacy is needed at all suggests an important difference between a standard learning problem and the *concurrent learning* problem that I study. From the agent's perspective, a standard learning problem is one in which her utility function $u^A(p, r, s)$ does not depend on the realized sample $S_1, \dots, S_n$. But when the policymaker is learning concurrently, the agent's utility function $u^A(P_n, r, s)$ depends on the sample through the realized policy P_n. Standard techniques for bounding the agent's regret may no longer apply. I give a concrete example of this in section 6.2.

4 Penalized Policy

This model formalizes a sense in which complex policies are undesirable. As policies become more complex, feasible regret bounds tend to loosen, and the agent's behavior becomes less predictable. I propose a *penalized policy* that balances the advantages of complexity with the disadvantages of unpredictability. I show that penalization can lead to new insights in four illustrative examples.

First, I restrict attention to regret bounds B that take on a particular form. I require a distribution-free upper bound on the Rademacher complexity, i.e.

$$\overline{\mathcal{RC}}_n^A(p) \geq \max_{\pi^s} \mathcal{RC}_n^A(p, \pi^s) \tag{7}$$

It follows immediately from proposition 2 that

$$B\left(\sigma_n^P, p, \pi^s\right) := e^{\epsilon} \cdot \overline{\mathcal{RC}}_n^A(p) + \delta \cdot \Delta^A(p)$$

is a feasible regret bound.

Remark 3. There are many well-known distribution-free upper bounds on the Rademacher complexity, based on measures like the VC dimension, Pollard's pseudo-dimension, and the covering number. For example, a useful bound due to Massart (2000) applies to finite response spaces with $n_{\mathcal{R}} < \infty$ elements. In that case,

$$\overline{\mathcal{RC}}_n^A(p) := \Delta^A(p) \cdot \sqrt{\frac{2 \ln n_{\mathcal{R}}}{n}}$$

is an upper bound on the Rademacher complexity.

These regret bounds are convenient because they only depend on the policy p and privacy parameters (ϵ, δ). Assuming I can guarantee sample privacy, this makes it possible to evaluate the policymaker's utility from a given policy p without having to consider the strategy σ_n^P that generated that policy. In particular, I can evaluate the policymaker's utility with respect to the worst-case mixed response π^r that satisfies the agent's regret bound:

$$\begin{aligned} \mathrm{WC}_n(p, \epsilon, \delta, \pi^s) = \min_{\pi^r} & \mathrm{E}_{\pi^s, \pi^r}\left[u^P(p, r, s)\right] \\ \text{s.t.} \quad \max_{r'} & \mathrm{E}_{\pi^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\pi^s, \pi^r}\left[u^A(p, r, s)\right] \leq 4e^{\epsilon} \cdot \overline{\mathcal{RC}}_n^A(p) + \delta \cdot \Delta^A(p) \end{aligned} \tag{8}$$

The mixed response π^r reflects the marginal distribution of the realized response R_n conditional on the realized policy $P_n = p$.[6] Intuitively, an agent may choose an optimal response with high probability, when the realized sample is representative, and a suboptimal response with low probability, when the realized sample is misleading, and still be nearly-optimal in expectation.

If the policymaker knew the true distribution π^s, he could solve for the *optimal penalized policy*. In this case, the agent is still learning from the dataset, but the policymaker can ignore the realized sample and guarantee $(0, 0)$-sample privacy.

Definition 7. *An* optimal penalized policy *is any policy p that solves*

$$\mathrm{OP}_n(\pi^s) = \max_p \mathrm{WC}_n(p, 0, 0, \pi^s) \tag{9}$$

In section 5, I develop a strategy for the policymaker that approximates the optimal penalized policy by learning from the available data. For now, I focus on the optimal penalized policy.

I call these penalized policies because the worst-case objective implicitly penalizes policies that the agent perceives as more complex, as measured by the Rademacher complexity. As in penalized regression and other forms of regularization in statistics, penalization biases the policymaker towards policies that are less complex. However, there is a key difference. In statistics, policies are considered complex if they make the policymaker's learning problem hard. Here, policies are considered complex because they make the agent's learning problem hard.

[6] Only the marginal distribution is relevant because the sample does not directly enter into the policymaker's utility. The sample only affects the distribution of the realized policy P_n, which I have already conditioned on.

From the perspective of microeconomic theory, penalization is interesting because it highlights a new tradeoff related to policy complexity. The more complex a policy, the looser the regret bound, and the less predictable the agent's behavior. If the policymaker is ambiguity-averse, he will tend to choose policies that are less complex than if we had assumed common knowledge. This is especially pronounced when the sample size n is small.

5 Data-Driven Penalized Policy

It is useful to study the optimal penalized policy to understand the qualitative effect of penalization on policymaking, but the optimal penalized policy is not feasible unless the policymaker knows the distribution π^s in advance. In this section, I show that it is possible to approximate the optimal penalized policy using the available data.

I construct a strategy $\hat{P}_n$ for the policymaker. This modifies the naive estimator that evaluates the policymaker's worst-case utility with respect to the empirical distribution $\hat{\pi}^s$, i.e.

$$\begin{aligned} \mathrm{WC}_n(p, \epsilon, \delta, \hat{\pi}^s) = & \min_{\pi^r} \mathrm{E}_{\hat{\pi}^s, \pi^r}\left[u^P(p, r, s)\right] \\ & \text{s.t.} \quad \max_{r'} \mathrm{E}_{\hat{\pi}^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\hat{\pi}^s, \pi^r}\left[u^A(p, r, s)\right] \leq 4e^{\epsilon} \cdot \overline{\mathcal{RC}}_n^A(p) + \delta \end{aligned} \tag{10}$$

For many problems in statistics, econometrics, and machine learning, replacing the true distribution with the empirical distribution is enough to come up with a good estimator. In this model, where the agent is learning in addition to the policymaker, two changes are needed.

First, I conservatively estimate the set of mixed responses that meet the agent's regret bound. The naive estimator assumes that the agent's empirical regret, i.e.

$$\max_{r'} \mathrm{E}_{\hat{\pi}^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\hat{\pi}^s, \pi^r}\left[u^A(p, r, s)\right]$$

satisfies the regret bound B. However, B is a bound on the agent's regret evaluated with respect to the true distribution π^s. If the sample is unrepresentative, mixed responses π^r that satisfy the regret bound may violate the constraint in equation (10) because their empirical regret overestimates their true regret. In that case, the constraint rules out mixed strategies that the agent might use.

I can address this by adding a *buffer* to the regret bound, so that the empirical regret minus the buffer is unlikely to overestimate the true regret. Let $\alpha \in (0, 1)$ be a tuning parameter. Define the buffer as follows:

$$\mathrm{BFR}_n = 8\sqrt{\frac{2\ln 4}{n}} + 8\sqrt{-\frac{2\ln\exp(-n^\alpha)}{n}}$$

Lemma 1. *With probability exceeding* $1 - n_{\mathcal{P}}\exp(-n^\alpha)$*, the buffer exceeds the difference between empirical and true regret. That is,*

$$4\mathcal{RC}_n^A(p, \pi^s) + \mathrm{BFR}_n \geq \left| \left(\max_{r'} \mathrm{E}_{\hat{\pi}^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\hat{\pi}^s, \pi^r}\left[u^A(p, r, s)\right] \right) \right.$$

$$- \left(\max_{r'} \mathrm{E}_{\pi^s} \left[u^A \left(p, r', s \right) \right] - \mathrm{E}_{\pi^s, \pi^r} \left[u^A \left(p, r, s \right) \right] \right) \Big|$$

for all mixed responses π^r and policies p,

Second, I introduce white noise into the objective $\mathrm{WC}_n(\cdot)$ order to control the sample privacy of $\hat{P}_n$. The challenge here is that there is an inherent circularity. Sample privacy is a property of the policymaker's strategy $\hat{\sigma}_n^P$, but in order to define $\hat{\sigma}_n^P$ I need to specify the privacy parameters (ϵ, δ) in the agent's regret bound. I address this challenge in theorem 1.

More concretely, I ensure sample privacy by adapting the exponential mechanism proposed by McSherry and Talwar (2007). Let $\beta \in (0, \alpha/2)$ be another tuning parameter. I add noise from the Gumbel distribution into the policymaker's objective function, i.e.

$$\nu_n(p) \sim \textsc{Gumbel} \left(0, n^{-\beta} \right)$$

For any given $\epsilon > 0$, this estimator will satisfy (ϵ, δ_n)-privacy, where δ_n depends on tuning parameters α, β, ϵ and is decreasing exponentially in n. More precisely,

$$\delta_n = n_{\mathcal{P}} \exp \left(-\frac{\epsilon^2}{2K^2} \cdot n^{\alpha - 2\beta} \right) \tag{11}$$

where the constant K is defined by

$$K := \max_p \max \left\{ \frac{2\Delta^P(p)^2}{8\sqrt{2}}, \Delta^P(p) \right\} \tag{12}$$

Incorporating this into equation (**??**) gives a noisy, conservative estimate of the worst-case utility.

$$\begin{aligned} \widehat{\mathrm{WC}}_n(p) = & \min_{\pi^r} \mathrm{E}_{\hat{\pi}^s, \pi^r} \left[u^P \left(p, r, s \right) \right] + \nu_n(p) \\ \text{s.t.} \quad & \max_{r'} \mathrm{E}_{\hat{\pi}^s} \left[u^A \left(p, r', s \right) \right] - \mathrm{E}_{\hat{\pi}^s, \pi^r} \left[u^A \left(p, r, s \right) \right] \\ & \leq (4e^{\epsilon} + 4) \cdot \overline{\mathcal{RC}}_n^A(p) + \delta_n \cdot \Delta^A(p) + \mathrm{BFR}_n \end{aligned} \tag{13}$$

Definition 8. *The data-driven penalized policy $\hat{P}_n$ solves*

$$\hat{P}_n \in \arg\max_p \widehat{\mathrm{WC}}_n(p) \tag{14}$$

To summarize, $\hat{P}_n$ depends on three tuning parameters. The parameter $\alpha \in (0, 1)$ controls how quickly the buffer BFR_n vanishes as n grows. The parameter $\beta \in (0, \alpha/2)$ controls how quickly the privacy-preserving noise vanishes as n grows. The parameter $\epsilon > 0$ controls how the two dimensions of sample privacy are balanced; decreasing ϵ means increasing δ_n, and vice-versa.

The next theorem verifies that this estimator obtains the privacy guarantees that were assumed in its definition. It holds for any fixed $\epsilon > 0$, with δ_n defined according to equation (11).

Theorem 1. *The estimator $\hat{P}_n$ guarantees $\left(\epsilon, \delta_n\right)$-sample privacy.*

This result holds under very weak assumptions on the underlying game (assumption 1), which makes it challenging to prove. I outline the proof in subsection 5.3.

5.1 Convergence

I show that, in the limit as the sample size grows, the policymaker's payoff under $\hat{P}_n$ converges to his optimal payoff in a model where the distribution is common knowledge. That is, common knowledge is the limiting phenomenon under $\hat{P}_n$, despite the challenges that arise when the policymaker and agent are learning simultaneously.

This requires a regularity condition that ensures that the agent and policymaker can learn the optimal response and policy in the easy case where their opponent is not also learning. The condition ensures that the Rademacher complexity of both the agent and policymaker's learning problem vanishes at a reasonable rate as the sample size grows.

Definition 9. *The policymaker's Rademacher complexity is defined over all policy-response pairs, i.e.*

$$\mathcal{RC}_n^P(\pi^s) = \mathrm{E}_{\pi^s}\left[\sup_{p,r} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \cdot u^P(p, r, S_i)\right]$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables.[7]

Assumption 2. *The agent and policymaker's Rademacher complexity vanish at the typical $\tilde{O}(n^{-1/2})$ rate, where the tilde in $\tilde{O}$ means "up to log factors". More precisely:*

1. *There exists a constant K^A such that*

$$\overline{\mathcal{RC}}_n^A(p) \leq K^A n^{-1/2} \log n$$

   for all policies p and sample sizes n, where K^A does not depend on p or n.

2. *There exists a constant K^P such that*

$$\mathcal{RC}_n^P(\pi^s) \leq K^P n^{-1/2} \log n$$

   for all distributions π^s and sample sizes n, where K^P does not depend on π^s or n.

This assumption is both easy to satisfy and stronger than necessary. It is easy to satisfy because it holds whenever the agent's response space is finite. It also holds whenever the relevant optimization

[7] I use this quantity to bound the generalization error of $\hat{P}_n$. More precisely, for any given policy p, the generalization error is the difference between the performance of policy p according to the empirical distribution $\hat{\pi}^s$ and its actual performance under the true distribution π^s. Note that this error will generally depend on the agent's response r. To account for this, the policymaker's Rademacher complexity takes a supremum over responses r in addition to policies p. This ensures that the bound on the generalization error holds regardless of what the agent's response is or how it is influenced by the data.

problems have a finite VC dimension, or a finite pseudo-dimension. And the assumption is stronger than necessary because the $\tilde{O}(n^{-1/2})$ rate is not needed for proposition 3. I do make use of this rate for theorem 1, in order to give concrete rates of convergence for my estimator.

Proposition 3 establishes convergence. It refers to three quantities of interest. First, the common knowledge benchmark $\mathrm{CK}(\pi^s)$ (equation 3) describes the policymaker's optimal payoff when the distribution π^s is common knowledge. Second, the optimal penalized benchmark $\mathrm{OP}_n(\pi^s)$ (equation 9) describes the policymaker's optimal payoff when he knows the distribution π^s but the agent is still learning. Third, the performance of the data-driven penalized policy $\hat{P}_n$ is given by

$$\mathrm{E}_{\pi^s}\left[\mathrm{WC}_n\left(\hat{P}_n, \epsilon, \delta_n, \pi^s\right)\right]$$

It is easy to see that

$$\mathrm{E}_{\pi^s}\left[\mathrm{WC}_n\left(\hat{P}_n, \epsilon_n, \delta_n, \pi^s\right)\right] \leq \mathrm{OP}_n(\pi^s) \leq \mathrm{CK}(\pi^s)$$

I show that all three quantities coincide in the limit as $n \to \infty$.

Proposition 3. *Both the performance of $\hat{P}_n$ and the optimal penalized benchmark converge to the common knowledge benchmark, as $n \to \infty$. That is,*

$$\lim_{n\to\infty} \mathrm{E}_{\pi^s}\left[\mathrm{WC}_n\left(\hat{P}_n, \epsilon, \delta_n, \pi^s\right)\right] = \lim_{n\to\infty} \mathrm{OP}_n(\pi^s) = \mathrm{CK}(\pi^s)$$

I outline the proof in subsection 5.4.

5.2 Rate of Convergence

Establishing convergence is not enough to argue that the data-driven penalized policy $\hat{P}_n$ is a practical solution to the policymaker's problem. At a minimum, $\hat{P}_n$ should have a reasonable rate of convergence. In this subsection, I show that the rate of convergence of $\hat{P}_n$ is approximately as good as the rate of convergence of the optimal penalized policy. To do this, I need a richness assumption that rules out cases where the agent is indifferent between all of her responses.

First, I observe that there are games in which $\hat{P}_n$ cannot possibly have a good rate of convergence. This is because even the optimal penalized policy has a poor rate of convergence.

Proposition 4. *For any $\gamma > 0$, there exists a game where the optimal penalized benchmark has at best an $n^{-\gamma}$ rate of convergence. That is,*

$$\mathrm{OP}_n(\pi^s) = \mathrm{CK}(\pi^s) - \Omega\left(n^{-\gamma}\right)$$

Furthermore, this game satisfies all of my regularity assumptions.

This result appears to be quite pessimistic, but it is not all that surprising. There happen to be games where the policymaker cannot guarantee the ideal outcome unless the agent has very precise distributional knowledge. In these cases, slow rates of convergence are inevitable. But that speaks to the fundamentals of the game itself, rather than to the quality of $\hat{P}_n$ as a strategy.

A more instructive question is to ask how $\hat{P}_n$'s rate of convergence compares to optimal penalized benchmark.

Definition 10. *Let $\gamma > 0$ be the largest real number such that the optimal penalized benchmark converges to the common knowledge benchmark at the rate $n^{-\gamma}$. That is,*

$$\mathrm{OP}_n(\pi^s) = \mathrm{CK}(\pi^s) - O(n^{-\gamma})$$

Loosely, the richness assumption requires that, for any distribution $\tilde{\pi}^s$ and policy p, the agent's best response r must be strictly better than her worst response r'.

Assumption 3. *For every distribution $\tilde{\pi}^s$ and policy p, there exist responses r, r' such that*

$$\mathrm{E}_{\tilde{\pi}^s}\left[\left(u^A\left(p, r, s\right) - u^A\left(p, r', s\right)\right)^2\right] \geq C^2$$

for some constant $C > 0$ that does not depend on $\tilde{\pi}^s, p, r, r'$.

Theorem 2 says that $\hat{P}_n$'s rate of convergence is approximately optimal. I characterize the rate of convergence in terms of its tuning parameters and then show how to optimize them.

Theorem 2. *The performance of $\hat{P}_n$ converges to the common knowledge benchmark at the rate:*

$$\mathrm{E}_{\pi^s}\left[\mathrm{WC}_n\left(\hat{P}_n, \epsilon, \delta_n, \pi^s\right)\right] = \mathrm{CK}(\pi^s) - O\left(n^{-\min(\gamma(1-\alpha),\beta)}\right)$$

I outline the proof in subsection 5.4.

The next corollary describes the rate of convergence when the tuning parameters are optimized. Generally, my estimator will only approximate the optimal rate of convergence $n^{-\gamma}$, with the difference reflecting the cost of sample privacy.

Corollary 1. *For any $\varepsilon > 0$, there exist parameter values α, β such that*

$$\mathrm{E}_{\pi^s}\left[\mathrm{WC}_n\left(\hat{P}_n, \epsilon, \delta_n, \pi^s\right)\right] = \mathrm{CK}(\pi^s) - O\left(n^{\frac{\gamma}{1+2\gamma}-\varepsilon}\right)$$

Proof. Set $\beta = \gamma/(1+2\gamma)$ and let α be slightly larger than 2β. □

For example, suppose that the optimal penalized benchmark converges at a typical $n^{-1/2}$ rate. Then $\hat{P}_n$'s rate of convergence can be set arbitrarily close to $n^{-1/4}$.

Remark 4. Theorem 2 can be understood as a possibility result. It says that $\hat{P}_n$ achieves a particular rate of convergence. But it does not claim that this rate of convergence is the best possible. It relies on explicit finite sample bounds, but does not evaluate whether they are tight enough to be useful in practice. It provides limited guidance on how to choose the tuning parameter α in finite samples, and no guidance on how to choose ϵ because it does not affect the rate of convergence.[8] Answering these questions effectively would likely require putting more structure on the underlying game.

[8] All else equal, it is better if ϵ is small, but it does not need to be small in order for the results to hold. It only needs

5.3 Proof Outline of Theorem 1

The proof of theorem 1 relies on four lemmas, some of which are applications of known results. The key challenge is that $\widehat{\mathrm{WC}}_n(\cdot)$ is not a friendly object. It is a constrained minimization problem where the empirical distribution enters into both the objective and the constraint. And it has little structure, because I made few assumptions on the game between the policymaker and the agent.

The first step is to show that the objective $\widehat{\mathrm{WC}}_n(\cdot)$ falls within a distance t of its mean, with high probability. This will be used to establish that $\widehat{\mathrm{WC}}_n(\cdot)$ satisfies a sample privacy property, which immediately implies that $\hat{P}_n$ satisfies that property. Intuitively, if $\widehat{\mathrm{WC}}_n(\cdot)$ varies substantially with the sample, then a substantial amount of noise $\nu_n(\cdot)$ will be needed to ensure privacy. This first step will limit how much $\widehat{\mathrm{WC}}_n(\cdot)$ varies with the sample.

I rely on a concentration inequality due to McDiarmid (1989), applied to this setting. It relies on a bounded differences property that I will substantiate later on in this proof.

Lemma 2 (McDiarmid 1989)**.** *Suppose that $\widehat{\mathrm{WC}}_n(p)$ has the bounded differences property, where changing the ith sample realization from s to s' will change its value by at most c. Formally,*

$$\widehat{\mathrm{WC}}_n\big(p \mid S_1, \ldots, S_{i-1}, s, S_{i+1}, S_n\big) - \widehat{\mathrm{WC}}_n\big(p \mid S_1, \ldots, S_{i-1}, s', S_{i+1}, S_n\big) \leq c \tag{15}$$

Then the following concentration inequality holds. For any $t > 0$,

$$\Pr\left[\widehat{\mathrm{WC}}_n(p) - \mathrm{E}\left[\widehat{\mathrm{WC}}_n(p)\right] \geq t\right] \leq \exp\left(-\frac{2t^2}{nc^2}\right)$$

where the probability and expectation are over the sampling process.

It follows from McDiarmid's inequality and the union bound that

$$\Pr\left[\exists p \in \mathcal{P}, \widehat{\mathrm{WC}}_n(p) - \mathrm{E}\left[\widehat{\mathrm{WC}}_n(p)\right] \geq t\right] \leq n_{\mathcal{P}} \exp\left(-\frac{2t^2}{nc^2}\right) \tag{16}$$

where $n_{\mathcal{P}}$ is the number of policies in $\mathcal{P}$.

Condition (16) is enough to guarantee sample privacy. This follows from the same reasoning that McSherry and Talwar (2007) use to establish differential privacy of the exponential mechanism.

Lemma 3. *Let c ensure the bounded differences property (15). For any $t > 0$, $\hat{P}_n$ satisfies (ϵ, δ)-sample privacy where*

$$\epsilon = 2tn^{\beta} \quad \text{and} \quad \delta = n_{\mathcal{P}} \exp\left(-\frac{2t^2}{nc^2}\right)$$

to be constant. The reason is that the empirical regret bound depends on ϵ through

$$e^{\epsilon} \cdot \overline{\mathcal{RC}}_n^A(p)$$

This term vanishes as n grows because the Rademacher complexity vanishes as n grows.

To establish the bounded differences property, I rely on the robustness lemma of Camara et al. (2020). Keep in mind that changing a sample realization S_i affects not only the objective in $\widehat{\mathrm{WC}}_n(p)$, but also the constraint. Since the game between the policymaker and agent is largely arbitrary, the impact of tightening or relaxing the constraint can be difficult to capture. However, it is possible to derive a loose bound that does not depend at all on the underlying structure of the game. It only relies on the fact that the worst case is being evaluated with respect to mixed strategies.

To state the robustness lemma, I need additional notation. Consider the policymaker's worst-case utility when the agent's regret is bounded by a constant $b \geq 0$, i.e,

$$\begin{aligned} \mathrm{WC}(p, b, \pi^s) = \min_{\pi^r} & \mathrm{E}_{\pi^s, \pi^r}\left[u^P(p, r, s)\right] \\ \text{s.t.} \quad & \max_{r'} \mathrm{E}_{\pi^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\pi^s, \pi^r}\left[u^A(p, r, s)\right] \leq b \end{aligned} \tag{17}$$

Lemma 4 (Camara et al. 2020)**.** *The worst-case utility* $\mathrm{WC}(p, b, \pi^s)$ *decreases smoothly in the bound* b*. That is, for any constants* $b' > b > 0$*,*

$$\mathrm{WC}\left(p, b', \pi^s\right) \geq \mathrm{WC}(p, b, \pi^s) - \Delta^A(p)\left(\frac{b' - b}{b}\right)$$

I use this result to establish and quantify the bounded differences property, as follows.

Lemma 5. *The random variable* $\widehat{\mathrm{WC}}_n(p)$ *satisfies the bounded differences property as long as*

$$c \geq \Delta^A(p)\left(\frac{2\Delta^P(p) \cdot n^{-1}}{(4e^{\epsilon} + 4) \cdot \overline{\mathcal{RC}}_n^A(p) + \delta + \mathrm{BFR}_n}\right) + \Delta^P(p) \cdot n^{-1}$$

Now, I can define the parameter c. Recall from lemma 3 that δ depends on c. To avoid a circular definition, c should not depend on δ. Moreover, c should not depend on the particular policy p, which depends on δ indirectly through the strategy $\hat{\sigma}_n^P$. By lemma 5 and simple inequalities, it is sufficient to set

$$c \geq \max_p \left(\Delta^A(p)\left(\frac{2\Delta^P(p) \cdot n^{-1}}{(4e^{\epsilon} + 4) \cdot \overline{\mathcal{RC}}_n^A(p) + \mathrm{BFR}_n}\right) + \Delta^P(p) \cdot n^{-1}\right)$$

Technically, I could define c as the right-hand side on this expression and be done. But I will use a slightly looser bound for the sake of interpretability. Recall the constant K defined in equation (12). It is sufficient to set

$$c := K n^{-\frac{1+\alpha}{2}}$$

The last step of the proof is to derive δ_n. Recall from lemma 3 that $\hat{P}_n$ satisfies (ϵ, δ)-sample

privacy where, after plugging in the value of c,

$$\epsilon = 2tn^{\beta} \quad \text{and} \quad \delta = n_{\mathcal{P}} \exp\left(-\frac{2t^2}{K^2} \cdot n^{\alpha}\right)$$

Since this holds for any value $t > 0$, I can invert the equation $\epsilon = 2tn^{\beta}$ to find the value $t = \epsilon/(2n^{\beta})$ that keeps the parameter ϵ constant. Plugging this value of t into δ gives me δ_n, as defined in equation (11). This completes the proof of theorem 1.

5.4 Proof Outline of Theorem 2

The proof of theorem 2 has to contend with the same high-level challenge as the proof of theorem 1: namely, $\widehat{\mathrm{WC}}_n(\cdot)$ is not a friendly object. First, I prove proposition 3, which establishes convergence. Then I build on this argument to prove theorem 2.

To prove proposition 3, I need some way to characterize $\hat{P}_n$'s performance. I do this by substituting its actual performance, which is hard to describe, with its estimated performance $\widehat{\mathrm{WC}}_n\big(\hat{P}_n\big)$. The estimated performance is easier to work with because that is what $\hat{P}_n$ is maximizing. The next lemma shows that this substitution is justified.

Lemma 6. *The estimated performance $\widehat{\mathrm{WC}}_n\big(\hat{P}_n\big)$ determines a lower bound on the actual performance of $\hat{P}_n$. More precisely,*

$$\mathrm{E}_{\pi^s}\left[\mathrm{WC}_n\big(\hat{P}_n, \epsilon, \delta_n, \pi^s\big)\right] \geq \widehat{\mathrm{WC}}_n\big(\hat{P}_n\big) - n_{\mathcal{P}}\sqrt{3} \cdot n^{-\beta} - 2\mathcal{R}\mathcal{C}_n^P(\pi^s) - n_{\mathcal{P}} \exp(-n^{\alpha}) \cdot \max_p \Delta^P(p)$$

This lower bound reflects three observations. First, by construction, $\widehat{\mathrm{WC}}_n(p)$ erred on the side of being too conservative with respect to the agent's empirical regret bound. All else equal, this would mean that $\widehat{\mathrm{WC}}_n(p)$ should lower bound $\mathrm{WC}_n(p)$ with high probability. Second, $\widehat{\mathrm{WC}}_n(p)$ involves sampling noise. This can lead to generalization error, where the policymaker expects a policy to perform better than it does. I can bound the generalization error using the policymaker's Rademacher complexity. Finally, $\widehat{\mathrm{WC}}_n(p)$ involves privacy-preserving noise. By construction, this is vanishing as the sample size grows.

In light of this lemma, in order to prove proposition 3 it is enough to show that

$$\lim_{n\to\infty} \mathrm{E}_{\pi^s}\left[\widehat{\mathrm{WC}}_n\big(\hat{P}_n\big)\right] = \mathrm{CK}(\pi^s)$$

This is true because the privacy-preserving noise is vanishing as $n \to \infty$, the policymaker's empirical utility is converging in probability to his expected utility according to the true distribution, and the regret bound is vanishing as $n \to \infty$. In particular, Berge's maximum theorem implies that the policymaker's worst-case utility is continuous with respect to the regret bound.

Next, I turn to theorem 2.

I just showed that the empirical regret bound in the definition of $\widehat{\mathrm{WC}}_n(p)$ (13) is vanishing as the sample size grows. Similarly, the regret bound in the definition of $\mathrm{WC}_n(p, 0, 0, \pi^s)$ is vanishing

as the sample size grows. For a given sample size n, these bounds will take on different values. In general, they will shrink at different rates. And one bound involves empirical regret while the other involves regret with respect to the true distribution.

Despite these differences, we need to compare $\widehat{\mathrm{WC}}_n(p)$ with $\mathrm{WC}_n(p, 0, 0, \pi^s)$ to prove a result. As in theorem 1, these objects are too abstract to characterize directly. However, I can compare one abstract object with another: $\widehat{\mathrm{WC}}_n(p)$ for sample size n with $\mathrm{WC}_m(p, 0, 0, \pi^s)$ for a smaller sample size m. The idea is that if m is sufficiently small compared to n, then the regret bound for $\mathrm{WC}_m(p, 0, 0, \pi^s)$ is more conservative than the empirical regret bound for $\widehat{\mathrm{WC}}_n(p)$, even though the latter would be much more conservative if $m = n$. After accounting for some other differences, I can show that $\widehat{\mathrm{WC}}_n(p)$ is comparable to $\mathrm{WC}_m(p, 0, 0, \pi^s)$. Since I know the rate of convergence for the latter in m, I can determine the rate of convergence for the former in n.

The next lemma formalizes this argument.

Lemma 7. *The estimated performance $\widehat{\mathrm{WC}}_n\left(\hat{P}_n\right)$ with sample size n is comparable to the strategically-regularized benchmark with sample size*

$$m = \Theta(n^{1-\alpha})$$

evaluated with respect to the true distribution. More precisely,

$$\mathrm{E}_{\pi^s}\left[\widehat{\mathrm{WC}}_n\left(\hat{P}_n\right)\right] \geq \mathrm{SR}_m(\pi^s) - n_{\mathcal{P}}\sqrt{3} \cdot n^{-\beta} - 2\mathcal{RC}_n^P(\pi^s) - n_{\mathcal{P}} \exp(-n^{\alpha}) \cdot \max_p \Delta^P(p)$$

This result relies critically on another lemma that may be of independent interest. It is a lower bound on Rademacher complexity that makes use of Khintchine's inequality. This is the only part of the proof that relies on assumption 3.

Lemma 8. *Recall the constant C in assumption 3. For any policy p and distribution π^s, the Rademacher complexity is bounded below by*

$$\mathcal{RC}_n^A(p, \pi^s) \geq \frac{C}{2\sqrt{2n}} \tag{18}$$

Combining lemma 7 with lemma 6, we have

$$\begin{aligned}
\mathrm{E}_{\pi^s}\left[\mathrm{WC}_n\left(\hat{P}_n, \epsilon, \delta_n, \pi^s\right)\right] &\geq \mathrm{SR}_m(\pi^s) - O\left(n^{-\beta}\right) - O\left(n^{-1/2}\right) \\
&= \mathrm{CK}(\pi^s) - O\left(m^{-\gamma}\right) - O\left(n^{-\beta}\right) - O\left(n^{-1/2}\right) \\
&= \mathrm{CK}(\pi^s) - O\left(m^{-\gamma}\right) - O\left(n^{-\beta}\right) \\
&= \mathrm{CK}(\pi^s) - O\left(n^{-\gamma(1-\alpha)}\right) - O\left(n^{-\beta}\right) \\
&= \mathrm{CK}(\pi^s) - O\left(n^{-\min(\gamma(1-\alpha),\beta)}\right)
\end{aligned}$$

The second line follows from assumption 10. The third line removes lower-order terms. The fourth line plugs in the value $m = \Theta\left(n^{2(1-\alpha)}\right)$. The last line takes the maximum over the two rightmost terms, and completes the proof of theorem 2.

6 Illustrative Examples

I argue that penalization can lead to new insights in four illustrative examples: vaccine distribution, prescription drug approval, performance-based pay, and product bundling. These examples are not intended to be as general or realistic as possible. Instead, they are meant to convey a core insight that motivates the use of penalized policies in similar applications.

6.1 Vaccine Distribution

I present a model of vaccine distribution where penalization can motivate a common practice: insisting on statistically-significant clinical trial results before delivering medical treatments.

Decision-making based on statistical significance is hard to justify (e.g. Wasserstein and Lazar 2016) and conflicts with the recommendations of the treatment choice literature (e.g. Manski 2019). Likewise, existing solution concepts (e.g. common knowledge, maximin, and minimax regret) do not support this practice. But penalization does. The intuition is that, when vaccine quality is not common knowledge, skepticism among the population can undermine a vaccine rollout. Since vaccine distribution involves fixed costs, it may be better to wait until clinical trial results are sufficiently persuasive before attempting to vaccinate the population.

Model. Consider a town of m agents that is afflicted by a disease. A new vaccine is being developed to treat this disease. However, in order to distribute this vaccine, the policymaker must invest in a treatment center at a fixed cost c. Given the treatment center, the policymaker can treat each agent at zero marginal cost. Therefore, the policymaker must decide whether to provide treatments ($T = 1$) at cost c, or not to treat ($T = 0$).

An agent's outcome Y depends on both whether she is treated, and whether she complies with the treatment. Let $C = 1$ indicate compliance, and $C = 0$ indicate noncompliance. Let Y_1 denote her outcome conditional on being successfully treated and let Y_0 denote her outcome otherwise. For simplicity, I assume that the agent has no private information about her outcome, so that compliance C is independent of the outcomes Y_0 and Y_1.

It remains to specify payoffs and the dataset. The agent tries to maximize her expected outcome:

$$\mathrm{E}[Y \mid C, T] = \mathrm{E}\big[Y_0 + C \cdot T \cdot (Y_1 - Y_0) \mid C, T\big] = \omega_0 + \omega_1 \cdot C \cdot T$$

The parameter ω_1 is called the average treatment effect (ATE). The policymaker wants to maximize the expected welfare minus costs, i.e.

$$m\mathrm{E}[Y \mid C, T] - c$$

Both the policymaker and the agents have access to clinical trial data where compliance is guaranteed. This includes n treated outcomes Y_1^i and n untreated outcomes Y_0^i. The key summary statistic

is the sample average treatment effect, i.e.

$$\hat{\omega}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_1^i - \frac{1}{n}\sum_{i=1}^{n} Y_0^i$$

This is a sufficient statistic for both the policymaker and the agent to optimize.

Existing Solution Concepts. To establish that penalization leads to a new insight, I first need to evaluate what existing solution concepts recommend.

Claim 1. *The optimal common knowledge policy treats iff the ATE exceeds the per-capita cost, i.e.*

$$\omega_1 \geq \frac{c}{m}$$

Proof. The agent complies with the treatment iff $\omega_1 \geq 0$ and, given compliance, the policymaker prefers to treat whenever the ATE exceeds the per-capita costs. □

Claim 2. *The maximin policy never treats.*

Proof. If the policymaker treats, he incurs a cost $c > 0$ and the agents may not comply anyways, leading to a negative payoff in the worst case. By not treating, he guarantees a payoff of zero. □

Claim 3. *The minimax regret policy never treats iff the ATE exceeds twice the per-capita cost, i.e.*

$$\omega_1 \geq \frac{2c}{m}$$

Proof. If the policymaker does not treat, the maximum regret $m\omega_1 - c$ occurs when the agents would have complied with treatment. If he does treat the population, the maximum regret c occurs when the agents decide not to comply. The policymaker treats iff $m\omega_1 - c \leq c$, or $\omega_1 \leq 2c/m$. □

None of these solutions justify statistically-significant clinical trial results as a precondition for treatment. Furthermore, this conclusion does not depend on the assumption that the policymaker knows the true distribution. The literature on treatment choice has studied problems like these, under the assumption that compliance is independent of sample size. It recommends statistical decision rules like empirical welfare maximization (e.g. Manski 2004b, Stoye 2009, Kitagawa and Tetenov 2018, Mbakop and Tabord-Meehan 2021). In this case, the empirical welfare maximizer under common knowledge would treat iff the sample ATE exceeds cost per-capita, i.e.

$$\hat{\omega}_1 \geq \frac{c}{m} \tag{19}$$

This rule follows the preponderence of the evidence but does not insist on statistical significance. In particular, the threshold for treatment does not vary with the sample size n.

Optimal Penalized Policy. Penalization will motivate a form of statistical significance because, in this model, a treatment is only successful if agents agree to comply with said treatment. When clinical trial results are not statististically significant, agents may be sufficiently uncertain that they decide not to comply, even if compliance is optimal under the true distribution.

To substantiate this intuition, I solve for the optimal penalized policy. To define a regret bound, I assume that the ATE is bounded, i.e. $\omega_1 \in [\underline{\omega}_1, \bar{\omega}_1]$ where $\bar{\omega}_1 > c/m > \underline{\omega}_1$, and I define

$$\overline{\mathcal{RC}}_n^A(T=0) := 0 \quad \text{and} \quad \overline{\mathcal{RC}}_n^A(T=1) := \frac{(\bar{\omega}_1 - \underline{\omega}_1)\sqrt{2\ln 2}}{\sqrt{n}}$$

This is a valid upper bound on the Rademacher complexity (see Massart's lemma in remark 3).

Claim 4. *The optimal penalized policy treats iff*

$$\omega_1 \geq \frac{c}{m} + O\left(n^{-1/2}\right)$$

Proof. The policymaker never treats if $\omega_1 < c/m$, since his payoff would be negative regardless of whether the agents comply. Suppose $\omega \geq c/m$. Given treatment, the maximum probability q of non-compliance that satisfies the regret bound solves

$$q \cdot \omega_1 + (1-q) \cdot 0 = \frac{(\bar{\omega}_1 - \underline{\omega}_1)\sqrt{2\ln 2}}{\sqrt{n}}$$

where the left-hand side is the agent's regret and the right-hand side is the regret bound. This leaves

$$q = \frac{(\bar{\omega}_1 - \underline{\omega}_1)\sqrt{2\ln 2}}{\omega_1\sqrt{n}}$$

The policymaker's worst-case payoff from treatment is $(1-q)m \cdot \omega_1 - c$. Plugging in the value of q, he prefers to treat iff

$$m\omega_1\left(1 - \frac{(\bar{\omega}_1 - \underline{\omega}_1)\sqrt{2\ln 2}}{\omega_1\sqrt{n}}\right) - c \geq 0$$

This simplifies to

$$\omega_1 \geq \frac{c}{m} + \frac{(\bar{\omega}_1 - \underline{\omega}_1)\sqrt{2\ln 2}}{\sqrt{n}}$$

□

The $O(n^{-1/2})$ term is analogous to a critical value. In that sense, the optimal penalized policy insists that the difference between the ATE and the per-capita cost exceeds that critical value. Of course, this requires the policymaker to know the true distribution. In general, he would have to

evaluate the sample ATE rather than the true ATE, in line with the data-driven penalized policy developed in section 5. But the essential intuition survives.

6.2 Prescription Drug Approval

In a model of prescription drug approval by a regulator, I show that strategic regularization restricts doctors' ability to prescribe drugs that have not been proven effective in clinical trials. The threshold for approval increases as more drugs are approved. This is similar to stepwise methods for multiple hypothesis testing (e.g. Holm 1979, Romano and Wolf 2005).

In contrast, in models with a common prior or rational expectations, the optimal policy is to approve all drugs. Essentially, this delegates the decision to doctors, who are better informed than the regulator. In my model, however, doctors may prescribe ineffective drugs if the clinical trial returns a "false positive" where an ineffective drug appears to be effective by random chance. Limiting the number of drugs approved can reduce the risk of false positives, and provide better welfare guarantees.

Model. A population of patients is afflicted by a disease. There are m treatments available, as well as a placebo. As in the previous example (section 6.1), let $\omega_j \in [\underline{\omega}, \bar{\omega}]$ be the average treatment effect (ATE) of treatment j. The placebo's treatment effect is normalized to zero. In addition, patients incur a private cost $c_j \in [0, \bar{c}]$ from treatment j, where the maximum cost $\bar{c} > \bar{\omega}$ exceeds the maximum treatment effect. For example, this could represent the patient's copay for a prescription drug. The patients are nonstrategic and accept whatever treatment is offered.

There is a regulator (policymaker) who approves treatments and a doctor (agent) who prescribes them. Formally, the regulator specifies a set $\mathcal{A} \subseteq \{1, \dots, m\}$ of approved treatments. Then the doctor either prescribes a treatment $j \in \mathcal{A}$ to a given patient, or prescribes the placebo $j = \emptyset$. Formally, the doctors response r is a choice function, mapping sets $\mathcal{A}$ to treatments $j \in \mathcal{A} \cup \{\emptyset\}$.

Both participants have access to clinical trial data with sample size n, where $\hat{\omega}_j$ is the sample ATE of treatment j. Both participants want to maximize the patient's expected outcome minus costs, i.e. $\omega_j - c_j$ for the chosen treatment j. But the doctor has an informational advantage. She knows patient costs c_j at the time of treatment choice, but the regulator does not.

There are many reasonable ways to accommodate the uncertainty in the costs c_j. I assume that the policymaker evaluates his worst-case regret with respect to these costs (not necessarily with respect to the agent's beliefs). That is,

$$u^P(\mathcal{A}, r, \vec{\omega}) = \max_{\vec{c}} \left(\max_{\mathcal{A}'} \left(\omega_{r(\mathcal{A}')} - c_{r(\mathcal{A}')} \right) - \left(\omega_{r(\mathcal{A})} - c_{r(\mathcal{A})} \right) \right)$$

Existing Solution Concepts. I first establish what existing solution concepts recommend, before moving onto the optimal penalized policy.

Claim 5. *The optimal common knowledge policy approves all treatments.*

Proof. If the ATEs are common knowledge, the doctor will choose the treatment j that maximizes outcome minus costs, i.e. $\omega_j - c_j$. The regulator prefers this to any other policy. If he excludes a treatment j, it is always possible that treatment j was the only treatment with low costs, e.g. where $c_j = 0$ and $c_i = \bar{c}$ for $i \neq j$. In that case, the regulator may regret excluding treatment j. □

These same conclusion holds if the regulator does not know the true disribution, as long as the doctor does. In that case, the regulator has even more incentive to defer to the doctor.

Claim 6. *The maximin policy approves no treatments.*

Proof. Suppose the approved set $\mathcal{A}$ is nonempty. The worst case outcome occurs when all treatments $j \in \mathcal{A}$ share the maximum cost, $c_j = \bar{c}$, and the doctor chooses the worst treatment $j \in \mathcal{A}$ because she believes it to be highly effective. The regulator's utility is

$$\min_{j \in \mathcal{A}} \omega_j - \bar{c}$$

This is always negative, since $\omega_j < \bar{c}$. It is better to not approve any treatments, since this at least guarantees non-negative utility for the regulator. □

Claim 7. *The minimax regret policy approves no treatments.*

Proof. Suppose that all treatments have zero cost and the doctor believes they have zero effectiveness. Let j^* be the treatment with the highest ATE.

Suppose the regulator deviates from approvals $\mathcal{A}$ to approvals $\mathcal{A}'$ that includes j^*. The doctor breaks her indifference in favor of the placebo when presented with $\mathcal{A}$, but breaks her indifference in favor of treatment j^* when presented with $\mathcal{A}'$. The regulator's regret is given by $\omega_{j^*} = \max_j \omega_j$. Therefore, his worst-case regret is bounded below by ω_{j^*}. In contrast, approving no treatments guarantees that the regulator's regret is bounded above by ω_{j^*}. Therefore, approving no treatments minimizes worst-case regret. □

Optimal Penalized Policy. Penalization will motivate policies that are less extreme and more realistic, where the regulator approves some treatments but not all. The intuition is that the doctor faces a multiple testing problem. The more treatments are approved, the greater the chances of a false positive – a treatment that appears to be effective but is not. Approving too many treatments can cause doctors to prescribe false positives with high probability.

To substantiate this intuition, I solve for the optimal penalized policy, where

$$\overline{\mathcal{RC}}_n^A(\mathcal{A}) := \frac{(\bar{\omega} - \underline{\omega})\sqrt{2 \ln (|\mathcal{A}| + 1)}}{\sqrt{n}}$$

As in the previous subsection, this bound follows from Massart's lemma (remark 3). Note that the regret bound is increasing in the $|\mathcal{A}|$ of approved treatments.[9]

[9]The agent's response space is technically larger than $|\mathcal{A}| + 1$ since it consists of maps from $\mathcal{A}$ to either a treatment $j \in \mathcal{A}$ or the placebo. But for a fixed policy $\mathcal{A}$, it is without loss to restrict attention to $|\mathcal{A}| + 1$ unique actions.

The optimal penalized policy begins by ordering treatments according to their ATE. Let $\omega_{(k)}$ denote the k^{th} highest ATE.

Claim 8. *The optimal penalized policy approves the k^{th} best treatment iff*

$$\omega_{(k)} \geq \frac{(\bar{\omega} - \underline{\omega})\sqrt{2\ln(k+1)}}{\sqrt{n}}$$

Proof. To derive the optimal penalized policy, I refer to two terms that reflect the regulator's regret with respect from the unknown cost. First, the regulator's regret from not approving treatment j will be ω_j in the worst case. This occurs when treatment j has zero cost and all other treatments have maximum cost. It follows that worst-case regret is at least $\max_{j \notin \mathcal{A}} \omega_j$.

Second, the regulator's regret from approving a treatment j will be $\overline{\mathcal{RC}}_n^{\mathcal{A}}(\mathcal{A})$ in the worst case. Suppose $\mathcal{A}$ is nonempty. This level of regret can be achieved by setting the cost of all treatments that the doctor does not choose to $\bar{c}$, and the cost of the treatment j that the doctor does choose to $c_j = \omega_j + \overline{\mathcal{RC}}_n^{\mathcal{A}}(\mathcal{A})$. The doctor will be indifferent between treatment j and the placebo. At the same time, the $\overline{\mathcal{RC}}_n^{\mathcal{A}}(\mathcal{A})$ is an upper bound on the regulator's regret, because that is an upper bound on the agent's regret and both participants care about welfare.

The optimal penalized policy minimizes the larger of these two regret terms. It begins by ordering treatments according to their ATE. The regulator approves the best treatment iff

$$\omega_{(1)} \geq \frac{(\bar{\omega} - \underline{\omega})\sqrt{2}}{\sqrt{n}}$$

Similarly, the regulator approves the kth best treatment iff

$$\omega_{(k)} \geq \frac{(\bar{\omega} - \underline{\omega})\sqrt{2\ln(k+1)}}{\sqrt{n}}$$

□

As in section 6.1, treatments are approved only if they reach a critical value. More treatments are approved as the sample size n grows, since the critical value decreases according to $O(n^{-1/2})$. Moreover, the critical value is increasing in the number of treatments already approved. As I mentioned earlier in this section, this is similar to stepwise methods for multiple hypothesis testing. That is not surprising, because the motivation for limiting the number of approved drugs is precisely to ensure that doctors are not misled by false positives.

Role of Sample Privacy. In Subsection 3.3, I claimed that bounds from statistical learning theory on the agent's regret may not be valid if the policymaker is also learning from data. This model of prescription drug approval makes that clear. I elaborate below.

The doctor's regret from expected utility maximization can vary significantly based on how the regulator uses the sample. Suppose the regulator approves a single treatment j independently of the sample. The doctor's regret will be on the order of $O\left(n^{-1/2}\right)$. That is the bound implied by Massart's lemma (remark 3). Alternatively, suppose that regulator only approves the empirical utility maximizer, i.e. $\arg\max_j \hat{\omega}_j$. In that case, the doctor's choice is the same as if all treatments had been approved, and her regret will be on the order of $O\left(n^{-1/2}\log m\right)$. Again, that is the bound implied by Massart's lemma (remark 3). The dependence on m reflects a multiple testing problem that arises even though the agent only has two choices: the approved treatment and the placebo. It arises because the choice presented to her is correlated with the sample.

Sample privacy means that, with some probability, the regulator is not selecting the k treatments whose sample ATE is highest. More generally, it is well-known that using data to determine the hypotheses one wants to test will threaten the validity of those tests. There are several ways to get around this. For example, the regulator could use his prior beliefs to fix the k most promising treatments that he wants to evaluate and approve. In contrast, the approach that I develop in Section 5.1 is data-driven, and uses sample privacy to control how aggressively the data is used.

6.3 Performance Pay

Third, I consider a model of performance pay. An employer incentivizes an employee to exert costly effort by paying wages contingent on observed performance. Here, penalization caps and flattens the wage schedule. Wages are zero until performance reaches some threshold. At that point, they jump and remain flat. The employee receives moderate pay, but with high probability.

In contrast, the common knowledge solution recommends very high pay, but only if the employee attains the best possible performance is obtained. In the penalized model, this does not work well. If historical data is limited, it may not be obvious to the employee that it is worth investing effort for a small chance of receiving the bonus. Flatter contracts may be less potent, but they provide clearer incentives.

This insight contributes to the recent literature in robust contract design (e.g. Carroll 2015; Carroll and Meng 2016a,b; Dütting et al. 2019) that, in turn, builds on previous attempts to explain the ubiquity of simple contracts (e.g. Holmstrom and Milgrom 1987, 1991). In particular, Holmstrom and Milgrom (1991) motivate fixed wages when the principal is unable to measure some dimensions of the agent's performance. The optimal penalized contract developed here suggests that learnability may be another motivation.[10]

Model. I define a simple principal-agent problem, along the lines of Sappington (1983). A principal wants to incentivize the agent to take desirable actions. The timing of the game is as follows:

[10]Relatedly, Valenzuela-Stookey (2020) proposes an axiomatic model where an agent's evaluation of lotteries may be imprecise. He motivates his axioms in a frequentist model of learning. He applies his representation towards a general principal-agent problem where, if the agent is suitably cautious, optimal contracts correspond to step functions. In the principal-agent model that I consider, the optimal contract is a step function even in the standard case where the agent knows the true distribution.

first, the principal commits to a wage schedule or contract w; second, the agent takes a hidden action a; third, the principal observes an outcome x and pays the agent a wage $w(x)$.

Let $\mathcal{X} = \{x_1, \dots, x_m\} \subseteq \mathbb{R}$ be a finite outcome space, in increasing order. Outcomes are determined, stochastically, by the agent's hidden action. Let $A = \{0, 1\}$ be a binary action space, where $a = 0$ indicates no effort and $a = 1$ indicates effort. The agent incurs a cost of effort $c > 0$. Let π_0^x be the outcome distribution conditional on no effort. Let π_1^x be the outcome distribution conditional on effort. I assume both distributions have full support. Let $X^a \sim \pi_a^x$ denote the realized outcome when the agent takes action a.

I restrict attention to conditional distributions that satisfy a monotone likelihood ratio property.

Assumption 4. *The likelihood ratio $\ell(x) = \pi_1^x(x)/\pi_0^x(x)$ is weakly increasing in x.*

Before the agent acts, the principal commits to a wage function $w : \mathcal{X} \to \mathbb{R}_+$.[11] By definition, the wage function satisfies *limited liability* (i.e. $w(x) \geq 0$ for all $x \in \mathcal{X}$). Both the agent and the principal are risk-neutral. Given action a and outcome x, the agent's utility is $w(x) - c \cdot a$ while the principal's utility is $x - w(x)$.

Let $X_1^a, \dots, Y_n^a \sim \pi_a^x$ be i.i.d. samples of outcomes for each action $a \in \{0, 1\}$. For example, this could be data that a manager collects through personal experimentation, or through costly monitoring of past employees, and shares with his current employees.

Existing Solution Concepts. I first establish what existing solution concepts recommend, before moving onto the optimal penalized policy. These results will refer to a *zero contract* w, which sets, for all $x_i \in \mathcal{X}$, $w(x_i) = 0$.

The optimal common knowledge contract makes a recommendation that seems extreme: pay the agent if and only if the realized performance is the highest possible.

Claim 9. *The optimal common knowledge contract is either the zero contract, or sets*

$$\forall x_i \in \mathcal{X} \quad w(x_i) = \begin{cases} \frac{c}{\pi_1^x(x_i) - \pi_0^x(x_i)} & i = m \\ 0 & i < m \end{cases}$$

For the the prior-free solution concepts, on the other hand, there is often no contract that guarantees effort by the agent.

Claim 10. *The maximin contract is the zero contract.*

Proof. Suppose the agent beliefs that the outcome distribution does not depend on effort, i.e. $\pi_0^x = x_1^x$. Then no contract can incentivize her to put in effort. Since the outcome distribution has full support, any non-zero contract would make positive payments for no change in effort. □

[11] The reader is welcome to treat wages as dollar-valued in order to ensure that the policy space is finite, as specified in assumption 1. This does not meaningfully affect the analysis.

Claim 11. *Let π_1^x be the uniform distribution. If the number of outcomes m is sufficiently large, then the minimax regret contract is the zero contract.*

Proof. Let w be the minimax regret contract. Suppose the agent believes that putting in effort guarantees outcome x_i and not putting in effort guarantees outcome $x_j < x_i$ (satisfying assumption 4). Let $q_i = \min_{i'} \Pr_{\pi_1^x}\left[x_{i'}\right]$ be the probability of outcome x_i The principal can incentivize the agent by paying her cost c conditional on outcome x_i. This costs him $q_i c$ in expectation, and he benefits from the surplus generated by the agent's effort.

If the contract w does not incentivize effort under these beliefs, then regret is the surplus $\mathrm{E}_{\pi^x}\left[X^1 - X^0\right]$ plus the expected wages $\mathrm{E}_{\pi^x}\left[w(X^1)\right]$ minus $q_i c$. I claim that the contract w cannot incentivize effort when m is large. In order to guarantee effort across all choices of x_i, the contract would need to ensure that $w(x_i) = w(x_j) + c$ for all $i, j < i$. In particular, set $j = i - 1$ and invoke limited liability to see that $w(x_i) = c \cdot i$. As $m \to \infty$, the wages paid at any $i > m/2$ grows to infinity as well. This contradicts the optimality of the contract w.

It follows that, for large m, regret is the surplus $\mathrm{E}_{\pi^x}\left[X^1 - X^0\right]$ plus the expected wages $\mathrm{E}_{\pi^x}\left[w(X^1)\right]$ minus $q_i c$. For the uniform distribution, $q_i = q$ does not depend on the contract, To minimize regret it suffices to minimize the expected wages, by letting w be the zero contract. □

Optimal Penalized Contract. The optimal penalized policy caps and flattens the optimal common knowledge contract. This makes the incentives for effort less potent, but it also makes them clearer. Given limited data, the agent is still able to determine that effort is optimal.

To substantiate this intuition, I solve for the optimal penalized policy, where

$$\overline{\mathcal{RC}}_n^A(w) := \left(\max_i w(x_i)\right) \cdot \sqrt{2 \ln 2}\sqrt{n}$$

As in the previous subsections, this bound follows from Massart's lemma (remark 3). Note that the regret bound is increasing in the *maximum wage*, i.e. $\max_i w(x_i)$.

Claim 12. *The optimal penalized contract is a threshold contract. More precisely, it optimizes the following contract across all maximum wages $\bar{w}$ and probabilities q of effort. Define*

$$\alpha_j = \frac{c + \bar{w}\left(\frac{4}{q} \cdot \sqrt{\frac{2 \log 2}{n}} - \sum_{i=j+1}^{m}\left(\pi_1^x(x_i) - \pi_0^x(x_i)\right)\right)}{\pi_1^x(x_j) - \pi_0^x(x_j)}$$

Let k be the largest j such that $\ell(x_j) > 1$ and $\alpha_j \leq \bar{w}$. If no such integer exists, set $w(x) = 0$ for all outcomes $x \in \mathcal{X}$. Otherwise, set $w(x_i)$ to equal $\bar{w}$ if $i > k$, α_k if $i = k$, and 0 if $i < k$.

The optimal penalized contract identifies the maximum wage as an hindrance to learnability. When wages conditional on high performance are very large, small changes in the perceived probability of high performance conditional on effort can have a large impact on the perceived utility of effort. For that reason, the principal would have to add a premium to the agent's wages in order to ensure that the agent puts in effort. This makes the contract actuarially less appealing. On the other

hand, if the maximum wage is capped, then the perceived utility of effort is less sensitive to beliefs. The premium needed to incentivize effort is smaller. The optimal penalized policy balances the benefits of higher maximum wages with the costs identified here.

6.4 Product Bundling

Finally, I consider a model of product bundling. A firm has several products for sale and wants to sell them in a way that maximizes expected profit. When the sample size is small, the penalized policy only offers large bundles of products, rather than selling them separately. This contrasts with prior work that recommends selling items separately (e.g. Carroll 2017).

In my model, the reason for bundling is that consumers learn about their value for the product through reviews. If there are many products, but few reviews per product, consumers can be confident in the value of a large bundle while being uncertain about the value of any given product. In that case, all else equal, it is easier to convince consumers to buy the bundle.

Model. There are n goods, one seller, and one buyer. The buyer has preferences over bundles $x \in \{0,1\}^n$ of goods. The seller can offer a menu M consisting of bundles x at prices p, i.e. $(x,p) \in M$. The seller cannot prevent the buyer from buying more than one good. If $(x,p),(x',p') \in M$ then for bundle

$$x'' = \left(\max\{x_1, x_1'\}, \dots, \max\{x_m, x_m'\}\right)$$

there exists a price $p'' \leq p + p'$ such that $(x'',p'') \in M$. The buyer's value for each of the goods is described by a vector $v \in [0,\bar{v}]^n$. Her utility from choosing a bundle $(x,p) \in M$ is

$$u^A(M,(x,p),v) = v \cdot x - p$$

The seller cares about profits and, for simplicity, has zero costs of production. His utility is p.

To relax the assumption that the buyer knows her value from product j, I assume that she knows her ranking relative to other agents. More precisely, she knows the quantile q_j of her value v_j in the marginal distribution π_j^v. Given menu M and a chosen bundle $(x,p) \in M$, the buyer at quantile $q = (q_1, \dots, q_m)$ receives a payoff of

$$U^A(q,M,(x,p),\pi^v) := \sum_{j=1}^{m} x_j \cdot \inf\left\{u \in \mathbb{R} \mid q \leq \Pr_{\pi^v}\left[v_j \leq u\right]\right\} - p$$

The quantile is her private type. If she knows π_j^v then she knows the value she derives from product j. If she is uncertain about π_j^v, then she faces uncertainty about her value. For example, a Netflix user might understand that she is particularly predisposed towards science fiction, but not know whether the quality of a particular science fiction movie.

The following assumption will be convenient later on.

Assumption 5. *The marginal distributions π_j^v have well-defined density functions f_j. There exists a constant $K > 0$ where $f_i(v_j) \geq K$ is bounded below by that constant on its support $[0,\bar{v}]$.*

I assume that there is a common dataset consisting of product reviews. Each review of product j is a single observation of the value v_j of the product to the user that reviewed it. Each product has n reviews. This data identifies the marginal distributions π_j^v, but not the joint distribution of the value profiles v. This is because I do not observe a n users observing m products, but rather nm users observing 1 product each. Arguably, this better reflects the kind of review data that would be available in practice.

Since only the marginal distributions are identified by the available data, I do not assume that the seller knows the joint distribution. Instead, I follow Carroll (2017) in assuming that the seller knows (at most) the marginal distributions. He evaluates his profits with respect to the worst-case joint distribution that is consistent with the true marginal distributions.

Existing Solution Concepts. I first establish what existing solution concepts recommend, before moving onto the optimal penalized policy. I rely on Carroll's (2017) characterization of the optimal menu when the marginal distributions are common knowledge.

Claim 13 (Carroll 2017)**.** *Selling each product separately is an optimal common knowledge menu. That is, each product j can be purchased individually at a price p_j. The price of a bundle x is the sum of prices $\sum_{j=1}^{m} x_j p_j$ of the constituent goods.*

The maxmin criterion is useless here because it fails to distinguish between any menus.

Claim 14. *Every menu is a maxmin menu.*

Proof. Suppose the buyer believes that the distribution π^{v^j} assigns probability one to $v^j = 0$. In that case, no menu can guarantee a positive payoff. □

The minimax regret criterion is only marginally more useful. It does not distinguish between menus that only sell the grand bundle and menus that also sell the goods separately. It specifies a price for the grand bundle that depends on the maximum value $\bar{v}$. If we allow $\bar{v} \to \infty$, the price of any bundle x in the menu grows to infinity, regardless of whether consumers are likely to have these extreme values.

Claim 15. *A menu is a minimax regret menu iff the price of the grand bundle is $m\bar{v}/2$.*

Proof. Let M be a menu. Suppose the buyer believes that the distributions π^{v^j} assigns probability one to $v^j = \bar{v}$. The seller regrets not choosing a menu M' that offers the grand bundle at price $m\bar{v}$. Specifically, his regret from M is $m\bar{v}$ minus the price of the grand bundle in M.

Suppose the buyer believes that the distributions π^{v^j} assigns probability one to values that make the buyer exactly indifferent between buying and not buying the grand bundle in M. In particular, by the definition of menus, I can set values such that the buyer is exactly indifferent between buying and not buying any bundle $(x, p) \in M$. She breaks indifferences in favor of not buying. The seller regrets not choosing a menu M' that offers the grand bundle at an ϵ discount, for arbitrarily small $\epsilon > 0$. This would generate profits arbitrarily close to the price of the grand bundle in M, whereas M generates zero profits. Therefore, the seller's regret from M is price of the grand bundle in M.

Altogether, the maximum regret is minimized when the menu M splits the difference and offers the grand bundle at price $m\bar{v}/2$. □

Optimal Penalized Menu. In contrast to the previous results, the optimal penalized policy will recommend bundling when the sample size is small. To formalize this, I need to define the penalized policy in a setting where the agent cares about quantile, rather than expected utility.

First, I redefine regret. Let x_n denote the bundle that the buyer purchases, as a function of the realized sample. Given menu M, the buyer's *quantile regret* is

$$\text{Q-Regret}(M, q, \pi^v) = \max_{(x,p)\in M} U^A(q, M, (x, p), \pi^v) - U^A\left(q, M, x_n, \pi^v\right)$$

Next, I specify a bound B on quantile regret. Let $H > 0$ be a constant and let K be defined as in assumption 5. Then

$$B(M, q, \pi^v) := 4HK^{-1}\sqrt{\frac{\ln 2 + \ln|M| + \ln n}{n}}$$

The following proposition implies that this regret bound is feasible in the spirit of Definition 3.

Claim 16. *There exists a constant $H > 0$ and a buyer's strategy x_n such that*

$$\text{Q-Regret}(M, q, \pi^v) \leq 4HK^{-1}\sqrt{\frac{\ln 2 + \ln|M| + \ln n}{n}}$$

The constant H varies with model parameters ($\bar{v}$, m, and K) but does not depend on π^s or M.

The key feature of the regret bound B is its dependence on the menu size $|M|$. This has immediate implications. If the seller only offers the grand bundle, then $|M| = 1$ and the term $\ln|M| = 0$ in the bound disappears. On the other hand, if the seller sells every item separately, then $|M| = 2^m$ and the term $\ln|M| = m$ in the bound is equal to the number of products. As such, if there are many products, selling separately could mean much less predictable buyer behavior compared to only offering the grand bundle.

The next claim formalizes this intuition with a limiting argument. It refers to the menu M^S in which all goods are sold separately (so $|M^S| = 2^m$).

Claim 17. *Let M be an optimal penalized menu. For every sample size n and fraction $\alpha \in (0, 1)$, there exists a sufficiently large number m of products such that the menu size $|M|$ is less than an α-fraction of the menu size from selling separately, i.e.*

$$\frac{|M|}{|M^S|} \leq \alpha$$

Proof. Suppose for contradiction that $|M| \geq \alpha|M^S|$, i.e. $|M| \geq \alpha 2^m$. As $m \to \infty$, the regret bound grows to infinity since $|M|$ grows to infinity. This causes the worst-case payoff to fall to zero. However, offering only the grand bundle at a suitable price sets $|M| = 1$ and ensures a positive profit (by assumption 5). This contradicts the optimality of M.

□

7 Related Literature

This work contributes to three research efforts. For robust mechanism design, it is a principled way to interpolate between two extremes: the common prior and prior-freeness. For learning in games, it provides a convenient behavioral assumption that does not rely on agents using a particular model or estimator. For data-driven mechanism design, it extends existing work to settings where the agent is learning, not just the policymaker. Below, I elaborate on each of these contributions.

Robust Mechanism Design. Robust mechanism design tries to relax the common prior assumption, as well as other knowledge assumptions used in mechanism design.

Initially, this literature focused on prior-free solution concepts that assumed no distributional knowledge whatsoever. Early on, Bergemann and Morris (2005) and Chung and Ely (2007) gave prior-free foundations for ex post incentive compatibility as a solution concept.[12] These papers worked with Harsanyi type spaces, where type profiles encode both the distribution of the state (or payoff type) as well as agents' higher-order beliefs. They sought to implement a social choice correspondence in any Bayes-Nash equilibrium of any type space.[13]

Prior-free solution concepts and the common prior assumption are two extreme cases, but there is a rich terrain that lies between them. For example, Oury and Tercieux (2012) propose *continuous implementation*. Given a type space that satisfies the common prior, the designer wants to implement a social choice correspondence in all type spaces that are arbitrarily close to the original one. Other researchers take a similar approach (e.g. Meyer-ter-Vehn and Morris 2011, Jehiel, Meyer-ter-Vehn, and Moldovanu 2012). Alternatively, Artemov et al. (2013) assume Δ-rationalizability (Battigalli and Siniscalchi 2003), where it is commonly known that the state distribution belongs to some pre-specified set Δ. In a similar spirit, Ollár and Penta (2017) assume that only pre-specified moments of the state distribution are common knowledge.

My work can also be seen as straddling the divide between prior-freeness on the one hand and the common prior on the other.[14] It is clearly inspired by the robustness literature, but its methods are different. Rather than specify a moment restriction or set Δ of plausible distributions, I assume that the agent has access to a dataset with sample size n. The parameter n controls how knowledgeable the policymaker and agent are supposed to be. It is a principled way to interpolate between prior-freeness ($n = 0$) and the common prior ($n = \infty$). This is true regardless of whether the dataset is interpreted literally (as in this paper), or as a stylized model of shared experience.

My model has three advantages relative to the nearest alternatives, namely Artemov et al. (2013) and Ollár and Penta (2017). First, it has few tuning parameters. The only parameter related to beliefs is the sample size n. Second, it makes it easier to decide "how much" robustness is required. I posit

[12] Later contributions moved beyond ex post incentive compatibility (e.g. Börgers and Smith 2014, Börgers 2017).

[13] Prior-free approaches also developed in algorithmic game theory (Goldberg et al. 2006). This work focused on prior-free approximations to Bayesian-optimal mechanisms, whereas the economics literature focused on worst-case optimal mechanisms and characterizing which social choice correspondences were implementable.

[14] Granted, I avoid some of the issues related to strategic uncertainty that the prior literature has to deal with, due to my focus on single-agent mechanism design problems. But my learning-theoretic approach to relaxing the common prior assumption also seems to be compatible with multi-agent settings (c.f. Liang 2020).

that researchers find it easier to gauge whether a sample size (or rate of convergence) is reasonable for their setting of interest, compared to an arbitrary set of beliefs or a set of moment restrictions. Third, it has a clear learning foundation. This is important because "robust" predictions can be quite sensitive to how one departs from the common prior assumption, so we need a good justification if we want to prioritize one over the other.[15]

Learning in Games. The literature on learning in games tries to replace prior knowledge or equilibrium assumptions with a more explicit process of learning from historical data. It is useful to divide this literature along two dimensions. First, whether data arises from repeated interaction (i.e. online learning) or random sampling (i.e. batch learning). Second, whether agents are learning about each other's strategies or about the state of nature. I am primarily concerned with models where agents learn about the state of nature through random sampling.

Liang (2020) is particularly relevant to my work. The author also studies incomplete information games where agents learn about the state through a finite dataset. In her model, agents adopt learning rules from a prespecified class of learning rules. If the learning rules are consistent, and converge uniformly, then predicted behavior is compatible with the common prior assumption in the limit as the sample size grows. In finite samples, predictions that hold under the common prior assumption (like the no-trade theorem) might not be necessarily true.

However, this paper differs from Liang (2020) in two respects. First, I commit to a particular class of learning rules: those that satisfy my regret bound. This class contains all learning rules that perform at least as well as empirical utility maximization, given the true distribution. By identifying a natural class of learning rules, I reduce the burden on researchers who want to use Liang's method. Second, my goal is policy design rather than predicting behavior. The new insights from my model do not come from agents learning per se. Instead, they come from the fact that policy choices can impact how quickly agents learn.

Researchers have also looked at the implications of statistical complexity for economic behavior. Some of this work considers the trade-off, from the agent's perspective, of choosing more or less complex statistical models to estimate (e.g. Al-Najjar and Pai 2014, Olea et al. 2021). Other work studies models of bounded rationality that can be motivated as a response to statistical complexity (e.g. Valenzuela-Stookey 2020, Jehiel 2005). In contrast, my work looks at how policy choices can make the agent's learning problem more or less complex. Furthermore, I do not assume that the agent is frequentist or that she relies on a particular statistical model.

Finally, researchers have also studied environments where agents' beliefs may not converge. This could be due to bounded rationality (Aragones et al. 2005; Haghtalab et al. 2021) or the fact that the environment is hopelessly complicated (Mailath and Samuelson 2020; Al-Najjar 2009). In particular, Al-Najjar (2009) relies on the notion of VC dimension, which is closely linked to the notion of Rademacher complexity used in this paper. Aside from this, however, the focus of

[15]For example, if departures are defined using the product topology on the universal type space, then even small departures from the common prior can lead to drastic changes in predicted behavior (Lipman 2003; Rubinstein 1989; Weinstein and Yildiz 2007). In contrast, under the strategic topology, small departures from the common prior lead to small changes in predicted behavior (Chen et al. 2010; Dekel et al. 2006).

these papers is different from my own. My assumptions imply that the agent can learn her optimal response to any policy, given sufficient data.

Data-driven Mechanism Design. The literature on data-driven mechanism design fuses robust mechanism design with learning in games. The goal is to combine microeconomic theory with data to provide more concrete policy recommendations. There is not much interaction between this literature and prior work in structural econometrics, presumably because it is driven by a community of computer scientists and microeconomic theorists, rather than empirical economists. As a result, the methods are somewhat different. The focus tends to be decision-theoretic, in line with Manski (2019), and there is less emphasis on estimating model parameters per se.

One prominent line of work studies the sample complexity of auction design. Here, the auctioneer lacks prior knowledge of the distribution of bidder values. Instead, he has access to a dataset, usually consisting of i.i.d. draws from the value distribution. A typical question is how many draws are needed in order for the auctioneer to guarantee near-optimal revenue with high probability (e.g. Balcan et al. 2008, Cole and Roughgarden 2014). Many of these papers rely on measures of learning complexity, like covering numbers (Balcan et al. 2008), pseudo-dimension (Morgenstern and Roughgarden 2015), and Rademacher complexity (Syrgkanis 2017). Gonçalves and Furtado (2020) use more familiar econometric methods towards a similar end.

These papers are focused on the auctioneer's learning problem, but ignore the bidders' learning problems. This is possible because of the application that they emphasize: auctions with dominant strategies, where agents have independent private values. In that context, there is no need for agents to learn about the value distribution. But this is not true in general. For example, in models with interdependent values, implementating reasonable outcomes in dominant strategies may be impossible (Jehiel, Meyer-ter-Vehn, Moldovanu, and Zame 2006). Alternatively, consider problems like monopoly regulation, contract design, or Bayesian persuasion. In these problems, the agent's optimal action depends on her beliefs over a hidden state of nature.

There is some prior work where both the policymaker and the agents are learning from data. For example, Camara et al. (2020) also study a single-agent policy design problem. Their model incorporates online learning, where data is generated over time through repeated interaction, and the data-generating process is arbitrary. In contrast, I consider batch learning, where data is generated from random sampling. Furthermore, Cummings et al. (2020) and Immorlica et al. (2020) consider agents that are learning from i.i.d. samples, respectively, in models of price discrimination and social learning. Both papers assume that agents' beliefs converge to the true distribution at a reasonable rate. In contrast, I assume that the agent's regret converges to zero at a reasonable rate. Although these assumptions should be mutually compatible, the advantage of regret bounds is that they make explicit how policies affect the complexity of the agent's learning problem.

8 Conclusion

I proposed a modeling assumption that replaces common knowledge with a common dataset. I studied this in the context of incomplete-information games where a policymaker commits to a policy, an agent responds, and both are able to learn from the available data. I formalized the modeling assumption using concepts like regret, Rademacher complexity, and sample privacy. I showed that policies that are too complex in a precise sense can be suboptimal because they lead to unpredictable behavior. I proposed penalized policies and motivated them through theoretical guarantees and illustrative examples.

The most important – and challenging – direction for future work is to turn this methodology towards real applications. One approach is to find highly-structured, data-rich settings where the data-driven penalized policy developed in this paper can actually be used. Particularly promising areas may lie in education or healthcare, where rich value-added measures have been used for policies like teacher compensation. Another approach is to treat penalization as desirable even in the absence of an explicit dataset. Here, the dataset would be seen as a metaphor for experiences that shape the agent's beliefs. Lab experiments could be used to determine whether policies that are more complex (in the sense of this paper) actually lead to suboptimal responses.

Bringing this method to any serious application will likely also require theoretical extensions. For example, there may be settings where the distribution is only partially identified, where there are multiple agents interacting strategically, or where more efficient estimators can take advantage of particular problem structure. These are all worthwhile open questions.

References

Aragones, E., Gilboa, I., Postlewaite, A., & Schmeidler, D. (2005, December). Fact-free learning. *American Economic Review*, *95*(5), 1355–1368.

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., 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.

Bartlett, P. L. & Mendelson, S. (2003, March). Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res. 3*, 463–482.

Battigalli, P. & Siniscalchi, M. (2003). Rationalization and Incomplete Information. *The B.E. Journal of Theoretical Economics*, *3*(1), 1–46.

Bergemann, D. & Morris, S. (2005). Robust mechanism design. *Econometrica*, *73*(6), 1771–1813.

Börgers, T. (2017, June). (no) foundations of dominant-strategy mechanisms: a comment on chung and ely (2007). *Review of Economic Design*, *21*(2), 73–82.

Börgers, T. & Smith, D. (2014, May). Robust mechanism design and dominant strategy voting rules. *Theoretical Economics*, *9*(2), 339–360.

Camara, M. K., Hartline, J. D., & Johnsen, A. (2020). Mechanisms for a no-regret agent: beyond the common prior. In *2020 ieee 61st annual symposium on foundations of computer science (focs)* (pp. 259–270).

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

Carroll, G. (2017). Robustness and separation in multidimensional screening. *Econometrica*, *85*(2), 453–488.

Carroll, G. & Meng, D. (2016a). Locally robust contracts for moral hazard. *Journal of Mathematical Economics*, *62*, 36–51.

Carroll, G. & Meng, D. (2016b). Robust contracting with additive noise. *Journal of Economic Theory*, *166*, 586–604.

Chen, Y.-C., di Tillio, A., Faingold, E., & Xiong, S. (2010). Uniform topologies on types. *Theoretical Economics*, *5*(3), 445–478.

Chung, K.-S. & Ely, J. C. (2007). Foundations of dominant-strategy mechanisms. *The Review of Economic Studies*, *74*(2), 447–476.

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.

Dekel, E., Fudenberg, D., & Morris, S. (2006). Topologies on types. *Theoretical Economics*, *1*(3), 275–309.

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.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi & T. Rabin (Eds.), *Theory of cryptography* (pp. 265–284). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dwork, C. & Roth, A. (2014, August). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci. 9*(3-4), 211–407.

Goldberg, A. V., Hartline, J. D., Karlin, A. R., Saks, M., & Wright, A. (2006). Competitive auctions. *Games and Economic Behavior*, *55*(2), 242–269. Mini Special Issue: Electronic Market Design.

Gonçalves, D. & Furtado, B. (2020, August). *Statistical mechanism design: robust pricing and reliable projections*.

Haagerup, U. (1981). The best constants in the khintchine inequality. *Studia Mathematica*, *70*(3), 231–283.

Haghtalab, N., Jackson, M. O., & Procaccia, A. D. (2021). Belief polarization in a complex world: a learning theory perspective. *Proceedings of the National Academy of Sciences*, *118*(19).

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Holmstrom, B. & Milgrom, P. (1987). Aggregation and linearity in the provision of intertemporal incentives. *Econometrica*, *55*(2), 303–328.

Holmstrom, B. & Milgrom, P. (1991). Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, *7*, 24–52.

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. (2005). Analogy-based expectation equilibrium. *Journal of Economic Theory*, *123*(2), 81–104.

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

Jehiel, P., Meyer-ter-Vehn, M., Moldovanu, B., & Zame, W. R. (2006). The limits of ex post implementation. *Econometrica*, *74*(3), 585–610.

Kitagawa, T. & Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, *86*(2), 591–616.

Liang, A. (2020, July). *Games of incomplete information played by statisticians*.

Lipman, B. L. (2003). Finite order implications of common priors. *Econometrica*, *71*(4), 1255–1267.

Mailath, G. J. & Samuelson, L. (2020, May). Learning under diverse world views: model-based inference. *American Economic Review*, *110*(5), 1464–1501.

Manski, C. F. (1993, January). Adolescent econometricians: how do youth infer the returns to schooling? In *Studies of supply and demand in higher education* (pp. 43–60). University of Chicago Press.

Manski, C. F. (2004a). Measuring expectations. *Econometrica*, *72*(5), 1329–1376.

Manski, C. F. (2004b). Statistical treatment rules for heterogeneous populations. *Econometrica*, *72*(4), 1221–1246.

Manski, C. F. (2019, December). *Econometrics for decision making: building foundations sketched by haavelmo and wald* (Working Paper No. 26596). National Bureau of Economic Research.

Massart, P. (2000). Some applications of concentration inequalities to statistics. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, *9*(2), 245–303.

Mbakop, E. & Tabord-Meehan, M. (2021). Model selection for treatment choice: penalized welfare maximization. *Econometrica*, *89*(2), 825–848.

McDiarmid, C. (1989). On the method of bounded differences. In J. Siemons (Ed.), *Surveys in combinatorics, 1989: invited papers at the twelfth british combinatorial conference* (pp. 148–188). London Mathematical Society Lecture Note Series. Cambridge University Press.

McSherry, F. & Talwar, K. (2007). Mechanism design via differential privacy. In *48th annual ieee symposium on foundations of computer science (focs'07)* (pp. 94–103).

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

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

Al-Najjar, N. I. (2009). Decision makers as statisticians: diversity, ambiguity, and learning. *Econometrica*, *77*(5), 1371–1401.

Al-Najjar, N. I. & Pai, M. M. (2014). Coarse decision making and overfitting. *Journal of Economic Theory*, *150*, 467–486.

Olea, J. L. M., Ortoleva, P., Pai, M. M., & Prat, A. (2021, February). Competing models.

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

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

Romano, J. P. & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.

Rubinstein, A. (1989). The electronic mail game: strategic behavior under "almost common knowledge". *The American Economic Review*, *79*(3), 385–391.

Salant, Y. & Cherry, J. (2020). Statistical inference in games. *Econometrica*, *88*(4), 1725–1752.

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

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, *151*(1), 7081.

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.

Valenzuela-Stookey, Q. (2020, September). *Subjective complexity under uncertainty*.

Wald, A. (1950). *Statistical decision functions*. Wiley: New York.

Wasserstein, R. L. & Lazar, N. A. (2016). The asa statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Weinstein, J. & Yildiz, M. (2007). A structure theorem for rationalizability with application to robust predictions of refinements. *Econometrica*, *75*(2), 365–400.

Wilson, R. (1987). Game-theoretic analyses of trading processes. In T. F. Bewley (Ed.), *Advances in economic theory: fifth world congress* (pp. 33–70). Econometric Society Monographs. Cambridge University Press.

A Omitted Proofs

A.1 Proof of Proposition 2

This proof will be slightly more general than the proposition statement. Let $f : \mathcal{S}^n \to \mathbb{R}_+$ be an arbitrary function with upper bound $\bar{f}$. Suppose that there is an upper bound

$$\mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right)\right] \leq B$$

The goal is to find a similar upper bound on

$$\mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) \mid P_n = p\right]$$

assuming that P_n is (ϵ, δ)-private.

First, I use the privacy property to show that bound the expected value of f conditional on the realized policy P_n and the event E in the definition of sample privacy. I show that this is not very different from the expected value of f conditioned on just the event E.

$$\begin{aligned}
\mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) \mid P_n = p, E\right] &= \sum_{(S_1,\ldots,S_n)\in E} \Pr_{\pi^s}\left[S_1, \ldots, S_n \mid P_n = p, E\right] \cdot f\left(S_1, \ldots, S_n\right) \\
&= \sum_{(S_1,\ldots,S_n)\in E} \frac{\Pr_{\pi^s}\left[S_1, \ldots, S_n \mid E\right] \cdot \Pr_{\pi^s}\left[P_n = p \mid S_1, \ldots, S_n\right]}{\Pr_{\pi^s}\left[P_n = p \mid E\right]} \cdot f\left(S_1, \ldots, S_n\right) \\
&\leq \sum_{(S_1,\ldots,S_n)\in E} \frac{\Pr_{\pi^s}\left[S_1, \ldots, S_n \mid E\right] \cdot e^{\epsilon} \cdot \Pr_{\pi^s}\left[P_n = p \mid E\right]}{\Pr_{\pi^s}\left[P_n = p \mid E\right]} \cdot f\left(S_1, \ldots, S_n\right) \\
&= \sum_{(S_1,\ldots,S_n)\in E} e^{\epsilon} \cdot \Pr_{\pi^s}\left[S_1, \ldots, S_n \mid E\right] \cdot f\left(S_1, \ldots, S_n\right) \\
&= e^{\epsilon} \cdot \mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) \mid E\right]
\end{aligned}$$

Next, I use the upper bound to show that

$$\begin{aligned}
B &\geq \mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right)\right] \\
&= (1-\delta) \cdot \mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) \mid E\right] + \delta \cdot \mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) \mid \neg E\right] \\
&\geq (1-\delta) \cdot \mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) \mid E\right] \\
&\geq (1-\delta) \cdot e^{-\epsilon} \cdot \mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) \mid E, P_n = p\right] \\
&\geq (1-\delta) \cdot e^{-\epsilon} \cdot \mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) \mid E, P_n = p\right] \\
&\quad + \delta \cdot e^{-\epsilon} \cdot \mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) - \bar{f} \mid \neg E, P_n = p\right] \\
&= e^{-\epsilon}\left(\mathrm{E}_{\pi^s}\left[f\left(S_1, \ldots, S_n\right) \mid P_n = p\right] - \delta \cdot \bar{f}\right)
\end{aligned}$$

Finally, I rearrange the lower bound on B to obtain the desired result, i.e.

$$\mathrm{E}_{\pi^s}\left[f\left(S_1,\ldots,S_n\right) \mid P_n = p\right] \leq e^{\epsilon}\cdot B + \delta\cdot\bar{f}$$

A.2 Proof of Lemma 1

I want to show that

$$\begin{aligned}4\mathcal{RC}_n^A(p,\pi^s) + \mathrm{BFR}_n \geq \Big|&\left(\max_{r'}\mathrm{E}_{\hat{\pi}^s}\left[u^A\left(p,r',s\right)\right] - \mathrm{E}_{\hat{\pi}^s,\pi^r}\left[u^A\left(p,r,s\right)\right]\right)\\ &-\left(\max_{r'}\mathrm{E}_{\pi^s}\left[u^A\left(p,r',s\right)\right] - \mathrm{E}_{\pi^s,\pi^r}\left[u^A\left(p,r,s\right)\right]\right)\Big|\end{aligned}$$

for all mixed responses π^r and policies p, with probability no less than $1-n_{\mathcal{P}}\exp(-n^{\alpha})$. For this purpose, it suffices to bound two quantities. First, observe that

$$\mathrm{E}_{\pi^s,\pi^r}\left[u^A\left(p,r,s\right)\right] - \mathrm{E}_{\hat{\pi}_n^s,\pi^r}\left[u^A\left(p,r,s\right)\right] \leq 2\overline{\mathcal{RC}}_n^A(p) + 4\sqrt{\frac{2\ln(4/\kappa)}{n}} \tag{20}$$

with probability $1-\kappa$. This is the typical way of expressing regret bounds based on the Rademacher complexity (see e.g. Bartlett and Mendelson 2003). Second, observe that

$$\max_{r'}\mathrm{E}_{\hat{\pi}_n^s}\left[u^A\left(p,r',s\right)\right] - \max_{r'}\mathrm{E}_{\pi^s}\left[u^A\left(p,r',s\right)\right] \leq 2\overline{\mathcal{RC}}_n^A(p) + 4\sqrt{\frac{2\ln(4/\kappa)}{n}} \tag{21}$$

This follows from the fact that

$$\mathrm{E}_{\hat{\pi}_n^s}\left[u^A\left(p,r^*,s\right)\right] - \mathrm{E}_{\pi^s}\left[u^A\left(p,r^*,s\right)\right] \leq 2\overline{\mathcal{RC}}_n^A(p) + 4\sqrt{\frac{2\ln(4/\kappa)}{n}}$$

where $r^* \in \arg\max_r \mathrm{E}_{\hat{\pi}_n^s}\left[u^A\left(p,r,s\right)\right]$, and

$$\mathrm{E}_{\pi^s}\left[u^A\left(p,r^*,s\right)\right] - \max_{r'}\mathrm{E}_{\pi^s}\left[u^A\left(p,r',s\right)\right] \leq 0$$

so

$$\mathrm{E}_{\hat{\pi}_n^s}\left[u^A\left(p,r^*,s\right)\right] - \mathrm{E}_{\pi^s}\left[u^A\left(p,r^*,s\right)\right] + \mathrm{E}_{\pi^s}\left[u^A\left(p,r^*,s\right)\right] - \max_{r'}\mathrm{E}_{\pi^s}\left[u^A\left(p,r',s\right)\right] \leq 2\overline{\mathcal{RC}}_n^A(p) + 4\sqrt{\frac{2\ln(4/\kappa)}{n}}$$

Adding together the last two inequalities gives inequality (21). Adding together inequalities (20) and (21) gives the desired result. Applying the union bound across all policies p, the result holds with probability $1-n_{\mathcal{P}}\kappa$. Furthermore, the probability is uniform over all responses r, and therefore uniform across all mixed responses π^r. All that remains is to derive the probability κ and buffer

BFR_n. Set $\kappa = \exp(-n^\alpha)$. Note that

$$\begin{aligned}
4\sqrt{\frac{2\ln(4/\kappa)}{n}} &= 4\sqrt{\frac{2\ln(4\exp(-n^\alpha))}{n}} \\
&= 4\sqrt{\frac{2\ln 4 - 2\ln(\exp(-n^\alpha))}{n}} \\
&\leq 4\sqrt{\frac{2\ln 4}{n}} + 4\sqrt{-\frac{2\ln(\exp(-n^\alpha))}{n}} \\
&\leq \text{BFR}_n
\end{aligned}$$

A.3 Proof of Lemma 3

Let $E \subseteq S^n$ be the set of all sample realizations $S_1, \ldots, S_n$ where

$$\widehat{\text{WC}}_n(p) - \text{E}\left[\widehat{\text{WC}}_n(p)\right] \leq t$$

By lemma 2, $\Pr_{\pi^s}[E] \geq 1 - \delta$ where

$$\delta = \exp\left(-\frac{2t^2}{nc^2}\right)$$

To establish sample privacy, I need to show that

$$\Pr_{\pi^s}\left[\hat{P}_n = p \mid S_1, \ldots, S_n\right] \leq e^\epsilon \cdot \Pr_{\pi^s}\left[\hat{P}_n = p, E\right]$$

for any sample realizations $\left(S_1, \ldots, S_n\right) \in E$. Let the sample $\left(S'_1, \ldots, S'_n\right) \in E$ minimize

$$\Pr_{\pi^s}\left[\hat{P}_n = p \mid S'_1, \ldots, S'_n\right]$$

so that it suffices to show

$$\Pr_{\pi^s}\left[\hat{P}_n = p \mid S_1, \ldots, S_n\right] \leq e^\epsilon \cdot \Pr_{\pi^s}\left[\hat{P}_n = p \mid S'_1, \ldots, S'_n\right] \tag{22}$$

I can characterize the distribution of $\hat{P}_n$ using standard results that link Gumbel error terms with exponential weights.

$$\Pr_{\pi^s}\left[\hat{P}_n = p \mid S_1, \ldots, S_n\right] = \frac{\exp\left(n^\beta \cdot \widehat{\text{WC}}_n\left(p \mid S_1 \ldots, S_n\right)\right)}{\sum_{p'} \exp\left(n^\beta \cdot \widehat{\text{WC}}_n\left(p' \mid S_1 \ldots, S_n\right)\right)}$$

Using this, I can rewrite equation (22) and manipulate it as follows.

$$\begin{aligned}
\frac{\Pr_{\pi^s}\left[\hat{P}_n = p \mid S_1, \ldots, S_n\right]}{\Pr_{\pi^s}\left[\hat{P}_n = p \mid S'_1, \ldots, S'_n\right]} &= \frac{\exp\left(n^\beta \cdot \widehat{\mathrm{WC}}_n\left(p \mid S_1 \ldots, S_n\right)\right)}{\exp\left(n^\beta \cdot \widehat{\mathrm{WC}}_n\left(p \mid S'_1 \ldots, S'_n\right)\right)} \cdot \frac{\sum_{p'} \exp\left(n^\beta \cdot \widehat{\mathrm{WC}}_n\left(p' \mid S'_1 \ldots, S'_n\right)\right)}{\sum_{p'} \exp\left(n^\beta \cdot \widehat{\mathrm{WC}}_n\left(p' \mid S_1 \ldots, S_n\right)\right)} \\
&\leq \exp\left(tn^\beta\right) \cdot \frac{\sum_{p'} \exp\left(n^\beta \cdot \widehat{\mathrm{WC}}_n\left(p' \mid S'_1 \ldots, S'_n\right)\right)}{\sum_{p'} \exp\left(n^\beta \cdot \widehat{\mathrm{WC}}_n\left(p' \mid S_1 \ldots, S_n\right)\right)} \\
&\leq \exp\left(tn^\beta\right) \cdot \exp\left(tn^\beta\right) \cdot \frac{\sum_{p'} \exp\left(n^\beta \cdot \widehat{\mathrm{WC}}_n\left(p' \mid S_1 \ldots, S_n\right)\right)}{\sum_{p'} \exp\left(n^\beta \cdot \widehat{\mathrm{WC}}_n\left(p' \mid S_1 \ldots, S_n\right)\right)} \\
&\leq \exp\left(2tn^\beta\right)
\end{aligned}$$

Therefore, $\hat{P}_n$ is (ϵ, δ)-private when $\epsilon = 2tn^\beta$.

A.4 Proof of Lemma 5

Recall the definition of $\mathrm{WC}(p, b, \pi^s)$ (17). This represents the policymaker's worst-case utility when the agent's regret is bounded by a constant $b \geq 0$. Note that

$$\widehat{\mathrm{WC}}_n(p) = \mathrm{WC}(p, b, \pi^s) \quad \text{where} \quad b = (4e^\epsilon + 4) \cdot \overline{\mathcal{RC}}^A_n(p) + \delta + \mathrm{BFR}_n$$

Let $\hat{\pi}^s$ be the empirical distribution. Let $\tilde{\pi}^s$ be a modified empirical distribution where $S_i = s'$ instead of $S_i = s$. As I shift from $\hat{\pi}^s$ to $\tilde{\pi}^s$, the agent's empirical regret changes by at most $2\Delta^P(p) \cdot n^{-1}$. In particular, for any mixed response π^r,

$$\max_{r'} \mathrm{E}_{\hat{\pi}^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\hat{\pi}^s, \pi^r}\left[u^A\left(p, r, s\right)\right] \leq b$$

implies

$$\max_{r'} \mathrm{E}_{\tilde{\pi}^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\tilde{\pi}^s, \pi^r}\left[u^A\left(p, r, s\right)\right] \leq b + 2\Delta^P(p) \cdot n^{-1}$$

Likewise, the policymaker's empirical utility changes by at most $\Delta^P(p) \cdot n^{-1}$. It follows from these two observations that

$$\widehat{\mathrm{WC}}_n(p \mid \hat{\pi}^s) \geq \mathrm{WC}\left(p, b + 2\Delta^P(p) \cdot n^{-1}, \tilde{\pi}^s\right) - \Delta^P(p) \cdot n^{-1}$$

where the notation $\widehat{\mathrm{WC}}_n(p \mid \hat{\pi}^s)$ is used to emphasize that $\widehat{\mathrm{WC}}_n(p)$ is being evaluated with respect to the empirical distribution $\hat{\pi}^s$. By the robustness lemma (4),

$$\mathrm{WC}\left(p, b + 2\Delta^P(p) \cdot n^{-1}, \tilde{\pi}^s\right) \geq \mathrm{WC}(p, b, \tilde{\pi}^s) - \Delta^A(p)\left(\frac{2\Delta^P(p) \cdot n^{-1}}{b}\right)$$

By definition, $\widehat{\mathrm{WC}}_n(p \mid \tilde{\pi}^s) = \mathrm{WC}(p, b, \tilde{\pi}^s)$. It follows that

$$\widehat{\mathrm{WC}}_n(p \mid \tilde{\pi}^s) - \widehat{\mathrm{WC}}_n(p \mid \hat{\pi}^s) \leq \Delta^A(p) \left(\frac{2\Delta^P(p) \cdot n^{-1}}{b} \right) + \Delta^P(p) \cdot n^{-1}$$

Therefore, $\widehat{\mathrm{WC}}_n(p)$ satisfies the bounded differences property as long as

$$c \geq \Delta^A(p) \left(\frac{2\Delta^P(p) \cdot n^{-1}}{b} \right) + \Delta^P(p) \cdot n^{-1}$$

A.5 Proof of Lemma 6

I begin by making some observations and introducing some notation. Recall the empirical regret bound in the definition of $\widehat{\mathrm{WC}}_n(p)$ (13).

$$\max_{r'} \mathrm{E}_{\hat{\pi}^s} \left[u^A \left(p, r', s \right) \right] - \mathrm{E}_{\hat{\pi}^s, \pi^r} \left[u^A \left(p, r, s \right) \right] \leq (4e^\epsilon + 4) \cdot \overline{\mathcal{RC}}_n^A(p) + \delta_n \cdot \Delta^A(p) + \mathrm{BFR}_n \tag{23}$$

By lemma 1, any mixed response that satisfies this bound also satisfies

$$\max_{r'} \mathrm{E}_{\pi^s} \left[u^A \left(p, r', s \right) \right] - \mathrm{E}_{\pi^s, \pi^r} \left[u^A \left(p, r, s \right) \right] \leq 4e^\epsilon \cdot \overline{\mathcal{RC}}_n^A(p) + \delta_n \cdot \Delta^A(p)$$

with probability $1 - n_{\mathcal{P}} \exp(-n^\alpha)$, where the expectations are evaluated with respect to the true distribution π^s. This gives an upper bound for $\widehat{\mathrm{WC}}_n(p)$ with high probability, i.e.

$$\begin{aligned} f(p, \hat{\pi}^s) = \min_{\pi^r} & \mathrm{E}_{\hat{\pi}^s, \pi^r} \left[u^P \left(p, r, s \right) \right] + \nu_n(p) \\ \text{s.t.} \quad & \max_{r'} \mathrm{E}_{\pi^s} \left[u^A \left(p, r', s \right) \right] - \mathrm{E}_{\pi^s, \pi^r} \left[u^A \left(p, r, s \right) \right] \leq 4e^\epsilon \cdot \overline{\mathcal{RC}}_n^A(p) + \delta_n \cdot \Delta^A(p) \end{aligned} \tag{24}$$

Let $\tilde{\pi}^r(p, \hat{\pi}^s)$ be the solution to the minimization problem (24). It is important to note that the set of feasible mixed responses no longer depends on the sample.

This proof consists of three parts. First, I want to bound the expected gap between $f(\hat{P}_n, \hat{\pi}^s)$ and $\widehat{\mathrm{WC}}_n\left(\hat{P}_n\right)$. It follows from the preceding discussion that

$$\mathrm{E}_{\pi^s} \left[f(\hat{P}_n, \hat{\pi}^s) - \widehat{\mathrm{WC}}_n\left(\hat{P}_n\right) \right] \geq -n_{\mathcal{P}} \exp(-n^\alpha) \cdot \max_p \Delta^P(p)$$

Next, I want to bound the expected gap between $f(p, \hat{\pi}^s)$ and $f(p, \pi^s)$, i.e.

$$\begin{aligned} \mathrm{E}_{\pi^s} \left[\max_p \left(f(p, \hat{\pi}^s) - f(p, \pi^s) \right) \right] &= \mathrm{E}_{\pi^s} \left[\max_p \left(\mathrm{E}_{\hat{\pi}^s, \tilde{\pi}^r(p, \hat{\pi}^s)} \left[u^P \left(p, r, s \right) \right] - \mathrm{E}_{\pi^s, \tilde{\pi}^r(p, \pi^s)} \left[u^P \left(p, r, s \right) \right] \right) \right] \\ &\leq \mathrm{E}_{\pi^s} \left[\max_p \left(\mathrm{E}_{\hat{\pi}^s, \tilde{\pi}^r(p, \hat{\pi}^s)} \left[u^P \left(p, r, s \right) \right] - \mathrm{E}_{\hat{\pi}^s, \tilde{\pi}^r(p, \pi^s)} \left[u^P \left(p, r, s \right) \right] \right) \right] \\ &\quad + \mathrm{E}_{\pi^s} \left[\max_p \left(\mathrm{E}_{\hat{\pi}^s, \tilde{\pi}^r(p, \pi^s)} \left[u^P \left(p, r, s \right) \right] - \mathrm{E}_{\pi^s, \tilde{\pi}^r(p, \pi^s)} \left[u^P \left(p, r, s \right) \right] \right) \right] \end{aligned}$$

$$
\begin{aligned}
&\leq \mathrm{E}_{\pi^s}\left[\max_p \left(\mathrm{E}_{\hat{\pi}^s,\tilde{\pi}^r(p,\pi^s)}\left[u^P(p,r,s)\right] - \mathrm{E}_{\pi^s,\tilde{\pi}^r(p,\pi^s)}\left[u^P(p,r,s)\right]\right)\right] \\
&\leq \mathrm{E}_{\pi^s}\left[\max_{p,\pi^r} \left(\mathrm{E}_{\hat{\pi}^s,\pi^r}\left[u^P(p,r,s)\right] - \mathrm{E}_{\pi^s,\pi^r}\left[u^P(p,r,s)\right]\right)\right] \\
&= \mathrm{E}_{\pi^s}\left[\max_{p,r} \left(\mathrm{E}_{\hat{\pi}^s}\left[u^P(p,r,s)\right] - \mathrm{E}_{\pi^s}\left[u^P(p,r,s)\right]\right)\right]
\end{aligned}
$$

At this point, it follows from the standard symmetrization argument that

$$
\mathrm{E}_{\pi^s}\left[\max_p \left(f(p,\hat{\pi}^s) - f(p,\pi^s)\right)\right] \leq 2\mathcal{R}C_n^P(\pi^s)
$$

Finally, I want to bound the expected gap between $\mathrm{WC}_n\left(p,\epsilon,\delta_n,\pi^s\right)$ and $f(\hat{P}_n,\pi^s)$. Note that $\mathrm{WC}_n\left(p,\epsilon,\delta_n,\pi^s\right) = f(p,\pi^s) - \nu_n(p)$. Furthermore, note that

$$
\begin{aligned}
\mathrm{E}_{\pi^s}\left[f(\hat{P}_n,\pi^s) - \mathrm{WC}_n\left(\hat{P}_n,\epsilon,\delta_n,\pi^s\right)\right] &= \mathrm{E}_{\pi^s}\left[\nu_n(\hat{P}_n)\right] \\
&\leq \mathrm{E}\left[\max_p \nu_n(p)\right] \\
&\leq \mathrm{E}\left[\sum_p |\nu_n(p)|\right] \\
&= n_{\mathcal{P}}\mathrm{E}\left[|\nu_n(p)|\right] \\
&\leq n_{\mathcal{P}}\sqrt{\mathrm{E}\left[|\nu_n(p)|^2\right]} \\
&\leq n_{\mathcal{P}}\sqrt{\mathrm{E}\left[\nu_n(p)\right]^2 + \mathrm{Var}\left[\nu_n(p)\right]} \\
&\leq n_{\mathcal{P}}\sqrt{n^{-2\beta} + 2n^{-2\beta}} \\
&\leq n_{\mathcal{P}}\sqrt{3}\cdot n^{-\beta}
\end{aligned}
$$

Combining these three steps gives us the desired result.

$$
\begin{aligned}
&\mathrm{E}_{\pi^s}\left[\mathrm{WC}_n\left(\hat{P}_n,\epsilon,\delta_n,\pi^s\right) - \widehat{\mathrm{WC}}_n\left(\hat{P}_n\right)\right] \\
&= \mathrm{E}_{\pi^s}\left[\mathrm{WC}_n\left(\hat{P}_n,\epsilon,\delta_n,\pi^s\right) - f(\hat{P}_n,\pi^s)\right] - \mathrm{E}_{\pi^s}\left[f(\hat{P}_n,\hat{\pi}^s) - f(\hat{P}_n,\pi^s)\right] + \mathrm{E}_{\pi^s}\left[f(\hat{P}_n,\hat{\pi}^s) - \widehat{\mathrm{WC}}_n\left(\hat{P}_n\right)\right] \\
&\geq -n_{\mathcal{P}}\sqrt{3}\cdot n^{-\beta} - \mathrm{E}_{\pi^s}\left[\max_p \left(f(p,\hat{\pi}^s) - f(p,\pi^s)\right)\right] - n_{\mathcal{P}}\exp(-n^\alpha)\cdot\max_p \Delta^P(p) \\
&\geq -n_{\mathcal{P}}\sqrt{3}\cdot n^{-\beta} - 2\mathcal{R}C_n^P(\pi^s) - n_{\mathcal{P}}\exp(-n^\alpha)\cdot\max_p \Delta^P(p)
\end{aligned}
$$

A.6 Proof of Proposition 4

This example will exhibit a simple game where

$$\mathrm{OP}_n(\pi^s) = \mathrm{CK}(\pi^s) - \Omega\left(n^{-\gamma}\right)$$

The agent faces an estimation problem and cares about her accuracy. The policy space is a singleton; it is irrelevant. The state space $\mathcal{S} = [0, 1]$ is the unit interval. The response space $\mathcal{R} = [0, 1]$ is also the unit interval. The agent's response $r \in [0, 1]$ is a prediction, subject to square loss, i.e.

$$u^A(p, r, s) = -(r - s)^2$$

The policymaker cares about the agent's accuracy with respect to a bliss point $s_0 \in [0, 1]$ that I will specify later. However, his sensitivity to inaccuracy is different from the agent, i.e.

$$u^P(p, r, s) = -|r - s_0|^{2\gamma}$$

I claim that there exists a distribution $\tilde{\pi}^s$ where the agent's regret bound is $\Omega(n^{-1})$. Let the bliss point $s_0 := \mathrm{E}_{\tilde{\pi}^s}[s]$ be the mean of s according to $\tilde{\pi}^s$. Let the distribution $\pi^s := \tilde{\pi}^s$. Existence follows from two observations. First, the mean square error of the maximum likelihood estimator is $O(n^{-1})$. Second, the maximum likelihood estimator is known be efficient.

To characterize the optimal penalized benchmark, I need to consider responses that satisfy the agent's regret bound. One such response is is $r_n = \mathrm{E}_{\tilde{\pi}^s}[s] + \Omega(n^{-1/2})$. The policymaker's expected utility under r_n must be at least as large as the optimal penalized benchmark, which is the worst case expected utility. That is,

$$\begin{aligned}
\mathrm{OP}_n(\pi^s) &\leq -|r_n - s_0|^{2\gamma} \\
&= -\left|\mathrm{E}_{\pi^s}[s] + \Omega(n^{-1/2}) - \mathrm{E}_{\tilde{\pi}^s}[s]\right|^{2\gamma} \\
&= -\left(\Omega(n^{-1/2})\right)^{2\gamma} \\
&= -\Omega(n^{-\gamma})
\end{aligned} \tag{25}$$

Next, consider the common knowledge benchmark. The agent will predict the mean, $r = \mathrm{E}_{\pi^s}[s]$, and the policymaker's expected utility will be

$$\mathrm{CK}(\pi^s) = \left|\mathrm{E}_{\pi^s}[s] - s_0\right|^{2\gamma} = \left|\mathrm{E}_{\tilde{\pi}^s}[s] - \mathrm{E}_{\tilde{\pi}^s}[s]\right|^{2\gamma} = 0 \tag{26}$$

I can combine equations (25) and (26) to show

$$\mathrm{OP}_n(\pi^s) \leq \mathrm{CK}(\pi^s) - \Omega(n^{-\gamma})$$

This completes the first part of the proof.

Next, I need to verify that this game satisfies (in particular) assumption 2. To do this, I need to

introduce the pseudodimension: a method for bounding the Rademacher complexity. The following definition is specialized to the agent's Rademacher complexity.

Definition 11. *A vector* $(w_1, \dots, w_n) \in \mathbb{R}^n$ *is a* witness *for a vector* $(S_1, \dots, S_n)$ *if, for any realizations* $(\sigma_1, \dots \sigma_n) \in \{-1, 1\}^n$*, there exists a response* r *such that*

$$\text{sign}\left(-\left(r - S_i\right)^2 - w_i\right) = \sigma_i \tag{27}$$

A vector $(S_1, \dots, S_n)$ *is* shattered *if it has a witness* $(w_1, \dots, w_n)$*. The* pseudo-dimension *is the largest integer* m *such that some vector* $(S_1, \dots, S_m)$ *is shattered.*

Claim 18. *The pseudo-dimension is at most 2.*

Since the pseudo-dimension is bounded, the agent's Rademacher complexity is $\tilde{O}(n^{-1/2})$.

Proof. For the sake of contradiction, suppose that the vector $S_1, \dots, S_n$ is shattered for $n > 2$. By condition (27), $\sigma_i = 1$ means that S_i is within some distance d_i of r, where d_i depends on w_i and γ. Define n intervals $I_1, \dots, I_n$ where $I_i = [S_i - d_i, S_i + d_i]$. Then $\sigma_i = 1$ means $r \in I_i$, and $\sigma_i = 0$ means $r \notin I_i$. Let $f(r)$ be the set of intervals I_i such that $r \in I_i$. Each vector σ corresponds to a unique element in the range of $f(r)$.

I claim that the range of $f(r)$ has at most $2n + 1$ elements. If we list the n left endpoints and the n right endpoints of intervals, in order, these define a different set of $2n + 1$ intervals. Within each interval J, we can move r from the left to the right, without entering or exiting any interval I_i. Therefore, f is invariant over each interval J. Since there are at most $2n + 1$ intervals J, the range of $f(r)$ must have at most $2n + 1$ elements.

However, this leads to a contradiction. There are 2^n distinct values of the vector σ. But each vector σ must correspond to a unique element in the range of r, and there are only $2n + 1$ such elements. When $n = 3$, $2^n = 8$ but $2n + 1 = 7$. When $n > 3$, the discrepancy is even larger. Therefore, the vector $S_1, \dots, S_n$ does not have a witness when $n > 3$. It follows from the definition that the pseudo-dimension is at most 2. □

Next, consider the policymaker's Rademacher complexity. Note that

$$\max_r \sum_{i=1}^{n} \sigma_i |r - s_0|^{2\gamma}$$

has only three possible solutions: $r = s_0$, $r = 0$, or $r = 1$. Without loss of generality, I can restrict the response space to $\{0, s_0, 1\}$. It follows from Massart's finite lemma that the policymaker's Rademacher complexity is $O(n^{-1/2})$.

A.7 Proof of Lemma 7

Recall the empirical regret bound in the definition of $\widehat{\text{WC}}_n(p)$ (13).

$$\max_{r'} \mathrm{E}_{\hat{\pi}^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\hat{\pi}^s, \pi^r}\left[u^A\left(p, r, s\right)\right] \leq (4e^\epsilon + 4) \cdot \overline{\mathcal{RC}}_n^A(p) + \delta_n \cdot \Delta^A(p) + \text{BFR}_n \tag{28}$$

I want to argue that every mixed response π^r that satisfies this empirical regret bound also satisfies the regret bound in the definition of $\text{WC}_m(p, 0, 0, \pi^s)$, i.e.

$$\max_{r'} \mathrm{E}_{\pi^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\pi^s, \pi^r}\left[u^A\left(p, r, s\right)\right] \leq 4 \cdot \mathcal{RC}_m^A(p, \pi^s) \tag{29}$$

At least, this should hold with high probability. Let π^r be a mixed response satisfying the empirical regret bound (28). By lemma 1, with probability $1 - n_{\mathcal{P}} \exp(-n^\alpha)$, we have

$$\max_{r'} \mathrm{E}_{\pi^s}\left[u^A\left(p, r', s\right)\right] - \mathrm{E}_{\pi^s, \pi^r}\left[u^A\left(p, r, s\right)\right] \leq (4e^\epsilon + 8) \cdot \overline{\mathcal{RC}}_n^A(p) + \delta_n \cdot \Delta^A(p) + 2\text{BFR}_n \tag{30}$$

This puts the agent's regret in terms of the true distribution. Next, I claim that the right-hand side of inequality (30) is $\Theta(m^{-1/2})$. There are three terms to consider. The first term is $\tilde{O}(n^{-1/2})$ by assumption 2. The second term is decreasing exponentially in n, since $\alpha < 2\beta$. The third term is $\Theta(n^{(\alpha-1)/2})$, and it is leading since $\alpha > 0$. Plugging in the value of m gives us $\Theta(m^{-1/2})$. Finally, note that as long as there is a sufficiently large constant in front of m, we have

$$\begin{aligned}(4e^\epsilon + 8) \cdot \overline{\mathcal{RC}}_n^A(p) + \delta_n \cdot \Delta^A(p) + 2\text{BFR}_n &\leq 4 \cdot \frac{C}{2\sqrt{2m}} \\ &\leq 4 \cdot \mathcal{RC}_m^A(p, \pi^s)\end{aligned} \tag{31}$$

where the last line follows from lemma 8. Combining inequalities (30) and (31) gives us the desired inequality (29), with probability $1 - n_{\mathcal{P}} \exp(-n^\alpha)$.

I have established that the set of mixed responses that $\widehat{\text{WC}}_n(p)$ minimizes over is, with high probability, a subset of the set of mixed responses that $\text{WC}_m(p, 0, 0, \pi^s)$ minimizes over. All that remains is to compare the policymaker's objective under $\widehat{\text{WC}}_n(p)$ with his objective under $\text{WC}_m(p, 0, 0, \pi^s)$. This compares expected utility under the empirical distribution, plus privacy-preserving noise, to expected utility under the true distribution. But this is precisely the situation we found ourselves in during the proof of lemma 6. I can apply the same bounds here to complete the proof.

A.8 Proof of Lemma 8

Recall the definition of Rademacher complexity:

$$\mathcal{RC}_n^A(p, \pi^s) = \frac{1}{n} \mathrm{E}_{\pi^s}\left[\max_r \sum_{i=1}^n \sigma_i \cdot u^A\left(p, r, S_i\right)\right]$$

$$= \frac{1}{n}\mathrm{E}_{\pi^s}\left[\mathrm{E}_{\pi^s}\left[\max_r \sum_{i=1}^n \sigma_i \cdot u^A\left(p, r, S_i\right) \mid S_1, \ldots, S_n\right]\right]$$

where the second equality follows from the law of iterated expectations. To bound the Rademacher complexity, it suffices to bound the interior expectation. Observe that

$$\begin{aligned}
&\mathrm{E}_{\pi^s}\left[\max_r \sum_{i=1}^n \sigma_i \cdot u^A\left(p, r, S_i\right) \mid S_1, \ldots, S_n\right] \\
&= \max_{r'} \mathrm{E}_{\pi^s}\left[\max_r \left(\sum_{i=1}^n \sigma_i \cdot u^A\left(p, r, S_i\right) - \sum_{i=1}^n \sigma_i \cdot u^A\left(p, r', S_i\right) + \sum_{i=1}^n \sigma_i \cdot u^A\left(p, r', S_i\right)\right) \mid S_1, \ldots, S_n\right] \\
&= \max_{r'} \mathrm{E}_{\pi^s}\left[\max_r \left(\sum_{i=1}^n \sigma_i \cdot \left(u^A\left(p, r, S_i\right) - u^A\left(p, r', S_i\right)\right)\right) + \sum_{i=1}^n \sigma_i \cdot u^A\left(p, r', S_i\right) \mid S_1, \ldots, S_n\right] \\
&= \max_{r'} \mathrm{E}_{\pi^s}\left[\max_r \left(\sum_{i=1}^n \sigma_i \cdot \left(u^A\left(p, r, S_i\right) - u^A\left(p, r', S_i\right)\right)\right) \mid S_1, \ldots, S_n\right] && (32) \\
&= \max_{r'} \mathrm{E}_{\pi^s}\left[\max_r \left(\sum_{i=1}^n \sigma_i \cdot \left(u^A\left(p, r, S_i\right) - u^A\left(p, r', S_i\right)\right)\right)^+ \mid S_1, \ldots, S_n\right] && (33) \\
&\geq \max_{r,r'} \mathrm{E}_{\pi^s}\left[\left(\sum_{i=1}^n \sigma_i \cdot \left(u^A\left(p, r, S_i\right) - u^A\left(p, r', S_i\right)\right)\right)^+ \mid S_1, \ldots, S_n\right] && (34) \\
&= \frac{1}{2}\max_{r,r'} \mathrm{E}_{\pi^s}\left[\left|\sum_{i=1}^n \sigma_i \cdot \left(u^A\left(p, r, S_i\right) - u^A\left(p, r', S_i\right)\right)\right| \mid S_1, \ldots, S_n\right] && (35) \\
&\geq \frac{1}{2\sqrt{2}}\max_{r,r'} \sqrt{\sum_{i=1}^n \left(u^A\left(p, r, S_i\right) - u^A\left(p, r', S_i\right)\right)^2} && (36) \\
&\geq \frac{C\sqrt{n}}{2\sqrt{2}}
\end{aligned}$$

The first two equalities follow from algebraic manipulations. Line (32) follows from the fact that

$$\mathrm{E}_{\pi^s}\left[\sum_{i=1}^n \sigma_i \cdot u^A\left(p, r', S_i\right) \mid S_1, \ldots, S_n\right] = 0$$

Line (33) follows from the fact that setting $r = r'$ ensures that the interior sum is zero, so that the maximum over all r is non-negative. Line (34) follows from Jensen's inequality. Line (35) follows from the fact that the sum inside the expectation is symmetrically distributed around zero. To see this, let X be a symmetric random variable with mean zero. Then

$$\mathrm{E}[|X|] = \Pr[X = 0] \cdot 0 + \Pr[X > 0] \cdot \mathrm{E}[X \mid X \geq 0] + \Pr[X < 0] \cdot \mathrm{E}[-X \mid X < 0]$$

$$
\begin{aligned}
&= \Pr[X > 0] \cdot \mathrm{E}[X \mid X > 0] + \Pr[X > 0] \cdot \mathrm{E}[X \mid X > 0] \\
&= 2 \cdot \Pr[X > 0] \cdot \mathrm{E}[X \mid X > 0] \\
&= 2 \cdot \mathrm{E}\left[X^{+}\right]
\end{aligned}
$$

Line (36) follows from Khintchine's inequality, with constants derived by Haagerup (1981). Finally, the last inequality follows from assumption 3.

A.9 Proof of Claim 9

This is a proof by contradiction. Let w be an effort-inducing contract that pays a positive wage $w(x_i) > 0$ for some outcome $i < m$. Consider a modified contract w' that pays $w'(x_i) = 0$ and

$$
w'(x_m) = w(x_m) + \frac{\pi_1^x(x_i)}{\pi_1^x(x_m)} w(x_i)
$$

Under contract w', the expected wages conditional on effort are

$$
\begin{aligned}
\sum_{j=1}^{m} \pi_1^x(x_j) w'(x_j) &= \pi_1^x(x_j) w'(x_j) + \pi_m^x w'(x_m) + \sum_{j \neq i,m} \pi_1^x(x_j) w'(x_j) \\
&= \pi_1^x(x_m) w(x_m) + \pi_1^x(x_m) \frac{\pi_1^x(x_i)}{\pi_1^x(x_m)} w(x_i) + \sum_{j \neq i,m} \pi_1^x(x_j) w'(x_j) \\
&= \sum_{j=1}^{m} \pi_1^x(x_j) w(x_j)
\end{aligned}
$$

That is, expected wages conditional on effort are the same for w and w'. However, expected wages conditional on effort are smaller for w' than for w. This follows from assumption 4, which implies that

$$
\pi_0^x(x_m) \frac{\pi_1^x(x_i)}{\pi_1^x(x_m)} \leq \pi_1^x(x_i)
$$

This inequality is strict unless $\pi_0^x = \pi_1^x$, in which case there is no effort-inducing contract anyways. Therefore, w' creates slack in the agent's incentive constraint without affecting the principal's utility. This allows the principal to slightly reduce wages, and be better off than under w.

A.10 Proof of Claim 12

Let w be an optimal contract with a maximum wage of $\bar{w}$. In the worst case, the agent will not put in effort if

$$
\mathrm{E}_{\pi^x}\left[w(X^1) - w(X^0)\right] - c \leq 4 \overline{\mathcal{RC}}_n^A(w)
$$

Otherwise, the agent will put in effort with probability $1 - q$ where

$$q = \frac{4\overline{\mathcal{RC}}_n^A(w)}{\mathrm{E}_{\pi^x}\left[w(X^1) - w(X^0)\right] - c}$$

This only depends on the contract through two quantities: $\bar{w}$ and $\mathrm{E}_{\pi^x}\left[w(X^1) - w(X^0)\right]$. Holding those quantities fixed, the agent's probability of effort is the same.

The rest of the proof essentially follows from arguments in the proof of Claim 9. There, I turned w into another contract w' that preserved the expected wages but increased $\mathrm{E}_{\pi^x}\left[w(X^1) - w(X^0)\right]$. That was the common knowledge case. Here, because increasing $\mathrm{E}_{\pi^x}\left[w(X^1) - w(X^0)\right]$ increases q, the principal's payoff under contract w' is actually better than under w. That does not adversely affect the argument.

For any given maximum wage $\bar{w}$, the previous argument shows that $w(x_m) < \bar{w}$ implies $w(x_i) = 0$ for all outcomes $i < m$. More generally, as long as $\ell(x_j) > 1$, this argument can be used to show that $w(x_j) < \bar{w}$ implies $w(x_i) = 0$ for all outcomes $i < j$. It follows from inspection that ceases to be true when $\ell(x_j) < 1$. In fact, a similar argument shows that $w(x_j) = 0$ for any outcome x_j where $\ell(x_j) < 1$.

This establishes the fact that w is a threshold function. Next, suppose I want to ensure that the agent puts in effort with probability q. I need

$$\mathrm{E}_{\pi^x}\left[w(X^1) - w(X^0)\right] - c = \frac{4\overline{\mathcal{RC}}_n^A(w)}{q} \tag{37}$$

Suppose $w(x_j) = \bar{w}$ for all $i > j$ and $w(x_j) = 0$ for all $j < i$. The expression in the statement of the claim ensures that I increase $w(x_i)$ as much as necessary to guarantee equation (37).

A.11 Proof of Claim 16

The buyer's strategy is the *empirical quantile maximizer*, i.e.

$$x_n \in \max_{x \in M} U^A(q, M, x, \hat{\pi}^s)$$

I will bound her quantile regret under this strategy. By the Dvoretzky-Kiefer-Wolfowitz inequality,

$$\Pr_{\pi^v}\left[\sup_u \left|\Pr_{\hat{\pi}^v}\left[u^A(M, x, v) \leq u\right] - \Pr_{\pi^v}\left[u^A(M, x, v) \leq u\right]\right| \geq t\right] \leq 2\exp\left(-2nt^2\right)$$

By assumption 5, this implies

$$\Pr_{\pi^v}\left[\left|U^A(q, M, x, \pi^v) - U^A(q, M, x, \pi^v)\right| \geq \frac{t}{K}\right] \leq 2\exp\left(-2nt^2\right)$$

By the union bound,

$$\Pr_{\pi^v}\left[\max_{x\in M}\left|U^A(q,M,x,\pi^v)-U^A(q,M,x,\pi^v)\right|\geq\frac{t}{K}\right]\leq 2|M|\exp\left(-2nt^2\right)$$

Set the right-hand side equal to a constant γ and solve for t. This yields

$$t=\sqrt{\frac{\ln\left(\frac{2|M|}{\gamma}\right)}{2n}}$$

It follows that

$$\text{Q-Regret}(M,q,\pi^v)\leq\sqrt{\frac{2\ln\left(\frac{2|M|}{\gamma}\right)}{K^2n}}+\bar{v}m\cdot\gamma$$

Setting $\delta=1/n$ to complete the proof.