Contracting on Value

Jonathan Bonham
University of Chicago
Booth School of Business

Amoray Riggs-Cragun
University of Chicago
Booth School of Business

November 9, 2022

Abstract

We revisit the optimal use of information under moral hazard while assuming that the agent can choose distributions nonparametrically. Under this assumption, optimal contracts behave as if the principal were making inferences about outcomes she values rather than about the agent's action. This has significant implications for what measures are included in contracts and how those measures are used. Most importantly, Holmström's (1979) informativeness principle changes. A performance measure is valuable if it improves inferences not about the agent's action, but about outcomes the principal values; and if those outcomes are contractible, additional measures have no value.

Keywords: Moral hazard, optimal contracts, value of information.

We appreciate helpful comments from Jeremy Bertomeu, Jörg Budde (discussant), Pierre Chaigneau, Alex Frankel, Robert Göx, Benjamin Hébert, Florian Heider, Thomas Hemmer, Bengt Holmström, Seung Lee, Christian Leuz, Valeri Nikolaev, Haresh Sapra, Abbie Smith, Lars Stole and participants in the DAR & DART Accounting Theory Seminar, the Chicago Booth Junior Theory meeting, the Accounting and Economics Society webinar, the 12th Accounting Research Workshop organized by the Universities of Basel and Zurich, and workshop participants at the University of Chicago, Rice University, Northwestern University and the University of Illinois Chicago. We gratefully acknowledge financial support from the Booth School of Business and research assistance from Junyoung Jeong and Daniel Leonard. The authors can be contacted at Jonathan.Bonham@chicagobooth.edu and amoray@chicagobooth.edu.

1 Introduction

Holmström (1979) shows that information is useful for contracting if it improves inferences about an agent's hidden action. He derives this result using the *parameterized distribution formulation* of the agency problem, wherein the agent chooses the parameter of a distribution.[1] In this paper, we return to Holmström's moral hazard setting and invoke the *generalized distribution formulation*, wherein the agent chooses a distribution nonparametrically.[2] We show that under this formulation of the problem, information is valuable for contracting if it improves inferences not about the agent's action, but about outcomes the principal values.

Assume that a principal values the random variables X and Y according to some benefit function $B(X, Y)$, where Y is contractible but X is not. Holmström's seminal result shows that if an agent chooses a distribution $p(X, Y; a)$ from a family of distributions parameterized by a, the optimal contract is characterized by the following expression, where $U'(\cdot)$ is the agent's marginal utility and λ and μ are positive constants.

$$\frac{1}{U'(s(y))} = \lambda + \mu \cdot \frac{p_a(y;a)}{p(y;a)} \tag{1}$$

The optimal contract s is a transformation of the *likelihood ratio* $\frac{p_a(y;a)}{p(y;a)}$, which indicates how likely it is that the agent took the desired action a given the observed outcome y. Therefore, the contract behaves as if the principal were making inferences about the agent's action (even though she knows what action the agent takes in equilibrium), and information is valuable for contracting if and only if it improves inferences about the agent's action.

The generalized distribution approach expands the agent's choice set relative to

[1]The parameterized distribution formulation was first used by Mirrlees ([1975] 1999). Its use was widely popularized by Holmström (1979), who made the problem more tractable by assuming the first-order approach to be valid.

[2]The terms *parameterized distribution formulation* and *generalized distribution formulation* were coined by Hart and Holmström (1987). We discuss the origins of the generalized approach in section 6.

the classic parametric approach; rather than selecting a distribution from an exogenously constrained parametric family via the "effort" parameter a, the agent can implement *any* distribution p from the probability simplex. There are many ways to conceptualize how an agent might "choose a distribution," and all of them involve giving the agent a rich opportunity set.[3] For example, the agent could choose his action conditional on a large set of private information in a static game (as suggested by Holmström and Milgrom 1987), could choose effort continuously throughout the contracting period (a microfoundation provided by Hébert 2018), or could have an action space that spans the probability simplex in a multi-tasking game (a microfoundation we provide in section 5.2). In all of these cases, the agent's opportunity set is sufficiently rich that he can implement any distribution through some action choice (e.g., a certain private-information-contingent effort strategy or a certain vector of task allocations), allowing the agent to effectively "choose" any distribution at some cost.

We assume that there is some distribution q that minimizes the agent's personal cost at zero; this zero-cost distribution can be interpreted as the distribution that arises when the agent shirks, as it is the distribution the agent implements absent contractual incentives to do something different. When the agent implements some distribution $p \neq q$, he incurs a cost equal to the divergence of p from q. Specifically, we represent the agent's cost as an f-divergence, a class of cost functions implied by the axioms of *decomposability* (i.e. additive separability in probabilities) and *invariance* (Amari 2016).[4]

These assumptions change Holmström's iconic characterization (1) in several ways. First, because the agent chooses the probability of every possible realization, the likelihood ratio is $\frac{1}{p(y)}$. Second, there is an incentive compatibility (IC) constraint and an associated multiplier $\mu(y)$ for each y, rather than a single IC constraint and multiplier μ. Moreover, due to the additive separability of the agent's cost function,

[3] See section 6.1 for more detailed discussion.

[4] The f-divergences are a broad class including all α-divergences such as the χ^2-divergence, the Hellinger distance, and the the Kullback Leibler divergence, which is the benchmark cost function employed by Hébert (2018). We discuss divergences further in section 6.2.

the IC constraint for each probability $p(y)$ is independent of all other probabilities, which allows us to easily solve for $\mu(y)$ in closed form. Given the agent's f-divergence cost function, this $\mu(y)$ absorbs the likelihood ratio such that the optimal contract is characterized by

$$\frac{1}{U'(s(y))} = \lambda + \frac{\mathbb{E}[B(X,Y)|y]-s(y)-\eta}{\check{f}(U(s(y)-\nu)}, \tag{2}$$

where $\check{f}(\cdot)$ is a transformation whose form depends on the particular f function specified in the divergence, and η and ν are constants. Notice that y enters the expression only through $s(y)$ and $\mathbb{E}[B(X,Y)|y]$; hence, the optimal contract is a transformation of the principal's expected payoff conditional on contractible information y. It follows that the signal y has contracting value if and only if it is informative about the principal's payoff, rather than about the agent's action.

The same result obtains in a binary setting where the agent chooses a single probability; we provide this binary example in section 3. In the binary setting, we need not define the agent's cost function as a divergence; a single-valued convex cost function is sufficient for the result. Section 4 is dedicated to deriving equation (2) in a general setting with many contractible and non-contractible signals. We show in section 5.1 that our approach is equivalent to a parametric multi-tasking model where the agent's action space is rich enough to make every probability distribution in the simplex accessible; that is, it is not the direct choice of probabilities that matters for our results, but rather that the agent's action space is sufficiently rich. In section 5.2 we generalize our results to settings in which some variables are exogenously distributed, and show that a signal is useful if and only if it is informative about the portion of value that the agent can control. We discuss our assumptions and their use in related literature in section 6. Much of this discussion relates our paper to Hébert (2018), who also pairs the generalized approach with an invariant divergence cost function in a principal-agent setting.

2 Notation and preliminary assumptions

A risk-neutral principal hires an agent to take hidden actions that stochastically improve *value*, which we define as the principal's payoff gross of compensation paid to the agent. We denote value by $B(X,Y)$, where $X = (X_1,\ldots,X_N)$ and $Y = (Y_1,\ldots,Y_M)$ are vectors of random variables with possible realizations $(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$ for some finite sets $\mathcal{X}$ and $\mathcal{Y}$. We denote a realization of X_i by $x_i \in \mathcal{X}_i$ and a realization of Y_j by $y_j \in \mathcal{Y}_j$, so that $\mathcal{X} \equiv \mathcal{X}_1\times\cdots\times\mathcal{X}_M$ and $\mathcal{Y} \equiv \mathcal{Y}_1\times\cdots\times\mathcal{Y}_N$. The distinction between X and Y is that Y is contractible whereas X is not. We denote a contract $s(Y)$ by $s : \mathcal{Y}\to\mathbb{R}$.

Denote the joint probability distribution over X and Y by $p_{XY} = p_Y p_{X|Y} \in \mathcal{P}(\mathcal{X},\mathcal{Y})$, where $\mathcal{P}(\mathcal{X},\mathcal{Y})$ is the probability simplex over $\mathcal{X}\times\mathcal{Y}$. Throughout we let $p(\mathbf{x},\mathbf{y})$, $p(\mathbf{x})$, $p(\mathbf{y})$, $p(\mathbf{x}\,|\,\mathbf{y})$, and $p(\mathbf{y}\,|\,\mathbf{x})$ denote the joint, marginal, and conditional probabilities of a specific outcome $(\mathbf{x},\mathbf{y})$, and we let p_{XY}, p_X, p_Y, $p_{X|Y}$, and $p_{Y|X}$ denote the joint, marginal, and conditional probability distributions over $\mathcal{X}\times\mathcal{Y}$. For notational ease, we will often refer to p_{XY} as simply p. We let $\sum_{\mathcal{X}\times\mathcal{Y}}$ denote summation over all $(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$; likewise, $\sum_{\mathcal{X}}$ and $\sum_{\mathcal{Y}}$ denote summation over all $\mathbf{x}\in\mathcal{X}$ and $\mathbf{y}\in\mathcal{Y}$, respectively. We let $\sum_{\mathcal{Y}\setminus\mathbf{y}_k}$ denote summation over all $\mathbf{y}\in\mathcal{Y}$ except for the outcome $\mathbf{y}_k$.

We say that Y_i is *useful* if there exist some $(\mathbf{y}_{-i}, y_i), (\mathbf{y}_{-i}, y_i') \in \mathcal{Y}$ such that an optimal contract necessarily sets $s(\mathbf{y}_{-i}, y_i) \neq s(\mathbf{y}_{-i}, y_i')$. We say that Y_i is *informative about value* if there exists some $(\mathbf{y}_{-i}, y_i), (\mathbf{y}_{-i}, y_i') \in \mathcal{Y}$ such that $\mathbb{E}[B(X,Y)|\,\mathbf{y}_{-i}, y_i] \neq \mathbb{E}[B(X,Y)|\,\mathbf{y}_{-i}, y_i']$. By contrast, we say that Y_i is *informative about action* p_{XY} if there exist some $(\mathbf{x},\mathbf{y}_{-i}, y_i), (\mathbf{x},\mathbf{y}_{-i}, y_i') \in \mathcal{X}\times\mathcal{Y}$ such that the likelihood ratios $\frac{1}{p(\mathbf{x},\mathbf{y}_{-i},y_i)}$ and $\frac{1}{p(\mathbf{x},\mathbf{y}_{-i},y_i')}$ are not equal. We define informativeness about action p_X, p_Y, $p_{X|Y}$, and $p_{Y|X}$ analogously. In all cases, we say that Y is informative if there exists at least one Y_i that is informative.

Let the agent's utility from compensation be given by $U(\cdot)$, where $U'(\cdot) > 0$ and $U''(\cdot) < 0$ so that the agent is risk averse, and assume that the agent's utility is additively separable in his compensation and personal cost. Let $\bar{U}$ denote the agent's utility from outside options.

3 Binary example

Assume that the only contractible performance measure is a binary random variable Y with possible realizations $y \in \{y_L, y_H\}$. The agent chooses the probability of the high outcome, $p_H \equiv p_Y(y_H)$, whereas $p_{X|Y}$ is equal to some exogenous distribution $q_{X|Y}$ such that the probability of the realization $(\mathbf{x}, y)$ is given by $p_{XY}(\mathbf{x}, y) = p_Y(y)q_{X|Y}(\mathbf{x}\,|y)$. That is, the agent influences X only through the related variable Y; in later sections, we allow the agent to directly choose the entire joint distribution p_{XY}.[5] Let the agent's cost of choosing p_H be given by $c(p_H)$, where $c : [0, 1] \to [0, \infty)$ is convex. The contract $s(Y)$ specifies the payments $s(y_L)$ and $s(y_H)$, depending on whether the high or low outcome is realized.

Faced with contract $s(Y)$, the agent chooses p_H to maximize his expected utility from compensation less his cost of effort.

$$\begin{aligned} \max_{a} \quad & (U(s(y_H)) - U(s(y_L)))p_H + U(s(y_L)) - c(p_H) \\ \text{s.t.} \quad & 0 \leq p_H \leq 1 \end{aligned} \tag{3}$$

Assume that the constraints do not bind.[6] Then the agent's first-order condition is as follows.

$$c'(p_H) \;\; = \;\; U(s(y_H)) - U(s(y_L)) \tag{4}$$

The principal's aim is to propose a contract and action pair $(s(Y), p_H)$ that maximizes her net payoff subject to three constraints. First, the proposed pair must make the agent's expected utility at least as great as her reservation utility $\bar{U}$; this individual rationality (IR) constraint ensures that the agent is willing to accept the contract. Second, the proposed pair must be incentive compatible (IC); that is, given the proposed scheme $s(Y)$, the agent chooses the proposed probability p_H voluntarily. Notice that the agent's expected utility (3) is concave in p_H for any $s(Y)$; then the agent's first-order condition (4) is sufficient to ensure incentive compatibility, and we can thus use (4) as the IC constraint in the principal's program. Finally, the proposed p_H must be between zero and one to ensure that p_Y is a probability mass function

[5]We show in section 4 that when faced with a contract $s(Y)$, in equilibrium the agent always sets $p_{X|Y} = q_{X|Y}$. We abstract away from this choice to simplify the binary example.

[6]We will consider boundary cases in the full analysis in section 4.

(p.m.f.). Then the principal's program can be written as follows, where we assume for simplicity that $\mathbb{E}[B(X,Y)|y_L] = 0$.

$$\max_{s(Y),p_H} \left(\sum_{\mathcal{X}} B(\mathbf{x}, y_H) q(\mathbf{x}\,|y_H) - s(y_H) \right) p_H - s(y_L)(1-p_H) \tag{5}$$

$$\text{s.t. } (U(s(y_H)) - U(s(y_L)))p_H + U(s(y_L)) - c(p_H) \geq \bar{U} \tag{IR}$$

$$U(s(y_H)) - U(s(y_L)) = c'(p_H) \tag{IC}$$

$$0 \leq p_H \leq 1 \tag{p.m.f}$$

Assume that the p.m.f. constraints are nonbinding, and let λ and μ be the Lagrange multipliers on the IR and IC constraints, respectively. The payment $s(y_L)$ is optimally chosen to bind the IR constraint; we show this formally in section 4. It remains to characterize the optimal high payment $s(y_H)$. Optimizing (5) over $s(y_H)$ and p_H gives the following first-order conditions.

$$\frac{1}{U'(s(y_H))} = \lambda + \mu \cdot \frac{1}{p_H} \tag{6}$$

$$\mu = \frac{\sum_{\mathcal{X}} B(\mathbf{x}, y_H) q(\mathbf{x}\,|y_H) - (s(y_H) - s(y_L))}{c''(p_H)}. \tag{7}$$

Because $c(\cdot)$ is strictly convex, $c'(\cdot)$ is strictly increasing and therefore has a strictly increasing inverse. Inverting (4) to obtain $p_H = c'^{-1}(U(s(y_H)) - U(s(y_L)))$ and substituting (7) into (6) yields the following characterization of $s(y_H)$, where $\eta(y_L) \equiv -s(y_L)$, $\nu(y_L) \equiv U(s(y_L))$, and $\check{c}(\cdot) \equiv c'^{-1}(\cdot)c''\left(c'^{-1}(\cdot)\right)$.

$$\frac{1}{U'(s(y_H))} = \lambda + \frac{\mathbb{E}[B(X,Y)|y_H] - s(y_H) - \eta(y_L)}{\check{c}(U(s(y_H)) - \nu(y_L))} \tag{8}$$

Notice that y_H enters the expression only through $s(y_H)$ and $\mathbb{E}[B(X,Y)|y_H]$; that is, the signal y is useful for contracting only to the extent that it is informative about value. In this binary setting where the agent chooses a single probability p_H, this result holds for any convex cost function. The following section shows that in nonbinary settings, the result holds for all divergence-based cost functions satisfying two axioms.

4 Main analysis

Assume now that the agent chooses the joint probability distribution p_{XY} and that $\mathcal{X}$ and $\mathcal{Y}$ are of any finite dimension. While there are many ways to model the agent's choice of distribution, we assume that the agent chooses the probability $p(\mathbf{x},\mathbf{y})$ for each $(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, subject to the p.m.f. constraints that $0 \leq p(\mathbf{x},\mathbf{y})$ for all $(\mathbf{x},\mathbf{y})$ and $\sum_{\mathcal{X}\times\mathcal{Y}} p(\mathbf{x},\mathbf{y}) = 1$.

4.1 Cost function assumptions

Let $C(p)$ denote the personal cost incurred by the agent when he implements distribution $p_{XY} = p$. We assume that $C : \mathcal{P}(\mathcal{X},\mathcal{Y}) \to [0,\infty)$ is strictly convex, twice differentiable, and that it attains a unique minimum at some $q = q_{XY} \in \mathcal{P}(\mathcal{X},\mathcal{Y})$ that satisfies $C(q) = 0$. It immediately follows that $C(p)$ is a divergence and can be written

$$C(p) \quad = \quad D(p||q). \tag{9}$$

A divergence is a weakened notion of the distance from one probability distribution to another. Like a distance, $D(p||q)$ is nonnegative and is equal to zero if and only if $p = q$; intuitively, the agent incurs zero cost only when he implements the cost-minimizing distribution q, and he incurs a positive cost for any $p \neq q$. Unlike a distance, a divergence need not satisfy the triangle inequality, nor need it be symmetric; that is, $D(p||q) \neq D(q||p)$ is permitted. This allows the cost of moving from q to p to differ from the cost of moving from p to q; for example, it may be more costly to increase the mean of a variable than to decrease it. We now place additional structure on the set of divergences we will consider, in the form of two axioms.

Axiom 1 (Invariance). *Let $Z = k(X,Y)$ be some transformation of the random variables (X,Y), where $\dim(\mathcal{Z}) \leq \dim(\mathcal{X},\mathcal{Y})$, and let $\tilde{p}_Z$ be the corresponding transformation of p_{XY}. Then*

$$D(\tilde{p}_Z||\tilde{q}_Z) \leq D(p_{XY}||q_{XY}), \tag{10}$$

where the equality holds if and only if Z is a sufficient statistic for (X, Y).

We adopt the invariance criterion from Amari (2016), who describes it as the "reasonable criterion" that is "needed for introducing a geometrical structure to a manifold of probability distributions" (p. 51). In information geometry, invariance ensures that the divergence between two probability distributions is the same when a random variable is transformed without losing information. Applied to our setting, it implies that the cost of implementing p_{XY} is weakly increasing in the number of variables he controls; giving the agent more variables to control cannot reduce his cost, ruling out strange cases in which "doing more costs less." Assume for example that $Z = k(X, Y) = Y$. Then invariance implies that the following relationships must hold for all $p, q \in \mathcal{P}(\mathcal{X}, \mathcal{Y})$.

$$D(p_Y||q_Y) = D(p_Y q_{X|Y}||q_{XY}) \leq D(p_{XY}||q_{XY}) \tag{11}$$

These relationships align with the definitions of *partial invariance* and *partial monotonicity* (in the X dimension) from Hébert and La'O (2022). Their significance will become clear when we analyze the agent's program, as will the usefulness of the following axiom.

Axiom 2 (**Decomposability**). *The cost of choosing the distribution p_{XY} is additively separable in the probabilities $p_{XY}(\mathbf{x}, \mathbf{y})$.*

Additive separability greatly enhances the problem's tractability and is a common assumption in information geometry and its applications (Amari 2016, p. 106). While tractability is our primary motivation for relying on this axiom, it is defensible for an agent with the flexibility to locally perturb the distribution of signals without considering the distribution's non-local properties. For example, decomposability implies that the cost of shifting mass above a salient threshold (e.g., a bonus target) does not depend on the distribution's tail risk.

It is well known that a divergence satisfies the above two axioms if and only if it is an *f-divergence*,[7] defined as follows.

$$C(p) \;=\; D_f(p||q) \;\equiv\; \sum_{\mathcal{X}\times\mathcal{Y}} q(\mathbf{x},\mathbf{y}) f\left(\tfrac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right) \tag{12}$$

We assume that $\text{supp}(q) = \mathcal{X}\times\mathcal{Y}$ to ensure that $p(\mathbf{x},\mathbf{y})/q(\mathbf{x},\mathbf{y}) < \infty$ for all p and $(\mathbf{x},\mathbf{y})$. This, along with our assumptions on $C(p)$, imply that $f : [0,\infty) \to [0,\infty)$ is convex, twice-differentiable, and satisfies $f(1) = 0$. One particular f-divergence we will refer to occasionally is the *Kullback-Leibler divergence*, which specifies that $f(0) = 0$ and $f(t) = t\ln(t)$ for $t > 0$.

4.2 The agent's problem

Because the agent's utility is additively separable in compensation and personal cost, his net utility is $U(s(Y)) - D_f(p||q)$. Faced with a contract $s(Y)$, the agent chooses p to maximize the following program.

$$\begin{aligned} \max_{p} \quad & \sum_{\mathcal{X}\times\mathcal{Y}} U(s(\mathbf{y}))p(\mathbf{x},\mathbf{y}) - D_f(p||q) \\ \text{s.t.} \quad & 1 = \sum_{\mathcal{X}\times\mathcal{Y}} p(\mathbf{x},\mathbf{y}),\; p(\mathbf{x},\mathbf{y}) \geq 0 \;\forall(\mathbf{x},\mathbf{y}) \end{aligned} \tag{13}$$

Suppose for the moment that the non-negativity constraints are non-binding, and for some arbitrary $(\mathbf{x}_0,\mathbf{y}_0) \in \mathcal{X}\times\mathcal{Y}$ rewrite the remaining constraint as $p(\mathbf{x}_0,\mathbf{y}_0) = 1 - \sum_{\mathcal{X}\times\mathcal{Y}\setminus(\mathbf{x}_0,\mathbf{y}_0)} p(\mathbf{x},\mathbf{y})$. Substituting this constraint into the objective function and taking the first-order condition with respect to $p(\mathbf{x},\mathbf{y})$ yields the following.

$$f'\left(\tfrac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right) - f'\left(\tfrac{p(\mathbf{x}_0,\mathbf{y}_0)}{q(\mathbf{x}_0,\mathbf{y}_0)}\right) \;=\; U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) \tag{14}$$

Because $f(\cdot)$ is a strictly convex function, $f'(\cdot)$ is strictly increasing and can therefore be inverted. We can therefore solve (14) for $p(\mathbf{x},\mathbf{y})$ and express the marginal and

[7] See for example Amari (2016), Theorem 3.1.

conditional distributions as follows, where $\nu(\mathbf{x}_0, \mathbf{y}_0) \equiv U(s(\mathbf{y}_0)) - f'\left(\frac{p(\mathbf{x}_0,\mathbf{y}_0)}{q(\mathbf{x}_0,\mathbf{y}_0)}\right)$.

$$\begin{aligned} p(\mathbf{y}) &= q(\mathbf{y}) f'^{-1}\left(U(s(\mathbf{y})) - \nu(\mathbf{x}_0, \mathbf{y}_0)\right) \\ p(\mathbf{x} \,|\, \mathbf{y}) &= q(\mathbf{x} \,|\, \mathbf{y}) \end{aligned} \tag{15}$$

In particular, the principal is unable to motivate any action for which $p_{X|Y} \neq q_{X|Y}$ using a contract written only on Y. This is because the cost-minimizing $p_{X|Y}$ is $q_{X|Y}$ regardless of the agent's choice of p_Y; it follows from (11) that $D_f(p_{XY}||q_{XY}) = D_f(p_Y q_{X|Y}||q_{XY}) = D_f(p_Y||q_Y)$. The following lemma extends these incentive compatibility conditions to cases in which the non-negativity constraints are permitted to bind.

Lemma 1 *For any contract $s(Y)$, a solution to program* (13) *exists. Moreover, for any $(\mathbf{x}_0, \mathbf{y}_0) \in supp(\mathcal{X} \times \mathcal{Y})$, the solution for $(\mathbf{x}, \mathbf{y}) \neq (\mathbf{x}_0, \mathbf{y}_0)$ satisfies $p(\mathbf{x}, \mathbf{y}) > 0$ if and only if $U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) > \lim_{t \to 0^+} f'(t) - f'\left(\frac{p(\mathbf{x}_0,\mathbf{y}_0)}{q(\mathbf{x}_0,\mathbf{y}_0)}\right)$, in which case $p(\mathbf{x}, \mathbf{y})$ satisfies equations* (14) *and* (15) *for all $(\mathbf{x}, \mathbf{y}) \neq (\mathbf{x}_0, \mathbf{y}_0)$.*

Permitting the non-negativity constraints to bind simply requires replacing $p(\mathbf{x}, \mathbf{y})$ with zero whenever (14) would imply a negative probability, which only occurs for payments below a threshold that is increasing in $\lim_{t \to 0^+} f'(t)$. Note that if $\lim_{t \to 0^+} f'(t) = -\infty$ as with, for example, the Kullback-Leibler divergence, then the agent always chooses a distribution with full support on $\mathcal{X} \times \mathcal{Y}$.

4.3 The principal's problem

Lemma 1 completely characterizes the agent's incentive compatible action under the contract $s(Y)$. The first-order approach is automatically valid under the generalized approach; to see this, notice that program (13) maximizes a concave function with linear constraints. Because Lemma 1 (specifically equation (15)) shows that $p_{X|Y} = q_{X|Y}$ for any $s(Y)$, we can write the the principal's program over the choice variables $s(Y)$ and p_Y. Then if we again for the moment rule out cases in which $\text{supp}(p) \neq$

$\mathcal{X} \times \mathcal{Y}$, the principal's program can be written as follows.

$$\max_{s(Y),p_Y} \sum_{\mathcal{Y}} \left(\sum_{\mathcal{X}} B(\mathbf{x},\mathbf{y})q(\mathbf{x}\,|\,\mathbf{y}) - s(\mathbf{y}) \right) p(\mathbf{y}) \tag{16}$$

$$\text{s.t.} \sum_{\mathcal{Y}} U(s(\mathbf{y}))p(\mathbf{y}) - \sum_{\mathcal{Y}} q(\mathbf{y}) f\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) \geq \bar{U} \tag{IR}$$

$$U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) = f'\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) - f'\left(\frac{p(\mathbf{y}_0)}{q(\mathbf{y}_0)}\right) \;\forall\, \mathbf{y} \neq \mathbf{y}_0 \tag{IC$_\mathbf{y}$}$$

$$p(\mathbf{y}_0) = 1 - \sum_{\mathcal{Y}\setminus\mathbf{y}_0} p(\mathbf{y}) \tag{p.m.f}$$

Notice that because the IC constraints depend only on the wage differentials $U(s(\mathbf{y})) - U(s(\mathbf{y}_0))$, any action that can be induced with a contract that does not bind the IR constraint can also be induced with one that does: simply reduce all wages by a constant in utility space until the IR constraint binds. Because this has no impact on the IC constraints, the two contracts motivate the same action while decreasing the expected wage, which is an unambiguous improvement from the principal's perspective. Thus the IR constraint binds in equilibrium, which we formalize in the following lemma.

Lemma 2 *The individual rationality constraint binds.*

Now substitute the p.m.f. constraint into the objective function and other constraints, and let λ and $\mu(\mathbf{y})$ be the Lagrange multipliers on the IR and the $\mathbf{y}^{th}$ IC constraints. Then the principal's first-order conditions with respect to $p(\mathbf{y})$ and $s(\mathbf{y})$ for $\mathbf{y} \neq \mathbf{y}_0$ are

$$0 = \mathbb{E}[B(X,Y)|\,\mathbf{y}] - s(\mathbf{y}) - \mu(\mathbf{y})\frac{1}{q(\mathbf{y})} f''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) - \eta(\mathbf{y}_0) \tag{17}$$

$$0 = -\,p(\mathbf{y}) + \lambda U'(s(\mathbf{y}))p(\mathbf{y}) + \mu(\mathbf{y})U'(s(\mathbf{y})), \tag{18}$$

where $\eta(\mathbf{y}_0) \equiv \mathbb{E}[B(X,Y)|\,\mathbf{y}_0] - s(\mathbf{y}_0) + \frac{1}{q(\mathbf{y}_0)} f''\left(\frac{p(\mathbf{y}_0)}{q(\mathbf{y}_0)}\right) \sum_{\mathcal{Y}\setminus\mathbf{y}_0} \mu(\tilde{\mathbf{y}})$ does not vary with $\mathbf{y} \in \mathcal{Y}\setminus\mathbf{y}_0$. The following lemma establishes conditions under which these first-order conditions are necessary and sufficient for an interior solution.

Lemma 3 *The principal's first-order conditions are necessary and sufficient for an* *interior solution if* $U^{-1}(f'(t)+\nu)\,t$ *and* $f(t)-tf'(t)$ *are convex in* t *for any* $\nu \in \mathbb{R}$ *and* $t \in (0,\infty)$.

The Kullback-Leibler divergence is an example of an f-divergence that satisfies the conditions in Lemma 3, as we will show at the end of this section. Returning to the principal's first-order conditions, we can solve (17) for $\mu(\mathbf{y})$ in closed form as follows.

$$\mu(\mathbf{y}) \quad = \quad \frac{\mathbb{E}[B(X,Y)|\,\mathbf{y}]-s(\mathbf{y})-\eta(\mathbf{y}_0)}{\frac{1}{q(\mathbf{y})}f''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)} \tag{19}$$

Finally, substituting this expression for $\mu(\mathbf{y})$ into (18) yields

$$\frac{1}{U'(s(\mathbf{y}))} \quad = \quad \lambda+\mu(\mathbf{y})\frac{1}{p(\mathbf{y})} \quad = \quad \lambda+\frac{\mathbb{E}[B(X,Y)|\,\mathbf{y}]-s(\mathbf{y})-\eta(\mathbf{y}_0)}{\frac{1}{q(\mathbf{y})}f''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)}\cdot\frac{1}{p(\mathbf{y})}. \tag{20}$$

Equation (20) suggests that a signal Y_i may be useful if it is informative about value, if it is informative about the action p_Y, or if it affects the higher-order properties of the agent's cost function through $\frac{\partial^2 D_f(p_Y||q_Y)}{\partial p(\mathbf{y})^2}=\frac{1}{q(\mathbf{y})}f''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)$. We now show that of these three potential criteria for a signal's usefulness, informativeness about value is the only one that matters.

4.4 The usefulness of information

Returning to equation (14), define $\nu(\mathbf{y}_0)\equiv U(s(\mathbf{y}_0))-f'\left(\frac{p(\mathbf{y}_0)}{q(\mathbf{y}_0)}\right)$ so that the agent's IC constraint can be written $f'\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)=U(s(\mathbf{y}))-\nu(\mathbf{y}_0)$. Again relying on the invertibility of $f'(\cdot)$, (14) can be rewritten

$$\frac{p(\mathbf{y})}{q(\mathbf{y})} \quad = \quad f'^{-1}\left(U(s(\mathbf{y}))-\nu(\mathbf{y}_0)\right). \tag{21}$$

Substituting (21) into (20) yields the paper's main result, given in the following proposition.

Proposition 1 *For any* $\mathbf{y}_0 \in$ *supp*(p_Y)*, the optimal contract is characterized on supp*(p_Y) *by*

$$\frac{1}{U'(s(\mathbf{y}))} \quad = \quad \lambda+\frac{\mathbb{E}[B(X,Y)|\,\mathbf{y}]-s(\mathbf{y})-\eta(\mathbf{y}_0)}{\check{f}(U(s(\mathbf{y}))-\nu(\mathbf{y}_0))}, \tag{22}$$

where $\check{f}(\cdot) \equiv f'^{-1}(\cdot) f''\left(f'^{-1}(\cdot)\right)$. *Moreover, this contract is weakly efficient for* $\mathbf{y} \notin$ *supp*(p_Y) *if* $\frac{tf'''(t)}{f''(t)} \geq -1\ \forall u > 0$.

Examining (22) reveals that the general solution takes the same form as the binary one in equation (8); in particular, $\mathbf{y}$ only enters the expression through $s(\mathbf{y})$ and $\mathbb{E}[B(X,Y)|\,\mathbf{y}]$. Because the equality must be maintained for every $\mathbf{y} \neq \mathbf{y}_0$, $s(\mathbf{y})$ can vary with $\mathbf{y}$ only if $\mathbb{E}[B(X,Y)|\,\mathbf{y}]$ also varies with $\mathbf{y}$. The implication is that a signal is useful if and only if it is informative about value, not actions.

Corollary 1 *A signal is useful if and only if it is informative about value.*

The proof of this corollary shows that the optimal contract characterized by (22) necessarily varies with a signal Y_i if and only if $\mathbb{E}[B(X,Y)|\,\mathbf{y}] = \mathbb{E}[B(X,Y)|\,\mathbf{y}_{-i}, y_i]$ varies with y_i. This is in contrast to the solution under the classic parametric approach, where a signal Y_i is valuable if and only if the likelihood ratio $\frac{p_a(\mathbf{y};a)}{p(\mathbf{y};a)} = \frac{p_a(\mathbf{y}_{-i},y_i;a)}{p(\mathbf{y}_{-i},y_i;a)}$ varies with y_i. An immediate corollary is that if value is itself contractible, then additional signals are not useful.

Corollary 2 *If* $B(X,Y) = Y_i$ *is contractible, then* Y_i *is the only useful signal.*

The corollary shows that if the principal values a single variable, Y_i, and if that variable is directly contractible, the principal will write the contract on that signal and will ignore all other available signals. This is in stark contrast to the classic parametric approach, where the optimal contract will incorporate any contractible signal that is incrementally informative about the agent's action.

4.5 Properties of the optimal contract and action

We conclude this section by establishing some additional properties of the optimal contract and action. First, we provide a sufficient condition on f for $s(Y)$ and $\frac{p_{XY}}{q_{XY}}$ to be monotone in the expected value given Y.

Corollary 3 *If* $\frac{tf'''(t)}{f''(t)} \geq -1\ \forall t > 0$, *then* $s(\mathbf{y})$ *and* $\frac{p(\mathbf{y})}{q(\mathbf{y})} = \frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}$ *are increasing in* $\mathbb{E}[B(X,Y)|\,\mathbf{y}]$.

When $B(X,Y) = Y_i$, the monotonicity of $\frac{p(y_i)}{q(y_i)}$ in y_i implies that p_Y first-order stochastically dominates q_Y. The condition on f requires that the convexity of f not be decreasing too quickly in u. The following corollary provides an example of a cost function that satisfies this condition as well as those in Lemma 3 that ensure the sufficiency of the principal's first-order conditions for a solution.

Corollary 4 *Let $f(t) = t\ln(t)$, so that $D_f(p||q)$ is the Kullback-Leibler divergence. Then f satisfies the conditions in Lemma 3 and Corollary 3.*

Under a Kullback-Leibler divergence cost function, the optimal contract from (22) takes the following simple form.

$$\frac{1}{U'(s(\mathbf{y}))} \;=\; \lambda + \mathbb{E}[B(X,Y)|\,\mathbf{y}] - s(\mathbf{y}) - \eta(\mathbf{y}_0) \tag{23}$$

With a Kullback-Leibler cost function, the denominator of $\mu(\mathbf{y})$ and the likelihood ratio $\frac{1}{p(\mathbf{y})}$ cancel exactly. There is a good reason for this based in information geometry: a non-parametric distribution belongs to the family of mixture distributions, and the Kullback-Leibler divergence is the canonical divergence for this family. All canonical divergences have the property that $\frac{\partial^2 D(p||q)}{\partial a_i \partial a_j} = g_{ij}(p)$ for coordinate system (i.e., parameterization) $\mathbf{a}$, where $g_{ij}(p)$ is ij^{th} entry of the Fisher information matrix. In our non-parametric setup, a canonical divergence satisfies $\frac{\partial^2 D(p||q)}{\partial a_i \partial a_j} = \frac{\partial^2 D(p||q)}{\partial p(\mathbf{y})\partial p(\tilde{\mathbf{y}})}$ and $g_{ij}(p_Y) = \frac{1}{p(\mathbf{y})} \cdot \mathbb{1}_{\tilde{\mathbf{y}}=\mathbf{y}}$. It follows that for the canonical (Kullback-Leibler) divergence cost function, the denominator of $\mu(\mathbf{y})$ is the likelihood ratio.[8]

Having established that the second-best contract and action are monotone under the Kullback-Leibler divergence, we now consider how the second-best solution compares to the first-best. The following lemma characterizes the first-best action; i.e., the one the principal would select absent moral hazard.

Lemma 4 *For any $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$ in supp$(\mathcal{X} \times \mathcal{Y})$, the first-best action satisfies*

$$f'\left(\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right) - f'\left(\frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}\right) \;=\; B(\mathbf{x},\mathbf{y}) - B(\mathbf{x}',\mathbf{y}'). \tag{24}$$

In particular, $\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})} - \frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}$ is increasing in $B(\mathbf{x},\mathbf{y}) - B(\mathbf{x}',\mathbf{y}')$.

There are two sources of departure from first best: the inability to contract on X and the agent's risk aversion. To examine how these frictions impact the equilibrium

[8]See Amari and Nagaoka (2000) for details.

action, it will be convenient to construct a measure of risk aversion based on secant approximations of the Arrow-Pratt coefficient of absolute risk aversion. Specifically, for any $(\mathbf{x},\mathbf{y})$ and $(\mathbf{x}',\mathbf{y}')$ in $\text{supp}(\mathcal{X}\times\mathcal{Y})$, define $u'(\mathbf{y},\mathbf{y}')$ as the slope of the secant line connecting the points $U(s(\mathbf{y}))$ and $U(s(\mathbf{y}'))$, and define $u''(\mathbf{y},\mathbf{y}')$ as the slope of the secant line connecting the points $U'(s(\mathbf{y}))$ and $U'(s(\mathbf{y}'))$. Note that $u'(\mathbf{y},\mathbf{y}')$ and $u''(\mathbf{y},\mathbf{y}')$ respectively converge to $U'(s(\mathbf{y}))$ and $U''(s(\mathbf{y}))$ as $s(\mathbf{y}')$ converges to $s(\mathbf{y})$. Our measure of risk aversion is then

$$-\frac{u''(\mathbf{y},\mathbf{y}')}{u'(\mathbf{y},\mathbf{y}')} \equiv -\frac{(U'(s(\mathbf{y}))-U'(s(\mathbf{y}'))/(s(\mathbf{y})-s(\mathbf{y}'))}{(U(s(\mathbf{y}))-U(s(\mathbf{y}'))/(s(\mathbf{y})-s(\mathbf{y}'))} = -\frac{U'(s(\mathbf{y}))-U'(s(\mathbf{y}'))}{U(s(\mathbf{y}))-U(s(\mathbf{y}'))}. \tag{25}$$

Given these definitions, the following proposition characterizes the second-best action.

Proposition 2 *Let $D_f(p||q)$ be the Kullback-Leibler divergence. Then for any $(\mathbf{x},\mathbf{y})$ and $(\mathbf{x}',\mathbf{y}')$ in supp$(\mathcal{X}\times\mathcal{Y})$, the second-best action satisfies*

$$f'\left(\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right) - f'\left(\frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}\right) = \frac{\mathbb{E}[B(X,Y)|\,\mathbf{y}]-\mathbb{E}[B(X,Y)|\,\mathbf{y}']}{\frac{U'(s(\mathbf{y}))U'(s(\mathbf{y}'))}{u'(\mathbf{y},\mathbf{y}')}-\frac{u''(\mathbf{y},\mathbf{y}')}{u'(\mathbf{y},\mathbf{y}')}} \cdot U'(s(\mathbf{y}))U'(s(\mathbf{y}')). \tag{26}$$

In particular, $\left|\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})} - \frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}\right|$ *is increasing in* $|\mathbb{E}[B(X,Y)|\,\mathbf{y}] - \mathbb{E}[B(X,Y)|\,\mathbf{y}']|$ *and is decreasing in* $-\frac{u''(\mathbf{y},\mathbf{y}')}{u'(\mathbf{y},\mathbf{y}')}$.

Intuitively, risk-aversion causes a reduction in $|U(s(\mathbf{y})) - U(s(\mathbf{y}'))|$, and therefore in $\left|\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})} - \frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}\right|$, relative to the first-best, thereby reducing the risk borne by the agent. However, risk neutrality is not sufficient to restore first-best production. Specifically, setting $U(s) = s$ causes (26) to reduce to

$$f'\left(\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right) - f'\left(\frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}\right) = \mathbb{E}[B(X,Y)|\,\mathbf{y}] - \mathbb{E}[B(X,Y)|\,\mathbf{y}'], \tag{27}$$

which is equivalent to (24) if and only if $\mathbb{E}[B(X,Y)|\,\mathbf{y}] - \mathbb{E}[B(X,Y)|\,\mathbf{y}'] = B(\mathbf{x},\mathbf{y}) - B(\mathbf{x}',\mathbf{y}')$. This is true for all $(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$ if and only if Y perfectly reveals $B(X,Y)$; i.e., if value or a perfect measure of value is contractible.

5 Extensions

5.1 Equivalence to a rich multi-tasking model

The generalized distribution approach expands the agent's action space relative to the classic approach; in Holmström (1979), the agent has a one-dimensional action, a, whereas in the generalized approach, the agent's action has the same dimensionality as $\mathcal{X} \times \mathcal{Y}$. It is therefore useful to relate the generalized approach to multi-tasking models, which also expand the dimensionality of the agent's control space by allowing the agent to choose a vector rather than a scalar.

In classic multi-tasking models (e.g., Holmström and Milgrom 1991), the agent chooses a vector of parameters, $\mathbf{a} = (a_1, ..., a_m)$ at some cost $V(\mathbf{a})$. We can interpret the generalized approach as a particular parameterization of a multi-tasking model in which $\mathbf{a} = \{p(\mathbf{x}, \mathbf{y})\}_{\mathcal{X} \times \mathcal{Y}}$ at cost $D_f(\mathbf{a}||q)$; that is, the choice of each probability is a "task." However, it turns out that setting $\mathbf{a} = \{p(\mathbf{x}, \mathbf{y})\}_{\mathcal{X} \times \mathcal{Y}}$ is not necessary for our results. As we show in this section, the generalized approach is equivalent to *any* parameterization of a multitasking model in which the agent's action space is rich enough to span the entire simplex $\mathcal{P}(\mathcal{X}, \mathcal{Y})$.

Let $\mathbf{a} = (a_1, \ldots, a_m) \in \mathcal{A} \subset \mathbb{R}^m$ be a vector of hidden actions chosen by the agent. Let $\mathbf{a}$ parameterize the distribution $p(\mathbf{a})$, and let $p(\mathbf{x}, \mathbf{y}; \mathbf{a})$ denote the probability of $(\mathbf{x}, \mathbf{y})$ under the distribution $p(\mathbf{a})$. Assume that for each $p \in \mathcal{P}(\mathcal{X}, \mathcal{Y})$, there exists at least one action $\mathbf{a} \in \mathcal{A}$ such that $p(\mathbf{a}) = p$; that is, the action space $\mathcal{A}$ is rich enough to make every distribution accessible at some cost. Let $V(\mathbf{a}) \equiv D_f(p(\mathbf{a})||q)$ be the agent's personal cost of taking action $\mathbf{a}$. Then faced with a contract $s(Y)$, the agent chooses $\mathbf{a}$ to maximize the following program.

$$\max_{\mathbf{a}} \quad \sum_{\mathcal{X} \times \mathcal{Y}} U(s(\mathbf{y}))p(\mathbf{x}, \mathbf{y}; \mathbf{a}) - D_f(p(\mathbf{a})||q) \tag{28}$$

The first-order condition with respect to a_i is as follows, where $p^i(\mathbf{x}, \mathbf{y}; \mathbf{a})$ denotes the partial derivative of $p(\mathbf{x}, \mathbf{y}; \mathbf{a})$ with respect to a_i and $\nu \in \mathbb{R}$ is an arbitrary real number.[9]

$$0 \;=\; \sum_{\mathcal{X}}\sum_{\mathcal{Y}}\left[U(s(\mathbf{y})) - \nu - f'\left(\tfrac{p(\mathbf{x},\mathbf{y};\mathbf{a})}{q(\mathbf{x},\mathbf{y})}\right)\right] p^i(\mathbf{x},\mathbf{y};\mathbf{a}) \tag{29}$$

Note that (14) implies (29) for $\nu = U(s(\mathbf{y}_0)) - f'\left(\frac{p(\mathbf{x}_0,\mathbf{y}_0)}{q(\mathbf{x}_0,\mathbf{y}_0)}\right)$; that is, the generalized approach yields a first-order condition in which (29) holds term-by-term. It is straightforward to show that this solution to (29) is globally optimal.

Lemma 5 *The solution to the agent's nonparametric program* (13) *characterized by Lemma 1 also solves the agent's parametric program* (28)*.*

Given Lemma 5, we can replace the first-order conditions (29) with (14) and (15), which allows the principal's program to be written analogously to (16).

$$\max_{s(Y),\mathbf{a}} \sum_{\mathcal{Y}}\left(\sum_{\mathcal{X}} B(\mathbf{x},\mathbf{y})q(\mathbf{x}\,|\,\mathbf{y}) - s(\mathbf{y})\right) p(\mathbf{y};\mathbf{a}) \tag{30}$$

$$\text{s.t.} \sum_{\mathcal{Y}} U(s(\mathbf{y}))p(\mathbf{y};\mathbf{a}) - \sum_{\mathcal{Y}} q(\mathbf{y}) f\left(\frac{p(\mathbf{y};\mathbf{a})}{q(\mathbf{y})}\right) \geq \bar{U} \tag{IR}$$

$$U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) = f'\left(\frac{p(\mathbf{y};\mathbf{a})}{q(\mathbf{y})}\right) - f'\left(\frac{p(\mathbf{y}_0;\mathbf{a})}{q(\mathbf{y}_0)}\right) \;\; \forall\, \mathbf{y} \neq \mathbf{y}_0 \tag{IC$_\mathbf{y}$}$$

$$p(\mathbf{y}_0;\mathbf{a}) = 1 - \sum_{\mathcal{Y}\setminus \mathbf{y}_0} p(\mathbf{y};\mathbf{a}) \tag{p.m.f}$$

Because $\mathbf{a}$ enters this program only through $p(\mathbf{a})$, the optimal distribution p and contract $s(Y)$ must be independent of the parameterization, and therefore identical to the nonparametric solution. More formally, after substituting (p.m.f) into the objective function and other constraints, the principal's first-order conditions with respect to a_i and $s(\mathbf{y})$ for $\mathbf{y} \neq \mathbf{y}_0$ are

$$0 = \sum_{\mathcal{Y}}\left[\mathbb{E}[B(X,Y)|\,\mathbf{y}] - s(\mathbf{y}) - \mu(\mathbf{y})\frac{1}{q(\mathbf{y})}f''\left(\frac{p(\mathbf{y};\mathbf{a})}{q(\mathbf{y})}\right) - \eta(\mathbf{y}_0;\mathbf{a})\right] p^i(\mathbf{y};\mathbf{a}) \tag{31}$$

$$0 = -\,p(\mathbf{y};\mathbf{a}) + \lambda U'(s(\mathbf{y}))p(\mathbf{y};\mathbf{a}) + \mu(\mathbf{y})U'(s(\mathbf{y})), \tag{32}$$

[9]Note that the right-hand side of (29) is identical for every real number ν because $\sum_{\mathcal{X}\times\mathcal{Y}} p(\mathbf{x},\mathbf{y};\mathbf{a}) = 1$ implies that $\sum_{\mathcal{X}\times\mathcal{Y}} p^i(\mathbf{x},\mathbf{y};\mathbf{a}) = 0$.

where $\eta(\mathbf{y}_0;\mathbf{a}) \equiv \mathbb{E}[B(X,Y)|\,\mathbf{y}_0] - s(\mathbf{y}_0) + \frac{1}{q(\mathbf{y}_0)} f''\left(\frac{p(\mathbf{y}_0;\mathbf{a})}{q(\mathbf{y}_0)}\right) \sum_{\mathcal{Y}\setminus\mathbf{y}_0} \mu(\tilde{\mathbf{y}})$. In particular, (18) and (32) are identical whereas (17) implies (31); that is, the generalized approach yields a first-order condition in which (31) holds term-by-term. It is straightforward to show that this solution to (31) is globally optimal.

Proposition 3 *The solution to the principal's nonparametric program* (16) *characterized by Proposition 1 also solves the principal's parametric program* (30)*.*

In sum, we can interpret the generalized approach as a reduced-form version of any parametric multi-tasking model in which the entire simplex $\mathcal{P}(\mathcal{X},\mathcal{Y})$ is accessible to the agent. Thus, it is not the agent's ability to directly choose probabilities that matters for our results, but rather that the agent's action space is sufficiently rich. In the following extension, we consider the robustness of our results when the agent's action space is constrained.

5.2 Exogenously-distributed signals

Up to this point, we have allowed the agent to choose the joint distribution over all contractible and non-contractible signals. In this section we investigate the robustness of our results to cases in which some signals are exogenously distributed.

Assume that there are two vectors of exogenously-distributed variables that may be valuable to the principal; $W \in \mathcal{W}$ that is not contractible and $Z \in \mathcal{Z}$ that is contractible. Let $q_{WZ|XY}$ be an exogenous conditional distribution over (W,Z) conditional on (X,Y); that is, the agent still chooses $p_{XY} \in \mathcal{P}(\mathcal{X},\mathcal{Y})$, but he is limited by $q_{WZ|XY}$ to a subset of the simplex $\mathcal{P}(\mathcal{W},\mathcal{X},\mathcal{Y},\mathcal{Z})$. We assume that $q_{WZ|XY}$ has full support, and we restrict our attention to cases in which an interior p_{XY} is optimal.

The derivation of the optimal contract and action closely follows the analysis in Section 4. For an arbitrary $(\mathbf{w}_0,\mathbf{x}_0,\mathbf{y}_0,\mathbf{z}_0)$, the optimal contract is characterized by the following proposition.

Proposition 4 *For all* $(\mathbf{y},\mathbf{z}) \neq (\mathbf{y}_0,\mathbf{z}_0)$*, the optimal contract satisfies*

$$\frac{1}{U'(s(\mathbf{y},\mathbf{z}))} = \lambda + \sum_{\mathcal{W}\times\mathcal{X}} \frac{\mathbb{E}_{WZ}[B(W,X,Y,Z)-s(Y,Z)|\,\mathbf{x},\mathbf{y}]-\eta_0}{\check{f}\left(\sum_{\mathcal{Z}} U(s(\mathbf{y},\tilde{\mathbf{z}}))q(\tilde{\mathbf{z}}|\,\mathbf{x},\mathbf{y})-\nu_0\right)} \cdot p(\mathbf{w},\mathbf{x}\,|\,\mathbf{y},\mathbf{z}), \tag{33}$$

where λ*,* ν_0*, and* η_0 *are constant in* $(\mathbf{y},\mathbf{z})$ *and* $\check{f}(\cdot) \equiv f'^{-1}(\cdot) f''\left(f'^{-1}(\cdot)\right)$*.*

Define *controllable residual value* as $\mathbb{E}_{WZ}[B(W,X,Y,Z) - s(Y,Z)|X,Y]$, which is the expected residual value conditional on the variables under the agent's control. We say that a signal Y_i or Z_i is *informative about controllable residual value* if

$$\begin{aligned} &\mathbb{E}_{WX}[\mathbb{E}_{WZ}[B(W,X,Y,Z)|X,Y]|Y,Z] \neq \mathbb{E}_{WX}[\mathbb{E}_{WZ}[B(W,X,Y,Z)|X,Y]|Y_{-i},Z] \\ \text{or } &\mathbb{E}_{WX}[\mathbb{E}_{WZ}[B(W,X,Y,Z)|X,Y]|Y,Z] \neq \mathbb{E}_{WX}[\mathbb{E}_{WZ}[B(W,X,Y,Z)|X,Y]|Y,Z_{-i}], \end{aligned}$$

respectively. The optimal contract in Proposition 4 is an expectation of weighted controllable residual values, where the weights are the reciprocal of the term in the denominator, $\check{f}\left(\sum_{\mathcal{Z}} U(s(\mathbf{y},\tilde{\mathbf{z}}))q(\tilde{\mathbf{z}}|\,\mathbf{x},\mathbf{y}) - \nu_0\right)$. If this term is constant in $\mathbf{x}$, then the weights are all equal and the optimal contract is a function of the expected controllable residual value conditional on the observable signals. A sufficient condition for this to be the case is that $D_f(p||q)$ is the Kullback-Leibler divergence.

Corollary 5 *If $D_f(p||q)$ is the Kullback-Leibler divergence, then a signal Y_i or Z_i is useful if and only if it is informative about controllable residual value.*

When $D_f(p||q)$ is the Kullback-Leibler divergence, the denominator on the right-hand side of (33) is equal to one, and we are left with the nested expectation $\mathbb{E}_{WX}[\mathbb{E}_{WZ}[B(W,X,Y,Z)|X,Y]|\,\mathbf{y},\mathbf{z}]$ minus a constant. The inner expectation gives the principal's expected residual conditioned on a particular realization of the controllable variables X and Y; it prevents noise in the uncontrollable variables W and Z from entering the contract, filtering out fluctuations in the principal's residual value that the agent cannot control. The outer expectation is required because X and W are not observed; it uses the observable variables Y and Z to give an estimate of the controllable residual value.

Proposition 4 is a generalization of our main result. When W and Z are the empty set, we are back to the setting in which the principal's objective is $B(X,Y)$; in that case, the controllable residual value is $\mathbb{E}[\mathbb{E}[B(X,Y) - s(Y)|X,Y]|\,\mathbf{y}] = \mathbb{E}[B(X,Y) - s(Y)|\,\mathbf{y}] = \mathbb{E}[B(X,Y)|\,\mathbf{y}] - s(\mathbf{y})$, which is what appears on the right-hand side of (22). That is, $\mathbb{E}[B(X,Y)|\,\mathbf{y}] - s(\mathbf{y})$ is an expectation over the controllable residual value, but where all variables are controlled.

6 Discussion and related literature

6.1 The generalized distribution approach

In a review of the early agency literature, Hart and Holmström (1987) identified three approaches to modeling the moral hazard problem. The first is the *state space formulation* of early agency work (e.g. Wilson 1968), where output is jointly determined by the agent's action and a random state of nature. The second is the *parameterized distribution formulation* of the agency problem, wherein the agent's action is represented as a parameter in the distribution over one or more random variables; this approach has dominated the literature since the seminal work of Holmström (1979). Finally, Hart and Holmström describe the *generalized distribution* approach as follows.

> The third, most abstract formulation is the following. Since the agent in effect chooses among alternative distributions, one is naturally led to take the distributions themselves as the actions, dropping the reference to a . . . Of course, the economic interpretation of the agent's action and the incurred cost is obscured in this *generalized distribution formulation*, but in return one gets a very streamlined model of particular use in understanding the formal structure of the problem. (Hart and Holmström 1987, pp.78-79.)

When Hart and Holmström (1987) wrote their review, the only paper to have used the generalized distribution approach was Holmström and Milgrom (1987), who provide two examples to motivate the assumption that the agent can choose distributions directly. First, imagine that an agent chooses a single action conditional on a rich set of private information in a static game; then the space of contingent effort strategies maps to the space of nonparametric unconditional distributions at the outset, before any private information is revealed. Second, imagine that an agent acts continuously throughout the contracting period, conditioning his action on a continuously observed state variable; Holmström and Milgrom (1987) argue that this setting can be represented in reduced form as the agent choosing the unconditional distribution at the outset.[10] Notice that both of these motivations involve giving

[10] Hébert (2018) shows this formally; he microfounds the generalized approach by

the agent more options, and that the generalized approach does greatly expand the dimensionality of the agent's action space relative to the classic approach.

Not many papers have employed the generalized approach since Holmström and Milgrom (1987), but there have been a few in recent years. Most closely related to ours is Hébert (2018), who studies a setting in which a seller (the agent) chooses the distribution over an asset and designs a security that gives a buyer (the principal) some claim on the asset's value realization. (In our paper it is the principal who designs the contract, but this does not change the fundamental nature of the agency problem.) Like us, Hébert (2018) uses divergences to model the agent's cost function. Our paper differs from Hébert (2018) in two important respects. First, we model a risk-averse agent with unlimited liability, whereas Hébert assumes a risk-neutral agent with limited liability. Second, the focus of our paper is different, and consequently, the set of cost functions under which our main result holds is different, as we discuss further in section 6.2.

Another recent paper, Bonham (2020), uses the generalized approach to study how measurement and contracts shape productive incentives. In that paper, an agent with limited liability has distributional control only over value, which is assumed to be non-contractible, and the relationship between value and the contractible signals is completely exogenous to the agent's choice; this is akin to letting $B(W, X, Y, Z) = X_i$ in our extended model in section 5.2, but where only the exogenous signal Z_i is contractible. Bonham (2020) focuses on how the exogenous mapping from X_i to Z_i (i.e. the measurement of value) shapes the equilibrium distribution and the optimal contract.

A concurrent paper, Georgiadis, Ravid, and Szentes (2022), also studies optimal contracting under the generalized approach. The premise of their paper is similar to ours – to study optimal contracting under moral hazard when the agent has a flexible action space – and they give a binary example with an IC constraint that

showing that a continuous time model in which an agent with quadratic cost controls the drift of a Brownian motion is equivalent to a static model wherein an agent with a Kullback Leibler divergence cost function chooses a probability distribution nonparametrically.

is virtually identical to ours.[11] The two papers diverge significantly in focus and in the specification of the general model. Georgiadis, Ravid, and Szentes (2022) assume a single contractible variable, and they are largely focused on the existence of contracts that implement each distribution as well as the monotonicity of the agent's wage. On the other hand, our paper is focused on the value of information, and we therefore study a multivariate setting in which value cannot be directly observed. To make this setting tractable, we model the agent's cost function as an f-divergence, whereas Georgiadis, Ravid, and Szentes (2022) do not restrict the agent's cost function to take a particular form, but instead impose the assumptions of smoothness and monotonicity.

Garrett, Georgiadis, Smolin, and Szentes (2022) also model the agent as choosing a distribution, but allow the agent to choose his cost function ex ante. In our model, this would be akin to the agent choosing the exogenous distribution q before interacting with the principal, except that Garrett et al. (2022) do not impose a functional form on the cost function. They show that the optimal cost function is one that makes all non-binary distributions infinitely costly.

In addition to the papers mentioned above, we know of a few others that use the generalized distribution approach. Hellwig (2007) uses it to extend Holmström and Milgrom (1987) to include boundary solutions. Bertomeu (2008) uses it to study risk management. Hemmer (2017) studies relative performance evaluation using a binary version of the generalized approach in which the agent directly chooses the probability of the principal's preferred outcome. Diamond (1998) takes a related but more restrictive approach in which the agent exerts costly effort to generate a set of distributions with equal means and then costlessly selects an element from that set. Carroll (2015) models actions as distribution-cost pairs to study *robust contracts*, i.e. those that give the best worst-case guarantee when the principal does not know what distribution-cost pairs are available to the agent; by contrast, we assign the agent's cost of choosing distributions a functional form (that of an f-divergence), and we assume that this cost function is common knowledge.

[11]See equation (4) of our paper and equation (1) in Georgiadis, Ravid, and Szentes (2022).

6.2 Divergence-based cost functions

The rational inattention literature uses divergences to model the cost of processing or paying attention to information. Much of the literature has followed Sims (1998, 2003) by modeling information costs as the expected reduction in *Shannon mutual information*, which is the Kullback-Leibler (KL) divergence of a joint distribution from the product of its marginals. However, many behavioral implications of the Shannon model were found to be inconsistent with experimental evidence (for example, see Dean and Neligh 2019), and consequently, recent papers have generalized to broader classes of cost functions. Variations of decomposability and invariance are prevalent in these generalizations. For example, Caplin, Dean, and Leahy (2022) study *uniformly posterior-separable* cost functions that are additively separable in both the prior and the posterior, and *invariant posterior-separable* cost functions, which specify that the least costly way to learn about an event is to learn nothing additional about the relative probabilities about the states that make up the event.

Like Hébert (2018), we use divergences to model the cost of an agent's hidden action in a moral hazard model. Hébert (2018) is primarily interested in the shape of the contract – specifically, the optimality of debt. He shows that debt contracts are *exactly* optimal when the agent's cost function is proportional to KL-divergence, and that debt is *approximately* optimal for all invariant divergences. By contrast, we focus on the contracting value of information. Our main result – that the contract only varies with information about the principal's expected payoff – exactly holds for all f-divergences, which are invariant divergences that are also additively separable. Thus, our main result exactly holds for a broader set of cost functions than Hébert's (the f-divergences rather than the KL), but Hébert's main result approximately holds for a broader class than we can show (all invariant divergences rather than the f-divergences).[12]

[12]It is possible that our main result holds for all invariant divergences; however, the analysis is intractable without the additional axiom of Decomposability. Under this axiom, the agent's cost of choosing the probability $p_{XY}(\mathbf{x}, \mathbf{y})$ is independent of $p_{XY}(\mathbf{x}', \mathbf{y}')$ for all $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$. Note that this is consistent with the notion of giving the agent a rich action space, as it allows the agent the flexibility to change

Some papers employing the generalized approach also use divergence-based cost functions, but the authors do not explicitly identify them as such. Bonham (2020) assumes a χ^2-divergence cost function, which is a special case of an f-divergence in which the separable cost of choosing each probability is quadratic. Bertomeu (2008) measures the distance between two distributions as the integrated squared error, which is a divergence but is not invariant.

7 Concluding remarks

Holmström's informativeness principle says that a signal is valuable for contracting if it is informative about the agent's action, and it is considered one of the most robust results in agency theory (Bolton and Dewatripont 2005). We return to the setting in Holmström's seminal 1979 paper and apply an assumption pioneered in his 1987 paper—the generalized distribution approach. Under this approach, a signal is useful for contracting if it changes inferences about outcomes the principal values, rather than about the agent's action.

This finding may help to explain executive compensation practice. The informativeness principle—derived in a model where the agent chooses a single parameter that is often interpreted as effort—predicts that all signals that are incrementally informative about a CEO's effort should be included in his compensation contract. There are many readily available signals that firms could use to measure effort (such as hours worked, meetings held, or emails sent), but in reality executive contracts tend to instead include measures that are informative about firm value.[13] Perhaps this is because CEO actions are better described as choosing distributions than as supplying a one-dimensional input like effort.

one probability without affecting the cost of the others.

[13]For example, De Angelis and Grinstein (2015) document that 92 percent of performance-based awards are contingent on accounting metrics or stock performance.

References

Amari, Shun-ichi. 2016. *Information geometry and its applications.* Springer.

Amari, Shun-ichi, and Hiroshi Nagaoka. 2000. *Methods of information geometry.* Vol. 191. American Mathematical Soc.

Bertomeu, Jeremy. 2008. Risk management and the design of efficient incentive schemes. PhD diss., Carnegie Mellon University.

Bolton, Patrick, and Mathias Dewatripont. 2005. *Contract theory.* MIT press.

Bonham, Jonathan. 2020. Shaping incentives through measurement and contracts. Working Paper.

Caplin, Andrew, Mark Dean, and John Leahy. 2022. Rationally inattentive behavior: Characterizing and generalizing Shannon entropy. *Journal of Political Economy* 130 (6): 1676–1715.

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

De Angelis, David, and Yaniv Grinstein. 2015. Performance terms in CEO compensation contracts. *Review of Finance* 19 (2): 619–651.

Dean, Mark, and Nate Leigh Neligh. 2019. Experimental tests of rational inattention. *Columbia mimeo.*

Diamond, Peter. 1998. Managerial incentives: On the near linearity of optimal compensation. *Journal of Political Economy* 106 (5): 931–957.

Garrett, Daniel, George Georgiadis, Alex Smolin, and Bálazs Szentes. 2022. Optimal technology design. Working paper.

Georgiadis, George, Doron Ravid, and Balázs Szentes. 2022. Flexible moral hazard problems. Working Paper.

Hart, Oliver, and Bengt Holmström. 1987. The theory of contracts. In *Advances in Economic Theory: Fifth World Congress,* vol. 1. Cambridge University Press.

Hébert, Benjamin. 2018. Moral hazard and the optimality of debt. *The Review of Economic Studies* 85 (4): 2214–2252.

Hébert, Benjamin, and Jennifer La'O. 2022. Information acquisition, efficiency, and non-fundamental volatility. Working paper.

Hellwig, Martin F. 2007. The role of boundary solutions in principal–agent problems of the Holmström–Milgrom type. *Journal of Economic Theory* 136 (1): 446–475.

Hemmer, Thomas. 2017. Optimal dynamic relative performance evaluation. Working Paper.

Holmström, Bengt. 1979. Moral hazard and observability. *The Bell Journal of Economics* 10 (1): 74–91.

Holmström, Bengt, and Paul Milgrom. 1987. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica* 55 (2): 303–328.

———. 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics and Organization* 7:24–52.

Mirrlees, James. (1975) 1999. The theory of moral hazard and unobservable behaviour. Part I. *The Review of Economic Studies* 66 (1): 3–21. Paper completed 1975.

Sims, Christopher A. 1998. Stickiness. In *Carnegie-rochester conference series on public policy,* 49:317–356. Elsevier.

———. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50 (3): 665–690.

Wilson, Robert. 1968. The theory of syndicates. *Econometrica* 36 (1): 119–132.

A Proofs

Proof of Lemma 1. Substitute the constraint $p(\mathbf{x}_0, \mathbf{y}_0) = 1 - \sum_{\mathcal{X} \setminus \mathbf{x}_0} \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\mathbf{x}, \mathbf{y})$ into the objective function and other constraints in (13) to obtain the following program.

$$\begin{aligned} \max_p \quad & \sum_{\mathcal{X} \times \mathcal{Y} \setminus (\mathbf{x}_0, y_0)} \left(U(s(\mathbf{y}))p(\mathbf{x}, \mathbf{y}) - q(\mathbf{x}, \mathbf{y}) f\left(\frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x}, \mathbf{y})}\right)\right) \\ & + U(s(\mathbf{y}_0)) \left(1 - \sum_{\mathcal{X} \times \mathcal{Y} \setminus (\mathbf{x}_0, \mathbf{y}_0)} p(\mathbf{x}, \mathbf{y})\right) \\ & - q(\mathbf{x}_0, \mathbf{y}_0) f\left(\frac{1 - \sum_{\mathcal{X} \times \mathcal{Y} \setminus (\mathbf{x}_0, \mathbf{y}_0)} p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x}_0, \mathbf{y}_0)}\right) \\ \text{s.t.} \quad & p(\mathbf{x}, \mathbf{y}) \geq 0 \quad \forall (\mathbf{x}, \mathbf{y}) \neq (\mathbf{x}_0, \mathbf{y}_0) \\ & 1 - \sum_{\mathcal{X} \setminus \mathbf{x}_0} \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\mathbf{x}, \mathbf{y}) \geq 0 \end{aligned} \tag{34}$$

Because $D_f(p||q)$ is convex whereas the other terms in this program are linear in p, it follows that the Lagrangian is concave and, therefore, that the first-order conditions are necessary and sufficient for a solution. Let $\delta(\mathbf{x}, \mathbf{y})$ and $\delta(\mathbf{x}_0, \mathbf{y}_0)$ denote the Lagrange multipliers on the constraints. Then the first-order condition with respect to $p(\mathbf{x}, \mathbf{y})$ is

$$f'\left(\frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x}, \mathbf{y})}\right) - f'\left(\frac{p(\mathbf{x}_0, \mathbf{y}_0)}{q(\mathbf{x}_0, \mathbf{y}_0)}\right) = U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) + \delta(\mathbf{x}, \mathbf{y}) - \delta(\mathbf{x}_0, \mathbf{y}_0). \tag{35}$$

Because $(\mathbf{x}_0, \mathbf{y}_0)$ is selected from the set $\{\mathcal{X} \times \mathcal{Y} \,|p(\mathbf{x}_0, \mathbf{y}_0) > 0\}$, it follows that $\delta(\mathbf{x}_0, \mathbf{y}_0) = 0$. There are then two remaining possibilities: either $p(\mathbf{x}, \mathbf{y}) = 0$ or $p(\mathbf{x}, \mathbf{y}) > 0$. If $p(\mathbf{x}, \mathbf{y}) = 0$, then (35) reduces to $f'(0) - f'\left(\frac{p(\mathbf{x}_0, \mathbf{y}_0)}{q(\mathbf{x}_0, \mathbf{y}_0)}\right) - U(s(\mathbf{y})) + U(s(\mathbf{y}_0)) = \delta(\mathbf{x}, \mathbf{y}) \geq 0$, where $f'(0) \equiv \lim_{u \to 0^+} f'(u)$. Adding $U(s(\mathbf{y})) - U(s(\mathbf{y}_0))$ to both sides of this inequality implies that $p(\mathbf{x}, \mathbf{y}) = 0 \iff U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) \leq f'(0) - f'\left(\frac{p(\mathbf{x}_0, \mathbf{y}_0)}{q(\mathbf{x}_0, \mathbf{y}_0)}\right)$.

On the other hand, if $p(\mathbf{x}, \mathbf{y}) > 0$ then $\delta(\mathbf{x}, \mathbf{y}) = 0$ and (35) reduces to $f'\left(\frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{x}, \mathbf{y})}\right) - f'\left(\frac{p(\mathbf{x}_0, \mathbf{y}_0)}{q(\mathbf{x}_0, \mathbf{y}_0)}\right) = U(s(\mathbf{y})) - U(s(\mathbf{y}_0))$. □

Proof of Lemma 2. Let $s^*(Y)$ be an optimal contract and suppose by way of contradiction that IR does not bind. Then there exists some $\epsilon > 0$ sufficiently small that $U(s^*(Y)) - \epsilon$ is also individually rational. Moreover, because U is strictly increasing, it is invertible in a neighborhood of $s^*(Y)$. Taking ϵ small enough to remain in this neighborhood and not violate the IR constraint, define an alternative contract $s(\mathbf{y};\epsilon) \equiv U^{-1}(U(s^*(\mathbf{y})) - \epsilon)$. By construction, this contract pays the agent less for every realization of Y. Furthermore, this contract satisfies

$$\begin{aligned} U(s(\mathbf{y};\epsilon)) - U(s(\mathbf{y}_0;\epsilon)) &= U(s^*(\mathbf{y})) - \epsilon - U(s^*(\mathbf{y}_0)) + \epsilon \\ &= U(s^*(\mathbf{y})) - U(s^*(\mathbf{y}_0)), \end{aligned} \tag{36}$$

which implies that $s(Y;\epsilon)$ implements the same p as does $s^*(Y)$. Because $s(Y)$ maintains individual rationality and incentive compatibility at a strictly lower cost to the principal, $s^*(Y)$ is strictly dominated by $s(Y;\epsilon)$. But then $s^*(Y)$ was not optimal in the first place, a contradiction. It follows that IR must bind. □

Proof of Lemma 3. Because $U(\cdot)$ is strictly increasing it can be inverted in a neighborhood of the $(p_Y, s(Y))$ satisfying (17) and (18). Rewrite IC

$$s(\mathbf{y}) \;=\; U^{-1}\left(f'\left(\tfrac{p(\mathbf{y})}{q(\mathbf{y})}\right) - f'\left(\tfrac{p(\mathbf{y}_0)}{q(\mathbf{y}_0)}\right) + U(s(\mathbf{y}_0))\right) \;=\; U^{-1}\left(f'\left(\tfrac{p(\mathbf{y})}{q(\mathbf{y})}\right) + \nu\right) \tag{37}$$

for $\nu \equiv U(s(\mathbf{y}_0)) - f'\left(\frac{p(\mathbf{y}_0)}{q(\mathbf{y}_0)}\right)$, and substitute this into the objective function and other constraints. Then the principal's program can be rewritten as a maximization problem over $(s(\mathbf{y}_0), p_Y)$:

$$\begin{aligned}
\max_{s(\mathbf{y}_0), p_Y} \quad & \sum_{\mathcal{Y}} \left(\sum_{\mathcal{X}} B(\mathbf{x},\mathbf{y}) q(\mathbf{x} \mid \mathbf{y}) - U^{-1}\left(f'\left(\tfrac{p(\mathbf{y})}{q(\mathbf{y})}\right) + \nu\right)\right) p(\mathbf{y}) \\
\text{s.t.} \quad & \nu + \sum_{\mathcal{Y}} f'\left(\tfrac{p(\mathbf{y})}{q(\mathbf{y})}\right) p(\mathbf{y}) - \sum_{\mathcal{Y}} q(\mathbf{y}) f\left(\tfrac{p(\mathbf{y})}{q(\mathbf{y})}\right) \geq \bar{U} \\
& p(\mathbf{y}_0) = 1 - \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\mathbf{y}) \\
& \nu = U(s(\mathbf{y}_0)) - f'\left(\tfrac{p(\mathbf{y}_0)}{q(\mathbf{y}_0)}\right)
\end{aligned} \tag{38}$$

The second-order conditions are difficult to sign, but sufficient conditions on U and f can be derived to ensure that the Lagrangian is globally concave. First, note that because U^{-1} is an increasing function, the principal prefers that ν be as small as possible and therefore chooses $U(s(\mathbf{y}_0))$ to bind IR (see Lemma 2).

Second, note that this maximization program is additively separable in p_Y, so concavity in $p(\mathbf{y})$ for each $\mathbf{y} \neq \mathbf{y}_0$ implies global concavity. The objective function is clearly concave in $p(\mathbf{y})$ if $U^{-1}\left(f'\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) + \nu\right) p(\mathbf{y})$ is convex in $p(\mathbf{y})$ for any $\nu \in \mathbb{R}$ and $q(\mathbf{y}) \in (0,1)$. Dividing this condition through by $q(\mathbf{y})$ allows it to be expressed as convexity of $U^{-1}(f'(t) + \nu)t$ in $t \in (0, \infty)$. Moreover, the left-hand side of IR is clearly concave in $p(\mathbf{y})$ if $f'\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) p(\mathbf{y}) - q(\mathbf{y}) f\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)$ is concave in $p(\mathbf{y})$ for all $q(\mathbf{y}) \in (0,1)$, which again dividing through by $q(\mathbf{y})$ is equivalent to the convexity of $f(t) - tf'(t)$ in t. Finally, the equality constraints are linear in $p(\mathbf{y})$. It follows that the stated conditions are sufficient to ensure that the Lagrangian is globally concave and, therefore, that the first-order conditions are necessary and sufficient for an interior solution to (38), which is equivalent to (16). □

Proof of Proposition 1. Define $f'(0) \equiv \lim_{u \to 0^+} f'(u)$. Choose some arbitrary $\mathbf{y}_0$ and assume without loss of generality that $p(\mathbf{y}_0) > 0$ is efficient (if it is not, choose a different $\mathbf{y}_0 \in \mathcal{Y}$ for which $p(\mathbf{y}_0) > 0$ is efficient). Restricting $p(\mathbf{y}_0) = 1 - \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\mathbf{y})$ allows the principal's program to be written as follows.

$$\max_{s(Y), p_Y} \sum_{\mathcal{Y} \setminus \mathbf{y}_0} \left(\sum_{\mathcal{X}} B(\mathbf{x}, \mathbf{y}) q(\mathbf{x} \,|\, \mathbf{y}) - s(\mathbf{y}) \right) p(\mathbf{y}) + \left(\sum_{\mathcal{X}} B(\mathbf{x}, \mathbf{y}_0) q(\mathbf{x} \,|\mathbf{y}_0) - s(\mathbf{y}_0) \right) \left(1 - \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\mathbf{y}) \right) \tag{39}$$

$$\text{s.t.} \sum_{\mathcal{Y} \setminus \mathbf{y}_0} \left(U(s(\mathbf{y})) p(\mathbf{y}) - q(\mathbf{y}) f\left(\frac{p(\mathbf{y})}{q(\mathbf{y})} \right) \right) + U(s(\mathbf{y}_0)) \left(1 - \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\mathbf{y}) \right) - q(\mathbf{y}_0) f\left(\frac{1 - \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\mathbf{y})}{q(\mathbf{y}_0)} \right) \geq \bar{U} \tag{IR}$$

$$U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) = f'\left(\frac{p(\mathbf{y})}{q(\mathbf{y})} \right) - f'\left(\frac{1 - \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\tilde{\mathbf{y}})}{q(\mathbf{y}_0)} \right) \ \forall p(\mathbf{y}) > 0 \tag{$\text{IC}_{\mathbf{y}}^{+}$}$$

$$U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) \leq f'(0) - f'\left(\frac{1 - \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\tilde{\mathbf{y}})}{q(\mathbf{y}_0)} \right) \ \forall p(\mathbf{y}) = 0 \tag{$\text{IC}_{\mathbf{y}}^{-}$}$$

$$p(\mathbf{y}) \geq 0 \ \forall \, \mathbf{y} \neq \mathbf{y}_0 \tag{$\text{NN}_{\mathbf{y}}$}$$

$$1 - \sum_{\mathcal{Y} \setminus \mathbf{y}_0} p(\mathbf{y}) \geq 0 \tag{$\text{NN}_{\mathbf{y}_0}$}$$

By our selection of $\mathbf{y}_0$, ($\text{NN}_{\mathbf{y}_0}$) does not bind. Moreover, notice that ($\text{IC}_{\mathbf{y}}^{+}$) and ($\text{IC}_{\mathbf{y}}^{-}$) take the same form, except that ($\text{IC}_{\mathbf{y}}^{-}$) has an inequality rather than an equality. Without loss of generality, we can use the contract that satisfies ($\text{IC}_{\mathbf{y}}^{-}$) with equality. To see why this is without loss of generality, note that when $p(\mathbf{y}) = 0$ the principal pays $s(\mathbf{y})$ with zero probability, and $s(\mathbf{y})$ disappears from the objective function and other constraints. Therefore, the principal gets the same utility for all $s(\mathbf{y})$ satisfying ($\text{IC}^-_{\mathbf{y}}$), so she may as well choose the $s(\mathbf{y})$ that satisfies ($\text{IC}^-_{\mathbf{y}}$) with equality. Under these conditions, we can combine ($\text{IC}^+_{\mathbf{y}}$) and ($\text{IC}^-_{\mathbf{y}}$) into a single constraint set ($\text{IC}_{\mathbf{y}}$): $U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) = f'\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) - f'\left(\frac{1-\sum_{\mathcal{Y}\setminus\mathbf{y}_0} p(\tilde{\mathbf{y}})}{q(\mathbf{y}_0)}\right)$ for all $\mathbf{y} \neq \mathbf{y}_0$. Let λ, $\mu(\mathbf{y})$, and $\delta(\mathbf{y})$ be the Lagrange multipliers on (IR), ($\text{IC}_{\mathbf{y}}$), and ($\text{NN}_{\mathbf{y}}$), respectively. The Lagrangian is given by the following equation.

$$
\begin{aligned}
\mathcal{L} \;=\; & \sum_{\mathcal{Y}\setminus\mathbf{y}_0}\left(\sum_{\mathcal{X}} B(\mathbf{x},\mathbf{y})q(\mathbf{x}\,|\,\mathbf{y}) - s(\mathbf{y})\right)p(\mathbf{y}) \\
& + \left(\sum_{\mathcal{X}} B(\mathbf{x},\mathbf{y}_0)q(\mathbf{x}\,|\,\mathbf{y}_0) - s(\mathbf{y}_0)\right)\left(1 - \sum_{\mathcal{Y}\setminus\mathbf{y}_0} p(\mathbf{y})\right) \\
& + \lambda\left[\sum_{\mathcal{Y}\setminus\mathbf{y}_0}\left(U(s(\mathbf{y}))p(\mathbf{y}) - q(\mathbf{y})f\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)\right)\right. \\
& \left. + U(s(\mathbf{y}_0))\left(1 - \sum_{\mathcal{Y}\setminus\mathbf{y}_0} p(\mathbf{y})\right) - q(\mathbf{y}_0)f\left(\frac{1-\sum_{\mathcal{Y}\setminus\mathbf{y}_0} p(\mathbf{y})}{q(\mathbf{y}_0)}\right) - \bar{U}\right] \\
& + \sum_{\mathcal{Y}\setminus\mathbf{y}_0}\mu(\mathbf{y})\left[U(s(\mathbf{y})) - U(s(\mathbf{y}_0)) - f'\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) + f'\left(\frac{1-\sum_{\mathcal{Y}\setminus\mathbf{y}_0} p(\tilde{\mathbf{y}})}{q(\mathbf{y}_0)}\right)\right] \\
& + \sum_{\mathcal{Y}\setminus\mathbf{y}_0}\delta(\mathbf{y})p(\mathbf{y})
\end{aligned}
\tag{40}
$$

First-order condition with respect to $p(\mathbf{y})$: Taking the first-order condition of (40) with respect to $p(\mathbf{y})$ for some $\mathbf{y} \neq \mathbf{y}_0$ yields

$$
\begin{aligned}
0 \;=\; & \mathbb{E}[B(X,Y)|\,\mathbf{y}] - s(\mathbf{y}) - (\mathbb{E}[B(X,Y)|\,\mathbf{y}_0] - s(\mathbf{y}_0)) \\
& + \lambda\left[U(s(\mathbf{y})) - f'\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) - U(s(\mathbf{y}_0)) + f'\left(\frac{1-\sum_{\mathcal{Y}\setminus\mathbf{y}_0} p(\tilde{\mathbf{y}})}{q(\mathbf{y}_0)}\right)\right] \\
& - \mu(\mathbf{y})\frac{1}{q(\mathbf{y})}f''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) + \sum_{\mathcal{Y}\setminus\mathbf{y}_0}\mu(\tilde{\mathbf{y}})\frac{1}{q(\mathbf{y}_0)}f''\left(\frac{1-\sum_{\mathcal{Y}\setminus\mathbf{y}_0} p(\hat{\mathbf{y}})}{q(\mathbf{y}_0)}\right) + \delta(\mathbf{y}).
\end{aligned}
\tag{41}
$$

The bracketed term that is pre-multiplied by λ is equal to zero by ($\text{IC}_{\mathbf{y}}$). Substituting $p(\mathbf{y}_0) = 1 - \sum_{\mathcal{Y}\setminus\mathbf{y}_0} p(\tilde{\mathbf{y}})$ back into the above expression, defining $\eta(\mathbf{y}_0) \equiv \mathbb{E}[B(X,Y)|\,\mathbf{y}_0] - s(\mathbf{y}_0) - \frac{1}{q(\mathbf{y}_0)}f''\left(\frac{p(\mathbf{y}_0)}{q(\mathbf{y}_0)}\right)\sum_{\mathcal{Y}\setminus\mathbf{y}_0}\mu(\tilde{\mathbf{y}})$, and solving for $\mu(\mathbf{y})$ yields

$$
\mu(\mathbf{y}) \;=\; \frac{\mathbb{E}[B(X,Y)|\,\mathbf{y}] - s(\mathbf{y}) - \eta(\mathbf{y}_0) + \delta(\mathbf{y})}{\frac{1}{q(\mathbf{y})}f''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)}. \tag{42}
$$

First-order condition with respect to $s(\mathbf{y})$: Taking the first-order condition of (40) with respect to $s(\mathbf{y})$ for some $\mathbf{y} \neq \mathbf{y}_0$ yields

$$0 = -p(\mathbf{y}) + \lambda U'(s(\mathbf{y}))p(\mathbf{y}) + \mu(\mathbf{y})U'(s(\mathbf{y})) \tag{43}$$

Substituting (42) into (43) and rearranging terms yields

$$\left(\tfrac{1}{U'(s(\mathbf{y}))} - \lambda\right) \tfrac{p(\mathbf{y})}{q(\mathbf{y})} f''\left(\tfrac{p(\mathbf{y})}{q(\mathbf{y})}\right) = \mathbb{E}[B(X,Y)|\,\mathbf{y}] - s(\mathbf{y}) - \eta(\mathbf{y}_0) + \delta(\mathbf{y}) \tag{44}$$

Now, define $\nu(\mathbf{y}_0) \equiv U(s(\mathbf{y}_0)) - f'\left(\frac{p(\mathbf{y}_0)}{q(\mathbf{y}_0)}\right)$ and write ($\text{IC}_\mathbf{y}$) as $f'\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) = U(s(\mathbf{y})) - \nu(\mathbf{y}_0)$. Because $f(\cdot)$ is twice differentiable and strictly convex, $f'(\cdot)$ is strictly increasing and can therefore be inverted. It follows that $\frac{p(\mathbf{y})}{q(\mathbf{y})} = f'^{-1}(U(s(\mathbf{y})) - \nu(\mathbf{y}_0))$. Substituting this into (44) yields

$$\frac{1}{U'(s(\mathbf{y}))} = \lambda + \frac{\mathbb{E}[B(X,Y)|\,\mathbf{y}] - s(\mathbf{y}) - \eta(\mathbf{y}_0) + \delta(\mathbf{y})}{f'^{-1}(U(s(\mathbf{y})) - \nu(\mathbf{y}_0)) f''\left(f'^{-1}(U(s(\mathbf{y})) - \nu(\mathbf{y}_0))\right)} \tag{45}$$

Efficiency of a contract that ignores $\delta(\mathbf{y})$: It remains only to deal with $\delta(\mathbf{y})$, which is the Lagrange multiplier on the constraint that $p(\mathbf{y}) \geq 0$. If $p(\mathbf{y}) > 0$, then $\delta(\mathbf{y}) = 0$ by construction. Suppose then that $p(\mathbf{y}) = 0$ so that $\delta(\mathbf{y}) \geq 0$.

We begin by showing that $s(\mathbf{y})$ is increasing in $\delta(\mathbf{y})$. First, notice that because $U''(s) < 0$, $U'(s)$ is decreasing in s and, therefore, the left-hand side of (45) is increasing in s. Moreover, partially differentiating the denominator on the right-hand side with respect to s yields

$$\frac{\partial\left(f'^{-1}(U(s(\mathbf{y})) - \nu(\mathbf{y}_0)) f''\left(f'^{-1}(U(s(\mathbf{y})) - \nu(\mathbf{y}_0))\right)\right)}{\partial s(\mathbf{y})} = U'(s(\mathbf{y})) \left(1 + \frac{\frac{p(\mathbf{y})}{q(\mathbf{y})} f'''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)}{f''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)}\right), \tag{46}$$

where $\left(f'^{-1}\right)' = \frac{1}{f''\left(f'^{-1}\right)}$ follows from the inverse function theorem applied to f'. Therefore, if $\frac{tf'''(t)}{f''(t)} \geq -1$ for all $t \in (0, \infty)$, the denominator of (45) is weakly increasing in s. Because the numerator is decreasing in s, this implies that the entire right-hand side of (45) is decreasing in s.

Now consider a positive variation in $\delta(\mathbf{y})$, which causes the right-hand side of (45) to increase. In order to maintain the equality, $s(\mathbf{y})$ must be adjusted to increase the left-hand side and/or decrease the right-hand side of (45). Because we have just established that the left-hand (right-hand) side of (45) is increasing (decreasing) in s, it immediately follows that $s(\mathbf{y})$ must be adjusted upward in response to the positive variation in $\delta(\mathbf{y})$ in order to maintain the equality. Therefore, $s(\mathbf{y})$ is increasing in $\delta(\mathbf{y})$.

To complete the proof, recall that we chose s such that $(\mathrm{IC}_{\mathbf{y}}^{-})$ holds with equality; however, we also established that the principal is indifferent among the set of contracts satisfying $(\mathrm{IC}_{\mathbf{y}}^{-})$. Therefore, any contract $\bar{s}$ satisfying $\bar{s}(\mathbf{y}) \leq s(\mathbf{y})$ for all $\mathbf{y} \notin \mathrm{supp}(p_Y)$ is also efficient. Consider, then, the contract $\bar{s}(Y)$ satisfying (22). Because $s(\mathbf{y})$ is increasing in $\delta(\mathbf{y}) \geq 0$ and is equal to $\bar{s}(Y)$ when $\delta(\mathbf{y}) = 0$, it follows that $\bar{s}(\mathbf{y}) \leq s(\mathbf{y})$ and is therefore efficient for all $\mathbf{y} \notin \mathrm{supp}(p_Y)$. Because $\bar{s}(Y)$ is identical to $s(Y)$ for $\mathbf{y} \in \mathrm{supp}(p_Y)$, (22) characterizes an efficient contract over $\mathcal{Y}$. □

Proof of Corollary 1. We prove that for some $y_i' \neq y_i$, $s(\mathbf{y}_{-i}, y_i') \neq s(\mathbf{y}_{-i}, y_i)$ is necessary (i.e., Y_i is useful) if and only if $\mathbb{E}[B(X,Y)|\,\mathbf{y}_{-i}, y_i'] \neq \mathbb{E}[B(X,Y)|\,\mathbf{y}_{-i}, y_i]$ (i.e., Y_i is informative about value). Let $\nu \equiv \nu(\mathbf{y}_0)$ and $\eta \equiv \eta(\mathbf{y}_0)$, and define a function W as follows.

$$W(s) \;\equiv\; \left(\tfrac{1}{U'(s)} - \lambda\right) f'^{-1}(U(s) - \nu) f''\left(f'^{-1}(U(s) - \nu)\right) + s + \eta \tag{47}$$

It follows that (22) can be rewritten

$$\mathbb{E}[B(X,Y)|\,\mathbf{y}] \;=\; W(s(\mathbf{y})). \tag{48}$$

First, suppose that $\mathbb{E}[B(X,Y)|\,\mathbf{y}_{-i}, y_i'] \neq \mathbb{E}[B(X,Y)|\,\mathbf{y}_{-i}, y_i]$. Then (48) requires that $W(s(\mathbf{y}_{-i}, y_i')) \neq W(s(\mathbf{y}_{-i}, y_i))$, which is only possible if $s(\mathbf{y}_{-i}, y_i') \neq s(\mathbf{y}_{-i}, y_i)$. Therefore informativeness implies usefulness.

Conversely, suppose that $\mathbb{E}[B(X,Y)|\,\mathbf{y}_{-i}, y_i'] = \mathbb{E}[B(X,Y)|\,\mathbf{y}_{-i}, y_i]$. Then (48) requires that $W(s(\mathbf{y}_{-i}, y_i')) = W(s(\mathbf{y}_{-i}, y_i))$, which is trivially satisfied for $s(\mathbf{y}_{-i}, y_i') = s(\mathbf{y}_{-i}, y_i)$; that is, $s(\mathbf{y}_{-i}, y_i) \neq s(\mathbf{y}_{-i}, y_i)$ is not necessary for optimality. Therefore uninformativeness implies uselessness, or equivalently, usefulness implies informativeness. Taken together, a signal is useful if and only if it is informative about value. □

Proof of Corollary 2. Note that if $B(X, Y) = Y_i$ then

$$\mathbb{E}[B(X,Y)|\,\mathbf{y}_{-j}, y_j] \;\;=\;\; y_i \;\;=\;\; \mathbb{E}[B(X,Y)|\,\mathbf{y}_{-j}, y_j'] \tag{49}$$

for all $\mathbf{y}_{-i}$. Therefore, Y_j is not informative about value. It follows immediately from Corollary 1 that Y_j is not useful. □

Proof of Corollary 3. We begin by showing that (22) implies that $s(\mathbf{y})$ is increasing in $\mathbb{E}[B(X,Y)|\,\mathbf{y}]$. First, notice that because $U''(s) < 0$, $U'(s)$ is decreasing in s and, therefore, the left-hand side of (22) is increasing in s. Moreover, partially differentiating the denominator on the right-hand side with respect to s yields

$$\frac{\partial\left(f'^{-1}(U(s(\mathbf{y}))-\nu(\mathbf{y}_0))f''\left(f'^{-1}(U(s(\mathbf{y}))-\nu(\mathbf{y}_0))\right)\right)}{\partial s(\mathbf{y})} \;=\; U'(s(\mathbf{y}))\left(1+\frac{\frac{p(\mathbf{y})}{q(\mathbf{y})}f'''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)}{f''\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right)}\right), \qquad (50)$$

where $\left(f'^{-1}\right)' = \frac{1}{f''\left(f'^{-1}\right)}$ follows from the inverse function theorem applied to f'. Therefore, if $\frac{tf'''(t)}{f''(t)} \geq -1$ for all $t \in (0,\infty)$, the denominator of (22) is weakly increasing in s. Because the numerator is decreasing in s, this implies that the entire right-hand side of (22) is decreasing in s.

Now consider a positive variation in $\mathbb{E}[B(X,Y)|\,\mathbf{y}]$, which causes the right-hand side of (22) to increase. In order to maintain the equality, $s(\mathbf{y})$ must be adjusted to increase the left-hand side and/or decrease the right-hand side of (22). Because we have just established that the left-hand (right-hand) side of (22) is increasing (decreasing) in s, it immediately follows that $s(\mathbf{y})$ must be adjusted upward in response to the positive variation in $\mathbb{E}[B(X,Y)|\,\mathbf{y}]$ in order to maintain the equality. Therefore, $s(\mathbf{y})$ is increasing in $\mathbb{E}[B(X,Y)|\,\mathbf{y}]$.

Turning to $\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}$, because $s(\mathbf{y})$ is increasing in $\mathbb{E}[B(X,Y)|\,\mathbf{y}]$, the function $W(s)$ from equations (47) and (48) is invertible. Therefore, we can write

$$s(\mathbf{y}) \;=\; W^{-1}(\mathbb{E}[B(X,Y)|\,\mathbf{y}]), \qquad (51)$$

where $W^{-1}(\cdot)$ is also an increasing function. By Lemmas **??** and **??**, we have

$$p(\mathbf{x}\,|\,\mathbf{y}) = q(\mathbf{x}\,|\,\mathbf{y}) \quad \text{and} \quad p(\mathbf{y}) = q(\mathbf{y})f'^{-1}\left(U(s(\mathbf{y}))-\nu(\mathbf{y}_0)\right). \qquad (52)$$

Multiplying these equalities, dividing by $q(\mathbf{x},\mathbf{y})$, and substituting in (51) yields

$$\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})} \;=\; \frac{p(\mathbf{y})}{q(\mathbf{y})} \;=\; f'^{-1}\left(U\left(W^{-1}\left(\mathbb{E}[B(X,Y)|\,\mathbf{y}]\right)\right)-\nu(\mathbf{y}_0)\right). \qquad (53)$$

Because f'^{-1}, U, and W^{-1} are all increasing functions, the right-hand side is increasing in $\mathbb{E}[B(X,Y)|\,\mathbf{y}]$, which completes the proof. □

Proof of Corollary 4. We must show that $U^{-1}(f'(t)+\nu)t$ and $f(t)-tf'(t)$ are convex in $t \in (0,\infty)$ for any $\nu \in \mathbb{R}$ and that $\frac{tf'''(t)}{f''(t)} \geq -1$. Differentiating $f(t) = t\ln(t)$ yields

$$\begin{aligned} f(t) &= t\ln(t) \\ f'(t) &= 1+\ln(t) \\ f''(t) &= \tfrac{1}{t} \\ f'''(t) &= -\tfrac{1}{t^2}, \end{aligned} \tag{54}$$

which implies that

$$\begin{aligned} U^{-1}(f'(t)+\nu)t &= U^{-1}(\ln(t)+1+\nu)t \\ f(t)-tf'(t) &= -t \\ \tfrac{tf'''(t)}{f''(t)} &= -1. \end{aligned} \tag{55}$$

The second and third expressions satisfy the conditions. Twice differentiating the first condition yields

$$-\tfrac{U''(U^{-1}(\ln(t)+\nu))}{tU'(U^{-1}(\ln(t)+\nu))^3} + \tfrac{1}{tU'(U^{-1}(\ln(t)+\nu))} > 0, \tag{56}$$

where the inequality follows from $U'' < 0$, $U' > 0$, and $t \in (0,\infty)$. It follows that $U^{-1}(\ln(t)+1+\nu)t$ is convex for any increasing concave utility function U, which completes the proof. □

Proof of Lemma 4. Suppose the principal chooses p_{XY} at cost $D_f(p||q)$, and let some $(\mathbf{x}_0, \mathbf{y}_0)$ be arbitrary. Define $\chi_0 \equiv \mathcal{X} \times \mathcal{Y} \setminus (\mathbf{x}_0, \mathbf{y}_0)$ Then the principal's maximization program is as follows.

$$\begin{aligned} \max_{p_{XY}} \quad & \sum_{\chi_0} \left(B(\mathbf{x},\mathbf{y}) p(\mathbf{x},\mathbf{y}) - q(\mathbf{x},\mathbf{y}) f\left(\tfrac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right)\right) \\ & + B(\mathbf{x}_0,\mathbf{y}_0)\left(1 - \sum_{\chi_0} p(\mathbf{x},\mathbf{y})\right) - q(\mathbf{x}_0,\mathbf{y}_0) f\left(\tfrac{1-\sum_{\chi_0} p(\mathbf{x},\mathbf{y})}{q(\mathbf{x}_0,\mathbf{y}_0)}\right) \\ \text{s.t.} \quad & p(\mathbf{x},\mathbf{y}) \geq 0 \ \forall (\mathbf{x},\mathbf{y}) \neq (\mathbf{x}_0,\mathbf{y}_0) \\ & 1 - \sum_{\chi_0} p(\mathbf{x},\mathbf{y}) \geq 0 \end{aligned} \tag{57}$$

Let $\delta(\mathbf{x},\mathbf{y})$ and $\delta(\mathbf{x}_0,\mathbf{y}_0)$ be the Lagrange multipliers on the constraints. Because $-D_f(p||q)$ is concave and all other terms in this program are linear, the following first-order conditions with respect to p at $(\mathbf{x},\mathbf{y}) \in \chi_0$ are necessary and sufficient for a solution.

$$0 \ = \ B(\mathbf{x},\mathbf{y}) - B(\mathbf{x}_0,\mathbf{y}_0) - f'\left(\tfrac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right) + f'\left(\tfrac{p(\mathbf{x}_0,\mathbf{y}_0)}{q(\mathbf{x}_0,\mathbf{y}_0)}\right) + \delta(\mathbf{x},\mathbf{y}) - \delta(\mathbf{x}_0,\mathbf{y}_0) \tag{58}$$

Because this equality must hold for every $(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ (note that if $(\mathbf{x},\mathbf{y}) = (\mathbf{x}_0,\mathbf{y}_0)$ then the equality holds trivially), take any two $(\mathbf{x},\mathbf{y}), (\mathbf{x}',\mathbf{y}') \in \text{supp}(\mathcal{X} \times \mathcal{Y})$ so that $\delta(\mathbf{x},\mathbf{y}) = \delta(\mathbf{x}',\mathbf{y}') = 0$. Subtracting the $(\mathbf{x}',\mathbf{y}')^{th}$ from the $(\mathbf{x},\mathbf{y})^{th}$ first-order condition yields

$$f'\left(\tfrac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right) - f'\left(\tfrac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}\right) \ = \ B(\mathbf{x},\mathbf{y}) - B(\mathbf{x}',\mathbf{y}'). \tag{59}$$

Finally, an increase in $B(\mathbf{x},\mathbf{y}) - B(\mathbf{x}',\mathbf{y}')$ causes the right-hand side to increase, which requires that the left-hand side also increase to maintain the equality. Because $f(\cdot)$ is convex, $f'(\cdot)$ is an increasing function; therefore $\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})} - \frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}$ must be increasing in $B(\mathbf{x},\mathbf{y}) - B(\mathbf{x}',\mathbf{y}')$. □

Proof of Proposition 2. We approximate $U'(\cdot)$ by the slope of the secant line connecting the points $U(s(\mathbf{y}))$ and $U(s(\mathbf{y}'))$. We denote this slope by $u'(\mathbf{y},\mathbf{y}') \equiv \frac{U(s(\mathbf{y}))-U(s(\mathbf{y}'))}{s(\mathbf{y})-s(\mathbf{y}')}$. Analogously, we denote $u''(\mathbf{y},\mathbf{y}') \equiv \frac{U'(s(\mathbf{y}))-U'(s(\mathbf{y}'))}{s(\mathbf{y})-s(\mathbf{y}')}$. Then $-\frac{u''(\mathbf{y},\mathbf{y}')}{u'(\mathbf{y},\mathbf{y}')} = -\frac{U'(s(\mathbf{y}))-U'(s(\mathbf{y}'))}{U(s(\mathbf{y}))-U(s(\mathbf{y}'))}$ is our measure of absolute risk aversion.

Differencing the $(\mathbf{y}')^{th}$ from the $\mathbf{y}^{th}$ expression for the optimal contract in (23), dividing through by $U(s(\mathbf{y}))-U(s(\mathbf{y}'))$, and rearranging terms yields

$$U(s(\mathbf{y}))-U(s(\mathbf{y}')) \quad = \quad \frac{\mathbb{E}[B(X,Y)|\,\mathbf{y}]-\mathbb{E}[B(X,Y)|\,\mathbf{y}]}{\frac{U'(s(\mathbf{y}))U'(s(\mathbf{y}'))}{u'(\mathbf{y},\mathbf{y}')}-\frac{u''(s(\mathbf{y})}{u'(\mathbf{y},\mathbf{y}')}} \cdot U'(s(\mathbf{y}))U'(s(\mathbf{y}')). \tag{60}$$

Substituting the IC constraint into the left-hand side yields

$$f'\left(\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right)-f'\left(\frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}\right) \quad = \quad \frac{\mathbb{E}[B(X,Y)|\,\mathbf{y}]-\mathbb{E}[B(X,Y)|\,\mathbf{y}']}{\frac{U'(s(\mathbf{y}))U'(s(\mathbf{y}'))}{u'(\mathbf{y},\mathbf{y}')}-\frac{u''(\mathbf{y},\mathbf{y}')}{u'(\mathbf{y},\mathbf{y}')}} \cdot U'(s(\mathbf{y}))U'(s(\mathbf{y}')). \tag{61}$$

Without loss of generality, assume that $\mathbb{E}[B(X,Y)|\,\mathbf{y}] \geq \mathbb{E}[B(X,Y)|\,\mathbf{y}']$. By Corollary 3, it follows that $\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})} \geq \frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}$, so the left- and right-hand sides of the above equation are positive because $f'(\cdot)$ is an increasing function.

It also follows from the convexity of f that the left-hand side is increasing in $\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})} - \frac{p(\mathbf{x}',\mathbf{y}')}{q(\mathbf{x}',\mathbf{y}')}$. Moreover, the right-hand side is increasing in $\mathbb{E}[B(X,Y)|\,\mathbf{y}] - \mathbb{E}[B(X,Y)|\,\mathbf{y}']$ and is decreasing in $-\frac{u''(\mathbf{y},\mathbf{y}')}{u'(\mathbf{y},\mathbf{y}')}$. In other words, the variation in $\frac{p}{q}$ is increasing in the informativeness of Y about value and is decreasing in the agent's risk aversion, which completes the proof. $\square$

Proof of Lemma 5. Faced with an arbitrary contract $s(Y)$, let $\mathbf{a}'$ be the action that implements the distribution characterized by Lemma 1. Suppose to the contrary that $\mathbf{a}'$ is not a global solution to (28); that is, there exists some action $\mathbf{a}''$ that awards the agent strictly greater expected net utility than $\mathbf{a}'$. Because $\mathbf{a}$ only enters (28) through $p(\mathbf{a})$, it is immediate that $p(\mathbf{a}') \neq p(\mathbf{a}'')$. Moreover, because $p(\mathbf{a}'') \in \mathcal{P}(\mathcal{X}, \mathcal{Y})$, $p(\mathbf{a}'')$ is in the agent's action opportunity set in the nonparametric problem. But then Lemma 1 does not characterize a solution to the nonparametric problem, a contradiction. Thus $\mathbf{a}'$ is a global solution to (28). $\square$

Proof of Proposition 3. Let $\mathbf{a}'$ be the action that implements the distribution characterized by Proposition 1. Suppose to the contrary that $\mathbf{a}'$ is not a global solution to (30); that is, there exists some action $\mathbf{a}''$ that awards the principal strictly greater expected net utility than $\mathbf{a}'$. Because $\mathbf{a}$ only enters (30) through $p_Y(\mathbf{a})$, it is immediate that $p_Y(\mathbf{a}') \neq p_Y(\mathbf{a}'')$. Moreover, because $p_Y(\mathbf{a}'') \in \mathcal{P}(\mathcal{Y})$, $p_Y(\mathbf{a}'')$ is in the agent's action opportunity set in the nonparametric problem and can be implemented with the same contract $s(Y)$ as in the parametric problem. But then Proposition 1 does not characterize a solution to the nonparametric problem, a contradiction. Thus $\mathbf{a}'$ is a global solution to (30). $\square$

Proof of Proposition 4. Faced with a contract $s(Y,Z)$, the agent chooses p_{XY} to solve the following program, where $(\mathbf{w}_0,\mathbf{x}_0,\mathbf{y}_0,\mathbf{z}_0)\in\mathcal{W}\times\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ is arbitrary and we denote $\chi_0\equiv\mathcal{X}\times\mathcal{Y}\backslash(\mathbf{x}_0,\mathbf{y}_0)$, $\Gamma_0\equiv\mathcal{W}\times\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}\backslash(\mathbf{w}_0,\mathbf{x}_0,\mathbf{y}_0,\mathbf{z}_0)$, and $s_0\equiv s(\mathbf{y}_0,\mathbf{z}_0)$.

$$\begin{aligned}\max_{p_{XY}}\quad & U(s_0)+\sum_{\Gamma_0}(U(s(\mathbf{y},\mathbf{z}))-U(s_0))q(\mathbf{w},\mathbf{z}\,|\,\mathbf{x},\mathbf{y})p(\mathbf{x},\mathbf{y})\\ & -\sum_{\chi_0}q(\mathbf{x},\mathbf{y})f\left(\tfrac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right)-q(\mathbf{x}_0,\mathbf{y}_0)f\left(\tfrac{1-\sum_{\chi_0}p(\mathbf{x},\mathbf{y})}{q(\mathbf{x}_0,\mathbf{y}_0)}\right)\end{aligned}\tag{62}$$

Taking the first-order condition with respect to $p(\mathbf{x},\mathbf{y})$ for $(\mathbf{x},\mathbf{y})\neq(\mathbf{x}_0,\mathbf{y}_0)$ yields

$$f'\left(\tfrac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right)-f'\left(\tfrac{p(\mathbf{x}_0,\mathbf{y}_0)}{q(\mathbf{x}_0,\mathbf{y}_0)}\right)\;=\;\sum_{\mathcal{W}\times\mathcal{Z}}U(s(\mathbf{y},\mathbf{z}))q(\mathbf{w},\mathbf{z}\,|\,\mathbf{x},\mathbf{y})-U(s_0)\tag{63}$$

Denoting $B_0\equiv B(\mathbf{w}_0,\mathbf{x}_0,\mathbf{y}_0,\mathbf{z}_0)$, the principal's program is given below.

$$\max_{s(Y,Z),p_{XY}} B_0-s_0+\sum_{\Gamma_0}\left(B(\mathbf{w},\mathbf{x},\mathbf{y},\mathbf{z})-s(\mathbf{y},\mathbf{z})-B_0+s_0\right)q(\mathbf{w},\mathbf{z}\,|\,\mathbf{x},\mathbf{y})p(\mathbf{x},\mathbf{y})\tag{64}$$

$$\begin{aligned}\text{s.t. }& U(s_0)+\sum_{\Gamma_0}(U(s(\mathbf{y},\mathbf{z}))-U(s_0))q(\mathbf{w},\mathbf{z}\,|\,\mathbf{x},\mathbf{y})p(\mathbf{x},\mathbf{y})\\ &-\sum_{\chi_0}q(\mathbf{x},\mathbf{y})f\left(\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right)-q(\mathbf{x}_0,\mathbf{y}_0)f\left(\frac{1-\sum_{\chi_0}p(\mathbf{x},\mathbf{y})}{q(\mathbf{x}_0,\mathbf{y}_0)}\right)\geq\bar{U}\end{aligned}\tag{IR}$$

$$\begin{aligned}&\sum_{\mathcal{W}\times\mathcal{Z}}U(s(\mathbf{y},\mathbf{z}))q(\mathbf{w},\mathbf{z}\,|\,\mathbf{x},\mathbf{y})-U(s_0)\\ &=f'\left(\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right)-f'\left(\frac{1-\sum_{\chi_0}p(\tilde{\mathbf{x}},\tilde{\mathbf{y}})}{q(\mathbf{x}_0,\mathbf{y}_0)}\right)\quad\forall(\mathbf{x},\mathbf{y})\in\chi_0\end{aligned}\tag{IC$_{\mathbf{x}\,\mathbf{y}}$}$$

Let λ and $\mu(\mathbf{x},\mathbf{y})$ be the Lagrange multipliers on the IR and the $(\mathbf{x},\mathbf{y})^{th}$ IC constraints. Then the principal's first-order conditions with respect to $p(\mathbf{x},\mathbf{y})$ and $s(\mathbf{y},\mathbf{z})$ for $(\mathbf{x},\mathbf{y})\neq(\mathbf{x}_0,\mathbf{y}_0)$ and $(\mathbf{y},\mathbf{z})\neq(\mathbf{y}_0,\mathbf{z}_0)$ are

$$0=\mathbb{E}_{WZ}[B(W,X,Y,Z)-s(Y,Z)|\,\mathbf{x},\mathbf{y}]-\mu(\mathbf{x},\mathbf{y})\frac{1}{q(\mathbf{x},\mathbf{y})}f''\left(\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right)-\eta_0\tag{65}$$

$$0=-p(\mathbf{y},\mathbf{z})+\lambda U'(s(\mathbf{y},\mathbf{z}))p(\mathbf{y},\mathbf{z})+U'(s(\mathbf{y},\mathbf{z}))\sum_{\mathcal{X}}\mu(\mathbf{x},\mathbf{y})q(\mathbf{z}\,|\,\mathbf{x},\mathbf{y}),\tag{66}$$

where $\eta_0\equiv B_0-s_0+\frac{1}{q(\mathbf{x}_0,\mathbf{y}_0)}f'\left(\frac{p(\mathbf{x}_0,\mathbf{y}_0)}{q(\mathbf{x}_0,\mathbf{y}_0)}\right)\sum_{\chi_0}\mu(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$. We can solve (65) for $\mu(\mathbf{x},\mathbf{y})$ in closed form as follows.

$$\mu(\mathbf{x},\mathbf{y}) \quad = \quad \frac{\mathbb{E}_{WZ}[B(W,X,Y,Z)-s(Y,Z)\,|\,\mathbf{x},\mathbf{y}]-\eta_0}{\frac{1}{q(\mathbf{x},\mathbf{y})}f''\left(\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{x},\mathbf{y})}\right)} \tag{67}$$

Finally, substituting this expression for $\mu(\mathbf{x},\mathbf{y})$ as well as IC into (66) yields

$$\frac{1}{U'(s(\mathbf{y},\mathbf{z}))} \quad = \quad \lambda + \sum_{\mathcal{W}\times\mathcal{X}} \frac{\mathbb{E}_{WZ}[B(W,X,Y,Z)-s(Y,Z)\,|\,\mathbf{x},\mathbf{y}]-\eta_0}{\check{f}\left(\sum_{\mathcal{Z}} U(s(\mathbf{y},\tilde{\mathbf{z}}))q(\tilde{\mathbf{z}}\,|\,\mathbf{x},\mathbf{y})-\nu_0\right)} \cdot p(\mathbf{w},\mathbf{x}\,|\,\mathbf{y},\mathbf{z}), \tag{68}$$

where $\nu_0 \equiv U(s_0) - f'\left(\frac{p(\mathbf{x}_0,\mathbf{y}_0)}{q(\mathbf{x}_0,\mathbf{y}_0)}\right)$ and $\check{f}(t) \equiv f'^{-1}(t)f''\left(f'^{-1}(t)\right)$.

□

Proof of Corollary 5. Suppose that $D_f(p||q)$ is the Kullback-Liebler divergence, so that

$$f(t) = t\ln(t) \implies f'(t) = 1 + \ln(t) \implies f''(t) = \tfrac{1}{t} \implies tf''(t) = 1. \tag{69}$$

Setting $t = f'^{-1}(u)$ implies that $\check{f}(u) = 1$ for all $s(Y, Z)$. Substituting $\check{f}(t) = 1$ into (33) yields

$$\tfrac{1}{U'(s(\mathbf{y},\mathbf{z}))} = \lambda + \mathbb{E}_{WX}\left[\mathbb{E}_{WZ}[B(W, X, Y, Z) - s(Y, Z)|X, Y]\right| \mathbf{y}, \mathbf{z}] - \eta_0 \tag{70}$$

Because $\mathbf{y}$ and $\mathbf{z}$ only enter this expression through $s(\mathbf{y}, \mathbf{z})$ and the expected controllable residual value $\mathbb{E}_{WX}\left[\mathbb{E}_{WZ}[B(W, X, Y, Z) - s(Y, Z)|X, Y]\right| \mathbf{y}, \mathbf{z}]$, it follows that Y_i or Z_i is useful if and only if it is informative about controllable residual value. □