Journal of Mathematical Economics 62 (2016) 36–51

Contents lists available at ScienceDirect

Journal of Mathematical Economics

journal homepage: www.elsevier.com/locate/jmateco

Locally robust contracts for moral hazard☆

Gabriel Carroll, Delong Meng*

Stanford University, United States

ARTICLE INFO

Article history:
Received 9 August 2015
Received in revised form
2 November 2015
Accepted 5 November 2015
Available online 14 November 2015

Keywords:
Contract
Principal–agent problem
Local robustness
Worst-case
Optimality-robustness tradeoff

ABSTRACT

We consider a moral hazard problem in which the principal has a slight uncertainty about how the agent's actions translate into output. An incentive contract can be made robust against an ϵ amount of uncertainty, at the cost of a loss to the principal on the order of $\sqrt{\epsilon}$, by refunding a small fraction of profit to the agent. We show that as ϵ goes to zero, this construction is essentially optimal, in the sense of minimizing the worst-case loss, among all modifications to the contract that do not depend on the details of the environment.

1. Introduction

Economic models tend to assume that agents have perfect knowledge of the environment in which they operate. What happens to the predictions, qualitatively, when a small amount of uncertainty is introduced? If models with perfectly specified environments are used to derive policy prescriptions, what is the best way to make these policies robust against a small amount of uncertainty that inevitably occurs in the real world?

We consider a standard principal–agent model, in which an agent can privately choose one of several effort levels, producing a stochastic output, and the principal can write a contract specifying payment as a function of output, in order to incentivize the agent to exert effort. The principal, in designing a contract, has a model in mind that describes the probability distribution over output that results from each of the agent's possible effort levels. But she also knows that her model might be mistaken by some small amount ϵ, meaning that the actual probability of any output, for any given effort level, might be up to ϵ larger or smaller than the model assumes.

If the principal simply evaluates a possible contract based on its performance in her model (for example, she might solve the model for the optimal contract), then its actual performance might be precipitously worse than she expects. For example, in a textbook model with just two effort levels (e.g. Hart and Holmström, 1987, Section 1.3), the optimal contract implementing high effort will have the incentive constraint exactly bind. Then, if the principal's model is off by an arbitrarily small amount, the incentive constraint may actually be violated and the agent may choose to exert low effort instead.

A moment's reflection suggests the principal can make the contract robust by adding slack to the incentive constraints. But how much slack should she add, and how would the resulting contract look different from the one derived in the model? To put it more sharply, if the principal evaluates a contract using the idealized model, what is a simple recipe for modifying the contract to make it robust to the ϵ uncertainty?

We show that the principal can make the contract robust by giving a share τ of her profit back to the agent, where τ is on the order of $\sqrt{\epsilon}$. We show that, in any actual environment that is within ϵ of the principal's model, the profit from this modified contract falls short of the model profit by at most an amount of order $\sqrt{\epsilon}$. Intuitively, the modified contract pads the incentive constraints in favor of effort levels that are more profitable for the principal; and the $\sqrt{\epsilon}$ factor optimally trades off the padding with the principal's desire not to have to make large extra payments to the agent. This construction draws on the work of Madarász and Prat (unpublished paper) who apply a similar construction to provide local robustness in a multidimensional screening problem. (A very similar approximation argument also appears in Chassang (2013, Lemma A.1).)

We then further show that this construction is optimal, for $\epsilon \to 0$: there is no other recipe for modifying a given contract that guarantees a significantly smaller worst-case loss relative to the model. More precisely, for any given contract, if the principal

☆ Authors are listed in random order.

* Corresponding author.
E-mail address: delong13@stanford.edu (D. Meng).

http://dx.doi.org/10.1016/j.jmateco.2015.11.001
0304-4068/

considers any "black-box" modification to make it robust – one that does not depend on the details of her model – then the construction above asymptotically attains the smallest possible value for the worst-case loss. If the principal does take into account the details of the model in modifying the contract, then it may be possible to do better, indeed attaining a loss of order ϵ rather than $\sqrt{\epsilon}$. However, there is no such guarantee that is uniform across all models; the best possible uniform guarantee is back to order $\sqrt{\epsilon}$ (although it may improve on the black-box construction by a constant factor). We also note that none of our results actually make any reference to the original contract being optimal for the model; the results start from any arbitrary contract, and compare the actual profit from the modified contract against the model profit from the original contract.

For notational simplicity, we state first in a model in which the principal and agent are risk-neutral, and payments to the agent are constrained from below by limited liability. We then show that the arguments carry over to a setting with risk-aversion and with a participation constraint; now the principal modifies the contract by increasing the agent's *utility level* by an amount equal to share τ of her profit. In Section 4, we also show how a version of our construction generalizes much more broadly beyond our simple static hidden-action model, although the optimality result does not generalize as readily.

This paper contributes to a small but growing literature on robust moral hazard contracting in uncertain environments, such as Chassang (2013), Garrett (2014), Carroll (2015), and Antić (unpublished paper). Much of that literature allows for a large space of uncertainty. Most closely related is the smaller literature on local robustness in mechanism design, in which the principal has only a small amount of uncertainty about the correctness of her model. Positive results in this area include the screening paper of Madarász and Prat (unpublished paper) mentioned above, which inspired our study; as well as Meyer-ter-Vehn and Morris (2011), on higher-order belief perturbations. Negative results include Jehiel et al. (2012) on Bayesian incentive compatibility with interdependent values, and Chung and Ely (2003) and Aghion et al. (2012), both on almost-complete-information implementation.

2. The basic model

2.1. Setup

We present here the basic model. This is a standard principal–agent model — the agent exerts costly effort, leading to a stochastic output, and only output can be contracted on. For this section, the principal and agent are both risk-neutral, and there is a limited liability constraint; the principal can never pay less than zero.

We consider a discrete setup: There are $K \geq 2$ possible levels of effort that the agent can exert, which we will simply call $1, \ldots, K$; and $N \geq 2$ possible values of output that may be realized, $y = (y_1, \ldots, y_N)$. We assume that the values of K and the y_i are commonly known. What is not perfectly known is the *environment*, which describes the cost of each level of effort and the corresponding probability distributions over output. Thus, an environment $\mathcal{E} = (c_1, \ldots, c_K; f_1, \ldots, f_K)$ consists of K real numbers representing the costs of effort, along with K probability distributions $f_1, \ldots, f_K$, each of which may be represented as a vector of N nonnegative numbers summing to 1.

A *contract* w is a vector of N nonnegative numbers, specifying what the agent is paid for each possible level of output. The nonnegativity requirement captures the limited liability constraint. If the environment is $\mathcal{E}$ and the agent is offered contract w, then his payoff from taking any action k is his expected payment minus the effort cost, which can be written using the dot product, as $f_k \cdot w - c_k$. Accordingly, the set of incentive-compatible actions is

$$\mathbf{k}^*(w|\mathcal{E}) = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} (f_k \cdot w - c_k).$$

The principal's corresponding profit is the expected output minus wage paid,

$$V(w|\mathcal{E}) = \max_{k \in \mathbf{k}^*(w|\mathcal{E})} f_k \cdot (y - w).$$

(The max operator reflects the possibility that the agent is indifferent among multiple optimal actions, in which case we assume he chooses the one that is best for the principal. This approach is consistent with having incentive constraints bind in the optimal contract for a particular $\mathcal{E}$.)

We will assume that the principal has some model environment $\mathcal{E}$ in mind when she contemplates a contract and predicts her resulting profit. But she allows that the model may be slightly misspecified, and knows only that the true environment $\widetilde{\mathcal{E}} = (\widetilde{c}_1, \ldots, \widetilde{c}_K; \widetilde{f}_1, \ldots, \widetilde{f}_K)$ is within ϵ of $\mathcal{E}$. For our purposes, this means that the effort costs are the same as in $\mathcal{E}$ but the output probabilities may be off by up to ϵ. (This is just one of numerous ways that we could specify the set of possible true environments. We follow this approach for simplicity, but many others would give qualitatively similar results, as we discuss later in Section 5.)

Accordingly, we write $B_\epsilon(\mathcal{E})$ for the set of all possible true environments satisfying this condition:

$$B_\epsilon(\mathcal{E}) = \{(\widetilde{c}_1, \ldots, \widetilde{c}_K; \widetilde{f}_1, \ldots, \widetilde{f}_K) \mid \widetilde{c}_k = c_k \text{ for each } k; \\ |\widetilde{f}_k(i) - f_k(i)| < \epsilon \text{ for each } k, i\}.$$

We are interested in how best to modify a given contract w to make it robust to the ϵ uncertainty. We express the principal's desire for robustness by assuming that she does not have a prior over the possible true environments $\widetilde{\mathcal{E}}$; instead, the modified contract $\widetilde{w}$ is evaluated by its worst-case performance over all such environments. We are concerned with how this performance compares to the "ideal" profit that the original contract w would give in the model environment; thus, we focus on the discrepancy

$$D(w, \widetilde{w}, \mathcal{E}, \epsilon) = V(w|\mathcal{E}) - \inf_{\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})} V(\widetilde{w}|\widetilde{\mathcal{E}}). \tag{2.1}$$

Throughout, we will treat K, N, y and w as fixed, and constant factors in the results below may depend on them (except where explicitly stated). However, we will not necessarily treat $\mathcal{E}$ as fixed, because we will sometimes be interested in "black-box" modifications to the contract, which can be performed without knowing the environment. (This can be motivated as a simplicity restriction on the possible modifications; they cannot depend too much on the principal's knowledge of the environment. Alternatively, we can imagine that the principal does not know the environment, and was handed the contract w by a designer who claims to know $\mathcal{E}$, but the principal is not entirely confident that the designer's model is correct.) Note that in this black-box approach it makes sense to take w as given – and N, y which are prerequisite for defining w – since otherwise the principal's problem is not well-defined.[1]

We also will not take ϵ as fixed; we are interested in the situation where uncertainty is small, which we represent by taking limits as $\epsilon \to 0$.

[1] The exception is K, which this motivation suggests treating as part of $\mathcal{E}$. For the black-box results the assumption of known K is actually never needed. It is only used to give stronger bounds in the non-black-box result, Theorem 2.5: if we wanted bounds valid for all K, then (2.2) would collapse to Corollary 2.3 since $h_K \to 2$ as $K \to \infty$; and (2.3) would disappear since $h_2 = 0$.