Locally Robust Contracts for Moral Hazard

Gabriel Carroll Delong Meng
Stanford University

April 7, 2015

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.

Keywords: contract, principal-agent problem, local robustness, worst-case, optimality-robustness tradeoff

JEL Classification: D81, D82, D86

Authors are listed in random order. We thank Ilya Segal, Yiqing Xing, Paul Milgrom, and Matthew Jackson for helpful comments.

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 uses the contract that is optimal for her model, then its actual performance might be precipitously worse than the model predicts. For example, in a textbook model with just two effort levels (e.g. [7, 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 computes a contract from 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 [10] who apply a similar construction to provide local robustness in a multidimensional screening problem. (A very similar approximation argument also appears in Chassang [4, 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 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 depend on the original contract being optimal for the model; the results simply compare the actual profit from the modified contract against the model profit from the original contract.

For notational simplicity, we state our results 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 [4], [6], [3], [2]. 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. This includes the screening paper of Madarász and Prat [10] mentioned above, which inspired our study; as well as Jehiel, Meyer-ter-Vehn, and Moldovanu [9] which considers local robustness to higher-order beliefs, and Chung and Ely [5] and Aghion, Fudenberg, Holden, Kunimoto, and Tercieux [1] on implementation with almost-complete information.

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

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

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 designs 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

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

Finally, we make some assumptions to simplify exposition. Notice that we may as well assume the initial contract w satisfies $\min_i w_i = 0$, since otherwise the principal can clearly reduce each w_i by the amount $\min_i w_i$ and save money without affecting the agent's incentives for effort. We also can reorder the output values to assume $w_1 = \min_i w_i$ and $y_1 = \min_{i:w_i=w_1} y_i$. Also, if we shift all y_i by a constant, the quantity $D(w, \widetilde{w}, \mathcal{E}, \epsilon)$ is invariant, so we can normalize $y_1 = 0$ without loss of generality. Thus, we henceforth assume that y and w satisfy two conditions: 1) $w_1 = y_1 = 0$ and 2) if $w_i = 0$, then $y_i \geq 0$.

2.2 Results

Consider a given contract w. The argument in the introduction shows that the performance of the contract may fail to be robust to arbitrarily small uncertainty: if we simply take $\widetilde{w} = w$, then for some $\mathcal{E}$, the quantity $D(w, \widetilde{w}, \mathcal{E}, \epsilon)$ stays large as $\epsilon \to 0$. In particular, this happens when $\mathcal{E}$ is such that an incentive constraint binds in w, so will typically happen if (for example) the principal came up with w by solving for the optimal contract in her model $\mathcal{E}$.

We show how to optimally modify the contract w to make it robust, for small ϵ. For any positive real number τ, define the contract $w(\tau)$ as $w + \tau \cdot (y - w)$: pay a fraction τ of the principal's profit back to the agent. The contract $w(\tau)$ satisfies the limited liability constraint for small enough τ (because $y_i \geq 0$ whenever $w_i = 0$). We show that the contract $w(\tau)$ is robust for appropriately chosen τ.

To state our results, we define a constant C^* as follows. Let $A = \{a \in [-1, 1]^N \mid a_1 + \cdots + a_N = 0\}$, and for any N-dimensional vector $v = (v_1, \ldots, v_N)$, define $\text{maxdiff}(v) = \max_{a \in A}(a \cdot v)$: it is not hard to check that $\text{maxdiff}(v)$ is the total difference between the $\lfloor N/2 \rfloor$ largest and $\lfloor N/2 \rfloor$ smallest components of v. Then, put

$$C^* = 2\sqrt{2 \cdot \text{maxdiff}(w) \cdot \max_{i=1,\ldots,N}(y_i - w_i)}.$$

This quantity is known to the principal, since y and w are given.

Theorem 2.1. *Let $C > C^*$ be fixed. For all sufficiently small ϵ, there exists a τ that depends only on w, y, and ϵ, such that for any environment $\mathcal{E}$, we have*

$$D(w, w(\tau), \mathcal{E}, \epsilon) < C\sqrt{\epsilon}.$$

Theorem 2.1 demonstrates that the contract $w(\tau)$ successfully approximates the ideal profit $V(w|\mathcal{E})$ in spite of the uncertainty about the environment, and the possible discrepancy in profit is on the order of $\sqrt{\epsilon}$. Intuitively, the discrepancy comes from two sources. The first source is the change in payment from contract w to $w(\tau)$, which is proportional to τ. The second source is the potential change in the agent's effort choice (from $\mathbf{k}^*(w|\mathcal{E})$ to $\mathbf{k}^*(w(\tau)|\widetilde{\mathcal{E}})$), and this effect turns out to be proportional to $\frac{\epsilon}{\tau}$. Therefore, if we pick τ on the order of $\sqrt{\epsilon}$, then we can attain a total discrepancy on the order of $\sqrt{\epsilon}$.

Note that the construction in Theorem 2.1 is a "black-box" recipe for modifying the contract w: it depends (unavoidably) on w, on the outputs y, and the degree of uncertainty ϵ, but it does not further depend on the specifics of the environment $\mathcal{E}$.

A natural follow-up question is whether our construction is optimal: Is there a different way of modifying the contract that would ensure a discrepancy significantly smaller than $C^*\sqrt{\epsilon}$? If we are interested specifically in black-box constructions, the next theorem shows that there is no such construction.

Theorem 2.2. *Let $C < C^*$ be fixed. Then, for small enough ϵ, for any contract $\widetilde{w}$, we have*

$$\sup_{\mathcal{E}} D(w, \widetilde{w}, \mathcal{E}, \epsilon) > C\sqrt{\epsilon}.$$

Theorem 2.2 indicates that our construction of $w(\tau)$ is optimally robust for small ϵ: it essentially guarantees a discrepancy of $C^*\sqrt{\epsilon}$, regardless of what the model environment was, and no alternative construction of $\widetilde{w}$ can give a better such guarantee. (More exact upper and lower bounds appear in the proofs of the theorems.)

However, we can also consider non-black-box constructions of $\widetilde{w}$; after all, the principal in our model presumably knows $\mathcal{E}$, so she might wish to use this knowledge in writing the robust contract $\widetilde{w}$. In this case, as $\epsilon \to 0$, the principal can attain a discrepancy of order ϵ instead of $\sqrt{\epsilon}$.

Proposition 2.3. *Let $\mathcal{E}$ be given. Then there exists a constant $C(\mathcal{E})$ such that, for all small enough $\epsilon > 0$, there is some $\tau > 0$ satisfying*

$$D(w, w(\tau), \mathcal{E}, \epsilon) < C(\mathcal{E}) \cdot \epsilon.$$

However, the constant factor $C(\mathcal{E})$ depends on the model environment $\mathcal{E}$. If we are interested in a quantitative bound that is uniform over environments, then we are back to order $\sqrt{\epsilon}$, even though the principal is allowed to make $\widetilde{w}$ depend on $\mathcal{E}$.

Theorem 2.4. *Assume that $w_i > 0$ for all $i > 1$. For every K, there exists a constant $h_K < 2$, depending only on K, with the following property. For any y and w:*

(a) If $C > h_K \sqrt{2 \cdot \text{maxdiff}(w) \cdot \max_i (y_i - w_i)}$, then for all sufficiently small ϵ, for any environment $\mathcal{E}$ there exists $\widetilde{w}$ such that

$$D(w, \widetilde{w}, \mathcal{E}, \epsilon) < C\sqrt{\epsilon}.$$

(b) If $C < h_K \sqrt{2 \cdot \max_i (w_i \cdot (y_i - w_i))}$, then for all sufficiently small ϵ, there exists an environment $\mathcal{E}$ such that for all $\widetilde{w}$,

$$D(w, \widetilde{w}, \mathcal{E}, \epsilon) > C\sqrt{\epsilon}.$$

Note that in Theorem 2.4, the principal can choose $\widetilde{w}$ based on the structure of $\mathcal{E}$, whereas in Theorem 2.1 the principal uses no information about $\mathcal{E}$. Thus, we should expect to attain a smaller optimal discrepancy than we had in the black-box case, and this is indeed the case, as the fact that $h_K < 2$ indicates.

Theorem 2.4 gives tight bounds if there are just two output levels; but for more than two outputs, there is a gap between our upper and lower bounds. The difficulty is that a tight lower bound rests on the premise that w is optimal or near-optimal for $\mathcal{E}$ — indeed, otherwise we could build $\widetilde{w}$ based on the optimal contract for $\mathcal{E}$, and the discrepancy might even be negative. Proving a lower bound thus requires constructing an environment $\mathcal{E}$ for which the given contract w is optimal. Our proof is based on a particular construction that focuses on just two output levels. We suspect a sharper bound may be obtainable by a subtler construction.

2.3 Arguments for the black-box results

We begin by proving Theorem 2.1. Here, and subsequently, we use standard "big O" notation: for any function $g(\epsilon)$, we write $O(g(\epsilon))$ to mean a quantity that is bounded above by $C \cdot g(\epsilon)$ as $\epsilon \to 0$, where C is a constant that depends only on the primitives (K, N, w, y).

The intuition for the theorem is as described previously: in going from contract w in environment $\mathcal{E}$ to $w(\tau)$ and $\widetilde{\mathcal{E}}$, the profit discrepancy due to increased payment is on the order of τ, and the discrepancy due to any possible change in effort is on the order of ϵ/τ; choosing τ on the order of $\sqrt{\epsilon}$ optimally balances these two.

Proof of Theorem 2.1. We claim that there exists a $\tau > 0$ such that for all $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$ the following inequality holds for sufficiently small ϵ:

$$V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) \leq C^*\sqrt{\epsilon} + O(\epsilon).$$

It suffices to prove this inequality because as ϵ goes to 0, the term $O(\epsilon)$ is dominated by $\sqrt{\epsilon}$.

We first explicitly write out $V(w|\mathcal{E})$ and $V(w(\tau)|\widetilde{\mathcal{E}})$. Suppose contract w induces effort k in environment $\mathcal{E}$. We have $V(w|\mathcal{E}) = f_k \cdot (y - w)$. Suppose $w(\tau)$ induces effort l in some environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$. Then $V(w(\tau)|\widetilde{\mathcal{E}}) = \widetilde{f}_l \cdot (y - w(\tau))$. Since $y - w(\tau) = (y - w)(1 - \tau)$, we have

$$\begin{aligned} V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) &= f_k \cdot (y - w) - \widetilde{f}_l \cdot (y - w(\tau)) \\ &= f_k \cdot (y - w) - \widetilde{f}_l \cdot (y - w)(1 - \tau) \\ &= (\widetilde{f}_k - \widetilde{f}_l) \cdot (y - w) + \tau \cdot \widetilde{f}_l \cdot (y - w) + (f_k - \widetilde{f}_k) \cdot (y - w). \end{aligned}$$

In the last expression, the second term is bounded above by $\tau \cdot \max_i(y_i - w_i)$, and the third term is bounded by $O(\epsilon)$ because all components of $(f_k - \widetilde{f}_k)$ lie in the interval $[-\epsilon, \epsilon]$. Therefore we get

$$V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) \leq (\widetilde{f}_k - \widetilde{f}_l) \cdot (y - w) + \tau \cdot \max_i(y_i - w_i) + O(\epsilon). \tag{2.2}$$

We are left with bounding the first term $(\widetilde{f}_k - \widetilde{f}_l) \cdot (y - w)$. We consider the incentive constraints in environments $\mathcal{E}$ and $\widetilde{\mathcal{E}}$. In environment $\mathcal{E}$, the agent prefers effort k over l, so we have

$$f_k \cdot w - c_k \geq f_l \cdot w - c_l. \tag{2.3}$$

In environment $\widetilde{\mathcal{E}}$, the agent prefers effort l over k, so we have

$$\widetilde{f}_l \cdot w(\tau) - c_l \geq \widetilde{f}_k \cdot w(\tau) - c_k. \tag{2.4}$$

Summing up inequalities (2.3) and (2.4), we obtain that

$$f_k \cdot w + \widetilde{f}_l \cdot w(\tau) - c_k - c_l \geq f_l \cdot w + \widetilde{f}_k \cdot w(\tau) - c_l - c_k,$$

which implies that

$$f_k \cdot w + \widetilde{f}_l \cdot w(\tau) \geq f_l \cdot w + \widetilde{f}_k \cdot w(\tau).$$

The above inequality rearranges into $(\widetilde{f}_k - \widetilde{f}_l) \cdot (w(\tau) - w) \leq (f_k - \widetilde{f}_k - f_l + \widetilde{f}_l) \cdot w$. The vector $(f_k - \widetilde{f}_k - f_l + \widetilde{f}_l)$ has components that sum to 0, and all components lie within $[-2\epsilon, 2\epsilon]$. Thus $(f_k - \widetilde{f}_k - f_l + \widetilde{f}_l) \cdot w \leq 2\epsilon \cdot \text{maxdiff}(w)$. Since $w(\tau) - w = \tau \cdot (y - w)$, we have $\tau \cdot (\widetilde{f}_k - \widetilde{f}_l) \cdot (y - w) = (\widetilde{f}_k - \widetilde{f}_l) \cdot (w(\tau) - w) \leq 2\epsilon \cdot \text{maxdiff}(w)$. It follows that

$$(\widetilde{f}_k - \widetilde{f}_l) \cdot (y - w) \leq \frac{2\epsilon}{\tau} \cdot \text{maxdiff}(w).$$

Plug this bound into expression (2.2), we deduce that

$$V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) \leq \frac{2\epsilon}{\tau} \cdot \text{maxdiff}(w) + \tau \cdot \max_i(y_i - w_i) + O(\epsilon). \tag{2.5}$$

If $C^* > 0$, then let $\tau = \sqrt{\frac{2 \cdot \text{maxdiff}(w)}{\max_i(y_i - w_i)}}\,\epsilon$. For small enough ϵ, contract $w(\tau)$ satisfies the limited liability constraint, and the upper bound given by (2.5) is equal to $C^*\sqrt{\epsilon} + O(\epsilon)$.

If $C^* = 0$, then either $\text{maxdiff}(w) = 0$ or $\max_i(y_i - w_i)$=0. If $\text{maxdiff}(w) = 0$, then let $\tau = \epsilon$. We have $\tau \cdot \max_i(y_i - w_i) = O(\epsilon)$. If $\max_i(y_i - w_i) = 0$, then let τ be a small enough constant such that $w(\tau)$ satisfies the limited liability constraint. We have $\frac{2\epsilon}{\tau} \cdot \text{maxdiff}(w) = O(\epsilon)$. In either case, we get $V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) = O(\epsilon)$. Note also that throughout, the value of τ depends only on w, y, and ϵ, as promised. □

For the corresponding lower bound on discrepancy, Theorem 2.2, the intuition is as follows. Given the contract w and the modified contract $\widetilde{w}$, we construct an environment $\mathcal{E}$ to satisfy two incentive constraints: the agent should prefer high effort under w and $\mathcal{E}$, but low effort under $\widetilde{w}$ and a nearby environment $\widetilde{\mathcal{E}}$. We would like to design $\mathcal{E}$ such that this change in effort harms the principal by making a high output level i^* less likely. This just requires having high effort be more productive than low effort. Note however that we are limited in this direction: If $\widetilde{w}$ pays much more for high output than w does, and the agent prefers high effort under w, then he strongly prefers high effort under $\widetilde{w}$, and our ϵ wiggle room in going from $\mathcal{E}$ to $\widetilde{\mathcal{E}}$ will not be enough to offset this. Following this reasoning, the maximum possible difference in output probabilities between high and low

effort (while satisfying the two incentive constraints) is at most order $\epsilon/(\widetilde{w}_{i^*} - w_{i^*})$.

Now the discrepancy between w and $\widetilde{w}$ comes from two main sources: the difference between payments w_{i^*} and $\widetilde{w}_{i^*}$ for a high output i^*, and the difference in the probability of high output due to the change in effort. The first term is on the order of $\widetilde{w}_{i^*} - w_{i^*}$; the second term, following the above calculations, is on the order of $(y_{i^*} - w_{i^*}) \cdot \epsilon/(\widetilde{w}_{i^*} - w_{i^*})$. Summing these two terms for an optimally chosen $\widetilde{w}$ gives the theorem.

The formal proof is a bit tedious, and is deferred to the appendix.

2.4 Arguments for non-black-box results

The proof techniques for the non-black-box results are similar to those for the black-box results. We describe the intuition of Proposition 2.3 and Theorem 2.4 and defer the proofs to the appendix.

Proposition 2.3 states that there exists a $w(\tau)$ for which the discrepancy is on the order of ϵ. Recall from the proof of Theorem 2.1 that there are two sources of discrepancy: the change in the payment amount and the change in the effort choice. If we hold the environment $\mathcal{E}$ fixed and only allow ϵ to vary, then we can pick τ on the order of ϵ, such that the contract $w(\tau)$ in any $\widetilde{\mathcal{E}}$ will induce the same effort choice k as in the original model. Consequently, the discrepancy coming from effort choice disappears, and only the change in payment — which is on the order of ϵ — remains.

Why does this work? Any incentive constraint that was non-binding in the original model will remain non-binding for small ϵ and τ, so we only need to worry about incentive constraints that originally were binding, where the agent was indifferent between effort choice k and some other l. In this case, adding ϵ uncertainty may make the incentive constraint be violated by an amount of order ϵ. On the other hand, adding a fraction $\tau(y - w)$ back to payment will relax the incentive constraint by an amount on the order of τ (and this does make the incentive constraint looser, not tighter; this is because k was better than l for the principal). Therefore, we only need τ on the same order as ϵ to undo the violation.

We illustrate this argument in the following example.

Example 2.5. Suppose there are $K = 3$ effort levels and $N = 2$ output levels. We have $0 = y_1 < y_2$. Let x be a positive real number smaller than y_2. Consider the following environment $\mathcal{E}$:

$$f_1 = \left(\frac{3}{4}, \frac{1}{4}\right), c_1 = 0;\ f_2 = \left(\frac{1}{2}, \frac{1}{2}\right), c_2 = \frac{1}{4} \cdot x;\ f_3 = (0, 1), c_3 = \frac{3}{4} \cdot x.$$

The optimal contract w for this environment is $w_1 = 0$ and $w_2 = x$. Under contract w, the agent is indifferent between all three effort levels, so he picks the one best for the principal. The agent picks effort 3, and the principal's profit is equal to $V(w|\mathcal{E}) = y_2 - w_2$.

Now suppose there is ϵ-uncertainty. The argument for Proposition 2.3 goes as follows. The principal uses a contract $w(\tau)$ that induces effort 3 in any environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$. To achieve this goal, the principal needs to consider the worst-case environment, namely, $\widetilde{f_2} = (\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon)$ and $\widetilde{f_3} = (\epsilon, 1 - \epsilon)$. This environment is indeed the most adversarial for the principal, because it maximally makes effort 3 both less productive than before, and also harder to incentivize than before, due to effort 2 being more appealing to the agent. (We ignore effort 1 for now.)

The principal wishes to construct $w(\tau)$ so as to incentivize effort 3 in the adversarial environment. She solves for $w(\tau)_2$ from the incentive constraint

$$(1 - \epsilon) \cdot w(\tau)_2 - c_3 \geq (\frac{1}{2} + \epsilon) \cdot w(\tau)_2 - c_2.$$

After solving, and then extracting τ using $w(\tau_2) = w_2 + \tau(y_2 - w_2)$, we get

$$\tau \geq \frac{2w_2 \cdot \epsilon}{(\frac{1}{2} - 2\epsilon)(y_2 - w_2)}.$$

The important observation is that this minimum τ is on the same order as ϵ. Therefore, choosing this τ ensures that the agent still chooses effort 3 in all possible environments $\widetilde{\mathcal{E}}$, and as outlined above, this ensures a discrepancy of order ϵ.

A similar argument applies even if there are multiple effort levels and multiple binding incentive constraints.

In Example 2.5, we note that, although τ varies linearly in ϵ, the coefficient depends on the parameters of $\mathcal{E}$, so that the upper bound on discrepancy also depends on $\mathcal{E}$. Moreover, this coefficient may be arbitrarily large. Indeed, if f_2 is close to f_3 (e.g. if we change the example so that $f_2 = (0.01, 0.99)$, and set c_2 so that the agent is again indifferent in the model environment), then the principal needs to make τ be a large multiple of ϵ in order to still induce effort 3 in every possible environment.

To get any upper bound on discrepancy that holds uniformly across all environments, we must instead consider contracts that may induce a lower effort choice. We know already from Theorem 2.1 that once we allow this, we can get an upper bound on the order of $\sqrt{\epsilon}$. In order to make this argument tight, we need to carefully examine all possibilities for such contracts: For each effort level l, we consider a modified contract $w(\tau)$ that may

induce effort l or higher. Picking the best of these contracts now gives us a discrepancy on the order of $\sqrt{\epsilon}$. Why $\sqrt{\epsilon}$? This works just as in Theorem 2.1, where the discrepancy may come from two sources (changed effort and increased payment), and the $\sqrt{\epsilon}$ comes from choosing τ to optimally balance them. Going through this argument carefully gives us the upper bound of Theorem 2.4(a), which has a better constant factor than Theorem 2.1.

Example 2.6. We modify the setting of Example 2.5 as follows: keep efforts 1 and 3 as in the example, but now suppose more generally that $f_2 = (\delta, 1-\delta)$ for some $\delta \in (0, 3/4)$, and $c_2 = (3/4 - \delta) \cdot x$. The optimal contract is still $w_1 = 0$ and $w_2 = x$. Under contract w, the agent is again indifferent among all three effort levels in the model environment. Example 2.5 shows how to construct a contract $w(\tau)$, with τ as small as possible, that always induces effort 3 in any $\widetilde{\mathcal{E}}$. The resulting value of τ depends linearly on ϵ, but with a coefficient that is proportional to $1/\delta$.

Suppose the principal also writes another contract $w(\tau')$ that induces effort 2 or better in any $\widetilde{\mathcal{E}}$.

If δ is small, i.e. efforts 2 and 3 are close together, then τ will need to be very large, since τ is of order ϵ/δ. On the other hand, this means that the discrepancy from $w(\tau')$ due to the change from effort 3 to 2 is small, so the principal is more willing to use $w(\tau')$ in this case.

This reasoning and a bit of calculation leads us to

$$D(w, w(\tau), \mathcal{E}, \epsilon) \leq \tau \cdot (y_2 - w_2) + O(\epsilon) = O\left(\frac{\epsilon}{\delta}\right) + O(\epsilon) \quad \text{(always induce effort 3)}$$

and

$$D(w, w(\tau'), \mathcal{E}, \epsilon) \leq O(\delta) + O(\epsilon) \quad \text{(induce effort 2 or 3)}$$

where the $O(\epsilon)$ terms are independent of δ.

Now, for any values of δ and ϵ, we can construct two contracts $w(\tau)$ and $w(\tau')$ as above. Between these two contracts, we pick whichever gives the smaller right-hand side above. We can see that this minimum is always at most $O(\sqrt{\epsilon}) + O(\epsilon) = O(\sqrt{\epsilon})$. A more careful computation gives the constant factor in Theorem 2.4(a).

We also see that the $O(\sqrt{\epsilon})$ upper bound cannot be improved: by choosing δ adversarially, namely on the order of $\sqrt{\epsilon}$, it is tight.

The proof of the lower bound (Theorem 2.4(b)) follows the observation in the last line of Example 2.6. We construct an environment $\mathcal{E}$ for which the bounds are tight

throughout the proof of the upper bound (part (a)). This may at first seem impossible, because the lower bound in the theorem statement is supposed to apply to any possible modified contract $\widetilde{w}$, whereas our reasoning above only considers particular contracts $w(\tau)$. However, we construct $\mathcal{E}$ in which only two output levels occur (with nonnegligible probability). When there are only two output levels, *any* candidate contract $\widetilde{w}$ is of the form $w(\tau)$ for some τ (ignoring peripheral cases), and the proof of (a) identifies all the potentially relevant values of τ. We then construct an adversarial environment $\mathcal{E}$ to make the smallest of these discrepancies as large as possible, as sketched in the example.

This also elucidates why our upper and lower bounds in Theorem 2.4 do not coincide: our proof of the lower bound depends on narrowing the possible modified contracts $\widetilde{w}$ to a one-parameter family, which we do by focusing on just two output levels, but we may be throwing out useful information in doing so.

3 Risk aversion and a participation constraint

We now turn our attention to risk averse agents. We will assume a formulation in which the agent's preferences are additively separable between money and the disutility of effort c_k. So, suppose the agent has utility u for money, which is twice continuously differentiable, strictly increasing, and strictly concave. Assume that $u(0) = 0$, and the range of u is the entire real line. For the ease of notation, let $u(w)$ denote the vector $(u(w_1), \ldots, u(w_N))$.

We also replace the limited liability constraint with a participation constraint. We redefine an environment as a collection of effort costs, probability distributions, and an outside option for the agent. We write $\mathcal{E} = (c_1, \ldots, c_K; f_1, \ldots, f_K; \bar{U})$, where $\bar{U}$ is the agent's outside option. Like in the baseline model, we assume that the principal is uncertain about the f_k's, but not about the c_k's or $\bar{U}$. An environment $\widetilde{\mathcal{E}}$ in $B_\epsilon(\mathcal{E})$ has the following form: $(c_1, \ldots, c_K; \widetilde{f}_1, \ldots, \widetilde{f}_K; \bar{U})$, where $|\widetilde{f}_k(i) - f_k(i)| < \epsilon$ for all k and i.

A wrinkle relative to our baseline model is that the principal may be unsure whether a given contract satisfies the participation constraint, since it may be satisfied in some possible environments but not others. Thus, we assume there is some exogenous value V_0 that the principal gets from the outside option. A contract w satisfies the participation constraint in $\mathcal{E}$ if $\max_k(f_k \cdot u(w) - c_k) \geq \bar{U}$. We define $\mathbf{k}^*(w|\mathcal{E}) = \operatorname{argmax}_k(f_k \cdot u(w) - c_k)$ and then $V(w|\mathcal{E}) = \max_{k \in \mathbf{k}^*(w|\mathcal{E})}(f_k \cdot (y - w))$ as before if w satisfies the participation constraint in $\mathcal{E}$, and $V(w|\mathcal{E}) = V_0$ otherwise. We then define the discrepancy by (2.1) as before.

We assume that the given contract w satisfies $V_0 < \min_i(y_i - w_i)$. This ensures that

if the contract w satisfies the participation constraint in $\mathcal{E}$, then for small ϵ, the principal wants the modified contract $\widetilde{w}$ to satisfy the participation constraint for all $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$: Otherwise, the discrepancy is at least $V(w|\mathcal{E}) - V_0$, which is a positive constant.

For risk averse agents, we can no longer normalize w, but we can still normalize y while keeping the discrepancy unchanged. We normalize y such that $\min_i(y_i - w_i) = 0$, and assume that $V_0 < 0$. After this normalization, define

$$C^* = 2 \cdot \sqrt{2 \cdot \text{maxdiff}(u(w)) \cdot \max_i \frac{y_i - w_i}{u'(w_i)}}.$$

Notice that C^* is well-defined because $u'(w_i) > 0$ for all i. Also note that this definition of C^* is consistent with the C^* from the previous section; setting $u(w) = w$ gives the C^* in the risk-neutral case. We can obtain the black-box results in Theorems 2.1 and 2.2 for risk-averse agents with this new definition of C^*.

Theorem 3.1. *We have the following upper and lower bounds for the discrepancy.*

(a) Let $C > C^*$ be fixed. For all sufficiently small ϵ, there exists a $\widetilde{w}$, depending only on w, y, ϵ, and u, such that for any environment $\mathcal{E}$, we have

$$D(w, \widetilde{w}, \mathcal{E}, \epsilon) < C\sqrt{\epsilon}.$$

(b) Let $C < C^*$ be fixed. Then, for small enough ϵ, for any contract $\widetilde{w}$, we have

$$\sup_{\mathcal{E}} D(w, \widetilde{w}, \mathcal{E}, \epsilon) > C\sqrt{\epsilon}.$$

The proof follows the same strategy as the risk-neutral case. For the upper bound, the principal considers contracts of the following form:

$$u(\widetilde{w}_i) = u(w_i) + \tau \cdot (y_i - w_i) + \epsilon \cdot \text{maxdiff}(u(w)) + \epsilon \cdot \text{maxdiff}(y - w).$$

This contract is similar to the one from the risk neutral case. The principal now increases the agent's utility level (instead of payment amount) by a fraction of the profit. Moreover, she attaches an extra term of order ϵ to the agent's utility level. This extra term ensures that the contract satisfies the participation constraint in all possible environments $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$.

Like in the risk-neutral case, the discrepancy comes from the change in payment and the change in effort choice. The change in effort choice contributes to the discrepancy by

an amount proportional to $\frac{\epsilon}{\tau}$, just like in the risk-neutral case. The change in payment requires a little extra work. The change in payment for each output level is equal to $\widetilde{w}_i - w_i$, which can be approximated by $\frac{1}{u'(w_i)} \cdot (u(\widetilde{w}_i) - u(w_i))$ for small enough τ. Since the principal treats the factor $\frac{1}{u'(w_i)}$ as a constant (it is determined by w), the change in payment is on the order of τ. Therefore, our approach from the risk-neutral case applies to this setting as well.

For the lower bound, we first show that $\widetilde{w}$ pays at least as much as w for any output. Indeed, if $\widetilde{w}$ ever pays less than w, then the principal can construct an environment $\mathcal{E}$ such that w satisfies the participation constraint, but $\widetilde{w}$ fails to satisfy it. As a result, we can focus on contracts $\widetilde{w}$ that always pay at least as much as w. The rest of the proof follows the same approach as Theorem 2.2: The principal constructs an environment $\mathcal{E}$ such that the agent prefers high effort under w and $\mathcal{E}$, and low effort under $\widetilde{w}$ and a nearby environment $\widetilde{\mathcal{E}}$. Everything goes through, as long as we approximate $\widetilde{w}_i - w_i$ with $\frac{1}{u'(w_i)} \cdot (u(\widetilde{w}_i) - u(w_i))$.

4 A general model

Given the ideas of this paper so far and of Madarász and Prat [10], it is natural to formulate an abstract, and much more general, framework, in which the idea of "returning a share of the principal's profit to the agent" can be expressed more broadly. This general framework allows us, first, to avoid having to take a specific stand on which environments are considered to be within ϵ of the model environment, and second, to envision applying the same ideas in other contexts beyond our simple one-shot hidden-action model, to settings that may be dynamic and may involve hidden actions, hidden information, or both in combination. The disadvantage is that only our upper bound result generalizes transparently; it is not clear how to formulate a lower bound result without putting more structure on the possible environments.

These points can be made most clearly after writing out the general framework, so let us do so now, and then we comment on the interpretation.

There is a set $\mathcal{S}$ of *strategies* the agent can follow. We assume nothing about $\mathcal{S}$ except that it is a compact subset of some metric space. There is a compact set $Y \subseteq \mathbb{R}$ of possible payoffs for the principal; assume $\max(Y) > 0$ so that it is possible for the principal to make a profit. An *environment* $\mathcal{E} = (u, f)$ consists of two things: a function $u : \mathcal{S} \to \mathbb{R}$ specifying the agent's expected utility from following each strategy; and a function $f : \mathcal{S} \to \Delta(Y)$ specifying a probability distribution over principal's payoffs from

each strategy. We assume that u is upper semi-continuous, and f is continuous (using the weak topology on $\Delta(Y)$).

Unlike the moral hazard framework of the previous sections, here we implicitly assume that the environment (u, f) already incorporates the effects of whatever model contract w the principal has been considering. Thus, the interpretation is that $u(s)$ is the agent's payoff from following strategy s, including whatever payment the principal is planning on making according to w; and likewise $f(s)$ is the distribution over principal's payoff net of her payments specified by w.

Consequently, we will write $\widetilde{w}$ here for possible *adjustments* to the contract. We focus on adjustments which are functions of the realized payoff $y \in Y$. (Thus, we effectively assume that the principal can observe her payoff. She may also be able to observe other things.) An adjustment, then, is a continuous function $\widetilde{w} : Y \to \mathbb{R}$.

Given $\widetilde{w}$, the agent's set of incentive-compatible actions is

$$\mathbf{s}^*(\widetilde{w}|\mathcal{E}) = \text{argmax}_{s\in\mathcal{S}}(u(s) + E_{f(s)}[\widetilde{w}(y)])$$

which is nonempty and compact due to the continuity assumptions. The principal's corresponding profit is

$$V(\widetilde{w}|\mathcal{E}) = \max_{s\in\mathbf{s}^*(\widetilde{w}|\mathcal{E})} E_{f(s)}[y - \widetilde{w}(y)].$$

We take as given a *model* environment $\mathcal{E}$, and for any $\epsilon > 0$, there is a set of possible true environments, $B_\epsilon(\mathcal{E})$. We make no explicit specification of what $B_\epsilon(\mathcal{E})$ is. Instead, we make only two assumptions about its "closeness" to the model $\mathcal{E}$: For any $\widetilde{\mathcal{E}} = (\widetilde{u}, \widetilde{f}) \in \mathcal{E}$, we have

(i) $|\widetilde{u}(s) - u(s)| \leq \alpha\epsilon$ for all $s \in \mathcal{S}$, and

(ii) $|E_{\widetilde{f}(s)}[y] - E_{f(s)}[y]| \leq \beta\epsilon$ for all $s \in \mathcal{S}$,

where α, β are exogenously given constants.

For any constant $\tau \geq 0$, let us write $\widetilde{w}[\tau]$ for the adjustment given by $\widetilde{w}(y) = \tau y$. (We use square brackets here, since $\widetilde{w}$ is a function in this general model, to make clear that τ is a parameter of $\widetilde{w}$ and not its argument.)

The principal's default, if there were no uncertainty about the environment, would be to use whatever contract is implicitly represented in the model environment $\mathcal{E}$, so that the adjustment is $\widetilde{w}[0]$. With uncertainty, she can create robustness to the environment by using $\widetilde{w}[\tau]$ for small τ. This is expressed in our general result below. Let $\overline{y} = \max(Y)$.

Theorem 4.1. *There exists a τ that depends on ϵ but not on the model $\mathcal{E}$, such that for any $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$, we have*

$$V(\widetilde{w}[\tau]|\widetilde{\mathcal{E}}) \geq V(\widetilde{w}[0]|\mathcal{E}) - (\sqrt{2\alpha\overline{y}\epsilon} + \beta\epsilon).$$

Thus, in this very general setup, the principal can still attain a discrepancy on the order of $\sqrt{\epsilon}$, for small ϵ. The proof is a straightforward adaptation of Theorem 2.1, and is given below.

We can now discuss how the specific applications fit into this general model.

- In the moral hazard model of Section 2, $\mathcal{S}$ consists of the K possible effort levels. For each effort level s, $u(s)$ would be the agent's net expected payoff under contract w — that is, the expected payment he receives minus the cost of effort. $f(s)$ would be the distribution over the principal's net payoff, which was called $y - w(y)$ in the original model. Our adjustment $\widetilde{w}$ here corresponds to the difference $\widetilde{w}(y) - w(y)$ in the original model.

  It is easy to see that our specification of the set of close environments, $B_\epsilon(\mathcal{E})$, satisfies conditions (i) and (ii) for suitable α, β.

- In the screening model of Madarász and Prat [10], the agent has a privately known type, which specifies his value for each of various objects ("alternatives") he can purchase. The principal offers a menu of alternatives and corresponding prices. The principal's model specifies a prior distribution over the agent's types, and the uncertainty is about this distribution: the principal may have misspecified each type's location in preference space by up to ϵ.

  To map this into our model, we assume the principal's model menu is given (analogous to w in the moral hazard model). A strategy $s \in \mathcal{S}$ is now a *function* mapping the agent's type to the alternative he chooses from the menu. Then $u(s)$ is the agent's ex-ante expected net utility from this strategy; utility is random because of the uncertainty about the agent's type. $f(s)$ is the distribution over the principal's revenue, which is random for the same reason.

  In this case, there is no uncertainty about the principal's revenue from any given strategy, i.e. $\beta = 0$, since each strategy specifies exactly the probability that any given menu item will be purchased. The uncertainty is all about the agent's utilities.

- One can extend this to imagine plenty of other applications: the agent may take several actions in succession, and may receive information (publicly or privately) at various stages along the way. In any such interpretation, we would assume that $\mathcal{S}$ is the space of all possible strategies for the agent in this game, and that some contract to which the principal has committed has been implicitly fixed, so that $u(s)$ and $f(s)$ represent the payoffs from strategy s net of any incentives specified in this contract. The important assumption for interpreting $\widetilde{w}$ is that the principal's total payoff $y \in Y$ can be contracted on.

We can also elaborate on the strengths and weaknesses of presenting our ideas in this general model, compared to a specific application, such as the moral hazard model we have used in the earlier sections. The main advantage of the general model is the obvious gain in generality. The main disadvantage is that it is not clear how to formulate a lower-bound result such as Theorem 2.2. Giving such a result requires some assumption about the set of possible environments $\mathcal{E}$ (as well as the corresponding set of possible true environments $B_\epsilon(\mathcal{E})$), since we would in general not want to assume that *every* possible pair of functions $u : \mathcal{S} \to \mathbb{R}$, $f : \mathcal{S} \to \Delta(Y)$ represents an allowable environment, but instead would want to define environments in a way that reflects the structure of the application we had in mind (moral hazard, screening, etc.). Similarly, the set $B_\epsilon(\mathcal{E})$ should be defined in a way that is tailored to the application. A lower bound result on the profit discrepancy evidently requires some richness assumption on the space of possible model environments $\mathcal{E}$ (and true environments $\widetilde{\mathcal{E}}$), and it is not clear how this richness assumption would be formulated in the abstract.

We complete this section by proving the general upper bound on discrepancy.

Proof of Theorem 4.1. Consider any positive number τ. Let s be the strategy chosen under $\widetilde{w}[0]$ and environment $\mathcal{E} = (u, f)$, and let $\widetilde{s}$ be the strategy chosen under $\widetilde{w}[\tau]$ and nearby environment $\widetilde{\mathcal{E}} = (\widetilde{u}, \widetilde{f})$. We have

$$\begin{aligned} V(\widetilde{w}[0]|\mathcal{E}) - V(\widetilde{w}[\tau]|\widetilde{\mathcal{E}}) &= E_{f(s)}[y] - (1-\tau)E_{\widetilde{f}(\widetilde{s})}[y] \\ &= (E_{\widetilde{f}(s)}[y] - E_{\widetilde{f}(\widetilde{s})}[y]) + (E_{f(s)}[y] - E_{\widetilde{f}(s)}[y]) + \tau \cdot E_{\widetilde{f}(\widetilde{s})}[y]. \end{aligned}$$

By assumption (ii), the second term on the right side is bounded by $\beta\epsilon$, and the third is clearly bounded by $\tau\overline{y}$. Thus

$$V(\widetilde{w}[0]|\mathcal{E}) - V(\widetilde{w}[\tau]|\widetilde{\mathcal{E}}) \leq (E_{\widetilde{f}(s)}[y] - E_{\widetilde{f}(\widetilde{s})}[y]) + \beta\epsilon + \tau\overline{y}. \tag{4.1}$$

It remains to bound the first term on the right side.

If $\alpha = 0$, then we can simply take $\tau = 0$, and we can assume $\widetilde{s} = s$, since the agent's payoff from each strategy is the same in the two environments. Then this right-side term is 0, and (4.1) simply collapses to $V(\widetilde{w}[0]|\mathcal{E}) - V(\widetilde{w}[0]|\widetilde{\mathcal{E}}) \leq \beta\epsilon$, the desired result. So henceforth we assume $\alpha > 0$.

We consider the agent's preferences over strategies in each of the two environments. Because s is preferred in $\mathcal{E}$ under $\widetilde{w}[0]$, we have

$$u(s) \geq u(\widetilde{s}). \tag{4.2}$$

Because $\widetilde{s}$ is preferred in $\widetilde{\mathcal{E}}$ under $\widetilde{w}[\tau]$, we have

$$\widetilde{u}(\widetilde{s}) + \tau \cdot E_{\widetilde{f}(\widetilde{s})}[y] \geq \widetilde{u}(s) + \tau \cdot E_{\widetilde{f}(s)}[y]$$

from which, using property (i),

$$u(\widetilde{s}) + \alpha\epsilon + \tau \cdot E_{\widetilde{f}(\widetilde{s})}[y] \geq u(s) - \alpha\epsilon + \tau \cdot E_{\widetilde{f}(s)}[y]. \tag{4.3}$$

Adding inequalities (4.2) and (4.3) and canceling, we get

$$\alpha\epsilon + \tau \cdot E_{\widetilde{f}(\widetilde{s})}[y] \geq -\alpha\epsilon + \tau \cdot E_{\widetilde{f}(s)}[y]$$

or

$$E_{\widetilde{f}(s)}[y] - E_{\widetilde{f}(\widetilde{s})}[y] \leq 2\alpha\epsilon/\tau.$$

Hence, (4.1) turns into

$$V(\widetilde{w}[0]|\mathcal{E}) - V(\widetilde{w}[\tau]|\widetilde{\mathcal{E}}) \leq \frac{2\alpha\epsilon}{\tau} + \tau\overline{y} + \beta\epsilon.$$

Now taking $\tau = \sqrt{2\alpha\epsilon/\overline{y}}$ gives the result in the theorem statement.

Note that the τ we obtain is indeed independent of any details of the environment $\mathcal{E}$. □

5 Discussion

We have considered the problem of how best to make contracts robust to a small amount of uncertainty about the environment, using a canonical principal-agent model of moral

hazard. We have seen that the simple approach of refunding a small, fixed fraction of the principal's profit to the agent can make a contract robust to an ϵ amount of uncertainty, and that among "black-box" adjustments that do not depend on the specifics of the model environment, this approach is essentially optimal for small ϵ, in terms of minimizing the possible loss in profits relative to the model. If the adjustment does make use of the details of the environment, we can do slightly better — the discrepancy is still on the order of $\sqrt{\epsilon}$, the same as in the black-box construction, but with an improved constant factor; and this is the best we can hope for if we want a bound that does not itself depend on the specific environment. We have also seen that at least our general upper bound on profit discrepancy holds in a much larger, abstract class of contracting models, although a corresponding lower bound does not seem to be obtainable without putting more structure on the model.

We have made a number of specific modeling choices and assumptions in order to formulate these results. This is as good a place as any to discuss the robustness of our findings to possible variations of the model.

- **Alternative definitions of nearby environments.** We defined $B_\epsilon(\mathcal{E})$ to consist of all environments with exactly the same effort costs as in $\mathcal{E}$, and all output probabilities off by at most ϵ. This definition reflects one possible metric on the space of possible environments. But one could equally well consider many other metrics. For example, we might measure the distance between two output distributions by the maximum *total* difference in probability of any two events. The set of possible true environments then consists of all $(\widetilde{c}_1, \ldots, \widetilde{c}_K; \widetilde{f}_1, \ldots, \widetilde{f}_K)$ such that $\widetilde{c}_k = c_k$ for each k, and

$$\left| \sum_{i \in I} \widetilde{f}_k(i) - \sum_{i \in I} f_k(i) \right| < \epsilon \quad \text{for all } k = 1, \ldots, K \text{ and } I \subseteq \{1, \ldots, N\}.$$

  We notate this set as $\widehat{B}_\epsilon(\mathcal{E})$.

  In this case, since $\widehat{B}_\epsilon(\mathcal{E}) \subseteq B_\epsilon(\mathcal{E})$, the upper-bound results on discrepancies (Theorems 2.1, 2.4(a)) still hold. For lower bounds, note that $B_{\epsilon/N}(\mathcal{E}) \subseteq \widehat{B}_\epsilon(\mathcal{E})$, so applying Theorems 2.2, 2.4(b) with ϵ/N gives corresponding bounds. Thus we can immediately see that the optimal discrepancy bound is still on the order of $\sqrt{\epsilon}$.

  Alternatively, we could allow any $\widetilde{f}_k$ to differ from f_k by up to ϵ probability for each output level, as before, but also allow each $\widetilde{c}_k$ to differ from c_k by up to ϵ.

In this case, since this new set of possible realities includes the original set $B_\epsilon(\mathcal{E})$, our existing lower bounds apply unchanged. On the other hand, our general result, Theorem 4.1, again gives us an upper bound on the order of $\sqrt{\epsilon}$.

It should be evident that there are thus many similar ways we could have written down the definition of nearby environments, each of which would give the same qualitative result — an optimal bound on discrepancy on the order of $\sqrt{\epsilon}$.

- **Screening.** We have constrained the principal to consider a single modified contract $\widetilde{w}$, and examine the worst case over possible environments. However, since the problem is one of asymmetric information — the agent knows the true environment — it is natural to instead consider screening the agent, by offering a *menu* of contracts, and letting the agent choose whichever one he prefers. The agent might then choose different contracts from the menu depending on the environment.

  Allowing screening of this form expands the principal's options, but, in general, it still cannot guarantee a discrepancy smaller than order $\sqrt{\epsilon}$. Indeed, assume $K \geq 3$, and for any given w, consider the environment $\mathcal{E}$ from the proof of Theorem 2.4 (b). Let $\widetilde{\mathcal{E}}$ denote the following alternative environment: $\widetilde{f}_1(i^*) = f_1(i^*) = 0$, $\widetilde{f}_2(i^*) = f_2(i^*) + \epsilon$, $\widetilde{f}_k(i^*) = f_k(i^*) - \epsilon$ for all $k > 2$; moreover, for all $k \in [1, K]$, $\widetilde{f}_k$ puts nonnegative weights only on y_1 and y_{i^*}. We claim that the optimal profit in this environment is $V(w|\mathcal{E}) - \Omega(\sqrt{\epsilon})$.[1] By a version of the calculation in the proof of Theorem 2.4 (b), if a contract $\widetilde{w}$ induces at least effort 2 in $\widetilde{\mathcal{E}}$, then $V(\widetilde{w}|\widetilde{\mathcal{E}}) = V(w|\mathcal{E}) - \Omega(\sqrt{\epsilon})$. If $\widetilde{w}$ induces effort 1, then $V(\widetilde{w}|\widetilde{\mathcal{E}}) \leq 0$. Thus $\max_{\widetilde{w}} V(\widetilde{w}|\widetilde{\mathcal{E}}) = V(w|\mathcal{E}) - \Omega(\sqrt{\epsilon})$. So for this environment $\widetilde{\mathcal{E}}$, no matter what menu of contracts the principal offers, the profit cannot exceed $V(w|\mathcal{E}) - \Omega(\sqrt{\epsilon})$.

- **Randomization.** We could also allow the principal to hedge her uncertainty about the environment by deliberately randomizing over contracts, instead of offering a single $\widetilde{w}$. Then, she would consider a *distribution* over μ contracts $\widetilde{w}$, and then define the resulting discrepancy $V(w|\mathcal{E}) - \inf_{\widetilde{\mathcal{E}}}(E_\mu[V(\widetilde{w}|\widetilde{\mathcal{E}})])$. However, in general, such randomization once again cannot guarantee a discrepancy smaller than order $\sqrt{\epsilon}$. The reason is the same as the argument for screening. Consider the example given in the discussion for screening. We know that there is an $\widetilde{\mathcal{E}}$ such that the optimal profit in $\widetilde{\mathcal{E}}$ is $V(w|\mathcal{E}) - \Omega(\sqrt{\epsilon})$. No matter what distribution of contracts the principal offers, for this environment we have $E_\mu[V(\widetilde{w}|\widetilde{\mathcal{E}})] \leq V(w|\mathcal{E}) - \Omega(\sqrt{\epsilon})$.

[1] Notation: $\Omega(\cdot)$ is the opposite of $O(\cdot)$. That is, $\Omega(g(\epsilon))$ means a function that is bounded below by $C \cdot g(\epsilon)$, for some positive constant C.

- **Regret minimization.** We have assumed that the principal evaluates any possible $\widetilde{w}$ by its worst-case performance over the set of possible environments, and this worst case is compared to the ideal profit from the original contract. An alternative is that the principal wishes to minimize regret: for each possible true environment, she considers the loss relative to the contract that would have been optimal for that environment. Thus, the new "discrepancy" resulting from using contract $\widetilde{w}$ is

$$\widehat{D}(\widetilde{w}, \mathcal{E}, \epsilon) = \sup_{\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})} \left(\max_{w'} (V(w'|\widetilde{\mathcal{E}})) - V(\widetilde{w}|\widetilde{\mathcal{E}}) \right).$$

(Note that this formulation no longer makes any reference to an original contract; see discussion below.)

In this case, unlike the variations discussed above, the results are very different. In particular, if $f_k(i) = 0$ for some k and i, then the minimal discrepancy may fail to converge to 0 as $\epsilon \to 0$. Consider the following example with 3 output levels and 2 effort levels. Assume $0 = y_1 < y_2 < y_3$ and $c_2 - c_1 < y_3/4$. Let

$$\mathcal{E} = (c_1, c_2; f_1 = (1/2, 0, 1/2), f_2 = (0, 0, 1))$$

and

$$\widetilde{\mathcal{E}} = (c_1, c_2; \widetilde{f}_1 = (1/2, 0, 1/2), \widetilde{f}_2 = (0, \epsilon, 1-\epsilon)).$$

The optimal contract for $\mathcal{E}$ is $w = (0, 0, 2(c_2 - c_1))$, and $V(w|\mathcal{E}) = y_3 - 2(c_2 - c_1)$. The optimal contract for $\widetilde{\mathcal{E}}$ is $w' = (0, (c_2 - c_1)/\epsilon, 0)$, and $V(w'|\widetilde{\mathcal{E}}) = (1-\epsilon)y_3 + \epsilon y_2 - (c_2 - c_1)$. Clearly, as $\epsilon \to 0$, we have $V(w'|\widetilde{\mathcal{E}}) \approx V(w|\mathcal{E}) + (c_2 - c_1)$. Now suppose the principal offers a robust contract $\widetilde{w}$. In order to have regret go to 0 for small ϵ, $\widetilde{w}$ must induce effort 2 in both $\mathcal{E}$ and $\widetilde{\mathcal{E}}$. The incentive constraint for $\mathcal{E}$ gives $\widetilde{w}_3 \geq \frac{1}{2}\widetilde{w}_1 + \frac{1}{2}\widetilde{w}_3 + c_2 - c_1$, which implies that $\widetilde{w}_3 \geq 2(c_2 - c_1)$. However, this means

$$V(\widetilde{w}|\widetilde{\mathcal{E}}) \leq (1-\epsilon)(y_3 - 2(c_2 - c_1)) + \epsilon y_2.$$

In comparison, the optimal contract w' gives $V(w'|\widetilde{\mathcal{E}}) = (1-\epsilon)y_3 + \epsilon y_2 - (c_2 - c_1)$. Therefore the regret in $\widetilde{\mathcal{E}}$ is at least $V(w'|\widetilde{\mathcal{E}}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) \geq (1 - 2\epsilon)(c_2 - c_1)$, which converges to a positive constant. The regret does not approach 0.

An alternative way to write down the model with a regret-minimization objective would be to look at regret relative to the original contract w, in the true environment: $\widehat{D}(w, \widetilde{w}, \mathcal{E}, \epsilon) = \sup_{\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})} \left(V(w|\widetilde{\mathcal{E}}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) \right)$. However, this approach

seems inappropriate: The whole reason for looking for robust contracts is that w may perform poorly in environments close to the model, so $V(w|\widetilde{\mathcal{E}})$ is not an appropriate benchmark. In any case, this alternative approach would give uninteresting results, since the principal could achieve zero regret by taking $\widetilde{w} = w$.

Finally, we close by discussing some possible directions in which it is not immediate how to extend our existing model, and which call for further study.

One natural direction of future study is to consider alternative instances of the general model from Section 4, and look for corresponding lower bounds on discrepancy, in particular to see how widely our simple profit-sharing construction remains black-box optimal. Another direction is to consider variants of our setup in which there is no particular contract that is robust. For example, if we allow the set of possible output levels to be unbounded above, then there can be environments where it is possible to obtain infinite expected profit, yet for arbitrarily small ϵ there are nearby environments for which every effort level yields finite expected profit; thus no robust guarantees in our sense are possible. (Imagine that each positive integer level of output y is generated with probability proportional to $1/y^2$.) This raises the question of what the appropriate notion of local robustness is in such an environment. More generally, it suggests that the model we have put forward might not be the right one for thinking quantitatively about robustness even when output is finite but potentially very large compared to the expectation.

One other direction to consider is what happens when the principal cannot contract on output y directly, but only on some other noisy signal z of the agent's effort choice (as first introduced in Holmström [8]). It is not clear how to formulate our quantitative bounds in such an environment, since contracts would now be a function of the signal z, and it is unclear how to reproduce our key construction — paying back a fraction τ of profit — in a way that makes it depend only on z, particularly if this was to be done in a black-box way. However, incentives in organizations often do depend on indirect signals (such as a boss's written evaluation of an employee's performance) rather than being a function of individual contributions to profit (which may be unobservable), and it is especially natural to believe that there is uncertainty about how these indirect signals relate to actual production, so this seems a particularly worthwhile setting in which to ask questions about robustness.

A Omitted proofs in Section 2

Proof of Theorem 2.2. Fix w, $\widetilde{w}$, and an $\epsilon < \frac{1}{N}$. We begin by introducing two notations. First, let i^* be an element of $\text{argmax}_i(y_i - w_i)$. Second, let $\vec{\epsilon}$ denote the vector in $\mathbb{R}^N$ for which the first $\lfloor N/2 \rfloor$ elements are equal to $-\epsilon$ and the last $\lfloor N/2 \rfloor$ elements are equal to ϵ (if N is odd, the middle element is equal to 0). Without loss of generality, assume that $w_1 \leq w_2 \leq \cdots \leq w_N$. We have $\vec{\epsilon} \cdot w = \epsilon \cdot \text{maxdiff}(w)$.

Now, we will construct environments $\mathcal{E}$ and $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$ such that

$$V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) \geq C^* \sqrt{\epsilon} - O(\epsilon).$$

It suffices to prove this inequality because $D(w, \widetilde{w}, \mathcal{E}, \epsilon) \geq V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}})$.

First, we consider the case of $C^* = 0$. We simply assume that the agent can only produce y_1 with probability 1. More precisely, let $\mathcal{E} = \widetilde{\mathcal{E}} = (c_k = 0 \ \forall k;\ f_k = f \ \forall k)$, where $f(1) = 1$. We have $V(w|\mathcal{E}) = y_1 - w_1 = 0$ and $V(\widetilde{w}|\widetilde{\mathcal{E}}) = y_1 - \widetilde{w}_1 = -\widetilde{w}_1$. The difference is $V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) = \widetilde{w}_1 \geq 0$, so the lower bound trivially holds.

From now on, we assume that $C^* > 0$, which means that $\text{maxdiff}(w) > 0$ and $\max_i(y_i - w_i) > 0$. We also assume that $K = 2$: the argument extends to $K > 2$ by simply defining f_1, f_2, c_1, c_2 as below and then taking $(f_k, c_k) = (f_1, c_1)$ for all $k > 2$.

We first construct $\widetilde{\mathcal{E}}$. Let δ be an arbitrarily small positive real number: we have $\delta > 0$ and $\delta \ll \epsilon$. (The role of δ is to make the incentive constraints non-binding in order to avoid any potential ambiguity on the agent's effort choice.) Let t be a real number, to be chosen later, which will satisfy the restriction

$$0 < t \leq \frac{1 - N\epsilon}{w_{i^*} + \delta}. \tag{A.1}$$

Define $\widetilde{f}_1$ and $\widetilde{f}_2$ as follows:

$$\widetilde{f}_1(i) = \begin{cases} \epsilon + t(w_{i^*} + \delta) & \text{if } i = 1 \\ 1 - (N-1)\epsilon - t(w_{i^*} + \delta) & \text{if } i = i^* \\ \epsilon & \text{otherwise} \end{cases}, \quad \widetilde{f}_2(i) = \begin{cases} 1 - (N-1)\epsilon & \text{if } i = i^* \\ \epsilon & \text{otherwise} \end{cases}.$$

Let $f_1 = \widetilde{f}_1 - \vec{\epsilon}$ and $f_2 = \widetilde{f}_2 + \vec{\epsilon}$. Restriction (A.1) ensures all the probabilities are between

0 and 1. Let $c_1 = 0$ and $c_2 = (f_2 - f_1) \cdot w - \delta$. Define the environments $\mathcal{E}$ and $\widetilde{\mathcal{E}}$ as follows:

$$\mathcal{E} = (c_1, c_2; f_1, f_2), \quad \widetilde{\mathcal{E}} = (c_1, c_2; \widetilde{f}_1, \widetilde{f}_2).$$

The agent chooses effort 2 in environment $\mathcal{E}$ because $f_2 \cdot w - c_2 = f_1 \cdot w - c_1 + \delta > f_1 \cdot w - c_1$. We next identify a t such that the agent chooses effort 1 in environment $\widetilde{\mathcal{E}}$: that is, $\widetilde{f}_1 \cdot \widetilde{w} - c_1 > \widetilde{f}_2 \cdot \widetilde{w} - c_2$. This condition is equivalent to $(\widetilde{f}_2 - \widetilde{f}_1) \cdot \widetilde{w} < c_2 - c_1 = (f_2 - f_1) \cdot w - \delta$, which rearranges into $(\widetilde{f}_2 - \widetilde{f}_1) \cdot (\widetilde{w} - w) < 2 \cdot \vec{\epsilon} \cdot w - \delta$. Plugging in the definition of $\widetilde{f}_k$, we get

$$t(w_{i^*} + \delta)(\widetilde{w}_{i^*} - w_{i^*} - \widetilde{w}_1 + w_1) < 2 \cdot \vec{\epsilon} \cdot w - \delta. \tag{A.2}$$

If condition (A.2) holds, then $\mathbf{k}^*(\widetilde{w}|\widetilde{\mathcal{E}}) = \{1\}$. It follows that

$$\begin{aligned}
V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) &= f_2 \cdot (y - w) - \widetilde{f}_1 \cdot (y - \widetilde{w}) \\
&= \widetilde{f}_2 \cdot (y - w) - \widetilde{f}_1 \cdot (y - \widetilde{w}) + \vec{\epsilon} \cdot (y - w) \\
&= (\widetilde{f}_2 - \widetilde{f}_1) \cdot (y - \widetilde{w}) + \widetilde{f}_2 \cdot (\widetilde{w} - w) + \vec{\epsilon} \cdot (y - w) \\
&= t(w_{i^*} + \delta)(y_{i^*} - y_1 - \widetilde{w}_{i^*} + \widetilde{w}_1) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) \\
&\quad + \epsilon \cdot \sum_i (\widetilde{w}_i - w_i) + \vec{\epsilon} \cdot (y - w) \\
&\geq t(w_{i^*} + \delta)(y_{i^*} - \widetilde{w}_{i^*}) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon).
\end{aligned}$$

(The last step is due to the fact that $y_1 = 0$ and $\widetilde{w}_1 \geq 0$.)

We now turn to the choice of t. There are two cases.

If $(1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*} - \widetilde{w}_1 + w_1) < 2 \cdot \vec{\epsilon} \cdot w - \delta$, then let $t = \frac{1 - N\epsilon}{w_{i^*} + \delta}$, which satisfies (A.1) and (A.2). In this case, the result of the theorem trivially holds because the discrepancy does not vanish as $\epsilon \to 0$. Indeed, our choice of t yields $V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) \geq (1 - N\epsilon)(y_{i^*} - w_{i^*}) - O(\epsilon)$, which converges to a positive constant $(y_{i^*} - w_{i^*})$, so for small enough ϵ it is obviously greater than the lower bound given by $C^* \sqrt{\epsilon}$.

The other case is when $(1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*} - \widetilde{w}_1 + w_1) \geq 2 \cdot \vec{\epsilon} \cdot w - \delta$. In this case, since δ was chosen very small, the right side is positive, so $\widetilde{w}_{i^*} - w_{i^*} - \widetilde{w}_1 + w_1 > 0$, which implies that $\widetilde{w}_{i^*} > w_{i^*}$ (because $w_1 = 0$ and $\widetilde{w}_1 \geq 0$). Then, we can define $t = \frac{2 \cdot \vec{\epsilon} \cdot w - 2\delta}{(w_{i^*} + \delta)(\widetilde{w}_{i^*} - w_{i^*})}$, and this value of t satisfies condition (A.2); it also satisfies (A.1) by assumption of this

case. The discrepancy becomes

$$\begin{aligned} V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) &\geq t(w_{i^*} + \delta)(y_{i^*} - \widetilde{w}_{i^*}) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\ &= \frac{2 \cdot \vec{\epsilon} \cdot w - 2\delta}{\widetilde{w}_{i^*} - w_{i^*}} \cdot (y_{i^*} - \widetilde{w}_{i^*}) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\ &= \frac{2 \cdot \vec{\epsilon} \cdot w - 2\delta}{\widetilde{w}_{i^*} - w_{i^*}} \cdot (y_{i^*} - w_{i^*}) - (2 \cdot \vec{\epsilon} \cdot w - 2\delta) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\ &= \frac{2 \cdot \vec{\epsilon} \cdot w - 2\delta}{\widetilde{w}_{i^*} - w_{i^*}} \cdot (y_{i^*} - w_{i^*}) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\ &\geq 2\sqrt{2(\vec{\epsilon} \cdot w - \delta)(y_{i^*} - w_{i^*})(1 - N\epsilon)} - O(\epsilon). \end{aligned}$$

(The last step is due to the inequality $x + y \geq 2\sqrt{xy}$.) Since $\sqrt{1 - N\epsilon} > 1 - N\epsilon$ and δ is arbitrarily small, we can further reduce the bound to

$$2\sqrt{2(\vec{\epsilon} \cdot w - \delta)(y_{i^*} - w_{i^*})(1 - N\epsilon)} > 2\sqrt{2(\vec{\epsilon} \cdot w)(y_{i^*} - w_{i^*})} \cdot (1 - N\epsilon) = C^* \sqrt{\epsilon} \cdot (1 - N\epsilon).$$

Therefore, we indeed have $V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) \geq C^* \sqrt{\epsilon} \cdot (1 - N\epsilon) - O(\epsilon) = C^* \sqrt{\epsilon} - O(\epsilon)$. □

Proof of Proposition 2.3. We say that efforts k and l are equivalent if both k and l belong to $\mathbf{k}^*(w|\mathcal{E})$, and $f_k \cdot (y - w) = f_l \cdot (y - w) = V(w|\mathcal{E})$. Suppose the agent chooses effort k^* in environment $\mathcal{E}$. We claim that for small enough ϵ, there is a τ on the order of ϵ (though the constant factor can depend on $\mathcal{E}$), such that the agent chooses an effort equivalent to k^*, under contract $w(\tau)$ in any environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$.

To define τ, let S be the set of all effort levels $k \in \mathbf{k}^*(w|\mathcal{E})$ that are *not* equivalent to k^*. For each ϵ, take $\tau > 0$ such that

$$\tau \cdot (\widetilde{f}_{k^*} - \widetilde{f}_k) \cdot (y - w) > 2 \cdot \epsilon \cdot \text{maxdiff}(w) \tag{A.3}$$

for each $k \in S$ and all environments $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$. Note that as long as ϵ is small, we can indeed do this with τ on the order of ϵ, since $(f_{k^*} - f_k) \cdot (y - w) > 0$ for each $k \in S$ (this is the assumption of tie-breaking in favor of the principal) and

$$(\widetilde{f}_{k^*} - \widetilde{f}_k) \cdot (y - w) = (f_{k^*} - f_k) \cdot (y - w) + O(\epsilon).$$

Now we check that this $w(\tau)$ works. For all k, if the agent strictly prefers k^* over k in environment $\mathcal{E}$ under contract w, then he will also do so in environment $\widetilde{\mathcal{E}}$ under contract $w(\tau)$, for small enough ϵ. Indeed, if $f_{k^*} \cdot w - c_{k^*} > f_k \cdot w - c_k$, then for small enough ϵ and τ, we have $\widetilde{f}_{k^*} \cdot w(\tau)$ close to $f_{k^*} \cdot w$ and $\widetilde{f}_k \cdot w(\tau)$ close to $f_k \cdot w$, so the incentive

constraint still holds: $\widetilde{f}_{k^*} \cdot w(\tau) - c_{k^*} > \widetilde{f}_k \cdot w(\tau) - c_k$.

Suppose the agent is indifferent between k and k^*, under w and $\mathcal{E}$. Suppose that k and k^* are not equivalent, i.e. $k \in S$. In any environment $\widetilde{\mathcal{E}}$, the agent prefers k^* over k if $\widetilde{f}_{k^*} \cdot w(\tau) - c_{k^*} > \widetilde{f}_k \cdot w(\tau) - c_k$. This constraint is equivalent to

$$\begin{aligned}
(\widetilde{f}_{k^*} - \widetilde{f}_k) \cdot w(\tau) &> c_{k^*} - c_k = (f_{k^*} - f_k) \cdot w \\
(\widetilde{f}_{k^*} - \widetilde{f}_k) \cdot (w + \tau \cdot (y - w)) &> (f_{k^*} - f_k) \cdot w \\
\tau \cdot (\widetilde{f}_{k^*} - \widetilde{f}_k) \cdot (y - w) &> (f_{k^*} - \widetilde{f}_{k^*} - f_k + \widetilde{f}_k) \cdot w.
\end{aligned}$$

The right-hand side in the last inequality is at most $2 \cdot \epsilon \cdot \text{maxdiff}(w)$. So, by (A.3), the constraint is indeed satisfied, and the agent will not choose effort k.

So, the agent chooses some effort k equivalent to k^*. Now, the profit in the original model is equal to $V(w|\mathcal{E}) = f_{k^*} \cdot (y - w) = f_k \cdot (y - w)$, and the profit in environment $\widetilde{\mathcal{E}}$ is equal to $V(w(\tau)|\widetilde{\mathcal{E}}) = \widetilde{f}_k \cdot (y - w(\tau)) = \widetilde{f}_k \cdot (y - w)(1 - \tau)$. Since τ is on the order of ϵ, the discrepancy is on the order of ϵ. □

To prove Theorem 2.4, we need to define a sequence of constants h_K. Let $h_2 = 0$. For $K > 2$, we define h_K as follows. Let $x_3, x_4, \ldots, x_K$ be positive real numbers that solve the following equations:

$$\frac{1}{x_3} = x_3 + \frac{1}{x_4} = x_4 + \frac{1}{x_5} = \cdots = x_{K-1} + \frac{1}{x_K} = x_K. \tag{A.4}$$

Lemma A.1. *The solution to (A.4) exists and is unique.*

Proof. Consider minimizing the function

$$g(x_3, \ldots, x_K) = \max\left\{\frac{1}{x_3}, x_3 + \frac{1}{x_4}, \ldots, x_{K-1} + \frac{1}{x_K}, x_K\right\}$$

over all $(K-2)$-tuples of positive numbers. The minimum exists, because g is continuous and goes to ∞ uniformly as any x_i goes to 0 or ∞; and it is unique, because g is strictly convex. At the minimum, the values of the $K-1$ components entering the max operator must all be equal: if not, it is easy to see that we could perturb $(x_3, \ldots, x_K)$ slightly to reduce the values of all components attaining the max, thereby reducing the value of g, a contradiction. Therefore, (A.4) is satisfied. □

Let h_K denote the value of x_K in the solution to (A.4) — or, equivalently, the minimum value of g in the proof above. Note that this proof implies that $(1, 1, \ldots, 1)$ is not the

minimum of g, hence $h_K < g(1,1,\ldots,1) = 2$, as claimed in Theorem 2.4.

Lemma A.2. *For any positive real numbers $a_3, a_4, \ldots, a_K$, we have $h_K \geq \min\{\frac{1}{a_3}, a_3 + \frac{1}{a_4}, \ldots, a_{K-1} + \frac{1}{a_K}, a_K\}$. Moreover, equality only holds if $a_k = x_k$ for each k.*

Proof. For the first statement, suppose the desired conclusion is false. We have $\frac{1}{a_3} > h_K = \frac{1}{x_3}$, so $a_3 < x_3$. Next, we have $a_3 + \frac{1}{a_4} > x_3 + \frac{1}{x_4}$. Since $a_3 < x_3$, we have $\frac{1}{a_4} > \frac{1}{x_4}$, so $a_4 < x_4$. Continuing this argument, we get $a_K < x_K$, but $x_K = h_K$, so $a_K < h_K$. However, a_K is in the set, contradiction.

For the second statement, an identical argument shows that $a_3 \leq x_3, a_4 \leq x_4, \ldots$, *and* if any one of these inequalities is strict then so are all subsequent inequalities. But the final inequality $a_K \leq x_K = h_K$ cannot be strict, so all inequalities are equalities. □

Lemma A.3. *We have h_K is strictly increasing in K.*

Proof. Suppose $\frac{1}{x_3} = x_3 + \frac{1}{x_4} = x_4 + \frac{1}{x_5} = \cdots = x_{K-1} + \frac{1}{x_K} = x_K = h_K$. Let $x_{K+1} = x_K$. By Lemma A.2, we have $h_{K+1} \geq \min\{\frac{1}{x_3}, x_3 + \frac{1}{x_4}, \ldots, x_K + \frac{1}{x_{K+1}}, x_{K+1}\}$. Since all terms in this set are at least x_K, we deduce that $h_{K+1} \geq x_K = h_K$. Moreover, equality cannot hold because not all terms in the preceding set are equal. Therefore, we have $h_{K+1} > h_K$. □

Proof of Theorem 2.4 (a). We claim that for all sufficiently small ϵ, for any environment $\mathcal{E}$, there exists a $\widetilde{w}$ such that

$$D(w, \widetilde{w}, \mathcal{E}, \epsilon) \leq h_K \cdot \sqrt{2 \cdot \text{maxdiff}(w) \cdot \max_i (y_i - w_i) \cdot \epsilon} + O(\epsilon). \tag{A.5}$$

Without loss of generality, assume that $f_1 \cdot (y-w) \leq f_2 \cdot (y-w) \leq \cdots \leq f_K \cdot (y-w)$. Let τ^* denote the maximal value of $\tau \leq 1$ for which $w(\tau) = w + \tau(y-w)$ satisfies the limited liability constraint. The value of τ^* is determined by y and w only. Let $L = \max \mathbf{k}^*(w|\mathcal{E})$. For now assume that

$$(f_L - f_{L-1}) \cdot (y-w) > \frac{2 \cdot \text{maxdiff}(w)}{\tau^*} \cdot \epsilon + 2 \cdot \text{maxdiff}(y-w) \cdot \epsilon + \epsilon. \tag{A.6}$$

(At the end of the proof, we will discuss the case when this assumption fails.) Define contracts $\widetilde{w}_l$ as follows. For $l = 2, 3, \ldots, L$, let

$$\tau_l = \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{(f_L - f_{l-1}) \cdot (y-w) - 2 \cdot \text{maxdiff}(y-w) \cdot \epsilon - \epsilon}.$$

Our assumption implies that $\tau_l < \tau^*$, so contracts $w(\tau_l)$ satisfies the limited liability constraint.

For $l = 2, 3, \ldots, L$, let $\widetilde{w}_l = w(\tau_l)$. Let $\widetilde{w}_1$ denote the zero-contract. We claim that one of these L contracts we defined gives us the upper bound. We proceed in four steps as follows. First, we show that contract $\widetilde{w}_l$ induces at least effort l. Second, we provide a bound for the discrepancy given by each of the L contracts. Third, we analyze $\widetilde{w}_1$ and discuss its implications for the upper bound (A.5). Finally we prove that one of these L discrepancies satisfies this upper bound.

Step 1 We first prove that all elements of $\mathbf{k}^*(\widetilde{w}_l|\widetilde{\mathcal{E}})$ are at least l for all $1 \leq l \leq L$ and for all $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$. We show that for all $k < l$, the agent would never choose effort k in any environment $\widetilde{\mathcal{E}}$. It suffices to show that for all $k < l$ we have

$$\widetilde{f}_k \cdot \widetilde{w}_l - c_k < \widetilde{f}_L \cdot \widetilde{w}_l - c_L$$

in all possible environments, because this constraint means that the agent would always prefer effort L to any $k < l$. Since $L \in \mathbf{k}^*(w|\mathcal{E})$, we have $(f_L - f_k) \cdot w \geq c_L - c_k$. We only need to show that $(\widetilde{f}_L - \widetilde{f}_k) \cdot \widetilde{w}_l > (f_L - f_k) \cdot w$, which is equivalent to

$$(\widetilde{f}_L - \widetilde{f}_k) \cdot \tau_l \cdot (y - w) > (f_L - \widetilde{f}_L - f_k + \widetilde{f}_k) \cdot w.$$

The right hand side is at most $2 \cdot \text{maxdiff}(w) \cdot \epsilon$. We are left to show that $\tau_l > \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{(\widetilde{f}_L - \widetilde{f}_k) \cdot (y-w)}$. This inequality holds because the denominator of the right hand side is greater than the denominator of τ_l. Indeed, we have

$$\begin{aligned}
(\widetilde{f}_L - \widetilde{f}_k) \cdot (y - w) &\geq (f_L - f_k) \cdot (y - w) - 2 \cdot \text{maxdiff}(y - w) \cdot \epsilon \\
&\geq (f_L - f_{l-1}) \cdot (y - w) - 2 \cdot \text{maxdiff}(y - w) \cdot \epsilon \\
&> (f_L - f_{l-1}) \cdot (y - w) - 2 \cdot \text{maxdiff}(y - w) \cdot \epsilon - \epsilon.
\end{aligned}$$

Therefore, under contract $\widetilde{w}_l$, the agent always prefers effort L over any $k < l$, so the agent always exerts an effort at least l.

Step 2 We now bound the discrepancy given by each of the L contracts. For $l = 2, \ldots, L$, let $D_l = D(w, \widetilde{w}_l, \mathcal{E}, \epsilon)$. We claim that

$$D_l \leq f_L \cdot (y - w) - f_l \cdot (y - \widetilde{w}_l) + O(\epsilon).$$

Since contract $\widetilde{w}_l$ induces at least effort l in any environment, we need to show that for all $k \geq l$, in any environment $\widetilde{\mathcal{E}}$, we have $\widetilde{f}_k \cdot (y - \widetilde{w}_l) \geq f_l \cdot (y - \widetilde{w}_l) - O(\epsilon)$. We have

$$\begin{aligned}\widetilde{f}_k \cdot (y - \widetilde{w}_l) &= [(\widetilde{f}_k - f_k) + (f_k - f_l)] \cdot (y - \widetilde{w}_l) + f_l \cdot (y - \widetilde{w}_l) \\ &= (\widetilde{f}_k - f_k) \cdot (y - w_l) \cdot (1 - \tau_l) + (f_k - f_l) \cdot (y - w_l) \cdot (1 - \tau_l) + f_l \cdot (y - \widetilde{w}_l).\end{aligned}$$

The first term on the right side is $\geq -O(\epsilon)$, and the second term is ≥ 0, so the needed inequality does indeed hold, and the claim is true for all $l > 1$.

For $l = 2, 3, \ldots, L$, we have

$$\begin{aligned}D_l &\leq f_L \cdot (y - w) - f_l \cdot (y - \widetilde{w}_l) + O(\epsilon) \\ &= f_L \cdot (y - w) - f_l \cdot (y - w) \cdot (1 - \tau_l) + O(\epsilon) \\ &= (f_L - f_l) \cdot (y - w) + \tau_l \cdot f_l \cdot (y - w) + O(\epsilon).\end{aligned}$$

Recall that $\tau_l = \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{(f_L - f_{l-1}) \cdot (y-w) - 2 \cdot \text{maxdiff}(y-w) \cdot \epsilon - \epsilon}$. For $l = 2, 3, \ldots, L-1$, we have

$$D_l \leq \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_{l+1}} + \tau_l \cdot f_l \cdot (y - w) + O(\epsilon).$$

Moreover, we have

$$D_L \leq \tau_L \cdot f_L \cdot (y - w) + O(\epsilon).$$

Finally, since $\widetilde{w}_1$ is the zero-contract, for some k we have

$$D_1 \leq f_L \cdot (y - w) - f_k \cdot y + O(\epsilon).$$

We are left to show that one of $\widetilde{w}_l$ satisfies the upper bound.

Step 3 We analyze D_1. We know that $D_1 \leq f_L \cdot (y - w) - f_k \cdot y + O(\epsilon)$ for some k. We claim that either $D_1 = O(\epsilon)$ or $\tau_{k+1} \cdot f_{k+1} \cdot (y - w) = O(\epsilon)$. If $k \geq L$, then $D_1 \leq O(\epsilon)$. Assume that $k < L$. To prove our claim, we observe that

$$\begin{aligned}\tau_{k+1} \cdot f_{k+1} \cdot (y - w) &= \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon \cdot f_{k+1} \cdot (y - w)}{(f_L - f_k) \cdot (y - w) - 2 \cdot \text{maxdiff}(y - w) \cdot \epsilon - \epsilon} \\ &\leq \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon \cdot f_L \cdot (y - w)}{(f_L - f_k) \cdot (y - w) - 2 \cdot \text{maxdiff}(y - w) \cdot \epsilon - \epsilon}.\end{aligned}$$

(The denominator is positive, due to (A.6).)

Now, for $d > 1$ that is close enough to 1, we know that the vector $(1-d) \cdot y + d \cdot w$ has all nonnegative components, because $w_i > 0$ for all $i > 1$. Fix some such d, independent of ϵ or $\mathcal{E}$.

For any ϵ and $\mathcal{E}$, one of the following two cases holds:

- Case (i): $f_k \cdot (y-w) < \frac{1}{d} \cdot f_L \cdot (y-w) - 2 \cdot \epsilon \cdot \text{maxdiff}(y-w) - \epsilon$. Then the denominator on the right side above is greater than $\left(1 - \frac{1}{d}\right) \cdot f_L \cdot (y-w)$, and therefore

$$\tau_{k+1} \cdot f_{k+1} \cdot (y-w) < \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{1 - \frac{1}{d}} = O(\epsilon).$$

- Case (ii): $f_k \cdot (y-w) \geq \frac{1}{d} \cdot f_L \cdot (y-w) - 2 \cdot \epsilon \cdot \text{maxdiff}(y-w) - \epsilon$. This implies that

$$\begin{aligned} D_1 &\leq f_L \cdot (y-w) - f_k \cdot y + O(\epsilon) \\ &\leq d \cdot (f_k \cdot (y-w) + 2 \cdot \epsilon \cdot \text{maxdiff}(y-w) + \epsilon) - f_k \cdot y + O(\epsilon) \\ &= d \cdot f_k \cdot (y-w) - f_k \cdot y + O(\epsilon) \\ &= -f_k \cdot ((1-d) \cdot y + d \cdot w) + O(\epsilon). \end{aligned}$$

  By construction of d, the first term on the right is ≤ 0. Hence, we have $D_1 = O(\epsilon)$ in this case.

Step 4 We are finally ready to prove the upper bound. We claim that there exists an $l \in \{1, 2, \ldots, L\}$ such that

$$D_l \leq h_K \cdot \sqrt{2 \cdot \text{maxdiff}(w) \cdot \max_i (y_i - w_i) \cdot \epsilon} + O(\epsilon).$$

If either the zero-contract induces effort at least L or case (ii) of step 3 holds, then we have shown that $D_1 \leq O(\epsilon)$ and we are done. Otherwise, case (i) of step 3 holds, and we know that $\tau_{k+1} \cdot f_{k+1} \cdot (y-w) = O(\epsilon)$, where k is the effort level identified in Step 3.

Apply the bounds from Step 2, we find that

$$\begin{aligned}
D_{k+1} &\leq \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_{k+2}} + O(\epsilon) \\
D_{k+2} &\leq \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_{k+3}} + \tau_{k+2} \cdot f_{k+2} \cdot (y - w) + O(\epsilon) \\
&\leq \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_{k+3}} + \tau_{k+2} \cdot \max_i(y_i - w_i) + O(\epsilon) \\
D_{k+3} &\leq \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_{k+4}} + \tau_{k+3} \cdot \max_i(y_i - w_i) + O(\epsilon) \\
D_{k+4} &\leq \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_{k+5}} + \tau_{k+4} \cdot \max_i(y_i - w_i) + O(\epsilon) \\
&\cdots \\
D_{L-1} &\leq \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_L} + \tau_{L-1} \cdot \max_i(y_i - w_i) + O(\epsilon) \\
D_L &\leq \tau_L \cdot \max_i(y_i - w_i) + O(\epsilon).
\end{aligned}$$

(These equations assume that $k+1 < L$. If $k+1 = L$, then we simply have $D_{k+1} = O(\epsilon)$. For example, if $K = L = 2$ and $k = 1$, then $D_2 = O(\epsilon)$, which is consistent with the desired upper bound as $h_2 = 0$.)

We now ignore the $O(\epsilon)$ terms and compare the other terms in $D_{k+1}, D_{k+2}, \ldots, D_L$. Applying Lemma A.2 with $a_l = \sqrt{\max_i(y_i - w_i)/2 \cdot \text{maxdiff}(w) \cdot \epsilon} \cdot \tau_{k-1+l}$, and then multiplying through by $\sqrt{2 \cdot \text{maxdiff}(w) \cdot \max_i(y_i - w_i) \cdot \epsilon}$, and applying Lemma A.3, we see that

$$\begin{aligned}
&\min\{\frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_{k+2}}, \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_{k+3}} + \tau_{k+2} \cdot \max_i(y_i - w_i), \ldots, \\
&\qquad \frac{2 \cdot \text{maxdiff}(w) \cdot \epsilon}{\tau_L} + \tau_{L-1} \cdot \max_i(y_i - w_i), \tau_L \cdot \max_i(y_i - w_i)\} \\
&\leq h_{L-k+1} \cdot \sqrt{2 \cdot \text{maxdiff}(w) \cdot \max_i(y_i - w_i) \cdot \epsilon} \leq h_K \cdot \sqrt{2 \cdot \text{maxdiff}(w) \cdot \max_i(y_i - w_i) \cdot \epsilon}.
\end{aligned}$$

Therefore, there exists an l such that

$$D_l \leq h_K \cdot \sqrt{2 \cdot \text{maxdiff}(w) \cdot \max_i(y_i - w_i) \cdot \epsilon} + O(\epsilon).$$

We have thus established the upper bound.

Final case Throughout this proof, we have maintained assumption (A.6). What if this assumption fails? We have two cases to consider. The first case is when $(f_L - f_1) \cdot (y - w) \leq$

$\frac{2\cdot\text{maxdiff}(w)}{\tau^*}\cdot\epsilon + 2\cdot\text{maxdiff}(y-w)\cdot\epsilon+\epsilon$. In this case, we have $(f_L - f_1)\cdot(y-w) = O(\epsilon)$, so the discrepancy is bounded by $O(\epsilon)$.

The second case to consider is the complementary case. We ignore all effort levels l for which $(f_L - f_l)\cdot(y-w) \le \frac{2\cdot\text{maxdiff}(w)}{\tau^*}\cdot\epsilon + 2\cdot\text{maxdiff}(y-w)\cdot\epsilon+\epsilon$. Suppose we have ignored n such effort levels, $l = L-n,\ldots,L-1$. We then define contracts $\widetilde{w}_1,\widetilde{w}_2,\ldots,\widetilde{w}_{L-n}$ as in this proof; we can check that each such contract is $w(\tau_l)$ with $0<\tau_l<\tau^*$. We then repeat the proof almost exactly as before. At the end of step 2, instead of showing an upper bound on D_L, we show an upper bound on D_{L-n}, namely

$$D_{L-n} \le \tau_{L-n}\cdot f_{L-n}\cdot(y-w) + O(\epsilon).$$

This follows in the same way as the bounds for the other D_l once we notice that $(f_L - f_{L-n})\cdot(y-w) = O(\epsilon)$ by assumption. At the start of step 3, we split into cases $k \ge L-n$ and $k < L-n$, instead of $k \ge L$ and $k < L$. Similarly at the start of step 4. Then we apply the procedure of step 4 to upper bounds on $D_{k+1}, D_{k+2},\ldots,D_{L-n}$ to conclude that the maximum discrepancy is at most $h_{L-n-k+1}\cdot\sqrt{2\cdot\text{maxdiff}(w)\cdot\max_i(y_i-w_i)\cdot\epsilon}$. From Lemma A.3, we know $h_{L-n-k+1} < h_K$, so the upper bound still holds. □

To prove Theorem 2.4 (b), we need another lemma on h_K.

Lemma A.4. *Suppose* $\frac{1}{x_3} = x_3 + \frac{1}{x_4} = x_4 + \frac{1}{x_5} = \cdots = x_{K-1} + \frac{1}{x_K} = x_K$. *Then we have* $x_3 < x_4 < \cdots < x_K$.

Proof. Since $x_K = x_{K-1} + \frac{1}{x_K}$, we have $x_K > x_{K-1}$. Next we see that $x_{K-1} + \frac{1}{x_K} = x_{K-2} + \frac{1}{x_{K-1}}$. Since $x_K > x_{K-1}$, we have $\frac{1}{x_K} < \frac{1}{x_{K-1}}$, and it follows that $x_{K-1} > x_{K-2}$. Continuing this argument, we get $x_K > x_{K-1} > \cdots > x_4 > x_3$. □

Proof of Theorem 2.4 (b). We claim that for all sufficiently small ϵ, there exists an environment $\mathcal{E}$, such that the following inequality holds for all contracts $\widetilde{w}$:

$$D(w,\widetilde{w},\mathcal{E},\epsilon) \ge h_K\cdot\sqrt{2\cdot\max_i(w_i\cdot(y_i-w_i))\cdot\epsilon} - O(\epsilon). \tag{A.7}$$

First off, if $\max_i(w_i\cdot(y_i-w_i)) = 0$, then construct $\mathcal{E}$ such that the agent can only produce y_1 with probability 1. For all k, we have $f_k(1) = 1$ and $f_k(i) = 0$ for all $i > 1$. In this case, we have $V(w|\mathcal{E}) = 0$ and $V(\widetilde{w}|\widetilde{\mathcal{E}}) = O(\epsilon)$ for all $\widetilde{w}$ and $\widetilde{\mathcal{E}}$. Thus the lower bound trivially holds.

Suppose that $\max_i(w_i\cdot(y_i-w_i)) > 0$. Let i^* denote an index in $\text{argmax}_i(w_i(y_i-w_i))$. We now construct positive numbers $\tau_3,\tau_4,\ldots,\tau_K$ to make the argument in Step 4 in the

proof of part (a) tight. Specifically, let x_k be as given by the solution to (A.4), and put $\tau_k = x_K \cdot \sqrt{2w_{i^*}\epsilon/(y_{i^*} - w_{i^*})}$. Then we have

$$\frac{2w_{i^*} \cdot \epsilon}{\tau_3} = \tau_3 \cdot (y_{i^*} - w_{i^*}) + \frac{2w_{i^*} \cdot \epsilon}{\tau_4} = \cdots = \tau_{K-1} \cdot (y_{i^*} - w_{i^*}) + \frac{2w_{i^*} \cdot \epsilon}{\tau_K} = \tau_K \cdot (y_{i^*} - w_{i^*})$$
$$= h_K \cdot \sqrt{2w_{i^*}(y_{i^*} - w_{i^*}) \cdot \epsilon}.$$

Each quantity in the above equation represents a lower bound for the discrepancy given by a contract that induces certain effort levels. We show that no matter what effort levels a contract induces, the discrepancy must be (asymptotically) bounded by $h_K \cdot \sqrt{2w_{i^*}(y_{i^*} - w_{i^*}) \cdot \epsilon}$.

In our construction, we focus on output levels 1 and i^* — effectively assuming that there are just two output levels. We construct an environment $\mathcal{E}$ as follows.

$$f_1(i) = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases}, \qquad f_K(i) = \begin{cases} 1 & \text{if } i = i^* \\ 0 & \text{otherwise} \end{cases}.$$

For $k = 2, \ldots, K-1$, let

$$f_k(i) = \begin{cases} \frac{2w_{i^*} \cdot \epsilon}{\tau_{k+1}(y_{i^*} - w_{i^*})} & \text{if } i = 1 \\ 1 - \frac{2w_{i^*} \cdot \epsilon}{\tau_{k+1}(y_{i^*} - w_{i^*})} & \text{if } i = i^* \\ 0 & \text{otherwise} \end{cases}.$$

(This construction of f_k gives a valid probability distribution for small enough ϵ. Indeed, since τ_{k+1} is on the order of $\sqrt{\epsilon}$ by definition, we know that $f_k(1)$ is on the order of $\sqrt{\epsilon}$, so for small enough ϵ, all components of f_k are between 0 and 1.)

By Lemma A.4, we know that $\tau_3 < \cdots < \tau_K$, so $f_1(i^*) < f_2(i^*) < \cdots < f_K(i^*)$. Next, we let $c_1 = 0$, and for all $k > 1$, let $c_k = (f_k - f_1) \cdot w$. The agent is indifferent between all effort levels, so he picks effort K under contract w. Therefore, we obtain that $V(w|\mathcal{E}) = y_{i^*} - w_{i^*}$.

Let $\widetilde{w}$ denote any contract. Let k^* denote the minimal effort the agent exerts in any environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$ under contract $\widetilde{w}$. In other words, the agent exerts at least effort k^* in every environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$, and there exists an environment in which the agent exerts k^*.

If $k^* = 1$, then there exists an $\widetilde{\mathcal{E}}$ in which the agent chooses effort 1 under contract $\widetilde{w}$. Consequently, we have $D(w, \widetilde{w}, \mathcal{E}, \epsilon) \geq (y_{i^*} - w_{i^*}) - O(\epsilon)$, which is greater than the

desired lower bound (A.7) for small enough ϵ (since the latter goes to 0). Thus, for small ϵ, we may assume $k^* > 1$. In particular, in the original environment $\mathcal{E}$, the agent chooses an effort l greater than 1. Hence, we deduce that for some $l > 1$ we have

$$(f_l - f_1) \cdot \widetilde{w} \geq c_l - c_1.$$

The left hand side is equal to $(f_l(i^*) - f_1(i^*)) \cdot (\widetilde{w}_{i^*} - \widetilde{w}_1)$, and the right hand side is equal to $(f_l - f_1) \cdot w = (f_l(i^*) - f_1(i^*)) \cdot w_{i^*}$. Since $f_l(i^*) > f_1(i^*)$, we obtain that $\widetilde{w}_{i^*} - \widetilde{w}_1 \geq w_{i^*}$. Since $\widetilde{w}_1 \geq 0$, we have $\widetilde{w}_{i^*} \geq w_{i^*}$.

From now on, we assume that $k^* > 1$ and $\widetilde{w}_{i^*} \geq w_{i^*}$. We claim that

$$D(w, \widetilde{w}, \mathcal{E}, \epsilon) \geq (1 - f_{k^*}(i^*)) \cdot (y_{i^*} - w_{i^*}) + (f_{k^*}(i^*) - \epsilon) \cdot (\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \qquad \text{(A.8)}$$

To see why this inequality holds, recall that there is an environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$ for which $V(\widetilde{w}|\widetilde{\mathcal{E}}) = \widetilde{f}_{k^*} \cdot (y - \widetilde{w})$. We deduce that

$$\begin{aligned}
D(w, \widetilde{w}, \mathcal{E}, \epsilon) &\geq (y_{i^*} - w_{i^*}) - \widetilde{f}_{k^*} \cdot (y - \widetilde{w}) \\
&= (y_{i^*} - w_{i^*}) - \widetilde{f}_{k^*} \cdot y + \widetilde{f}_{k^*} \cdot \widetilde{w} \\
&\geq (y_{i^*} - w_{i^*}) - f_{k^*} \cdot y + \widetilde{f}_{k^*}(i^*) \cdot \widetilde{w}_{i^*} - O(\epsilon) \\
&= (y_{i^*} - w_{i^*}) - f_{k^*}(i^*) \cdot y_{i^*} + \widetilde{f}_{k^*}(i^*) \cdot \widetilde{w}_{i^*} - O(\epsilon) \\
&= (1 - f_{k^*}(i^*)) \cdot (y_{i^*} - w_{i^*}) + \widetilde{f}_{k^*}(i^*) \cdot (\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\
&\geq (1 - f_{k^*}(i^*)) \cdot (y_{i^*} - w_{i^*}) + (f_{k^*}(i^*) - \epsilon) \cdot (\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon).
\end{aligned}$$

Therefore, we have established inequality (A.8). Since $f_{k^*}(i^*)$ converges to 1 as $\epsilon \to 0$ for all $k^* > 2$, we can assume that $f_{k^*}(i^*) - \epsilon > 0$. If $k^* = 2$, then we have

$$\begin{aligned}
D(w, \widetilde{w}, \mathcal{E}, \epsilon) &\geq (1 - f_2(i^*)) \cdot (y_{i^*} - w_{i^*}) + (f_2(i^*) - \epsilon) \cdot (\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\
&\geq (1 - f_2(i^*)) \cdot (y_{i^*} - w_{i^*}) - O(\epsilon) \\
&= \frac{2w_{i^*} \cdot \epsilon}{\tau_3} - O(\epsilon) \\
&= h_K \cdot \sqrt{2w_{i^*}(y_{i^*} - w_{i^*}) \cdot \epsilon} - O(\epsilon).
\end{aligned}$$

Therefore the desired lower bound is valid for $k^* = 2$. If $k^* > 2$, then we need to bound the term $\widetilde{w}_{i^*} - w_{i^*}$. We know that the agent exerts effort at least k^* in every environment,

so consider the following modifications to $\mathcal{E}$: Let

$$\widetilde{f}_{k^*-1}(i) = \begin{cases} f_{k^*-1}(1) - \epsilon & \text{if } i = 1 \\ f_{k^*-1}(i^*) + \epsilon & \text{if } i = i^* \\ 0 & \text{otherwise} \end{cases}.$$

(This probability distribution is well-defined for small enough ϵ because $f_{k^*-1}(1)$ is on the order of $\sqrt{\epsilon}$. For small enough ϵ we have $f_{k^*-1}(1) - \epsilon > 0$.) Moreover, for all $k \geq k^*$, let

$$\widetilde{f}_k(i) = \begin{cases} f_k(1) + \epsilon & \text{if } i = 1 \\ f_k(i^*) - \epsilon & \text{if } i = i^* \\ 0 & \text{otherwise} \end{cases}.$$

(These probability distributions are well-defined because $f_k(i^*)$ converges to 1, so for small enough ϵ we have $f_k(i^*) - \epsilon > 0$.) And for $k < k^* - 1$, let $\widetilde{f}_k = f_k$.

Suppose the agent exerts effort l in the above environment. We know that $l > k^* - 1$, so we have $(\widetilde{f}_l - \widetilde{f}_{k^*-1}) \cdot \widetilde{w} \geq c_l - c_{k^*-1} = (f_l - f_{k^*-1}) \cdot w$. This inequality is equivalent to

$$\begin{aligned} (f_l(i^*) - f_{k^*-1}(i^*) - 2\epsilon) \cdot (\widetilde{w}_{i^*} - \widetilde{w}_1) &\geq (f_l(i^*) - f_{k^*-1}(i^*)) \cdot w_{i^*} \\ \widetilde{w}_{i^*} - \widetilde{w}_1 - w_{i^*} &\geq \frac{2w_{i^*} \cdot \epsilon}{f_l(i^*) - f_{k^*-1}(i^*) - 2\epsilon} \\ \widetilde{w}_{i^*} - \widetilde{w}_1 - w_{i^*} &> \frac{2w_{i^*} \cdot \epsilon}{1 - f_{k^*-1}(i^*)} \\ \widetilde{w}_{i^*} - \widetilde{w}_1 - w_{i^*} &> \tau_{k^*} \cdot (y_{i^*} - w_{i^*}). \end{aligned}$$

We now finish the proof as follows. Inequality (A.8) tells us that

$$\begin{aligned} D(w, \widetilde{w}, \mathcal{E}, \epsilon) &\geq (1 - f_{k^*}(i^*)) \cdot (y_{i^*} - w_{i^*}) + (f_{k^*}(i^*) - \epsilon) \cdot (\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\ &\geq (1 - f_{k^*}(i^*)) \cdot (y_{i^*} - w_{i^*}) + (f_{k^*}(i^*) - \epsilon) \cdot \tau_{k^*} \cdot (y_{i^*} - w_{i^*}) - O(\epsilon) \\ &= (1 - f_{k^*}(i^*)) \cdot (y_{i^*} - w_{i^*}) + \tau_{k^*} \cdot (y_{i^*} - w_{i^*}) \\ &\qquad -(1 - f_{k^*}(i^*) + \epsilon) \cdot \tau_{k^*} \cdot (y_{i^*} - w_{i^*}) - O(\epsilon). \end{aligned}$$

Let's analyze the third term in the above expression. We know that τ_{k^*} is on the order of $\sqrt{\epsilon}$. We also know at $(1 - f_{k^*}(i^*) + \epsilon)$ is at most on the order of $\sqrt{\epsilon}$: if $k^* = K$, then $1 - f_{k^*}(i^*) + \epsilon = \epsilon$; if $k^* < K$, then $1 - f_{k^*}(i^*) + \epsilon$ is on the order of $\sqrt{\epsilon}$. Therefore the term $(1 - f_{k^*}(i^*) + \epsilon) \cdot \tau_{k^*} \cdot (y_{i^*} - w_{i^*})$ is bounded by $O(\epsilon)$. We can drop this term and

further reduce the bound to

$$D(w, \widetilde{w}, \mathcal{E}, \epsilon) \geq (1 - f_{k^*}(i^*)) \cdot (y_{i^*} - w_{i^*}) + \tau_{k^*} \cdot (y_{i^*} - w_{i^*}) - O(\epsilon).$$

If $k^* = K$, then the first term is 0; the second term is equal to $h_K \cdot \sqrt{2w_{i^*}(y_{i^*} - w_{i^*}) \cdot \epsilon}$. If $k^* < K$, then the first term is equal to $\frac{2w_{i^*} \cdot \epsilon}{\tau_{k^*+1}}$, so the sum of the first two terms (by the definition of τ_{k^*}) is equal to $h_K \cdot \sqrt{2w_{i^*}(y_{i^*} - w_{i^*}) \cdot \epsilon}$. Therefore, the lower bound (A.7) holds for all possible values of k^*. □

B Omitted proofs in Section 3

First, a note on notation: In Section 2, the notation $O(g(\epsilon))$ was used for a quantity bounded by $C \cdot g(\epsilon)$ where the constant C could depend on (K, N, w, y); we now allow C to depend on u as well.

Now, we prove Theorem 3.1 (a) by considering contracts $w(\tau)$ of the following form:

$$u(w(\tau)_i) = u(w_i) + \tau \cdot (y_i - w_i) + \epsilon \cdot (\text{maxdiff}(u(w)) + \text{maxdiff}(y - w)).$$

We use this notation for convenience. We first show that as long as $\tau \in (0, 1)$, the contract $w(\tau)$ satisfies the participation constraint in all possible environments.

Lemma B.1. *Fix $\mathcal{E}$. If w satisfies the participation constraint for $\mathcal{E}$, then for all $\tau \in (0, 1)$, the contract $w(\tau)$ satisfies the participation constraint in every environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$.*

Proof. Suppose w induces effort k^* in environment $\mathcal{E}$. We have $f_{k^*} \cdot u(w) - c_{k^*} \geq \bar{U}$. We show that for every environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$, the agent's maximal expected utility $\max_k \widetilde{f}_k \cdot u(w(\tau)) - c_k$ is at least $\bar{U}$. It suffices to prove that $\widetilde{f}_{k^*} \cdot u(w(\tau)) \geq f_{k^*} \cdot u(w)$, because under this condition the agent gets at least $\bar{U}$ by exerting effort k^*. We observe that

$$\begin{aligned}
\widetilde{f}_{k^*} \cdot u(w(\tau)) &= \widetilde{f}_{k^*} \cdot u(w) + \tau \cdot \widetilde{f}_{k^*} \cdot (y - w) + \epsilon \cdot (\text{maxdiff}(u(w)) + \text{maxdiff}(y - w)) \\
&= [\widetilde{f}_{k^*} \cdot u(w) + \epsilon \cdot \text{maxdiff}(u(w))] + \tau \cdot \widetilde{f}_{k^*} \cdot (y - w) + \epsilon \cdot \text{maxdiff}(y - w) \\
&\geq f_{k^*} \cdot u(w) + \tau \cdot \widetilde{f}_{k^*} \cdot (y - w) + \epsilon \cdot \text{maxdiff}(y - w).
\end{aligned}$$

We know that $f_{k^*} \cdot (y - w) = V(w|\mathcal{E}) \geq \min_i(y_i - w_i) = 0$. Since $0 < \tau < 1$, we deduce

that

$$\tau \cdot \widetilde{f}_{k^*} \cdot (y - w) \geq \tau \cdot (\widetilde{f}_{k^*} - f_{k^*}) \cdot (y - w) \geq -\epsilon \cdot \text{maxdiff}(y - w).$$

Therefore, we indeed have $\widetilde{f}_{k^*} \cdot u(w(\tau)) \geq f_{k^*} \cdot u(w)$, and the participation constraint holds in every environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$ under contract $w(\tau)$. □

For the proof of Theorem 3.1 (a), we have already established the result on the participation constraint. We only need to choose the appropriate τ that gives the upper bound.

Proof of Theorem 3.1 (a). We will assume that w satisfies the participation constraint in $\mathcal{E}$. At the end, we will address the possibility that this participation constraint may not be satisfied.

Fix $C > C^*$. We claim that there exists a $\tau > 0$ as in the theorem statement, such that for all $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$ the following inequality holds for sufficiently small ϵ:

$$V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) \leq \sqrt{C \cdot C^*} \cdot \sqrt{\epsilon} + O(\epsilon).$$

We first explicitly write out $V(w|\mathcal{E})$ and $V(w(\tau)|\widetilde{\mathcal{E}})$. Suppose contract w induces effort k in environment $\mathcal{E}$. We have $V(w|\mathcal{E}) = f_k \cdot (y - w)$. Suppose $w(\tau)$ induces effort l in some environment $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$. Then $V(w(\tau)|\widetilde{\mathcal{E}}) = \widetilde{f}_l \cdot (y - w(\tau))$. We have

$$\begin{aligned} V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) &= f_k \cdot (y - w) - \widetilde{f}_l \cdot (y - w(\tau)) \\ &= (\widetilde{f}_k - \widetilde{f}_l) \cdot (y - w) + \widetilde{f}_l \cdot (w(\tau) - w) + O(\epsilon) \\ &\leq (\widetilde{f}_k - \widetilde{f}_l) \cdot (y - w) + \max_i (w(\tau)_i - w_i) + O(\epsilon). \end{aligned}$$

We claim that the first term is bounded by $\frac{2 \cdot \text{maxdiff}(u(w)) \cdot \epsilon}{\tau}$. We consider the incentive constraints in environments $\mathcal{E}$ and $\widetilde{\mathcal{E}}$. In environment $\mathcal{E}$, the agent prefers effort k over l, so we have

$$f_k \cdot u(w) - c_k \geq f_l \cdot u(w) - c_l. \tag{B.1}$$

In environment $\widetilde{\mathcal{E}}$, the agent prefers effort l over k, so we have

$$\widetilde{f}_l \cdot u(w(\tau)) - c_l \geq \widetilde{f}_k \cdot u(w(\tau)) - c_k. \tag{B.2}$$

Summing up inequalities (B.1) and (B.2), we obtain that

$$f_k \cdot u(w) + \widetilde{f}_l \cdot u(w(\tau)) \geq f_l \cdot u(w) + \widetilde{f}_k \cdot u(w(\tau)).$$

The above inequality rearranges into $(\widetilde{f}_k - \widetilde{f}_l) \cdot (u(w(\tau)) - u(w)) \leq (f_k - \widetilde{f}_k - f_l + \widetilde{f}_l) \cdot u(w)$. The vector $(f_k - \widetilde{f}_k - f_l + \widetilde{f}_l)$ has components that sum to 0, and all components lie within $[-2\epsilon, 2\epsilon]$. Thus $(f_k - \widetilde{f}_k - f_l + \widetilde{f}_l) \cdot u(w) \leq 2\epsilon \cdot \text{maxdiff}(u(w))$. As for the left-hand side, recall that

$$u(w(\tau)_i) - u(w_i) = \tau \cdot (y_i - w_i) + \epsilon \cdot (\text{maxdiff}(u(w)) + \text{maxdiff}(y - w)).$$

We know that $\sum_{i=1}^N \widetilde{f}_k(i) = \sum_{i=1}^N \widetilde{f}_l(i) = 1$, so we obtain that

$$\sum_{i=1}^N (\widetilde{f}_k(i) - \widetilde{f}_l(i)) \cdot \epsilon \cdot (\text{maxdiff}(u(w)) + \text{maxdiff}(y - w)) = 0.$$

As a result, we have

$$\tau \cdot (\widetilde{f}_k - \widetilde{f}_l) \cdot (y - w) = (\widetilde{f}_k - \widetilde{f}_l) \cdot (u(w(\tau)) - u(w)) \leq 2\epsilon \cdot \text{maxdiff}(u(w)).$$

It follows that

$$V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) \leq \frac{2\epsilon}{\tau} \cdot \text{maxdiff}(u(w)) + \max_i (w(\tau)_i - w_i) + O(\epsilon).$$

We now need to bound $\max_i (w(\tau)_i - w_i)$. Let i^* denote an index in $\text{argmax}_i (w(\tau)_i - w_i)$. If $C^* > 0$, let $\tau = \left(\max_i \frac{y_i - w_i}{u'(w_i)}\right)^{-1/2} \cdot \sqrt{\frac{C^*}{C} \cdot 2 \cdot \text{maxdiff}(u(w)) \cdot \epsilon}$. For small enough ϵ, we have

$$\begin{aligned}
w(\tau)_{i^*} - w_{i^*} &< \frac{C}{C^*} \cdot \frac{u(w(\tau)_{i^*}) - u(w_{i^*})}{u'(w_{i^*})} \\
&= \tau \cdot \frac{C}{C^*} \cdot \frac{y_{i^*} - w_{i^*} + \epsilon \cdot (\text{maxdiff}(u(w)) + \text{maxdiff}(y - w))}{u'(w_{i^*})} \\
&= \tau \cdot \frac{C}{C^*} \cdot \frac{y_{i^*} - w_{i^*}}{u'(w_{i^*})} + O(\epsilon) \\
&\leq \tau \cdot \frac{C}{C^*} \cdot \max_i \frac{y_i - w_i}{u'(w_i)} + O(\epsilon).
\end{aligned}$$

We conclude that

$$V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) \leq \frac{2\epsilon}{\tau} \cdot \text{maxdiff}(u(w)) + \tau \cdot \frac{C}{C^*} \cdot \max_i \frac{y_i - w_i}{u'(w_i)} + O(\epsilon) = \sqrt{C \cdot C^*} \cdot \sqrt{\epsilon} + O(\epsilon).$$

If $C^* = 0$, then either $\text{maxdiff}(u(w)) = 0$ or $\max_i (y_i - w_i) = 0$. If $\text{maxdiff}(u(w)) = 0$,

set $\tau = \epsilon$. If $\max_i(y_i - w_i) = 0$, set $\tau = 0.5$. In either case, for small ϵ, we have

$$\begin{aligned} w(\tau)_{i^*} - w_{i^*} &< 2 \cdot \frac{u(w(\tau)_{i^*}) - u(w_{i^*})}{u'(w_{i^*})} \\ &= 2\tau \cdot \frac{y_{i^*} - w_{i^*} + \epsilon \cdot (\text{maxdiff}(u(w)) + \text{maxdiff}(y - w))}{u'(w_{i^*})} \\ &= 2\tau \cdot \frac{y_{i^*} - w_{i^*}}{u'(w_{i^*})} + O(\epsilon) \\ &\leq 2\tau \cdot \max_i \frac{y_i - w_i}{u'(w_i)} + O(\epsilon). \end{aligned}$$

We conclude that

$$V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) \leq \frac{2\epsilon}{\tau} \cdot \text{maxdiff}(u(w)) + 2\tau \cdot \max_i \frac{y_i - w_i}{u'(w_i)} + O(\epsilon).$$

If $\text{maxdiff}(u(w)) = 0$, then $\tau = \epsilon$, so $2\tau \cdot \max_i \frac{y_i - w_i}{u'(w_i)} = O(\epsilon)$. If $\max_i(y_i - w_i) = 0$, then $\tau = 0.5$, so $\frac{2\epsilon}{\tau} \cdot \text{maxdiff}(u(w)) = O(\epsilon)$. In either case, the discrepancy is bounded by $O(\epsilon)$. And in every case, the value of τ depends on ϵ, w, y, u but not on the environment $\mathcal{E}$, as promised.

Finally, all of the above analysis assumed that w satisfied the participation constraint in $\mathcal{E}$ (and hence $w(\tau)$ did so in $\widetilde{\mathcal{E}}$, by Lemma B.1). What if w violates the participation constraint in $\mathcal{E}$? We have $w(\tau)$ still defined as above. If $w(\tau)$ also violates the participation constraint in $\widetilde{\mathcal{E}}$, then $V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) = V_0 - V_0 = 0$. This leaves only the possibility that w violates the participation constraint in $\mathcal{E}$ but $w(\tau)$ satisfies it in $\widetilde{\mathcal{E}}$.

Suppose this occurs. Our computations above showed that $\max_i(w(\tau)_i - w_i)$ is bounded above by a quantity that tends to 0 as $\epsilon \to 0$. Consequently, the assumption $V_0 < 0 = \min_i(y_i - w_i)$ implies $V_0 < \min_i(y_i - w(\tau)_i)$ when ϵ is small enough. Hence, for small ϵ, the principal's ex-post payoff under $w(\tau)$ for *any* output realization is higher than V_0. So we get $V(w|\mathcal{E}) - V(w(\tau)|\widetilde{\mathcal{E}}) \leq V_0 - V_0 = 0$, which again satisfies the desired upper bound. □

For the proof of Theorem 3.1 (b), we assume that $y_1 - w_1 = \min_i(y_i - w_i) = 0$. We also redefine two notations. First, let i^* denote an element of $\text{argmax}_i \frac{y_i - w_i}{u'(w_i)}$. Second, let $\vec{\epsilon}$ be the vector in $\mathbb{R}^N$ whose ith component is $-\epsilon$ if w_i is one of the $\lfloor N/2 \rfloor$ lowest components of w, and is ϵ if w_i is one of the $\lfloor N/2 \rfloor$ highest components (and is 0 if N is odd and w_i is the middle component); thus $\vec{\epsilon} \cdot u(w) = \epsilon \cdot \text{maxdiff}(u(w))$.

Proof of Theorem 3.1 (b). Fix w, $\widetilde{w}$, and $C < C^*$. Assume that ϵ is small enough so that

$C < (1-N\epsilon)\cdot C^*$. To prove the lower bound, we construct environments $\mathcal{E}$ and $\widetilde{\mathcal{E}} \in B_\epsilon(\mathcal{E})$ such that

$$V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) \geq \sqrt{C \cdot C^*} \cdot \sqrt{\epsilon} - O(\epsilon).$$

First, we show that we can assume $w_i \leq \widetilde{w}_i \leq w_i + 1$ for all i. Suppose this condition fails for some i. Let $\mathcal{E} = \widetilde{\mathcal{E}} = (c_k = 0 \ \forall k;\ f_k = f \ \forall k; \bar{U} = w_i)$, where $f(i) = 1$. The participation constraint binds for w and $\mathcal{E}$, so we get $V(w|\mathcal{E}) = y_i - w_i$. If $w_i > \widetilde{w}_i$, then the participation constraint fails for $\widetilde{w}$ and $\widetilde{\mathcal{E}}$, so $v(\widetilde{w}|\widetilde{\mathcal{E}}) = V_0$, and the discrepancy is a positive constant. If $\widetilde{w}_i > w_i + 1$, then $V(w|\mathcal{E}) = y_i - \widetilde{w}_i$, so the discrepancy is $\widetilde{w}_i - w_i$, which is greater than 1. In both cases, the discrepancy does not vanish as $\epsilon \to 0$, so the lower bound that we are trying to prove trivially holds.

Next, suppose that $C^* = 0$. We use the same construction as before: pick any i, and let $\mathcal{E} = \widetilde{\mathcal{E}} = (c_k = 0 \ \forall k;\ f_k = f \ \forall k; \bar{U} = w_i)$, where $f(i) = 1$. Since $w_i \leq \widetilde{w}_i$, we have $V(w|\mathcal{E}) = y_i - w_i$ and $V(\widetilde{w}|\mathcal{E}) = y_i - \widetilde{w}_i$, so the difference is equal to $\widetilde{w}_i - w_i$, which is at least 0.

From now on, we assume that $C^* > 0$ and $w_i \leq \widetilde{w}_i \leq w_i + 1$ for all i. We also assume that $K = 2$. (If $K > 2$, we can use exactly the same argument: we make $f_k = f_1$ and $c_k = c_1$ for all $k > 2$.) We construct environments $\mathcal{E}$ and $\widetilde{\mathcal{E}}$ such that the agent exerts high effort in $\mathcal{E}$ and low effort in $\widetilde{\mathcal{E}}$. The agent's change of effort engenders a large discrepancy in profit.

We first construct $\widetilde{\mathcal{E}}$. Let δ be a small positive real number such that $2\cdot\vec{\epsilon}\cdot u(w) - 2\delta > 2\cdot\vec{\epsilon}\cdot u(w)\cdot\frac{C}{C^*\cdot(1-N\epsilon)}$. In particular, we have $2\cdot\vec{\epsilon}\cdot u(w) - 2\delta > 0$. (The role of δ is to make the incentive constraints non-binding in order to avoid any potential ambiguity on the agent's effort choice.) Let x be a real number, to be chosen later, which will satisfy the restriction

$$0 < x \leq 1 - N\epsilon. \tag{B.3}$$

Define $\widetilde{f}_1$ and $\widetilde{f}_2$ as follows:

$$\widetilde{f}_1(i) = \begin{cases} \epsilon + x & \text{if } i = 1 \\ 1 - (N-1)\epsilon - x & \text{if } i = i^* \\ \epsilon & \text{otherwise} \end{cases}, \quad \widetilde{f}_2(i) = \begin{cases} 1 - (N-1)\epsilon & \text{if } i = i^* \\ \epsilon & \text{otherwise} \end{cases}.$$

Let $f_1 = \widetilde{f}_1 - \vec{\epsilon}$ and $f_2 = \widetilde{f}_2 + \vec{\epsilon}$. Because of (B.3), all the probabilities are between 0 and

1. Let $c_1 = 0$ and $c_2 = (f_2 - f_1) \cdot u(w) - \delta$. Define the environments $\mathcal{E}$ and $\widetilde{\mathcal{E}}$ as follows:

$$\mathcal{E} = (c_1, c_2; f_1, f_2; \bar{U}), \quad \widetilde{\mathcal{E}} = (c_1, c_2; \widetilde{f}_1, \widetilde{f}_2; \bar{U}),$$

where $\bar{U}$ is small enough such that participation constraints always hold for w and $\widetilde{w}$ in $\mathcal{E}$ and $\widetilde{\mathcal{E}}$.

The agent chooses effort 2 in environment $\mathcal{E}$ because $f_2 \cdot u(w) - c_2 = f_1 \cdot u(w) - c_1 + \delta > f_1 \cdot u(w) - c_1$. We next identify an x such that the agent chooses effort 1 in environment $\widetilde{\mathcal{E}}$: that is, $\widetilde{f}_1 \cdot u(\widetilde{w}) - c_1 > \widetilde{f}_2 \cdot u(\widetilde{w}) - c_2$. This condition is equivalent to $(\widetilde{f}_2 - \widetilde{f}_1) \cdot u(\widetilde{w}) < c_2 - c_1 = (f_2 - f_1) \cdot u(w) - \delta$, which rearranges into $(\widetilde{f}_2 - \widetilde{f}_1) \cdot (u(\widetilde{w}) - u(w)) < 2 \cdot \vec{\epsilon} \cdot u(w) - \delta$. Plugging in the definition of $\widetilde{f}_k$, we get

$$x \cdot (u(\widetilde{w}_{i^*}) - u(w_{i^*}) - u(\widetilde{w}_1) + u(w_1)) < 2 \cdot \vec{\epsilon} \cdot u(w) - \delta. \tag{B.4}$$

If condition (B.4) holds, then $\mathbf{k}^*(\widetilde{w}|\widetilde{\mathcal{E}}) = \{1\}$. It follows that

$$\begin{aligned}
V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) &= f_2 \cdot (y - w) - \widetilde{f}_1 \cdot (y - \widetilde{w}) \\
&= \widetilde{f}_2 \cdot (y - w) - \widetilde{f}_1 \cdot (y - \widetilde{w}) + \vec{\epsilon} \cdot (y - w) \\
&= (\widetilde{f}_2 - \widetilde{f}_1) \cdot (y - \widetilde{w}) + \widetilde{f}_2 \cdot (\widetilde{w} - w) + \vec{\epsilon} \cdot (y - w) \\
&= x \cdot (y_{i^*} - y_1 - \widetilde{w}_{i^*} + \widetilde{w}_1) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) \\
&\quad + \epsilon \cdot \sum_i (\widetilde{w}_i - w_i) + \vec{\epsilon} \cdot (y - w) \\
&\geq x \cdot (y_{i^*} - \widetilde{w}_{i^*}) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon).
\end{aligned}$$

(The last step is due to the fact that $y_1 = w_1 \leq \widetilde{w}_1$ and $\widetilde{w}_i - w_i \geq 0$ for all i.)

We now turn to the choice of x. There are two cases.

If $(1 - N\epsilon) \cdot (u(\widetilde{w}_{i^*}) - u(w_{i^*}) - u(\widetilde{w}_1) + u(w_1)) < 2 \cdot \vec{\epsilon} \cdot u(w) - \delta$, then let $x = 1 - N\epsilon$, which satisfies (B.3) and (B.4). In this case, the lower bound of $\sqrt{C \cdot C^*} \cdot \sqrt{\epsilon} - O(\epsilon)$ trivially holds because the discrepancy does not vanish as $\epsilon \to 0$. Indeed, our choice of x yields $V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) \geq (1 - N\epsilon)(y_{i^*} - w_{i^*}) - O(\epsilon)$, which converges to a positive constant $(y_{i^*} - w_{i^*})$, so for small enough ϵ it is obviously greater than the lower bound we are trying to prove.

Now suppose that $(1 - N\epsilon) \cdot (u(\widetilde{w}_{i^*}) - u(w_{i^*}) - u(\widetilde{w}_1) + u(w_1)) \geq 2 \cdot \vec{\epsilon} \cdot w - \delta$. By the assumptions on δ, we know that $u(\widetilde{w}_{i^*}) - u(w_{i^*}) - u(\widetilde{w}_1) + u(w_1) > 0$. We have $u(\widetilde{w}_{i^*}) - u(w_{i^*}) > u(\widetilde{w}_1) - u(w_1) \geq 0$. Therefore, we derive that $\widetilde{w}_{i^*} > w_{i^*}$. We define $x = \frac{2 \cdot \vec{\epsilon} \cdot u(w) - 2\delta}{(u(\widetilde{w}_{i^*}) - u(w_{i^*}))}$, and this value of x satisfies condition (B.4); it also satisfies (B.3) by

assumption of this case. Thus, the discrepancy becomes

$$\begin{aligned}
V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) &\geq x(y_{i^*} - \widetilde{w}_{i^*}) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\
&= \frac{2 \cdot \vec{\epsilon} \cdot u(w) - 2\delta}{u(\widetilde{w}_{i^*}) - u(w_{i^*})} \cdot (y_{i^*} - \widetilde{w}_{i^*}) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\
&= \frac{2 \cdot \vec{\epsilon} \cdot u(w) - 2\delta}{u(\widetilde{w}_{i^*}) - u(w_{i^*})} \cdot (y_{i^*} - w_{i^*}) - (2 \cdot \vec{\epsilon} \cdot u(w) - 2\delta) \cdot \frac{\widetilde{w}_{i^*} - w_{i^*}}{u(\widetilde{w}_{i^*}) - u(w_{i^*})} \\
&\quad + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon).
\end{aligned}$$

We claim that the term $(2 \cdot \vec{\epsilon} \cdot u(w) - 2\delta) \cdot \frac{\widetilde{w}_{i^*} - w_{i^*}}{u(\widetilde{w}_{i^*}) - u(w_{i^*})}$ is on the order of ϵ. Indeed, we know that $0 < \widetilde{w}_{i^*} - w_{i^*} \leq 1$, so the ratio $\frac{\widetilde{w}_{i^*} - w_{i^*}}{u(\widetilde{w}_{i^*}) - u(w_{i^*})}$ is bounded. Therefore we can reduce the discrepancy to

$$\begin{aligned}
V(w|\mathcal{E}) - V(\widetilde{w}|\widetilde{\mathcal{E}}) &\geq \frac{2 \cdot \vec{\epsilon} \cdot u(w) - 2\delta}{u(\widetilde{w}_{i^*}) - u(w_{i^*})} \cdot (y_{i^*} - w_{i^*}) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\
&\geq \frac{2 \cdot \vec{\epsilon} \cdot u(w)}{u(\widetilde{w}_{i^*}) - u(w_{i^*})} \cdot \frac{C}{C^* \cdot (1 - N\epsilon)} \cdot (y_{i^*} - w_{i^*}) + (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*}) - O(\epsilon) \\
&\geq 2 \cdot \sqrt{\frac{2 \cdot \vec{\epsilon} \cdot u(w)}{u(\widetilde{w}_{i^*}) - u(w_{i^*})} \cdot \frac{C}{C^* \cdot (1 - N\epsilon)} \cdot (y_{i^*} - w_{i^*}) \cdot (1 - N\epsilon)(\widetilde{w}_{i^*} - w_{i^*})} - O(\epsilon) \\
&= 2 \cdot \sqrt{2 \cdot \vec{\epsilon} \cdot u(w) \cdot \frac{C}{C^*} \cdot (y_{i^*} - w_{i^*}) \cdot \frac{\widetilde{w}_{i^*} - w_{i^*}}{u(\widetilde{w}_{i^*}) - u(w_{i^*})}} - O(\epsilon) \\
&\geq 2 \cdot \sqrt{2 \cdot \vec{\epsilon} \cdot u(w) \cdot \frac{C}{C^*} \cdot (y_{i^*} - w_{i^*}) \cdot \frac{1}{u'(w_{i^*})}} - O(\epsilon) \\
&= \sqrt{C \cdot C^*} \cdot \sqrt{\epsilon} - O(\epsilon).
\end{aligned}$$

The second to last step is to due the fact that $\frac{\widetilde{w}_{i^*} - w_{i^*}}{u(\widetilde{w}_{i^*}) - u(w_{i^*})} > \frac{1}{u'(w_{i^*})}$. We have thus established the desired lower bound stated in the beginning of the proof. □

References

[1] Philippe Aghion, Drew Fudenberg, Richard Holden, Takashi Kunimoto, and Olivier Tercieux (2012), "Subgame-Perfect Implementation under Information Perturbations," *Quarterly Journal of Economics* 127 (4), 1843–1881.

[2] Nemanja Antić (2014), "Contracting with Unknown Technologies," unpublished paper, Princeton University.

[3] Gabriel Carroll (2015), "Robustness and Linear Contracts," *American Economic Review* 105 (2), 536–563.

[4] Sylvain Chassang (2013), "Calibrated Incentive Contracts," *Econometrica* 81 (5), 1935–1971.

[5] Kim-Sau Chung and Jeffrey C. Ely (2003), "Implementation with Near-Complete Information," *Econometrica* 71 (3), 857–871.

[6] Daniel Garrett (2014), "Robustness of Simple Menus of Contracts in Cost-Based Procurement," *Games and Economic Behavior* 87, 631–641.

[7] Oliver Hart and Bengt Holmström (1987), "The Theory of Contracts," in T. F. Bewley (ed.), *Advances in Economic Theory: Fifth World Congress of the Econometric Society* (Cambridge: Cambridge University Press), 71–155.

[8] Bengt Holmström (1979), "Moral Hazard and Observability," *Bell Journal of Economics* 10 (1), 74–91.

[9] Philippe Jehiel, Moritz Meyer-ter-Vehn, and Benny Moldovanu (2012), "Locally Robust Implementation and its Limits," *Journal of Economic Theory* 147 (6), 2439–2452.

[10] Kristóf Madarász and Andrea Prat (2014), "Sellers with Misspecified Models," unpublished paper, Columbia University.