Monotonicity of optimal contracts without the first-order approach

Rongzhu Ke* Christopher Thomas Ryan†

April 21, 2017

Abstract

We develop a simple sufficient condition for an optimal contract of a moral hazard problem to be monotone in the output signal. Existing results on monotonicity require conditions on the output distribution (namely, the monotone likelihood ratio property (MLRP)) and additional conditions to ensure that agent's decision is approachable via the *first-order approach* of replacing that problem with its first-order conditions. We know of no positive monotonicity results in the setting where the first-order approach does not apply. Indeed, it is well-documented that when there are *finitely-many* possible outputs, and the first-order approach does not apply, the MLRP alone is insufficient to guarantee monotonicity. However, we show that when there is an *interval* of possible output signals, the MLRP does suffice to establish monotonicity under additional technical assumptions that do not ensure the validity of the first-order approach. To establish this result we examine necessary optimality conditions for moral hazard problems using a novel penalty function approach. We then manipulate these conditions and provide sufficient conditions for when they coincide with a simple version of the moral hazard problem with only two constraints. In this two-constraint problem, monotonicity is established directly via a strong characterization of its optimal solutions.

Keywords. Moral hazard problems, monotonicity, optimality conditions

*Department of Economics, The Chinese University of Hong Kong, E-mail: rzke@cuhk.edu.hk
†Booth School of Business, University of Chicago, E-mail: chris.ryan@chicagobooth.edu

1 Introduction

We study the analytical properties of optimal solutions to the classic principal-agent moral hazard problem in economics (for detailed background see Laffont and Martimort (2009)). We focus on the base version where the agent has a single action and the output is single-dimensional. An agent chooses an action $a \in \mathbb{A} \subseteq \mathbb{R}$ that is unobservable to a principal. This action influences the random outcome $X \in \mathcal{X} \subseteq \mathbb{R}$ through the probability density function $f(x, a)$. The principal chooses a wage contract $w : \mathcal{X} \to [\underline{w}, \infty)$ that is a function of the output, where $\underline{w}$ is an exogenously given minimum wage. The value generated by the output is given by the function $\pi : \mathcal{X} \to \mathbb{R}$.

Given an outcome realization $x \in \mathcal{X}$, the principal and agent derive the following utilities. The agent's utility under action a is separable in the associated wage $w(x)$ and the cost $c(a)$ of taking the action. In particular, he derives utility $u(w(x)) - c(a)$ where $u : [\underline{w}, \infty) \to \mathbb{R}$ and $c : \mathbb{A} \to \mathbb{R}$. The principal's utility for outcome x is a function of the net value $\pi(x) - w(x)$ and is denoted by $v(\pi(x) - w(x))$ where $v : \mathbb{R} \to \mathbb{R}$. We can now express the expected utilities of both players, given the action a and the contract w. The principal's expected utility is $V(w, a) = \int v(\pi(x) - w(x)) f(x, a) dx$ and the agent's expected utility is $U(w, a) = \int u(w(x)) f(x, a) dx - c(a)$. The agent has an outside alternative that earns him utility $\underline{U}$.

The principal chooses the contract w to maximize her expected utility subject to the optimizing behavior of the agent. In other words, she solves

$$\max_{w \geq \underline{w}, a \in \mathbb{A}} \quad V(w, a) \tag{1.1a}$$

$$\text{subject to} \quad a \in \arg\max_{a' \in \mathbb{A}} U(w, a') \tag{1.1b}$$

$$U(w, a) \geq \underline{U}. \tag{1.1c}$$

where (1.1b) ensures the agent responds optimally and (1.1c) guarantees the agent earns at least his reservation utility $\underline{U}$. In principle, this is an infinite-dimensional bilevel optimization problem and has been studied by several authors in the bilevel optimization community (see for instance Monahan and Vemuri (1996), Nasri (2016), Ye and Zhu (2010)).

Following numerous others, we make the standard assumption that the output distribution f satisfies the *monotone likelihood ratio property* (MLRP) where for any a, $\frac{\partial \log f(\cdot, a)}{\partial a}$ is nondecreasing. Milgrom (1981) gives an interpretation of this ratio in terms of statistical inference. If a principal uses maximum likelihood estimation methods to infer to effort of the agent given the outcomes, the ratio $\frac{\partial \log f(\cdot, a)}{\partial a}$ appears in the calculation. Roughly speaking, the MLRP condition says that higher effort from the agent gives rise to a stochastically improved outcome distribution.

Our over-arching goal is to study the analytical properties of optimal solutions to (1.1) and, in particular, *monotonicity* of an optimal solution (also called an optimal contract), given the MLRP. Precisely, we are interested in the following question: when does there exist an optimal solution w to (1.1) such that $w(x)$ is an nondecreasing function of x?

Monotone contracts enjoy the following "natural" logic: grant higher pay to agents whose efforts result in more valuable output. In practice it is relatively rare to find a contract that is *not* monotone. Establishing monotonicity of optimal contracts is thus a central issue in the study of moral hazard problems and the subject of considerable study (see for instance Grossman and Hart (1983), Lambert (2001), Milgrom (1981)). Indeed, even when assuming the MLRP an optimal monotone contracts need not exist. A troubling counter-example discovered early on by Grossman and Hart (1983) (and analyzed further by Monahan and Vemuri (1996)) shows that the MLRP is insufficient to guarantee monotonicity when the output set is finite. Typically, quite strong additional analytical assumptions (discussed below) are needed. The fact that the "monotonicity" of the relationship between actions and outcomes via the MLRP does not directly translate to monotonicity of the optimal contract is one of the great "puzzles" of agency theory (Brosig et al. 2010).

Known monotonicity holds under a variety of different assumptions (see for instance Grossman and Hart (1983), Holmstrom (1979), Jewitt (1988), Rogerson (1985)). However, all known results boil down to requiring a monotonicity assumption on the output distribution and assumptions that ensure the agent's problem lower level problem in (1.1b) is a convex optimization problem. These assumptions facilitate the *first-order approach* (FOA) to the problem where (1.1b) is replaced (without loss) by its first-order conditions. This idea is common in the bilevel optimization literature where it is sometimes referred to as the Karush-Kuhn-Tucker approach (Ye and Zhu 1995). The classical assumption that ensures the validity of the FOA is the convexity of the cumulative distribution function condition (CDFC) proposed by Rogerson (1985). This condition is thought to be restrictive and much later work is in search of relaxations that still guarantee the validity of the FOA.

Unfortunately, the first-order approach is well-documented to fail in many natural settings, as first pointed out in Mirrlees (1999) (a paper that originally appeared in 1975). For example, if the agent has constant relative risk averse (CRRA) utility and output is exponentially distributed, the first-order approach is invalid (Jewitt et al. (2008)). Despite this, numerous authors have mounted rigorous defenses of the general validity of the first-order approach, which simultaneously attest to the challenges of proceeding when it is invalid (see Conlon (2009), Jung and Kim (2015), Kirkegaard (2017b), Sinclair-Desgagné (1994)).

We seek a novel monotonicity result under weak assumptions that do not ensure the validity of the first-order approach. Our main result, informally stated, is as follows:

Theorem 1.1. Under additional assumptions (specified below) that do not ensure the validity of the first-order approach, if the output distribution f satisfies the MLRP then there exists an optimal contract that is nondecreasing in the output x.

This result is established below in Theorem 2.1, which details formal conditions for when MLRP implies monotonicity, and in Example 5.2, which gives an example where the first-order approach fails but, nonetheless, our conditions hold.

Our result does not contradict the counter-examples of Grossman and Hart (1983) and Monahan and Vemuri (1996) described above. Our theorem only applies to settings where there is an interval of (infinitely many) possible outcomes (see Assumption (A1.1) below). At an intuitive level, the source of non-monotonicity in the finite setting is due to an inherent inflexibility in designing a contract to recover the monotonicity properties of the output distribution (coming from the MLRP). This complication disappears with a continuum of possible outputs. Indeed, the principal has greater flexibility in designing a contract to capture underlying structure. Sections 4 and 5 gives precise realization to this high-level intuition.

Comparison with existing approaches when the first-order approach fails

Analyzing (1.1) when the first-order approach is not valid has also spurred several studies, although work in this direction is still relatively nascent. For instance, a result of the type of Theorem 1.1 is not known in the literature. Two recent efforts include Kirkegaard (2017a) and Renner and

Schmedders (2015)). Kirkegaard (2017a) considers a tractable moral hazard environment within a special class of output distributions where the first-order approach nonetheless fails and examines the implications. Renner and Schmedders (2015) assume the data can be modeled or approximated by polynomials and provide an algorithmic approach to determining optimal contracts.

We modify and extend a more classical approach due to Mirrlees (1999) (which first appeared in 1975), and later developed by Araujo and Moreira (2001), that remains an analytical method that applies generically under analytical assumptions. In the method of Mirrlees (1999), the lower level problem is replaced by an appropriately chosen subset of constraints of the form: for a given $\hat{a} \in \mathbb{A}$

$$U(w, a) - U(w, \hat{a}) \geq 0, \tag{1.2}$$

called the *no-jumping* constraint at $\hat{a}$. The name comes from the fact that if a contract violates the no-jumping constraint (1.2) then its does not implement a, since an optimizing agent can improve her expected utility by "jumping" from action a to $\hat{a}$.

The weakness of Mirrlees's approach is that it may require *many* (possibly infinitely many) no-jumping constraints, one corresponding to each stationary point of the agent's utility function at the proposed contract. Araujo and Moreira (2001) refine Mirrlees's approach using second-order information but also suffer from producing many no-jumping contracts. The characterizations of optimal contracts that result from such analysis also suffer from this complexity, making it difficult to establish analytical properties.

Another avenue that tackles the situation where the first-order approach fails is the bilevel literature, a class of problems that has moral hazard as a special case. As an example of work in this direction, Ye and Zhu (1995) study optimality conditions using the value function of the follower's (agent's) problem to define an equivalent single-level optimization problem. They give constraint qualifications for when Fritz-John- and Karush-Kuhn-Tucker-like necessary optimality conditions hold. One notable condition is *partial calmness* which allows the value function to be handled in the objective of the resulting single-level problem rather than in the constraints. This yields clean optimality conditions that apply to a variety of cases. Later in Ye and Zhu (2010), Ye and Zhu leverage a combination of the first-order approach and the value function approach to yield new constraint qualifications and optimality conditions referred to as *weak calmness* that apply even more broadly. Other researchers have built further on these methods (for instance, Dempe and Zemkoho (2011), Dempe et al. (2007)).

Common to all of these approaches is turning bilevel problems into single-level optimization problems. The resulting single-level optimization problems have additional complexity beyond a standard nonconvex optimization problem. When first-order conditions are used, complementarity constraints are considered. In the value function approach of Ye and Zhu (1995), a nonsmooth function are introduced. Known results on complementarity and nonsmooth optimization problems are adapted to the bilevel setting to derive optimality conditions. Unfortunately, these complexities typically give rise to complex optimality conditions, which like the approach of Mirrlees (1999) and Araujo and Moreira (2001), and involve Lagrangian multipliers for *many* alternate best responses.

Finally, we mention one study by Nasri (2016) that tackles the moral-hazard using a semi-infinite programming duality approach that is unique in the literature. Under the assumption of finitely-many outcomes $\mathcal{X} = [\underline{x} = x_1, x_2, \ldots, x_n = \bar{x}]$ and that $f(\bar{x}, a)$ is concave in a, Nasri shows that there exists an optimal contract with a very simple form of only giving a positive wage for outcome with the highest value to the principal. Trivially, such a contract is monotone. However, the assumption that $f(\bar{x}, a)$ is concave is quite restrictive, particularly when it is used to discretely approximate a continuous distribution where the probability of an outcome vanishes in the right tail. For example, a binomial distribution fails Nasri's condition. Our approach does not require such restrictions. Indeed, the classical example of Holmstrom (1979) (see also Example 5.2 below) fails Nasri's assumption when discretely approximated. Moreover, the counter-example due to Grossman and Hart (1983) discussed above is quite natural but does not fit the setting of Nasri (clearly, since that example does not admit monotone optimal contracts), despite having finitely many outcomes.

Our major point of departure is to show that there exists a *single* no-jumping constraint that suffices to characterize the optimal contract under the MLRP (and additional technical assumptions described below). Our approach relies on an alternate reformulation of the moral hazard problem to a single-level (Max-Min) structure. Analyzing this reformulation to reduce to a single no-jump constraint is a key step in establishing Theorem 1.1.

The significance of deriving a characterization that involves only a single no-jump constraint is the similarly of our characterization of an optimal contract to that of the FOA. Indeed, Holmstrom (1979) gives the following characterization (known as the Mirrlees-Holmstrom (MH) condition) of an optimal contract:

$$\frac{v'(\pi(x)-w(x))}{u'(w(x)))} = \lambda + \mu \frac{f_a(x,a)}{f(x,a)} \tag{MH}$$

for almost all $x \in \mathcal{X}$ where f_a represents a partial derivative and λ and μ are Lagrangian multipliers. Rogerson (1985) provided justification for this characterization under the appropriate assumptions, including the MLRP and a strong condition on the convexity of the cumulative distribution function of the outcome. We provide a strikingly similar characterization in (3.5) below, which we reproduce here and slightly simplify under the appropriate assumptions, which includes the MLRP but no strong convexity conditions:

$$\frac{v'(\pi(x)-w(x))}{u'(w(x))} = \lambda + \delta \left(1 - \frac{f(x,\hat{a})}{f(x,a)}\right) \tag{1.3}$$

for almost all $x \in \mathcal{X}$, where $\hat{a}$ is the alternate best response associated with the identified no-jump constraint and δ the associated Lagrangian multiplier. As mentioned earlier, other characterizations in the literature, including that of Araujo and Moreira (2001) and Ye and Zhu (1995), potentially give rise to *many* Lagrange multipliers, creating far greater distance from the elegance of (MH). The similarity of (MH) and (1.3) provide hope for further leveraging our theory to cases where the first-order approach is invalid and other approaches to this case fail because a lack of parsimony in their characterizations.

Our approach does impose that the set of outcomes be a continuum, but this is a common assumption (see, for instance, Carlier and Dana (2005), Innes (1990), Jewitt (1988), Oyer (2000)). In applications, a continuum of outcomes may represent the fact that the "quality" of an outcome resulting from an action, after some random realization, may not be representable by a discrete set or with finitely-many values. In applied theory, a continuum of outcomes can be assumed for purposes of tractability in deriving analytical results whose structure may fall apart in the discrete setting. Indeed, analytical development that uses integration-by-parts (such as in Jewitt (1988)) requires a continuum of outputs. We contend that assuming a continuum of outcomes may, in many situations, be a less economically strenuous condition than imposing the validity of the first-order approach. Indeed, assuring the validity of the first-order approach involves structured first- and second-order properties that influence economic tradeoffs and marginal reasoning. Moreover, these properties are endogenous to the contract that this is offered.

We leave for future work the implications of the similarity of (MH) and (1.3) for a variety of applied moral hazard problems, but discuss briefly one or two possibilities below. The focus of this paper is to use (1.3) to establish monotonicity.

Related literature in OR/MS

The previous section provided motivation for our study from a technical perspective, referring to a selection of relevant papers largely from the theoretical economics and bilevel optimization literatures. In this discussion we provide additional discussion of the background and significance of moral hazard problems, and in particular, the relevance of this problem to the broader operations research community.

Moral-hazard models are by now a standard tool in management literatures, including marketing (e.g., Coughlan (1993), Lal (1990)), finance (e.g., Innes (1990), Zhang (1997)), and accounting (e.g., Kwon (2005), Lambert (2001)). Operations management is particularly well-suited to models of this kind. This is well-stated by Plambeck and Zenios (2000):

> Operations Management (OM) is a natural area of application for the principal-agent paradigm . . . Most of the problems that we study in OM involve such delegated control, although classical OM models often suppress this feature.

Indeed, leveraging insights from principal-agent theory to enrich "classical OM models" with issues of asymmetric information, hidden actions, and lack of truthful revelation has become a mainstay of research in OM since the 1990s. For a thorough overview of agency models in OM we refer the reader to Krishnan and Winter (2012).

For purposes of illustration, we mention one specific thread of research in the operations management literature that concerns the design of salesforce contracts. Salesforce compensation is a classical topic in marketing science that has been the subject of much study (for a survey of research leading up to the early 1990s see Coughlan (1993)). Part of the early debate in that field concerned the value of agency models in the study of salesforce compensation, but agency theory eventually prevailed by researcher's demonstrating its ability to explain the prevalence of certain sales contracts widely seen in practice as being optimal contract designs in a principal-agent framework. An influential example of this is Oyer (2000), which illustrates the optimality of sales quota contracts with bonuses, a common salesforce contract seen in practice by analyzing a specific moral hazard model.

Oyer (2000) states his results by *assuming* that the optimal contract is monotone (nondecreasing in output) and discusses optimality with respect to this class. However, the only justification provided for concentrating on this class is the validity of the first-order approach (see Footnote 6 of Oyer (2000)) which Oyer himself acknowledges is not a consequence of his problem setup. Example 5.3 below is an example that fits the set-up of Oyer (2000) and fails the first-order approach, but nonetheless our approach produces an monotone optimal contract. In this sense, our results can be seen to generalize the arguments of Oyer (2000) (in particular, providing weaker conditions to verify his Proposition 4). Moreover, a key point of Oyer's paper is the structure of optimal *binary* contracts that take on two values, some minimum wage and then a "bonus" when a sales "quota" is met. However, Oyer's approach can only guarantee this structure in the case where the agent is risk neutral. By contrast, Example 5.3 reveals the optimality of a binary contract in the risk-averse case and suggests a more general approach. We do not pursue this in detail here, as it falls outside our scope.

The work of Oyer (2000) provides inspiration for several (including very recent) studies in the OM literature (Chu and Lai (2013), Dai and Jerath (2013, 2016), Wang et al. (2016)) that enrich the classical salesforce compensation problem by adding inventory considerations and capacity constraints. These papers build on Oyer (2000) as a foundation and implicitly or explicitly assume the validity of the first-order approach in their analysis. As discussed above, the approach of this paper may provide an alternate foundation for OM models built on Oyer (2000) with weaker assumptions. We leave a careful treatment of these issues for future work.

Despite the demonstrated value of the standard moral hazard problem to OM theory, this may be overshadowed by the potential for analyzing situations of *dynamic contracting*. Continuing the quote of Plambeck and Zenios (2000) cited earlier

> Unfortunately, the classical economic models for the principal-agent problem are of limited use to OM researchers, because they focus either on one-shot static problems or else on "repeated" problems involving a simple kind of multi-period replications, whereas even stylized OM models typically require a richer dynamic structure.

The standard-bearer of theory in dynamic contracting in the OM literature is to adapt the first-order approach to the dynamic setting. This is the approach of the influential study by Plambeck and Taylor (2006) that adapts the conditions of Rogerson (1985) discussed above to a dynamic operational setting. Our method of characterizing optimal contracts provides hope for developing new methodologies for studying dynamic contracting settings. Indeed, requiring the first-order approach provides a strong restriction that may not mesh with the "richer dynamic structure" of OM problems. This potentially promising future direction lies beyond the scope of this paper.

Lastly, we want to clarify a connection between the current paper and another by the same authors on a related model and question (Ke and Ryan 2015). That paper also provides a characterization of optimal contracts using an alternate method of establishing a strong duality theory for infinite dimensional optimization problems. There are two important distinctions between this characterization and the one provided here. First, the characterization in Ke and Ryan (2015) involves many (in fact, infinitely many) Lagrange multipliers, similar to other approaches to when the first-order approach fails. Second, the characterization in Ke and Ryan (2015) applies to general moral hazard problems, not necessarily those where the output distribution satisfies the MLRP. The characterization in the current paper needs to leverage the MLRP condition in its construction and cannot be seen as a special case of the characterization in Ke and Ryan (2015).

Overview of analytical approach

The following is the logical sequence for the development of our approach, which also serves as an outline of the rest of the paper. In Section 2 we set out our basic assumptions and initial observations. Section 3 looks at a family of relaxations of our moral hazard problem that involves a single no-jumping constraint derived from one alternate action of the agent. Each relaxation in this family admits a strong and simple characterization of its optimal solutions. The work here is to establish a strong duality result for these relaxed problems and establish the uniqueness of their primal and dual solutions. This allows us to derive monotonicity properties of the optimal solutions that are eventually leveraged in the full problem.

The main task of the remainder of the paper is to establish conditions for when a relaxation from this family is tight; that is, the full problem is equivalent to a relaxed problem with a single no-jumping constraint. Section 4 takes up this task. The work here is to derive necessary optimality conditions for (1.1) and manipulate those conditions to resemble those that characterize the optimal solutions of a relaxed problem. This is achieved using a penalty function that focuses attention on a single alternate best response with desirable properties. Penalty function methods allow for tremendous flexibility in designing optimality conditions by introducing additional penalty terms. We use a penalty function with a term that penalizes deviations away from a *single* alternate best response (denoted below by $\hat{a}^*$ and defined in (4.31)). This penalization reduces the required number of Lagrange multipliers to characterize an optimal contracts to a *single* multiplier associated with a optimization problem (1.1b). This is how we yield (1.3). We are unaware how non-penalty function methods for deriving optimality conditions can be adapted to provide this level of specificity. Indeed, showing how a penalty function can be defined to achieve this is one of the key technical results of this paper. Our penalty function method is inspired by the technique described in Chapter 3 of Bertsekas (1999) and draws partial inspiration from existing penalty function methods for finite-dimensional convex bilevel problems (for instance Liu et al. (2001), Marcotte and Zhu (1996)).

After deriving necessary optimality conditions, the resulting conditions are still complex, but we are able to analyze them using variational arguments. Through this analysis we show that assuming the output distribution satisfies the MLRP is a sufficient condition for transforming our complicated first-order conditions into simple conditions that precisely characterize a contract with a single no-jumping constraint. Assuming the set of possible outputs is an interval in the real line is essential for our argument to proceed. We demonstrate how our arguments fail when the output space is discrete. Finally, Section 5 summarizes our results and provides a formal statement and proof of Theorem 1.1.

2 Model assumptions

Turning now to details, this section provides the basic assumptions used in our development.

Assumption 1. The following hold:

(A1.1) The outcome set $\mathcal{X}$ is the interval $[\underline{x}, \overline{x}]$, with the possibility that $\underline{x} = -\infty$ or $\overline{x} = +\infty$ and the action set is the bounded interval $\mathbb{A} := [\underline{a}, \bar{a}]$,

(A1.2) the random outcome X is a continuous random variable and $f(x, a)$ is continuous in x and twice continuously differentiable in $a \in \mathbb{A}$,

(A1.3) for $a, a' \in \mathbb{A}$ with $a \neq a'$, there exists a positive measure subset of x in $\mathcal{X}$ such that $f(x, a) \neq f(x, a')$,

(A1.4) the support of $f(\cdot, a)$ does not depend on a, and hence (without loss of generality) the support is all of $\mathcal{X}$ for all a,

(A1.5) w is a measurable function on $\mathcal{X}$,

(A1.6) the value function π is an increasing, continuous, and almost everywhere differentiable function,

(A1.7) the expected value of output, given action a, $\int \pi(x) f(x, a) dx$ is bounded for all a, (A1.8) the agent's utility for wage function u is continuously differentiable, increasing and strictly concave,

(A1.9) the agent's cost function c is increasing and continuously differentiable in a, and

(A1.10) the principal's utility function v is continuously differentiable, increasing and concave.

These assumptions are largely standard in the moral hazard literature. For instance, Assumption (A1.3) says that every two actions can be distinguished in terms of providing differing output distributions. From a statistical inference point-of-view it says that the actions are identifiable from the data. This assumption is used in a proof of uniqueness of a subproblem in Theorem 3.5 where the ability to distinguish actions is important for "breaking ties". We also make some additional technical assumptions that are less standard but required for our development.

Assumption 2. We make the following additional technical assumptions:

(A2.1) either $\lim_{y\to\infty} u(y) = \infty$ or $\lim_{y\to-\infty} v(y) = -\infty$, and

(A2.2) the minimum wage $\underline{w}$ and reservation utility $\underline{U}$ and least costly action $\underline{a}$ for the agent are such that $u(\underline{w}) - c(\underline{a}) < \underline{U}$.

Assumption (A2.1) is mild, but required for solvability of the Lagrangian dual studied in Section 3. Assumption (A2.2) ensures that paying the minimum wage is insufficient to compensate the agent above his reservation utility $\underline{U}$ even when the agent gives his lowest possible effort $\underline{a}$. This assumption is reasonable and useful in analyzing our relaxed problem in Section 3.

Following Grossman and Hart (1983), we simplify (1.1) by assuming a target action a^* is given and exploring properties of the optimal contract where the target action is a best response of the agent. Thus, our problem of interest is to find an optimal solution to the following problem (P):

$$
\begin{aligned}
\max_{w\geq\underline{w}} \quad & V(w, a^*) && (P)\\
\text{subject to} \quad & U(w, a^*) - U(w, \hat{a}) \geq 0 \quad \text{for all } \hat{a} \in \mathbb{A} && \text{(IC)}\\
& U(w, a^*) \geq \underline{U}, && \text{(IR)}
\end{aligned}
$$

given a^*. Note that the (abused) notation $w \geq \underline{w}$ means that $w(x) \geq \underline{w}$ for almost all x. Following standard terminology, the (IC) constraints is termed "incentive compatibility" and the (IR) constraint is termed "individual rationality". When an action a satisfies (IC) and (IR) for a given w we say a is a best response to w. The set of all best responses to the contract w is denoted $a^{\mathrm{BR}}(w)$. Any feasible solution w to (P) is said to *implement* action a^*.

Assumption 3. There exists an optimal contract w^{a^*} to (P) such that:

(A3.1) there exists at least one $\hat{a}^* \neq a^*$ such that $U(w^{a^*}, a^*) = U(w^{a^*}, \hat{a}^*)$; i.e., the (IC) constraint cannot be dropped in (P), and

(A3.2) $U(w^{a^*}, a^*) = \underline{U}$; i.e., the (IR) constraint is binding in (P).

The main strength of this assumption is the existence of an optimal solution. Existence is not a core focus of our study, instead we are interested in the structure of optimal solutions. Existing studies have paid careful attention to the issue of existence. For instance, Kadan et al. (2014)

provide weak sufficient conditions that guarantee the existence of an optimal solution. These conditions far from guarantee the validity of the first-order approach and thus do not jeopardize proving a result in the form of Theorem 1.1.

Assumptions (A3.1) and (A3.2) can be made without loss (once an optimal contract is known to exist). Indeed, if (A3.1) does not hold, then there is a unique best response to w^{a^*}, and in this setting the first-order approach applies (Mirrlees 1986). If the first-order approach applies then monotonicity of the optimal contract was already established by Rogerson (1985), and so we can ignore this case. Moreover, if (A3.2) does not hold we may simply redefine $\underline{U} = U(w^{a^*}, a^*)$ without loss of generality, making (IR) binding in (P). This does not change the optimal value or optimal solution of (P).

Assumption 3 is critical in establishing the validity of our approach. For its use in the proof of two key results see the proof of Corollary 4.3 and Lemma 4.14. Also see Remark 4.16 for further discussion.

A formal statement of our main result can now be made as follows:

Theorem 2.1. Suppose Assumptions 1–3 hold. If the output distribution f satisfies the MLRP then there exists an optimal contract that is nondecreasing in x.

The work of the remainder of the paper is to provide a proof of this result. Of course, insights and ideas that have relevance outside of the context of this result arise throughout and are discussed as appropriate.

Without further comment, Assumptions 1–3 are taken throughout. Any additional assumptions are written explicitly in the statements of results.

3 A relaxation and its desirable properties

In this section we define a family of relaxations of (P) that involves selecting (and making tight) a single no-jumping constraint. We establish strong analytical properties for these relaxed problems, including a necessary and sufficient optimality condition, as well as the continuity and monotonicity of optimal solutions that are central to later development.

For any $\hat{a} \in \mathbb{A}$ not equal to the target action a^*, define the problem

$$\max_{w \geq \underline{w}} \{V(w, a^*) : U(w, a^*) \geq \underline{U} \text{ and } U(w, a^*) - U(w, \hat{a}) = 0\}. \qquad (P|\hat{a})$$

We derive a characterization of optimal solutions to ($P|\hat{a}$) by studying the Lagrangian:

$$\mathcal{L}(w, \lambda, \delta|\hat{a}) = V(w, a^*) + \lambda[U(w, a^*) - \underline{U}] + \delta[U(w, a^*) - U(w, \hat{a})], \qquad (3.1)$$

where $\lambda \geq 0$ and δ (unsigned) are Lagrangian multipliers with respect to constraints $U(w, a^*) \geq \underline{U}$ and $U(w, a^*) - U(w, \hat{a}) = 0$, respectively. The Lagrangian dual of ($P|\hat{a}$) is

$$\inf_{\lambda \geq 0, \delta} \sup_{w \geq \underline{w}} \mathcal{L}(w, \lambda, \delta|\hat{a}). \qquad (3.2)$$

Our first step in analyzing the dual is to examine the inner maximization problem of (3.2) over w. By Assumption (A1.4) we can express the Lagrangian (3.1) as

$$\mathcal{L}(w, \lambda, \delta|\hat{a}) = \int L(w(x), \lambda, \delta|x, \hat{a}) f(x, a^*) dx$$

where $L(\cdot,\cdot,\cdot|x,\hat{a})$ is a function from $\mathbb{R}^3 \to \mathbb{R}$ with

$$\begin{aligned} L(y,\lambda,\delta|x,\hat{a}) &= v(\pi(x)-y) + \lambda(u(y)-c(a)-\underline{U}) + \delta\left[u(y)\left(1-\frac{f(x,\hat{a})}{f(x,a)}\right) - c(a) + c(\hat{a})\right] \\ &= v(\pi(x)-y) + \left[\lambda + \delta\left(1-\frac{f(x,\hat{a})}{f(x,a)}\right)\right]u(y) - \lambda(c(a)+\underline{U}) - \delta(c(a)-c(\hat{a})) \end{aligned} \tag{3.3}$$

where the ratio $1-\frac{f(x,\hat{a})}{f(x,a^*)}$ comes from factoring out $f(x,a^*)$ from the terms involving u. Note that we can divide by $f(x,a^*)$ since all of the f have the same support by (A1.4).

The inner maximization of $\mathcal{L}(w,\lambda,\delta|\hat{a})$ over w in (3.2) can be done pointwise at x through solving

$$\max_{y \geq \underline{w}} L(y,\lambda,\delta|x,\hat{a}) \tag{3.4}$$

for each x and setting $w(x) = y$ where y is an optimal solution to (3.4). There are two cases to consider. If $\lambda + \delta\left(1-\frac{f(x,\hat{a})}{f(x,a^*)}\right) \leq 0$ then $L(y,\lambda,\delta|x,\hat{a})$ is decreasing function of y since v is decreasing by Assumption (A1.10) and u is increasing by Assumption (A1.8). In this case the unique optimal solution to (3.4) is $y = \underline{w}$. On the other hand, if $\lambda + \delta\left(1-\frac{f(x,\hat{a})}{f(x,a^*)}\right) > 0$ then $L(y,\lambda,\delta|x,\hat{a})$ is strictly concave in y since v is concave and u is strictly concave (again by Assumptions (A1.10) and (A1.8)). Furthermore, if $\frac{\partial}{\partial y}L(\underline{w},\lambda,\delta|x,\hat{a}) \leq 0$ then the corner solution $y = \underline{w}$ is optimal, otherwise there exists a unique y such that the first-order condition $\frac{\partial}{\partial y}L(y,\lambda,\delta|x,\hat{a}) = 0$ holds, by strict concavity. In both cases (3.4) has a unique optimal solution that we denote by $w(x)$.

Hence, we can determine an optimal solution $w : \mathcal{X} \to \mathbb{R}$ to the inner maximization of (3.2) via the condition:

$$w(x)\begin{cases} \text{solves } \frac{\partial}{\partial y}L(w(x),\lambda,\delta|x,\hat{a}) = 0 & \text{if } \quad \lambda + \delta\left(1-\frac{f(x,\hat{a})}{f(x,a^*)}\right) > 0 \text{ and } \frac{\partial}{\partial y}L(\underline{w},\lambda,\delta|x,\hat{a}) > 0 \\ = \underline{w} & \text{otherwise.} \end{cases}$$

Expressing the derivatives (we may divide by $u'(w(x))$ since $u'(\cdot) > 0$ by (A1.8)) this is precisely

$$w(x)\begin{cases} \text{solves } \frac{v'(\pi(x)-w(x))}{u'(w(x))} = \lambda + \delta\left(1-\frac{f(x,\hat{a})}{f(x,a^*)}\right) & \text{if } \quad \frac{v'(\pi(x)-\underline{w})}{u'(\underline{w})} < \lambda + \delta\left(1-\frac{f(x,\hat{a})}{f(x,a^*)}\right) \\ = \underline{w} & \text{otherwise.} \end{cases} \tag{3.5}$$

Since v' and u' are both positive, the condition $\frac{v'(\pi(x)-\underline{w})}{u'(\underline{w})} < \lambda + \delta\left(1-\frac{f(x,\hat{a})}{f(x,a^*)}\right)$ implies that $\lambda + \delta\left(1-\frac{f(x,\hat{a})}{f(x,a^*)}\right) > 0$ and so this correctly handles both cases discussed in the previous paragraph.

Condition (3.5) allows us to partition the set of outcomes $\mathcal{X}$ into two sets:

$$\begin{aligned} \mathcal{X}_{\underline{w}} &:= \{x \in \mathcal{X} : w(x) = \underline{w}\} \\ &= \left\{x \in \mathcal{X} : \frac{v'(\pi(x)-\underline{w})}{u'(\underline{w})} \geq \lambda + \delta\left(1-\frac{f(x,\hat{a})}{f(x,a^*)}\right)\right\} \end{aligned} \tag{3.6}$$

and its complement in $\mathcal{X}$, denoted $\overline{\mathcal{X}_{\underline{w}}}$.

Contracts that satisfy (3.5) play a central role in our analysis, so we make a formal definition.

Definition 1. A contract satisfying (3.5) for a given $\hat{a}$ with parameters λ, δ is called a *generalized Mirrlees-Holmstrom* (GMH). We denote such a contract by $w_{\lambda,\delta}(\cdot|\hat{a})$.

If the action is binary, GMH contracts are the classical Mirrlees-Holmstrom contracts. GMH contracts have several desirable properties that we detail over the next few results. These properties are straightforward to show, but central to our development throughout the paper. First, GMH contracts are continuous. This follows from (3.5) and the continuity of v', u', π and $f(\cdot|\hat{a})$ given by Assumption 1.

Proposition 3.1. Every GMH contract $w_{\lambda,\delta}(x|\hat{a})$ is a continuous function of x, λ and δ.

Second, GMH contracts are monotone under certain conditions. The following result is standard, but included here for ease of reference.

Lemma 3.2. For any output distribution f that satisfies the MLRP, (i) if $a^* > \hat{a}$ then $1 - \frac{f(x,\hat{a})}{f(x,a^*)}$ is nondecreasing in x and (ii) if $a^* < \hat{a}$ then $1 - \frac{f(x,\hat{a})}{f(x,a^*)}$ is nonincreasing in x.

Proposition 3.3. Suppose the output distribution f satisfies the MLRP and let $w_{\lambda,\delta}(\cdot|\hat{a})$ be a GMH contract. Then w is a monotone function of x. In particular, if $\delta > 0$ and $a^* > \hat{a}$ then w is nondecreasing function of x.

Proof. Under MLRP, $1 - \frac{f(x,\hat{a})}{f(x,a^*)}$ is monotone in x (whether it is nondecreasing or nonincreasing depends on the $\hat{a}$). Thus, the ratio on the right-hand side of (3.5) is also either nondecreasing or nonincreasing. Thus by (3.5) and that fact π is an increasing function and v' and u' are decreasing functions, w is itself monotone. For the second statement in the proposition, note that if $\delta > 0$ then the right-hand side of (3.5) is increasing in the ratio $1 - \frac{f(x,\hat{a})}{f(x,a^*)}$. If $a^* > \hat{a}$ then the ratio $1 - \frac{f(x,\hat{a})}{f(x,a^*)}$ itself is nondecreasing in x by Lemma 3.2(ii). Together this implies w is a nondecreasing function of x. □

The following is a straightforward consequence of Proposition 3.3 and the fact a monotone function is almost everywhere differentiable.

Proposition 3.4. Suppose the output distribution f satisfies the MLRP and let w be a GMH contract. Then w is almost everywhere differentiable.

The main result of this section is to show that, under the MLRP, there is a unique choice of λ and δ such that $w_{\lambda,\delta}(\cdot|\hat{a})$ solves $(P|\hat{a})$.

Theorem 3.5. Given the target action a^* and alternate action $\hat{a}$ there exists unique Lagrangian multipliers $\lambda^*(\hat{a})$ and $\delta^*(\hat{a})$ and associated unique GMH contract $w^*_{\hat{a}} := w_{\lambda^*(\hat{a}),\delta^*(\hat{a})}(\cdot|\hat{a})$ such that (i) $w^*_{\hat{a}}$ satisfies (3.5) and is an optimal solution to $(P|\hat{a})$, and (ii) the following complementary slackness condition holds:

$$\lambda^*(\hat{a})[U(w^*_{\hat{a}}, a^*) - \underline{U}] = 0. \tag{3.7}$$

A detailed proof is found in Appendix A. The essential argument is to establish a strong duality result between $(P|\hat{a})$ and (3.2) and establish the uniqueness of the Lagrangian multipliers. Duality gives complementary slackness (3.7) and the uniqueness of Lagrange multipliers yields uniqueness of the optimal contract through (3.5).

The above provides the following necessary and sufficient optimality conditions for $(P|\hat{a})$.

Corollary 3.6. Suppose the output distribution f satisfies the MLRP. Then a feasible solution w to $(P|\hat{a})$ is an optimal solution to $(P|\hat{a})$ if and only if there exists a $\lambda \geq 0$ and δ such that w satisfies (3.5).

The plan for the next section is as follows. We have established two important properties of the family of relaxations $(P|\hat{a})$:

(a) Corollary 3.6: there exist necessary and sufficient conditions for a contract to be an optimal solution of $(P|\hat{a})$ given by (3.5), and

(b) Proposition 3.3: under the MLRP, an optimal solution to $(P|\hat{a})$ is monotone in x.

The task ahead is to develop necessary optimality conditions for optimal solutions of the original problem (P). Then, we establish sufficient conditions for when those necessary conditions boil down to (3.5) for some constants λ and δ. Then from (a) we conclude that this contract is an optimal solution to $(P|\hat{a})$ for some appropriately chosen $\hat{a}$. Finally, (b) provides sufficient conditions for that optimal contract to be monotone.

4 Manipulating first-order conditions

Our necessary optimality conditions are based on the following equivalent formulation of (P):

$$
\begin{aligned}
&\max_{w \geq \underline{w}} && V(w, a^*) \\
&\text{subject to} && \inf_{\hat{a} \in \mathbb{A}} \{U(w, a^*) - U(w, \hat{a})\} \geq 0 && (4.1) \\
& && U(w, a^*) \geq \underline{U}. && (4.2)
\end{aligned}
$$

We pull the minimization operator out from the constraint (4.1) and behind the objective function. This requires handling the possibility that a choice of w does not implement a^*, in which case (4.1) is violated. We handle this issue as follows. Define the set

$$
\mathcal{W}(\hat{a}) \equiv \{(w, a) : U(w, a) \geq \underline{U} \text{ and } U(w, a) - U(w, \hat{a}) \geq 0\},
$$

and the characteristic function

$$
V^I(w, a|\hat{a}) \equiv \begin{cases} V(w, a) & \text{if } (w, a) \in \mathcal{W}(\hat{a}) \\ -\infty & \text{otherwise.} \end{cases} \tag{4.3}
$$

This is constructed so that when maximizing $V^I(w, a|\hat{a})$ over (w, a) results in a finite objective value then $(w, a) \in \mathcal{W}(\hat{a})$. Similarly, if maximizing $\inf_{\hat{a} \in \mathbb{A}} V^I(w, a|\hat{a})$ over (w, a) results in a finite objective value then we know (w, a) lies in $\mathcal{W}(\hat{a}, b)$ for all $\hat{a} \in \mathbb{A}$. This implies (w, a) is feasible to (P) and demonstrates the equivalence of (P) and the problem

$$
\max_{a \in \mathbb{A}} \max_{w \geq \underline{w}} \inf_{\hat{a} \in \mathbb{A}} V^I(w, a|\hat{a}). \tag{Max-Min}
$$

Under conditions stated below, we determine an alternate best response that achieves the infimum in the definition of (Max-Min). See the discussion surrounding (4.31) and Remark 4.17 below.

To analyze (Max-Min) we use a variational argument. We fix an optimal solution w^{a^*} to (P). Define the family of variations

$$\mathcal{H} \equiv \{h(x) : h(x) = 0 \text{ for } x \in \{x : w^{a^*}(x) = \underline{w}\} \text{ and } 0 \leq h(x) \leq \min\{w^{a^*}(x) - \underline{w}, \bar{h}\}\} \tag{4.4}$$

where $\bar{h}$ is some positive real number. Every $h \in \mathcal{H}$ yields a family of contracts $w^{a^*} + zh$ for $z \in Z \equiv [-1, 1]$ that are "close" to w^{a^*}. Since (P) is equivalent to (Max-Min), $V(w^{a^*}, a^*) = \max_{w \geq \underline{w}} \inf_{\hat{a}} V^I(w|\hat{a}) = \max_{z \in Z} \inf_{\hat{a}} V^I(w^{a^*} + zh|\hat{a})$ where the last equality holds for any $h \in \mathcal{H}$ since $z = 0$ is an optimal choice of $z \in Z$. This observation drives our derivation of necessary optimality conditions for (P).

Our focus is the following single-dimensional optimization problem for a given $h \in \mathcal{H}$:

$$\begin{aligned} \max_{z \in [-1,1]} \quad & V(w^{a^*} + zh, a^*) \\ \text{subject to} \quad & U(w^{a^*} + zh, a^*) - U(w^{a^*} + zh, \hat{a}) \geq 0 \quad \text{for all } \hat{a} \in \mathbb{A} \\ & U(w^{a^*} + zh, a^*) \geq \underline{U}. \end{aligned} \tag{4.5}$$

As argued, $z^* = 0$ is an optimal solution to (4.5) and we seek to to uncover necessary optimality conditions for this solution. We put (4.5) into a more standard form for bilevel optimization and lighten the notation as:

$$\begin{aligned} \max_{z \in [-1,1]} \quad & B(z, a^*) \\ \text{subject to} \quad & a^* \in \arg\max_a b(z, a), \\ & b(z, a^*) - \underline{U} \geq 0, \end{aligned} \tag{4.6}$$

where

$$B(z, a) = V(w^{a^*} + zh, a), \text{ and} \tag{4.7}$$

$$b(z, a) = U(w^{a^*} + zh, a). \tag{4.8}$$

For reasons that will be clear in the proofs below (particularly in Theorem 4.7), we also apply a further restriction on variations to satisfy

$$b_z(0, a^*) = \int u'(w^{a^*}(x))h(x)f(x, a^*)dx > 0, \text{ and} \tag{4.9}$$

$$b_z(0, a^*) - b_z(0, \hat{a}^*) = \int u'(w^{a^*}(x)) \left(1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)}\right) h(x)f(x, a^*)dx > 0. \tag{4.10}$$

An h satisfying (4.9) and (4.10) certainly exists. For example, $h(x) \geq 0$ and $h(x)$ positively correlated with $\frac{1}{u'(w^{a^*})}(1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)})$ works. More concretely,

$$h(x) = \min\left\{\bar{h}, \frac{\alpha}{u'(w^{a^*})}(1 - \exp(-\left(1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)}\right)))\right\}$$

with some $\alpha > 0$ works. ◀

4.1 Penalty function approach

We now define a penalty function for (4.6) to derive optimality conditions. We are inspired by the development in Bertsekas (1999), but there are complications to that standard method. First, (4.6) has an "argmax" constraint that needs care to handle. Second, we want to design the penalty function to involve a single alternate best response. Our solution is the following penalty function involving five penalty terms. Let $\hat{a}^*$ be a given alternate best response (more on how to choose $\hat{a}^*$ below) and define the penalty function

$$\begin{aligned} B^k(z, \hat{a}|\hat{a}^*) = B(z, a^*) - \underbrace{\tfrac{k}{2} \min\{0, b(z, a^*) - \underline{U}\}^2}_{(i)} - \underbrace{\tfrac{\alpha}{2} |z|^2}_{(ii)} + \underbrace{\tfrac{k^{3/4}}{2} (\hat{a} - \hat{a}^*)^2}_{(iii)} \\ - \underbrace{\tfrac{k}{2} \min\{0, b(z, a^*) - b(z, \hat{a})\}^2}_{(iv)} - \underbrace{\tfrac{\sqrt{k}}{2} \min\{0, -z\}^2}_{(v)}. \end{aligned} \tag{4.11}$$

We may assume that $\hat{a}^* \neq a^*$ by Assumption (A3.1).

Let $(z^k, \hat{a}^k)$ denote an optimal solution to $\max_z \min_{\hat{a}} B^k(z, \hat{a}|\hat{a}^*)$. These solution exist since z and $\hat{a}$ both lie in compact sets and $B^k(z, \hat{a}|\hat{a}^*)$ is a continuous function. These optimal solutions form a sequence as $k \to \infty$.

The essence of our penalty function method is to relate the first-order conditions of the original optimization problem (4.6) to the limit of the first order conditions of $\max_z \min_{\hat{a}} B^k(z, \hat{a}|\hat{a}^*)$ as $k \to \infty$. The complication here is that we would like to "evacuate" any conditions involving the derivative of $\hat{a}$ to recover optimality conditions solely in z, the decision variable in (4.6).

To proceed we define the function

$$\varphi^k(z) = \min_{\hat{a}} B^k(z, \hat{a}|\hat{a}^*) \tag{4.12}$$

and observe that z^k is a maximizer of φ^k. It is not initially clear that φ^k is differentiable. We must understand how the optimal choice of $\hat{a}$ acts as a function of the choice z. To do so we examine the properties of the following set-valued function:

$$\zeta^k(z) = \operatorname{argmin}_{\hat{a}} B^k(z, \hat{a}|\hat{a}^*). \tag{4.13}$$

A key result below (Corollary 4.5) is that ζ is in fact a function of z in a neighborhood sufficiently close to z^k when k is large. This is a key result since if ζ was merely a set-valued function it would make it difficult to derive optimality conditions for z^k. This property allows to write that z^k as a local maximizer of

$$\varphi^k(z) = B^k(z, \zeta^k(z)|\hat{a}^*)$$

where we have now handled the minimization operation that was complicating the definition of φ^k in (4.12). The next key result (Proposition 4.6), using this new expression for φ^k, is to show that φ^k is a *differentiable* function on a sufficiently small neighborhood of z^k. At this point we can give a relatively straightforward optimality condition for z^k to the penalized problem:

$$\tfrac{d}{dz} \varphi^k(z^k) = 0.$$

The final remaining step is to observe that $\frac{d}{dz}\varphi^k(z^k) = B_z^k(z^k, \zeta^k(z^k)|\hat{a}^*)$ for z^k sufficiently close to 0. This is also achieved in Proposition 4.6 below. Finally, we argue that

$$\lim_{k\to\infty} B_z^k(z^k, \zeta^k(z^k)|\hat{a}^*) = 0 \tag{4.14}$$

provides necessary optimality conditions for (4.5). The final discussion of this subsection is to elaborate on (4.14) to develop clean optimality conditions to the original problem (P).

To establish the above results we need to lay some groundwork over a series of intermediate results. All results in this section stated without proof are located in Appendix B.

Proposition 4.1 (Exactness)**.** $\lim_{k\to\infty} B^k(z^k, \hat{a}^k\,|\hat{a}^*) = B(z^*, a^*) = V(w^{a^*}, a^*)$. In other words, the optimal value of penalized problem as a function of k converges to the optimal value of the original problem (P).

Exactness yields the following two corollaries.

Corollary 4.2. Any convergent subsequence of z^k converges to $z^* = 0$.

Corollary 4.3. Let $\{\hat{a}^k\}_{k=1}^{\infty}$ be any sequence where $\hat{a}^k \in \zeta^k(z^k)$. Then $\hat{a}^k \to \hat{a}^*$.

We are now ready to argue that $\zeta^k(z^k)$ (as defined in (4.13)) is a singleton. This is intuitive because there are only two terms that involve $\hat{a}$ in the penalty function: terms (iii) and (iv). Observe that term (iii) is strictly convex in $\hat{a}$ suggesting there is a unique minimizer. The work is to show that term (iv) is dominated by term (iii) for k sufficiently large.

Lemma 4.4. $\zeta^k(z^k)$ is a singleton for sufficiently large k.

In fact, the argument in this proof generalizes to yield the following corollary. Details are nearly identical except replacing z^k above with z sufficiently near z^k, and thus omitted.

Corollary 4.5. For sufficiently large k, $\zeta^k(r)$ is a singleton for every r in the neighborhood $\mathcal{N}_{1/k}(z^k)$ of z^k, where $\mathcal{N}_{1/k}(z^k) := \{z : ||z - z^k|| < 1/k\}$.

We are now ready to state the main result to leverage the penalty function approach. Given the previous results the proof follows a similar development to the standard envelope theorem. Details of the proof are in Appendix B.

Proposition 4.6. For k sufficiently large, $\varphi^k(z)$ is differentiable in z for all $z \in \mathcal{N}_{1/k}(z^k)$ with derivative $B_z^k(z, \zeta^k(z)|\hat{a}^*)$ where $\zeta^k(z)$ is the unique optimal solution to $\min_{\hat{a}} B^k(z, \hat{a}|\hat{a}^*)$.

The last result provides a first-order condition for z^k as a maximizer of $\varphi^k(z)$ in (4.12) for k sufficiently large:

$$0 = \tfrac{d}{dz}\varphi^k(z^k) = B_z^k(z^k, \hat{a}^k|\hat{a}^*) \tag{4.15}$$

where $\hat{a}^k$ is $\zeta^k(z^k)$. Hence, the optimal solution $z^* = 0$ has first-order condition

$$0 = \lim_{k\to\infty} B_z^k(z^k, \hat{a}^k|\hat{a}^*). \tag{4.16}$$

The next (and main) result of this subsection works with this expression to develop a sufficient condition that, under certain restrictions, begins to resemble (3.5).

Theorem 4.7. Let w^{a^*} be an optimal solution to (*P*) and $\hat{a}^*$ an alternate best response with $\hat{a}^* \neq a^*$. Let $h \in \mathcal{H}$ (where $\mathcal{H}$ is defined in (4.4)) satisfy (4.9) and (4.10). Then there exist strictly positive multipliers

$$\lambda_h := \theta_h \int u'(w^{a^*}(x))h(x)f(x,a^*)dx > 0, \tag{4.17}$$

$$\delta_h := \theta_h \int u'(w^{a^*}(x)) \left(1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)}\right) h(x)f(x,a^*)dx > 0, \tag{4.18}$$

where

$$\theta_h := -\lim_{n\to\infty} k_n \min\left\{0, z^{k_n}\right\} \tag{4.19}$$

denotes the limit of a convergent subsequence of $k \min\left\{0, z^{k_n}\right\}$, that satisfy the following necessary optimality condition for w^{a^*}:

$$\int \left(-v'(\pi(x) - w^{a^*}(x)) + u'(w^{a^*}(x)) \left[\lambda_h + \delta_h \left(1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)}\right)\right]\right) h(x)f(x,a^*)dx = 0. \tag{4.20}$$

Proof. The starting point is writing out (4.15) across the terms of the penalty function:

$$\begin{aligned} 0 = {} & B_z(z^k,a^*) - k\min\{0, b(z^k,a^*) - \underline{U}\}b_z(z^k,a^*) - \alpha z^k \\ & - k\min\left\{0, b(z^k,a^*) - b(z^k,\hat{a}^k)\right\}(b_z(z^k,a^*) - b_z(z^k,\hat{a}^k)) + \sqrt{k}\min\left\{0, -z^k\right\}. \end{aligned} \tag{4.21}$$

We develop (4.21) by making repeated use of Taylor's expansions and leveraging the convergence of $z^k \to z^* = 0$ and $\hat{a}^k \to \hat{a}^*$ from Corollaries 4.2 and 4.3. We are assisted by the following claim, which compares the rate of the convergence of these two sequences.

Claim 1. $\hat{a}^k - \hat{a}^*$ is $o(z^k)$

THe proof is in Appendix B. We now develop the first-order condition already established in (4.21). We write the first term as:

$$\begin{aligned} B_z(z^k,a^*) &= B_z(0,a^*) + z^k B_{zz}(0,a^*) + h.o.t. \\ &= B_z(0,\hat{a}^*) + O(z^k) \\ &= \int -v'(\pi(x) - w^{a^*}(x)h(x)f(x,a^*)dx + O(z^k) \end{aligned} \tag{4.22}$$

by taking the Taylor's expansion with respect to z^k about $z^* = 0$ and, in the last step, recalling the definition of B in (4.7). Using identical reasoning we can also write:

$$\begin{aligned} b_z(z^k,a^*) &= b_z(0,a^*) + O(z^k) \\ &= \int u'(w^{a^*}(x))h(x)f(x,a^*)dx + O(z^k) \end{aligned} \tag{4.23}$$

and

$$\begin{aligned} b_z(z^k,a^*) - b_z(z^k,\hat{a}^k) &= b_z(0,a^*) - b_z(0,\hat{a}^k) + O(z^k) \\ &= \int u'(w^{a^*}(x)) \left(1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)}\right) h(x)f(x,a^*)dx + O(z^k) \end{aligned} \tag{4.24}$$

recalling the definition of b in (4.8). Next, we develop the expressions in the "mins" in (4.21) by leveraging our assumptions (4.9) and (4.10). Taking the Taylor expansion with respect to z^k around $z^* = 0$ yields

$$\begin{aligned} b(z^k, a^*) - \underline{U} &= b(0, a^*) + z^k b_z(0, a^*) + o(z^k) - \underline{U} \\ &= z^k \int u'(w^{a^*}(x))h(x)f(x, a^*)dx + o(z^k) \end{aligned} \tag{4.25}$$

where the second line follows from (4.10) and the fact $b(0, a^*) = U(w^{a^*}, a^*) = \underline{U}$ by Assumption (A3.2). Similarly,

$$b(z^k, a^*) - b(z^k, \hat{a}^k) = b(0, a^*) - b(0, \hat{a}^k) + z^k(b_z(0, a^*) - b_z(0, \hat{a}^k)) + o(z^k)$$

by the Taylor's expansion with respect to z^k around $z^* = 0$. We then take the Taylor expansion with respect to $\hat{a}^k$ around $\hat{a}^*$ in the terms above involving $\hat{a}^k$ to yield:

$$b(z^k, a^*) - b(z^k, \hat{a}^k) = z^k(b_z(0, a^*) - b_z(0, \hat{a}^k)) + o(z^k) + O(\hat{a}^k - \hat{a}^*) \tag{4.26}$$

where we can cancel $b(0, a^*) - b(0, \hat{a}^*)$ since $U(w^{a^*}, a^*) = U(w^{a^*}, \hat{a}^*)$ because a^* and $\hat{a}^*$ are both best responses, and the fact that $b_a(0, a^*) = 0$ eliminates the first-order term $(\hat{a}^k - \hat{a}^*)b_a(0, \hat{a}^*)$ in the Taylor expansion. However, from Claim 1 we know $\hat{a}^k - \hat{a}^*$ is $o(z^k)$ and so we conclude from (4.26) that

$$b(z^k, a^*) - b(z^k, \hat{a}^k) = z^k(b_z(0, a^*) - b_z(0, \hat{a}^k)) + o(z^k). \tag{4.27}$$

Plugging (4.22)–(4.25) and (4.27) into (4.21) yields:

$$\begin{aligned} 0 \;=\; & \int -v'(\pi(x) - w^{a^*}(x))h(x)f(x, a^*)dx + O(z^k) - \alpha z^k \\ & -k\left(\min\{0, z^k \int u'(w^{a^*})h(x)f(x, a^*)dx + o(z^k)\}\right)(\int u'(w^{a^*})h(x)f(x, a^*)dx + O(z^k)) \\ & -k\left(\min\{0, z^k \int u'(w^{a^*})(1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)})h(x)f(x, a^*)dx + o(z^k)\}\right) \times \\ & \left[\int u'(w^{a^*})(1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)})h(x)f(x, a^*)dx + O(z^k)\right] + \sqrt{k}\min\{0, -z^k\} \end{aligned} \tag{4.28}$$

which by collecting terms amounts to:

$$\begin{aligned} 0 = & \int -v'(\pi(x) - w^{a^*}(x))h(x)f(x, a^*)dx \\ & - k\min\{0, z^k\}[(\int u'(w^{a^*})h(x)f(x, a^*)dx)^2 + (\int u'(w^{a^*})(1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)})h(x)f(x, a^*)dx)^2] \\ & + \sqrt{k}\min\{0, -z^k\} + O(z^k). \end{aligned} \tag{4.29}$$

To simplify this expression further, we argue that kz^k is bounded as $k \to \infty$. We first claim that the sequence $-k\min\{0, z^k\}$ is bounded. Suppose not. It follows that $kz^k \to -\infty$. When dividing both sides of (4.29) by $-\lim_{k\to\infty} k\min\{0, z^k\}$, and taking advantage of (4.9) and (4.10) then (4.29) becomes $0 = (\int u'(w^{a^*})h(x)f(x,a^*)dx)^2 + \left(\int u'(w^{a^*})(1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)})h(x)f(x,a^*)dx\right)^2$, a contradiction.

Now, since kz^k is bounded from below by the boundedness of $-k\min\{0, z^k\}$, it remains to show that kz^k is bounded from above. Suppose $kz^k \to \infty$, then $-k\min\{0, z^k\} = 0$. The first order condition (4.29) becomes $0 = \int -v'(\pi - w^{a^*})h(x)f(x,a^*)dx + \lim_{k\to\infty} \sqrt{k}\min\{0, -z^k\} < 0$, a contradiction. Therefore, kz^k is bounded and so the final penalty term $\sqrt{k}\min\{0, -z^k\} \to 0$. This allows us to drop the lower order terms in (4.29) and also from the boundedness of $-k\min\{0, z^k\}$ we may restrict k to a subsequence such that the limit

$$\theta_h := -\lim_{n\to\infty} k_n \min\{0, z^{k_n}\}$$

exists. We can thus take $n \to \infty$ in (4.29) to get

$$\begin{aligned} 0 = \int -v'(\pi - w^{a^*})h(x)f(x,a^*)dx + \lambda_h \int u'(w^{a^*})h(x)f(x,a^*)dx \qquad (4.30) \\ + \delta_h \int u'(w^{a^*})(1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)})h(x)f(x,a^*)dx \end{aligned}$$

where

$$\lambda_h \equiv \theta_h \int u'(w^{a^*})h(x)f(x,a^*)dx,$$

and

$$\delta_h \equiv \theta_h \int u'(w^{a^*})(1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)})h(x)f(x,a^*)dx,$$

as required in the statement of the theorem (equations (4.17) and (4.18)).

From (4.9) and (4.10) it suffices to show that $\theta_h > 0$ to establish inequalities in (4.17) and (4.18). This follows since if $\theta_h = 0$ then $\lambda_h = \delta_h = 0$, which violates (4.30) because $v'(\cdot) > 0$. Collecting terms in (4.30) we get (4.20), which finishes the proof. □

Remark 4.8. We remark that a key reason we designed a customized penalty function method to construct first-order conditions for our problem is the structure provided in (4.17) and (4.18). The connection of λ_h and δ_h via θ_h is critical in our development. See, for instance, the proofs of Lemma 4.10 and 4.14 below. ◀

We now specify a specific alternate best response $\hat{a}^*$ to form our penalty function as follows:

$$\hat{a}^* = \begin{cases} \min a^{BR}(w^{a^*}) & \text{if } a^* \neq \min a^{BR}(w^{a^*}) \\ \max a^{BR}(w^{a^*}) & \text{otherwise} \end{cases}. \qquad (4.31)$$

Note that the min and max of the set $a^{BR}(w^{a^*})$ both exist since that set is closed, following from the fact $U(w^{a^*}, a)$ is a continuous function of a. Also, reiterating Assumption (A3.1) we know that $a^{BR}(w^{a^*})$ is not a singleton and so $\hat{a}^* \neq a^*$ under this choice.

The reason for this choice of $\hat{a}^*$ is to make unambiguous the direction of the monotonicity of the ratio term $1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)}$ for $\hat{a}^*$ via Lemma 3.2. If $\hat{a}^* = \min a^{BR}(w^{a^*})$ then $1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)}$ is nondecreasing, otherwise it is nonincreasing. This clarity is important for establishing the monotonicity of the optimal contract later in the paper. In Section 5 we show that under MLRP, $\hat{a}^* = \min a^{BR}(w^{a^*})$ without loss. For now we need to work with the generality expressed in (4.31).

4.2 Deriving a GMH contract

Before continuing our development we introduce some convenient notation that we will employ for the remainder of the paper:

$$T(x) := \frac{v'(\pi(x)-w^{a^*}(x))}{u'(w^{a^*}(x))} \tag{4.32}$$

and

$$R(x) := 1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)}. \tag{4.33}$$

We did not introduce this notation in Section 3 because in the ratio term $1 - \frac{f(x,\hat{a})}{f(x,a^*)}$ in (4.33) we allowed any choice of $\hat{a}$. We work with a fixed $\hat{a}^*$ (as defined in (4.31)) for the rest of the paper, hence the notation $T(x)$ and $R(x)$ are not indexed by w^{a^*}, a^* or $\hat{a}^*$, all of which we now fix. We collect a few properties of the functions T and R that will prove useful. The proof is straightforward and thus omitted.

Proposition 4.9. The following hold: (i) $R(x)$ is a continuous function, (ii) $R(x)$ is not a constant function, (iii) without loss we may assume that $T(x)$ is not a constant function, (iv) $R(x)$ is a monotone function under MLRP, nonincreasing when $\hat{a}^* > a^*$ and nondecreasing when $\hat{a}^* < a^*$.

Returning to our argument (and using our new notation), (4.20) amounts to

$$\int_{\overline{\mathcal{X}^*_{\underline{w}}}} \left(-v'(\pi(x) - w^{a^*}(x)) + u'(w^{a^*}(x))\left[\lambda_h + \delta_h R(x)\right]\right) h(x) f(x,a^*)dx = 0. \tag{4.34}$$

where

$$\mathcal{X}^*_{\underline{w}} = \left\{x \in \mathcal{X} : w^{a^*}(x) = \underline{w}\right\} \tag{4.35}$$

since for all variations in $\mathcal{H}$, $h(x) = 0$ for $x \in \mathcal{X}^*_{\underline{w}}$. The next step is to show that if (4.34) holds for every $h \in \mathcal{H}$ satisfying (4.9) and (4.10) we can conclude for some fixed λ and δ and almost all $x \in \overline{\mathcal{X}^*_{\underline{w}}}$:

$$-v'(\pi(x) - w^{a^*}(x)) + u'(w^{a^*}(x))[\lambda + \delta R(x)] = 0, \tag{4.36}$$

and $w^{a^*}(x) = \underline{w}$ for $x \in \mathcal{X}^*_{\underline{w}}$. This results in precisely condition (3.5), once dividing through by $u'(w^{a^*}(x)) > 0$. Thus if (4.36) holds, we know w^{a^*} is a GMH contract with given alternate best response $\hat{a}^*$. That is, $w^{a^*} = w^*_{\hat{a}^*}$, since λ and δ are unique given $\hat{a}^*$ (the notation $w^*_{\hat{a}^*}$ comes from Theorem 3.5).

Conditions to ensure this logic holds involve the following definition. Two functions φ and ψ with shared domain $\mathcal{X}$ are *comonotone on the set* $S \subseteq \mathcal{X}$ if φ and ψ are either both nondecreasing or both nonincreasing on S.

Lemma 4.10. Let w^{a^*} be an optimal solution to (P) and $\hat{a}^*$ satisfy (4.31). If both $T(x)$ and $R(x)$ are comonotone functions of x on $\overline{\mathcal{X}^*_{\underline{w}}}$ and $T(x)$ is continuous on $\overline{\mathcal{X}^*_{\underline{w}}}$ then (4.36) holds for some constants $\lambda > 0$ and $\delta > 0$.

Proof. The proof treats the case where both $T(x)$ and $R(x)$ are nondecreasing functions of x on $\overline{\mathcal{X}^*_{\underline{w}}}$. The case where both are nonincreasing can be handled analogously with a change of sign in certain locations. Details are not included for the sake of brevity. Also, for simplicity of the argument we will assume that $\overline{\mathcal{X}^*_{\underline{w}}}$ is all of $\mathcal{X}$. The more general case is easily adapted but requires a denser notation we prefer to avoid. Moreover, the main ideas of the proof can be understood when assuming $\frac{1}{u'(w^{a^*}(x))}$ is bounded for all x. This is relaxed in Appendix C. This allows us to normalize any given $h \in \mathcal{H}$ that satisfies (4.9) and (4.10) to $h(x)/u'(w^{a^*}(x))$. In this setting, (4.20) becomes (using the notation $T(x)$ and $R(x)$ and rearranging):

$$\int [T(x) - R_h(x)]h(x)f(x, a^*)dx = 0 \tag{4.37}$$

where

$$R_h(x) = \lambda_h + \delta_h R(x). \tag{4.38}$$

$$T(x) = R_h(x) \tag{4.39}$$

for almost all x, then by the uniqueness of Lagrangian multipliers established in Theorem 3.5, this implies λ_h and δ_h are constant in h. Thus (4.36) holds and we are done.

Suppose, by way of contradiction to (4.39), there exists an $h_0 \in \mathcal{H}$ that satisfies (4.9) and (4.10) such that $T(x) \neq R_{h_0}(x)$ for x in a positive measure subset. We construct (see below) an alternate variation h_1 that satisfies the following properties:

$$\int h_1(x)f(x, a^*)dx = \int h_0(x)f(x, a^*)dx, \tag{4.40}$$

$$\int R(x)h_1(x)f(x, a^*)dx = \int R(x)h_0(x)f(x, a^*)dx, \tag{4.41}$$

$$\int T(x)h_1(x)f(x, a^*)dx = \int T(x)h_0(x)f(x, a^*)dx, \tag{4.42}$$

which together with (4.37) for $h = h_0$ implies

$$\int [T(x) - R_{h_0}(x)]h_1(x)f(x, a^*)dx = 0. \tag{4.43}$$

We will then argue that under the assumption that $T(x) \neq R_{h_0}(x)$ for x in a positive measure subset, that (perversely)

$$\int [T(x) - R_{h_0}(x)]h_1(x)f(x, a^*)dx > 0, \tag{4.44}$$

a contradiction. Thus it remains to construct an h_1 that satisfies (4.40)–(4.42). Note that these conditions do not require that h_1 lie in $\mathcal{H}$ nor satisfy (4.9) or (4.10).

Our construction of h_1 relies on the following sets:

$$\begin{aligned} \mathcal{X}^+ &:= \{x \in \mathcal{X} : R_{h_0}(x) > T(x)\}, \\ \mathcal{X}^- &:= \{x \in \mathcal{X} : R_{h_0}(x) < T(x)\}, \\ \mathcal{X}^{h_0+} &:= \{x \in \mathcal{X} : T(x) > C_{h_0}\}, \\ \mathcal{X}^{h_0-} &:= \{x \in \mathcal{X} : T(x) < C_{h_0}\}, \\ L_1 &:= \{x \in \mathcal{X} : R_{h_0}(x) < C_{h_0}\}, \text{ and} \\ L_2 &:= \{x \in \mathcal{X} : R_{h_0}(x) > C_{h_0}\}, \end{aligned} \tag{4.45}$$

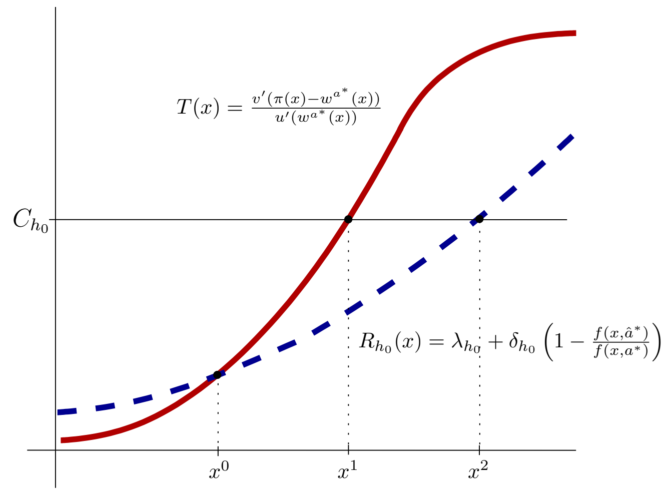


Figure 1: Illustrating the sets defined in the proof of Lemma 4.10.

where

$$C_{h_0} := \frac{\int R_{h_0}(x) h_0(x) f(x,a^*) dx}{\int h_0(x) f(x,a^*) dx} = \lambda_{h_0} + \delta_{h_0}\left(1 - \frac{\int h_0(x) f(x,\hat{a}^*) dx}{\int h_0(x) f(x,a^*) dx}\right), \tag{4.46}$$

which is a weighted-average of the values of R_{h_0} so that the coefficient on λ_{h_0} is 1 in the right-hand side of (4.46). By the first order condition (4.37), we also have

$$C_{h_0} = \frac{\int T(x) h_0(x) f(x,a^*) dx}{\int h_0(x) f(x,a^*) dx} = \frac{\int \frac{v'(\pi(x) - w^{a^*}(x))}{u'(w^{a^*}(x))} h_0(x) f(x,a^*) dx}{\int h_0(x) f(x,a^*) dx}.$$

Note that $T(x)$ is continuous (by assumption of the lemma) and $R_{h_0}(x)$ is continuous by Proposition 4.9(i), and so both $T(x)$ and $R_{h_0}(x)$ must intersect C_{h_0}. Also, by Proposition 4.9(iii), $T(x)$ is not a constant function and so the sets $\mathcal{X}^{h_0+}$ and $\mathcal{X}^{h_0-}$ have positive measure. Similarly, from Proposition 4.9(ii) the sets L_1 and L_2 have positive measure. Also, by the definition of h_0 we know $T(x)$ differs from $R_{h_0}(x)$ on a set of positive measure. Moreover, (4.37) implies that $T(x)$ cannot be almost everywhere greater (or less) than R_{h_0} and so $\mathcal{X}^+$ and $\mathcal{X}^-$ both have positive measure. We have thus established the existence of the three values of x – x^0, x^1, and x^2 – where these curves intersect. The point x^0 is the intersection point of R_{h_0} and T, x^1 is where T crosses C_{h_0}, and x^2 is where R_{h_0} crosses C_{h_0}. Because each of the sets in (4.45) have positive measure, this implies that x^0, x^1 and x^2 are all distinct.

To illustrate the above sets, consider the scenario illustrated in Figure 1. The variation h_0 is such that T crosses R_{h_0} at x^0 from below with $T(x^0) < C_{h_0}$. In this setting, $\mathcal{X}^+ = [0, x^0)$, $\mathcal{X}^- = (x^0, \infty)$, $\mathcal{X}^{h_0+} = (x^1, \infty)$, $\mathcal{X}^{h_0-} = [0, x^1)$, $L_1 = [0, x^2)$, and $L_2 = (x^2, \infty)$ (in the picture we have $\underline{x} = 0$ and $\overline{x} = +\infty$).

Claim 2. Suppose $T(x) \neq R_{h_0}(x)$ with positive probability in $\overline{\mathcal{X}^*_{\underline{w}}}$. Then (i) there exists an alternate variation h_1 (not necessarily in $\mathcal{H}$) that satisfies (4.40)–(4.44) if either

$$\Pr\left((L_1^- \cup L_2^-) \cap \mathcal{X}^{h_0-}\right) > 0 \text{ and } \Pr\left((L_1^- \cup L_2^-) \cap \mathcal{X}^{h_0+}\right) > 0 \text{ with } \Pr(L_i^-) > 0 \quad \text{or} \tag{4.47}$$

$$\Pr\left((L_1^+ \cup L_2^+) \cap \mathcal{X}^{h_0-}\right) > 0 \text{ and } \Pr\left((L_1^+ \cup L_2^+) \cap \mathcal{X}^{h_0+}\right) > 0 \text{ with } \Pr(L_i^+) > 0 \tag{4.48}$$

for $i = 1, 2$ where $L_i^j \equiv \mathcal{X}^j \cap L_i$, for $i \in \{1, 2\}, j \in \{+, -\}$ and (ii) there exists an $h_1 \in \mathcal{H}$ that satisfies (4.40) and (4.41) if (4.47) or (4.48) hold.

The main idea of the proof of the claim is that when either (4.47) or (4.48) hold there is sufficient flexibility to construct an h_1 to satisfy (4.40)–(4.42) by adjusting its values in the appropriate subregions of $[\underline{x}, \overline{x}]$. The proof of Claim 2 is straightforward but technical and so is included in Appendix C.

It remains to show that either (4.47) or (4.48) hold for our offending variation h_0. A detailed proof of this is in Appendix C and exhausts the different crossing patterns for T, R_{h_0} and C_{h_0}. Figure 1 illustrates Case 1, Subcase 1 (in the terminology of the proof in the appendix) where it is easy to see graphically that (4.47) holds. Indeed, $L_1^- = [x^0, x^2]$ and since x^0 and x^2 are distinct, this implies $\Pr(L_1^-) > 0$. Also, $[x_0, x_1] \subseteq (L_1^- \cup L_2^-) \cap \mathcal{X}^{h_0-}$ and since x^0 and x_1 are distinct this implies $\Pr\left((L_1^- \cup L_2^-) \cap \mathcal{X}^{h_0-}\right) > 0$. Similarly, $[x_1, x_2] \subseteq (L_1^- \cup L_2^-) \cap \mathcal{X}^{h_0+}$ and thus $\Pr\left((L_1^- \cup L_2^-) \cap \mathcal{X}^{h_0+}\right) > 0$. This establishes (4.47). We note that Claim 2(ii) is used in the degenerate case where the curves T and R_h intersect at C_{h_0} (see appendix for details).

It only remains to argue that the resulting constants λ and δ are strictly positive. This follows immediately by how they arise as constants λ_h and δ_h in (4.17) and (4.18) of Theorem 4.7, where strict positivity was previously established. □

Remark 4.11. The condition that $T(x)$ be continuous on $\overline{\mathcal{X}_{\underline{w}}^*}$ in Lemma 4.10 can also be assumed without loss. Indeed, it is known that there exists an optimal contract w^{a^*} that is continuous in our setting (thus establishing $T(x)$ is continuous on $\overline{\mathcal{X}_{\underline{w}}^*}$ under Assumption 1). This appears as Corollary 1 of Ke and Ryan (2015). For this reason, the caveat that $T(x)$ be continuous on $\overline{\mathcal{X}_{\underline{w}}^*}$ is dropped in all remaining theorem statements in the paper. ◀

Remark 4.12. An essential assumption for Lemma 4.10 to hold is that $\mathcal{X}$ is an interval. Under this assumption, if for some choice of h_0, $T(x) \neq R_{h_0}(x)$ with positive probability then Claim 2 holds. We now show that this need not be the case when $\mathcal{X}$ is discrete.

The simplest case is when there are two states of nature, $\mathcal{X} = \{x^0, x^1\}$. In this case, both $\mathcal{X}^-$ and $\mathcal{X}^+$ must contain exactly one element for (4.37) to be satisfied. Note also that L_1 and L_2 must have different elements so either $L_1^+ = \emptyset$ or $L_2^- = \emptyset$. Hence, there is no possibility of satisfying (4.47) or (4.48).

We now examine the phenomenon in three states. This same basic reasoning can apply to situations where $\mathcal{X}$ is even countably infinite and discrete. Consider the following setting exemplified by Figure 2 where there are three states of nature $\mathcal{X} = \left\{x^0, x^1, x^2\right\}$ and both $T(x)$ and R_{h_0} are nondecreasing. Observe that $\mathcal{X}^+ = \left\{x^0\right\}$, $\mathcal{X}^- = \left\{x^1, x^2\right\}$, $\mathcal{X}^{h_0-} = \left\{x^0\right\}$, $\mathcal{X}^{h_0+} = \left\{x^1, x^2\right\}$, $L_1 = \left\{x^0, x^1\right\}$, and $L_2 = \left\{x^2\right\}$. It is easy to check that neither (4.47) nor (4.48) hold. Hence, we cannot conclude that an optimal contract must satisfy (4.39) with some fixed Lagrange multipliers λ and δ using the reasoning provided above. ◀

Lemma 4.10 yields the immediate corollary:

Corollary 4.13. Let w^{a^*} be an optimal solution to (P) and $\hat{a}^*$ satisfy (4.31). If $T(x)$ and $R(x)$ are comonotone on $\overline{\mathcal{X}_{\underline{w}}^*}$ (the complement in $\mathcal{X}$ of the set $\mathcal{X}_{\underline{w}}^*$ defined in (4.35)) then w^{a^*} is equal to the unique optimal solution $w_{\hat{a}^*}^*$ to $(P|\hat{a}^*)$. In particular, w^{a^*} is a GMH contract with $\lambda^*(\hat{a}^*)$ and $\delta^*(\hat{a}^*)$ (as defined in Lemma 3.5) strictly positive.

Proof. This follows from Corollary 3.6 and Lemma 4.10. □

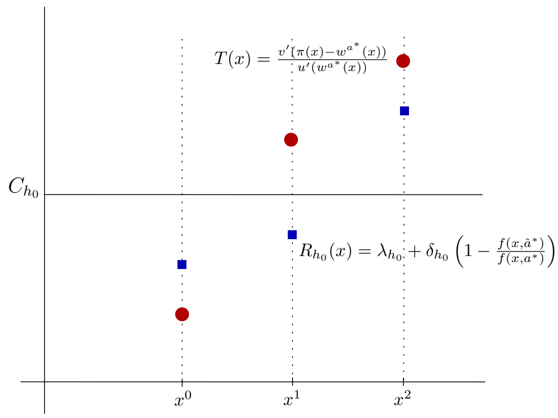


Figure 2: Illustrating the sets defined in the proof of Lemma 4.10.

4.3 Implications of the MLRP

In this subsection we will make repeated reference to the following function related to $T(x)$ (as defined in (4.32)):

$$\hat{T}(x) := \frac{v'(\pi(x) - w^*_{\hat{a}^*}(x))}{u'(w^*_{\hat{a}^*}(x))} \tag{4.49}$$

where $w^*_{\hat{a}^*}$ is the unique optimal solution to $(\mathrm{P}|\hat{a}^*)$ guaranteed by Theorem 3.5.

The goal of this subsection is uncover sufficient conditions for $T(x)$ and $R(x)$ to be comonotone, as required in Corollary 4.13. As the following lemma illustrates, the output distribution f satisfying the MLRP is one such sufficient condition. The proof is quite technical and so is included in Appendix C. An essential idea in the proof is to relate the properties of the optimal contract w^{a^*} to the GMH contract $w^*_{\hat{a}^*}$, which is monotone by Proposition 3.3. This is facilitated by the MLRP conditions.

Lemma 4.14. Let w^{a^*} be an optimal solution to (P) and $\hat{a}^*$ satisfy (4.31). If the output distribution f satisfies the MLRP then $T(x)$ and $R(x)$ are comonotone on $\overline{\mathcal{X}^*_{\underline{w}}}$.

The above lemmas establish the key result of Section 4.

Theorem 4.15. Let w^{a^*} be an optimal solution to (P) and $\hat{a}^*$ satisfy (4.31). If the MLRP holds then w^{a^*} is equal to the optimal solution $w^*_{\hat{a}^*}$ of $(P|\hat{a}^*)$. In particular, w^{a^*} is a GMH contract with $\lambda^*(\hat{a}^*), \delta^*(a^*) > 0$ (using the notation of Theorem 3.5).

Remark 4.16. We finish this section with some discussion of the importance of Assumption 3 to our development. A careful reading of the proof of Lemma 4.14 reveals that these assumptions are essential to establish comonotonicity. In particular, (C.14) is critical in connecting the monotonicity of the GMH contract $w^*_{\hat{a}^*}$ (via Proposition 3.3) to the optimal contract w^{a^*} and the crossing of T and $\hat{T}$.

At a high level, Assumption 3 plays a conceptual role in the execution of our approach. In order to connect the first-order conditions of (P) to the necessary and sufficient conditions for $(\mathrm{P}|\hat{a}^*)$ we cannot afford to send either λ_h or δ_h to zero in Theorem 4.7. Indeed, since $(\mathrm{P}|\hat{a}^*)$ has only two constraints, dropping to a single constraint to connect the optimality conditions of the original problem and the relaxed problem is insufficient for our characterization to go through.

The task of keeping both Lagrange multipliers positive is precisely why we only consider variations that satisfy (4.9) and (4.10), which ensures $\lambda_h > 0$ and $\delta_h > 0$ in Theorem 4.7 and (ultimately) $\lambda, \delta > 0$ in Theorem 4.15. Restricting attention to such variations suffices as long as Assumption 3 holds. Indeed, Corollary 4.3, a central result for the validity of the penalty function approach, requires Assumption 3(i) at a critical step.

Finally, maintaining $\delta > 0$ is critical in establishing our main Theorem 1.1. See the proof of Theorem 2.1 below. ◀

Remark 4.17. In a related remark, observe that Theorem 4.15 implies that the infimum in the definition of (Max-Min) is attained at $\hat{a} = \hat{a}^*$. Assumption 3 also plays a critical role here, ensuring that in (4.31) we can take $\hat{a}^* \neq a^*$. ◀

5 Monotonicity of optimal contracts

The main result of the previous section, Theorem 4.15, gives sufficient conditions for an optimal contract to our original problem (P) to be a GMH contract (as defined in Section 3). The final step of the paper is leverage the properties of GMH contracts (in particular, the monotonicity result in Proposition 3.3) to establish our main result.

Recall in Proposition 3.3 that a GMH contract $w_{\lambda,\delta}(\cdot|\hat{a})$ is nondecreasing if the associated $\delta > 0$ and $a^* > \hat{a}$. From Theorem 4.15 we already know that under the MLRP assumption, w^{a^*} is a GMH contract for alternate action $\hat{a}^*$ with $\delta^*(\hat{a}^*) > 0$. However, up until now we do not know if $\hat{a}^* < a^*$, only that $\hat{a}^*$ satisfies (4.31). The next result shows that if the MLRP holds then indeed $\hat{a}^* < a^*$. The proof appears in Appendix D. The result follows from the comonotonicity of $R(x)$ and $T(x)$ in Lemma 4.14 and how the monotonicity of $T(x)$ translates to the definition of $\hat{a}^*$.

Lemma 5.1. If the output distribution f satisfies the MLRP then $\hat{a}^*$ chosen via (4.31) must satisfy $\hat{a}^* < a^*$. In other words, $a^* \neq \min a^{BR}(w^{a^*})$.

We are now ready to prove the main result of the paper, Theorem 2.1.

Proof of Theorem 2.1. Given a target action a^* and an alternate best response $\hat{a}^*$ given by (4.31) there exists an optimal GMH contract w^{a^*} with multiplier $\delta^*(\hat{a}^*) > 0$ by Theorem 4.15. Lemma 5.1 implies $\hat{a}^* < a^*$ and so by Proposition 3.3, w^{a^*} is a nondecreasing function of x. □

The following example fits the setting of Theorem 2.1 but nonetheless the first-order approach is invalid. This example is adapted from a classical problem due to Holmstrom (1979) that fails the first-order approach but nonetheless satisfies the assumptions of Theorem 2.1.

Example 5.2. Consider the following principal-agent problem. The distribution of output X is exponential with $f(x,a) = \frac{1}{a}e^{-\frac{x}{a}}$, for $x \in \mathbb{R}_+$ and $a \in [1/10, 1/2]$ on $\mathcal{X} = \mathbb{R}$. The principal is risk-neutral (and so $v(y) = y$), the value of output is $\pi(x) = x$, the agent's utility is $u(y) = 2\sqrt{y}$, the agent's cost of effort $c(a) = 1 - (a - 1/2)^2$. The minimum wage $\underline{w} = 1/16$. It is straightforward to check that Assumptions 1 and 2 are satisfied. Existence of an optimal solution is guaranteed by

Kadan et al. (2014) and so Assumption 3 can also be satisfied. Hence Theorem 2.1 applies and an optimal monotone contract exists. Indeed, the reader may verify that

$$w^{a^*}(x) = \left[\tfrac{1}{2} + \tfrac{1}{16}(1 - (2+\sqrt{2})e^{-2x(1+\sqrt{2})})\right]^2.$$

with $a^* = 1/2$ is an optimal solution to (P). Clearly, w^{a^*} is nondecreasing.

However, if one uses the first order approach, using the first order condition $U_a(w, a) = 0$ to replace the original IC constraint, the resulting solution is $a^{\text{foa}} = 1/2$ and $w^{\text{foa}}(x) = 1/4$. Clearly, $w^{\text{foa}}(x)$ is a constant function and under $w^{\text{foa}}(x)$, the agent's optimal choice is $a = 1/10$, not $a^{\text{foa}} = 1/2$. Hence, the first-order approach fails. ◀

Our final example looks at a problem that fits the set up of Oyer (2000) and provides an optimal binary (two-value) contract but nonetheless fails the first-order approach, which is assumed in the development of Oyer (2000).

Example 5.3. Consider the following principal-agent problem. The distribution of output X is Pareto distribution $f(x, a) = 2a^2/x^3$ for $x \in [a, \infty)$, where $a \in [1/2, 1]$. The principal is risk-neutral (and so $v(y) = y$), the value of output is $\pi(x) = x$, the agent's utility is $u(y) = 2\sqrt{y}$. The agent's cost of effort is $c(a) = 3a$. The minimum wage $\underline{w} = 0$ and the reservation utility $\underline{U} = -1$.

First, we note that in this example, the first best contract is not implementable. The first-best contract w^{fb} when not equal to 0 (the minimum wage) satisfies (adapting (MH) with $\mu = 0$):

$$\tfrac{v'(\pi(x)-w(x))}{u'(w(x)))} = \lambda, \tag{5.1}$$

which after isolating for w and plugging into the IR constraint to solve for λ yields

$$w(x|a) = \begin{cases} \frac{(c(a)+\underline{U})^2}{4} & \text{if } x > a \\ 0 & \text{otherwise} \end{cases}$$

Given this contract structure, the first-best effort is

$$a^{fb} = \tfrac{7}{9} \in \arg\max_a \mathbb{E}X - \tfrac{1}{4}(c(a) + \underline{U})^2$$

where the argument of the arg max is the objective of the principal. However, this action is not implementable by the first best contract. Indeed, $w^{fb} = w(\cdot|a^{fb})$ has a best response of $a = 1/2$.

Next, we show that our approach can be adapted to solve for an optimal monotone contract. It is straightforward to check that Assumptions 1 and 2 are satisfied. We do remark that the distribution violates support independence assumption (A1.4), but it is straightforward to check that our approach is still applicable because of the simplicity of the structure of the problem, namely since $\underline{w} = 0$ and $u(\underline{w}) = 0$. Existence of an optimal solution is guaranteed by Kadan et al. (2014). We now verify that $a^* = 1$ with $\hat{a}^* = 1/2$ is such that the GMH contract $w^*_{\hat{a}^*}$ implements $a^* = 1$ with $U(w^*_{\hat{a}^*}, a^*) = \underline{U}$. It is outside of the scope of this paper to determine a^*, for the purposes of this example we take it as given.

The contract $w^*_{\hat{a}^*}$ has the characterization (3.5), which yields

$$w^*_{\hat{a}^*}(x) = \begin{cases} [\lambda + \delta(1 - \frac{(\hat{a}^*)^2}{(a^*)^2})]^2 & \text{if } x > a \\ 0 & \text{otherwise} \end{cases}$$

where plugging into the tight IR constraint yields

$$\lambda + \delta(1 - \frac{(\hat{a}^*)^2}{(a^*)^2}) = \frac{1}{2}(c(a^*) + \underline{U}).$$

The agent's utility under this contract is given by

$$U(w^*_{\hat{a}^*}, \tilde{a}) = \frac{2\tilde{a}^2}{(a^*)^2}\frac{1}{2}(c(a^*) + \underline{U}) - c(\tilde{a})$$

for any action $\tilde{a}$. Note that this is a convex function of $\tilde{a}$ and so a best response is on the boundary. In fact, both boundary points $1/2$ and 1 are optimal, justifying the definition of $a^* = 1$ and $\hat{a}^* = 1/2$. We have thus shown that $w^*_{\hat{a}^*}$ implements $a^* = 1$ and is therefore an optimal contract.

However, we can verify that $a^* = 1$ cannot be implemented by the contract derived from using the first-order approach. The first-order approach will pick the minimum of the agent's expected utility since, similar to the above case, and one can show that the agent's expected utility is convex in his effort. Suppose $a = 1$ is implemented by the contract determined by the (MH). This yields

$$w^{foa}(x|a) = \begin{cases} (\lambda + \frac{2\mu}{a})^2 & \text{if } x > a \\ 0 & \text{otherwise} \end{cases}$$

Plugging into the first order condition $U_a(w^{foa}(\cdot), a) = 0$ yields

$$(\lambda + \frac{2\mu}{a}) = \frac{1}{4}c'(a)a = \frac{3}{4},$$

which contradicts the IR constraint since

$$(\lambda + \frac{2\mu}{a}) = \frac{3}{4} < 1 = \frac{1}{2}(c(a) + \underline{U}).$$

Hence the first order approach fails. ◀

6 Conclusion

This paper provides sufficient conditions for the monotonicity of optimal contracts in the absence of the first-order approach. The key conditions are that the output distribution is defined over an interval of the real line and satisfies the MLRP. The connectedness of the output space is essential for our construction, which fails when the output can only take on discrete values.

Throughout the paper, the goal was to establish analytical properties of the optimal contract as a function of a given target action a^*. The question remains how, if at all, we can leverage the machinery here to determine an optimal pair (w^{a^*}, a^*) to the full moral-hazard problem. This is the subject of a future paper and requires additional insights into the structure of the single no-jumping constraint relaxations studied in Section 3.

This paper develops several novel optimization techniques to approach this problem that we believe have the potential for use in more general optimization problems. For instance, the penalty function approach can potentially be adapted to more general bilevel optimization problems besides the moral hazard setting. Also, our variational could have implications for deriving optimality conditions in other optimization settings.

Acknowledgments

We thank an anonymous associate editor and two anonymous referees for providing valuable feedback that greatly improved this manuscript. The second author thanks the University of Chicago Booth School of Business for its generous research support and the Department of Economics at the Chinese University of Hong Kong for hosting him during two extended research stays.

References

A. Araujo and H. Moreira. A general Lagrangian approach for non-concave moral hazard problems. *Journal of Mathematical Economics*, 35(1):17–39, 2001.

D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

J. Brosig, C. Lukas, and T. Riechmann. The monotonicity puzzle: an experimental investigation of incentive structures. *BuR-Business Research*, 3(1):8–35, 2010.

G. Carlier and R-A. Dana. Existence and monotonicity of solutions to moral hazard problems. *Journal of Mathematical Economics*, 41(7):826–843, 2005.

L.Y. Chu and G. Lai. Salesforce contracting under demand censorship. *Manufacturing & Service Operations Management*, 15(2):320–334, 2013.

J.R. Conlon. Two new conditions supporting the first-order approach to multisignal principal–agent problems. *Econometrica*, 77(1):249–78, 2009.

A.T. Coughlan. Salesforce compensation: A review of MS/OR advances. *Handbooks in Operations Research and Management Science*, 5:611–651, 1993.

T. Dai and K. Jerath. Salesforce compensation with inventory considerations. *Management Science*, 59(11): 2490–2501, 2013.

T. Dai and K. Jerath. Impact of inventory on quota-bonus contracts with rent sharing. *Operations Research*, 64(1):94–98, 2016.

S. Dempe and A.B. Zemkoho. The generalized Mangasarian-Fromowitz constraint qualification and optimality conditions for bilevel programs. *Journal of Optimization Theory and Applications*, 148(1):46–68, 2011.

S. Dempe, J. Dutta, and B.S. Mordukhovich. New necessary optimality conditions in optimistic bilevel programming. *Optimization*, 56(5-6):577–604, 2007.

S.J. Grossman and O.D. Hart. An analysis of the principal-agent problem. *Econometrica*, 51(1):7–45, 1983.

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

R.D. Innes. Limited liability and incentive contracting with ex-ante action choices. *Journal of Economic Theory*, 52(1):45–67, 1990.

I. Jewitt. Justifying the first-order approach to principal-agent problems. *Econometrica*, 56(5):1177–90, 1988.

I. Jewitt, O. Kadan, and J.M. Swinkels. Moral hazard with bounded payments. *Journal of Economic Theory*, 143(1):59–82, 2008.

J.Y. Jung and S.K. Kim. Information space conditions for the first-order approach in agency problems. *Journal of Economic Theory*, 160:243–279, 2015.

O. Kadan, P Reny, and J.M. Swinkels. Existence of optimal contract in the moral hazard problem. *Working paper*, 2014.

R. Ke and C.T. Ryan. Infinite-dimensional duality and moral hazard: Characterizing optimal contracts. *Working paper*, 2015.

R. Kirkegaard. Moral hazard and the spanning condition without the first-order approach. *Games and Economic Behavior*, 2017a.

R. Kirkegaard. A unifying approach to incentive compatibility in moral hazard problems. *Theoretical Economics*, 12(1):25–51, 2017b.

H. Krishnan and R.A. Winter. The economic foundations of supply chain contracting. *Foundations and Trends in Technology, Information and Operations Management*, 5(3–4):147–309, 2012.

Y.K. Kwon. Accounting conservatism and managerial incentives. *Management Science*, 51(11):1626–1632, 2005.

J.J. Laffont and D. Martimort. *The Theory of Incentives: The Principal-Agent model*. Princeton University Press, 2009.

R. Lal. Improving channel coordination through franchising. *Marketing Science*, 9(4):299–318, 1990.

Richard A Lambert. Contracting theory and accounting. *Journal of accounting and economics*, 32(1):3–87, 2001.

G.S. Liu, J.Y. Han, and J.Z. Zhang. Exact penalty functions for convex bilevel programming problems. *Journal of Optimization Theory and Applications*, 110(3):621–643, 2001.

P. Marcotte and D.L. Zhu. Exact and inexact penalty methods for the generalized bilevel programming problem. *Mathematical Programming*, 74(2):141–157, 1996.

P.R. Milgrom. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, pages 380–91, 1981.

J.A. Mirrlees. The theory of optimal taxation. *Handbook of Mathematical Economics*, 3:1197–1249, 1986.

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

G.E. Monahan and V. Vemuri. Monotone second-best optimal contracts. *European Journal of Operational Research*, 90(3):625–637, 1996.

M. Nasri. Characterizing optimal wages in principal-agent problems without using the first-order approach. *Optimization*, 65(2):467–478, 2016.

P. Oyer. A theory of sales quotas with limited liability and rent sharing. *Journal of Labor Economics*, 18 (3):405–426, 2000.

E. Plambeck and T.A. Taylor. Partnership in a dynamic production system with unobservable actions and noncontractible output. *Management Science*, 52(10):1509–1527, 2006.

E.L. Plambeck and S.A. Zenios. Performance-based incentives in a dynamic principal-agent model. *Manufacturing & Service Operations Management*, 2(3):240–263, 2000.

P. Renner and K. Schmedders. A polynomial optimization approach to principal–agent problems. *Econometrica*, 83(2):729–769, 2015.

W.P. Rogerson. The first-order approach to principal-agent problems. *Econometrica*, 53(6):1357–67, 1985.

B. Sinclair-Desgagné. The first-order approach to multi-signal principal-agent problems. *Econometrica*, 62 (2):459–65, 1994.

J. Wang, R. Zhao, and G. Sun. Selling strategy selection considering both the way of payment and inventory limitation. *Working paper*, 2016.

J.J. Ye and D.L. Zhu. Optimality conditions for bilevel programming problems. *Optimization*, 33(1):9–27, 1995.

J.J. Ye and D.L. Zhu. New necessary optimality conditions for bilevel programs by combining the MPEC and value function approaches. *SIAM Journal on Optimization*, 20(4):1885, 2010.

G. Zhang. Moral hazard in corporate investment and the disciplinary role of voluntary capital rationing. *Management Science*, 43(6):737–750, 1997.

A Appendix: Proofs for Section 3

A.1 Proof of Theorem 3.5

A.1.1 Existence

Here we will prove strong duality of ($P|\hat{a}$) and (3.2); that is, there exists an optimal dual solution to (3.2) that gives zero duality gap. This, in turn, establishes complementary slackness (3.7).

Let $\psi(\lambda, \delta) = \max_{w \geq \underline{w}} \mathcal{L}(w, \lambda, \delta|\hat{a})$. By the theorem of maximum and the fact that $\mathcal{L}$ is single-peaked (as we argued in the main text), ψ is a continuously differentiable function in λ and δ. Taking the derivative of ψ with respect to λ yields:

$$\frac{d\psi(\lambda,\delta)}{d\lambda} = \frac{\partial \mathcal{L}(w_{\lambda,\delta},\lambda,\delta|\hat{a})}{\partial \lambda} = U(w_{\lambda,\delta}, a^*) - \underline{U} \tag{A.1}$$

z by the envelope theorem where $w_{\lambda,\delta}$ is the unique optimal solution to $\max_{w \geq \underline{w}} \mathcal{L}(w, \lambda, \delta|\hat{a})$ for fixed λ and δ. Similarly,

$$\frac{d\psi(\lambda,\delta)}{d\delta} = U(w_{\lambda,\delta}, a^*) - U(w_{\lambda,\delta}, \hat{a}). \tag{A.2}$$

Since ψ is a convex function of λ and δ (it is the maximum of affine functions of λ and δ), a necessary and sufficient optimality condition for an interior optimal solution to (3.2) is setting $\frac{d\psi}{d\lambda} = 0$ and $\frac{d\psi}{d\delta} = 0$, which from (A.1) and (A.2) implies both constraints in ($P|\hat{a}$) are tight, ensuring zero duality gap. Thus, if there exists an interior point solution to the dual then we have strong duality.

For corner solutions the possibilities are $\lambda = 0$ or simply that $\lambda \to \infty$ or $\delta \to \pm\infty$ (we are more precise below). If $\lambda = 0$ then we again have complementary slackness. So it remains to consider scenarios where the "inf" defining the dual problem (3.2) corresponds to a divergent sequence of λ's or δ's. We show that this case cannot happen by deriving a contradiction.

To be more precise, by the definition of inf and the assumption that there is no finite λ or δ that solves the Lagrangian dual we know

$$\inf_{\lambda,\delta} \max_{w \geq \underline{w}} \mathcal{L}(w, \lambda, \delta|\hat{a}) = \lim_{k\to\infty} \min_{0\leq\lambda\leq k, |\delta|\leq k} \max_{w\geq\underline{w}} \mathcal{L}(w, \lambda, \delta|\hat{a}).$$

Let $(\lambda^k, \delta^k) \in \arg\min_{0\leq\lambda\leq k,|\delta|<k} \max_{w\geq\underline{w}} \mathcal{L}(w, \lambda, \delta|\hat{a})$ (the argmin is nonempty because the feasible region is compact and $\mathcal{L}$ is continuous in λ and δ) and by assumption at least one of λ^k and δ^k diverge. Construct the real sequence $\eta_k := \sqrt{(\lambda^k)^2 + (\delta^k)^2}$ where $\eta_k \to \infty$ as $k \to \infty$. If we divide (λ^k, δ^k) by η_k, the sequence $(1/\eta_k)(\lambda^k, \delta^k)$ is bounded and so there must exist a convergence subsequence indexed by k_n as $n \to \infty$. We denote the limit of that sequence by (λ', δ'); that is, $(\lambda', \delta') = \lim_{n\to\infty}(1/\eta_{k_n})(\lambda^{k_n}, \delta^{k_n})$.

The next step is to characterize the optimal solution $w_{\lambda',\delta'}$ to the inner maximization of the Lagrangian dual; that is, solve $\max_{w\geq\underline{w}} \mathcal{L}(w, \lambda', \delta'|\hat{a})$. The contradiction will come from an absurdity derived from characterizing $w_{\lambda',\delta'}$.

An intermediate step is to establish the following technical claims, that use the notation $\tilde{\mathcal{L}}(w, \lambda', \delta'|\hat{a}) = \mathcal{L}(w, \lambda', \delta'|\hat{a}) - V(w, a^*)$.

Claim 3. $\tilde{\mathcal{L}}(w_{\lambda^{k_n},\delta^{k_n}}, a; \lambda^{k_n}, \delta^{k_n}|\hat{a}) \leq 0$ for n sufficiently large.

This follows from the definition and differentiability of ψ defined at the outset of the proof. From (A.1) and the fact λ^{k_n} is not bounded we must have $\frac{d\psi(\lambda^{k_n},\delta^{k_n})}{d\lambda} = U(w_{\lambda^{k_n},\delta^{k_n}}, a^*) - \underline{U} < 0$. This drives $\lambda^{k^n}(U(w_{\lambda^{k_n},\delta^{k_n}}, a^*) - \underline{U}) \leq 0$ for n sufficiently large. Similarly from δ and hence the claim holds.

Claim 4. The following holds:

$$\max_{w \geq \underline{w}} \tilde{\mathcal{L}}(w, \lambda', \delta' | \hat{a}) = \lim_{n\to\infty} \max_{w \geq \underline{w}} \frac{\mathcal{L}(w, \lambda^{k_n}, \delta^{k_n} | \hat{a})}{\eta_{k_n}}. \tag{A.3}$$

The "$\leq$" direction of (A.3) follows by observing

$$\lim_{n\to\infty} \max_{w \geq \underline{w}} \frac{\mathcal{L}(w, \lambda^{k_n}, \delta^{k_n} | \hat{a})}{\eta_{k_n}} = \lim_{n\to\infty} \max_{w \geq \underline{w}} \frac{\tilde{\mathcal{L}}(w, \lambda^{k_n}, \delta^{k_n} | \hat{a}) + V(w, a^*)}{\eta_{k_n}} \geq \max_{w \geq \underline{w}} \tilde{\mathcal{L}}(w, \lambda', \delta' | \hat{a})$$

by taking the limit and noting that the $V(w_{\lambda^{k_n}, \delta^{k_n}}, a^*)$ is bounded below by Claim 3 and so $V(w_{\lambda^{k_n}, \delta^{k_n}}, a^*)/\eta_{k_n}$ converges to a number greater than or equal to 0 as $n \to \infty$. Now we turn to $\geq$ direction of (A.3). First, by weak duality, $\max_{w \geq \underline{w}} \mathcal{L}(w, \lambda^{k^n}, \delta^{k_n} | \hat{a})$ is bounded below by the optimal value of $(P|\hat{a})$ and so the right-hand side of (A.3) has a convergence subsequence. To abuse notation, we keep the same indices to index that convergent subsequence. Second, we can rewrite this right-hand side as

$$\begin{aligned} &\lim_{n\to\infty} \min_{\lambda \leq k_n, |\delta| \leq k_n} \max_{w \geq \underline{w}} \frac{\mathcal{L}(w, \lambda, \delta | \hat{a})}{\eta_{k_n}} \\ =\ &\lim_{n\to\infty} \min_{\tilde{\lambda} \leq \frac{k_n}{\eta_{k_n}}, |\tilde{\delta}| \leq \frac{k_n}{\eta_{k_n}}} \max_{w \geq \underline{w}} \left(\tilde{\mathcal{L}}(w, \tilde{\lambda}, \tilde{\delta} | \hat{a}) + \frac{V(w, a^*)}{\eta_{k_n}} \right), \end{aligned}$$

where $(\tilde{\lambda}, \tilde{\delta}) = \frac{1}{\eta_{k_n}}(\lambda, \delta)$. There are now two cases to consider.

Case 1: k_n/η_{k_n} is bounded. We can take a further subsequence k_{n_j} of the k_n such that $k_{n_j}/\eta_{k_{n_j}}$ converges to a constant $\bar{K}$. It follows that

$$\begin{aligned} &\lim_{n\to\infty} \min_{\tilde{\lambda} \leq \frac{k_n}{\eta_{k_n}}, |\tilde{\delta}| \leq \frac{k_n}{\eta_{k_n}}} \max_{w \geq \underline{w}} \left(\tilde{\mathcal{L}}(w, \tilde{\lambda}, \tilde{\delta} | \hat{a}) + \frac{V(w, a^*)}{\eta_{k_n}} \right) \\ =\ &\lim_{j\to\infty} \min_{\tilde{\lambda} \leq \frac{k_{n_j}}{\eta_{k_{n_j}}}, |\tilde{\delta}| \leq \frac{k_{n_j}}{\eta_{k_{n_j}}}} \max_{w \geq \underline{w}} \left(\tilde{\mathcal{L}}(w, \tilde{\lambda}, \tilde{\delta} | \hat{a}) + \frac{V(w, a^*)}{\eta_{k_{n_j}}} \right) \\ =\ &\min_{\tilde{\lambda} \leq \bar{K}, |\tilde{\delta}| \leq \bar{K}} \max_{w \geq \underline{w}} \tilde{\mathcal{L}}(w, \tilde{\lambda}, \tilde{\delta} | \hat{a}) \end{aligned}$$

where the first step is by the fact that any further subsequence k_{n_j} has the same limit as the original sequence. Note that $\lambda' = \lim_{n\to\infty} \frac{\lambda^{k_n}}{\eta_{k_n}} = \lim_{j\to\infty} \frac{\lambda^{k_{n_j}}}{\eta_{k_{n_j}}} \leq \lim_{j\to\infty} \frac{k_{n_j}}{\eta_{k_{n_j}}} = \bar{K}$, and similarly, $|\delta'| \leq \lim_{j\to\infty} \frac{k_{n_j}}{\eta_{k_{n_j}}} = \bar{K}$. Therefore, we have

$$\min_{\tilde{\lambda} \leq \bar{K}, |\tilde{\delta}| \leq \bar{K}} \max_{w \geq \underline{w}} \tilde{\mathcal{L}}(w, \tilde{\lambda}, \tilde{\delta} | \hat{a}) \leq \max_{w \geq \underline{w}} \tilde{\mathcal{L}}(w, \lambda', \delta' | \hat{a}),$$

since (λ', δ') is a feasible solution to the minimization on the left-hand side. Tracing back the equalities above, this establishes the "$\geq$" direction of (A.3).

Case 2: $\frac{k_n}{\eta_{k_n}}$ is unbounded. If $\frac{k_n}{\eta_{k_n}}$ is unbounded, we denote the set $B_n \equiv \{(\tilde{\lambda}, \tilde{\delta}) : \tilde{\lambda} \leq \frac{k_n}{\eta_{k_n}}, |\tilde{\delta}| \leq \frac{k_n}{\eta_{k_n}}\}$. The limit of the sequence of set B_n exists because the following fact:

$$\bigcup_{j=1}^{\infty} \left(\bigcap_{n=j}^{\infty} B_n \right) = \bigcap_{j=1}^{\infty} \left(\bigcup_{n=j}^{\infty} B_n \right) = \{(\tilde{\lambda}, \tilde{\delta}) : \tilde{\lambda} \in \mathbb{R}_+, |\tilde{\delta}| \in \mathbb{R}_+\}.$$

Therefore, passing to the limit, we have

$$\lim_{n\to\infty} \min_{\tilde{\lambda}\le \frac{k_n}{\eta_{k_n}}, |\tilde{\delta}|\le \frac{k_n}{\eta_{k_n}}} \max_{w\ge \underline{w}} \left(\tilde{\mathcal{L}}(w,\tilde{\lambda},\tilde{\delta}|\hat{a}) + \frac{V(w,a^*)}{\eta_{k_n}} \right) = \min_{\tilde{\lambda}\le \lim_{n\to\infty} \frac{k_n}{\eta_{k_n}}, |\tilde{\delta}|\le \lim_{n\to\infty} \frac{k_n}{\eta_{k_n}}} \max_{w\ge \underline{w}} \tilde{\mathcal{L}}(w,\tilde{\lambda},\tilde{\delta}|\hat{a}).$$

Recall $\lambda' = \lim_{n\to\infty} \lambda^{k_n}/\eta_{k_n} \le \lim_{n\to\infty} k_n/\eta_{k_n}$ and $|\delta'| \le \lim_{n\to\infty} k_n/\eta_{k_n}$, we obtain

$$\min_{\tilde{\lambda}\le \lim_{n\to\infty} \frac{k_n}{\eta_{k_n}}, |\tilde{\delta}|\le \lim_{n\to\infty} \frac{k_n}{\eta_{k_n}}} \max_{w\ge \underline{w}} \tilde{\mathcal{L}}(w,\tilde{\lambda},\tilde{\delta}|\hat{a}) \le \max_{w\ge \underline{w}} \tilde{\mathcal{L}}(w,\lambda',\delta'|\hat{a}),$$

again by the definition of the minimum. This yields the "$\ge$" direction of (A.3). Finally, this establishes the claim.

We use Claim 4 to characterize the optimal solution to $\max_{w\ge\underline{w}} \tilde{\mathcal{L}}(w,\lambda',\delta'|\hat{a})$. Observe that this optimization problem has the same optimal solution set as

$$\lim_{n\to\infty} \max_{w\ge\underline{w}} \frac{\mathcal{L}(w,\lambda^{k_n},\delta^{k_n}|\hat{a})}{\eta_{k_n}}$$

by writing out the definition of λ' and δ' and taking the limit out front by continuity. Since η_{k_n} is a constant, this is the same as optimizing over simply $\mathcal{L}(w,\lambda^{k_n},\delta^{k_n}|\hat{a})$. Then by (A.3) in Claim 4, we see that this is equivalent optimization problem as $\max_{w\ge\underline{w}} \tilde{\mathcal{L}}(w,\lambda',\delta'|\hat{a})$. As in the main body of the paper (see discussion surrounding (3.4)), we can solve this problem pointwise by maximizing over y for each x the following (3.3):

$$\lambda'(u(y) - c(a^*) - \underline{U}) + \delta'\left[u(y)\left(1 - \frac{f(x,\hat{a})}{f(x,a^*)}\right) - c(a^*) + c(\hat{a})\right].$$

However, this problem has a very simple structure so that its optimal solution w' satisfies

$$w'(x) = \begin{cases} \underline{w} & \text{if } \lambda' + \delta'\left(1 - \frac{f(x,\hat{a})}{f(x,a^*)}\right) \le 0 \\ \infty & \text{if } \lambda' + \delta'\left(1 - \frac{f(x,\hat{a})}{f(x,a^*)}\right) > 0 \end{cases}. \tag{A.4}$$

In other words, the assumption that the sequence (λ^k, δ^k) is unbounded (the start of our contradiction proof) forces the optimal solution to the $\max_{w\ge\underline{w}} \mathcal{L}(w,\lambda',\delta'|\hat{a})$ to have the "strange" form (A.4).

The last step is to observe that characterization of the optimal solution provides a contradiction. We now leverage the two assumptions (A2.1) and (A2.2). According to (A2.1) there are two cases to consider.

Case 1: $\lim_{y\to\infty} u(y) = \infty$. Suppose $\lambda' + \delta'(1 - \frac{f(x,\hat{a})}{f(x,a^*)}) > 0$ with positive measure. Since $\lim_{y\to\infty} u(y) = \infty$, we have

$$\int u(w')f(x,a^*)dx - c(a^*) > \underline{U}$$

since $\Pr(\{w' \to \infty\}) > 0$. It follows by reasoning similar to the outset of the proof (the differentiability and convexity of ψ) that $\lambda' = 0$. Therefore, $\delta'(1 - \frac{f(x,\hat{a})}{f(x,a^*)}) > 0$ with positive measure and $\delta' > 0$ and $(1 - \frac{f(x,\hat{a})}{f(x,a^*)}) > 0$ with positive measure. Hence

$$\int u(w')(1 - \frac{f(x,\hat{a})}{f(x,a^*)})f(x,a^*)dx - [c(a^*) - c(\hat{a})] > 0,$$

since $\Pr(\{w' \to \infty\} \cap \{(1 - \frac{f(x,\hat{a})}{f(x,a^*)}) > 0\}) > 0$. This is a contradiction, since again according to the logic of the outset of the proof this would drive $\delta' \to -\infty$ but, in fact, $\delta' > 0$. So the only possibility is $\lambda' + \delta'(1 - \frac{f(x,\hat{a})}{f(x,a^*)}) \leq 0$ a.e., which implies $w' = \underline{w}$. However, this is ruled out by Assumption (A2.2).

Case 2: $\lim_{y\to-\infty} v(y) = -\infty$. This case is aided by Claim 3.

Now, returning to our characterization of w' in (A.4), let us assume that there is a set of positive measure where $\lambda' + \delta'(1 - \frac{f(x,\hat{a})}{f(x,a^*)}) > 0$. We have:

$$\begin{aligned}\inf_{\lambda,\delta}\max_{w\geq\underline{w}} \mathcal{L}(w,\lambda,\delta|\hat{a}) &= \lim_{n\to\infty} \eta_{k_n} \min_{\lambda\leq k_n, |\delta|\leq k_n} \max_{w\geq \underline{w}} \tfrac{\mathcal{L}(w,\lambda,\delta|\hat{a})}{\eta_{k_n}} \\ &= \lim_{n\to\infty}\eta_{k_n}\left(\frac{V(w_{\lambda^{k_n},\delta^{k_n}},a^*)}{\eta_{k_n}} + \frac{1}{\eta_{k_n}}\tilde{\mathcal{L}}(w_{\lambda^{k_n},\delta^{k_n}},\lambda^{k_n},\delta^{k_n}|\hat{a})\right) \leq \lim_{n\to\infty} V(w_{\lambda^{k_n},\delta^{k_n}},a^*),\end{aligned}$$

where in the third step we utilizes Claim 3. By equivalence (A.3), as $n \to \infty$, $w_{\lambda^{k_n},\delta^{k_n}}$ converges to w' pointwise and so $\lim_{n\to\infty} V(w_{\lambda^{k_n},\delta^{k_n}},a^*) < V(w^{a^*},a^*)$ since $\lim_{y\to-\infty} v(y) = -\infty$, where $V(w^{a^*},a^*)$ is the optimal value of the original problem. However,

$$\inf_{\lambda,\delta}\max_{w\geq\underline{w}} \mathcal{L}(w,\lambda,\delta|\hat{a}) \leq \lim_{n\to\infty} V(w_{\lambda^{k_n},\delta^{k_n}},a^*) < V(w^{a^*},a^*) \leq \mathrm{val}(P|\hat{a}).$$

where $\mathrm{val}(P|\hat{a})$ is the optimal value of $(P|\hat{a})$. This contradicts weak duality. Therefore, the only remaining possibility is $w'(x) = \underline{w}$ almost everywhere. However, this contradicts (A2.2). This establishes strong duality.

A.1.2 Uniqueness

We now turn to the question of uniqueness. We argued above that for a given λ and δ there is a unique optimal solution to the inner minimization in (3.2) given by (3.5) that we have denoted $w_{\lambda,\delta}(\cdot|\hat{a})$. Suppose the Lagrangian dual has two optimal solutions (λ,δ) and (λ',δ'). By strong duality, $(\lambda,\delta,w_{\lambda,\delta}(\cdot|\hat{a}))$ and $(\lambda',\delta',w_{\lambda',\delta'}(\cdot|\hat{a}))$ are both saddle points of the Lagrangian function (3.1), and so by saddle point optimality $w_{\lambda,\delta}(\cdot|\hat{a})$ and $w_{\lambda',\delta'}(\cdot|\hat{a})$ are both optimal to $(P|\hat{a})$. Moreover, we claim that $w_{\lambda,\delta}(\cdot|\hat{a})$ and $w_{\lambda',\delta'}(\cdot|\hat{a})$ are equal. Indeed, we know by feasibility that

$$\mathcal{L}(w_{\lambda,\delta}(\cdot|\hat{a}),\lambda',\delta'|\hat{a}) \leq \mathcal{L}(w_{\lambda',\delta'}(\cdot|\hat{a}),\lambda',\delta'|\hat{a}) = V^*, \tag{A.5}$$

where V^* denotes the optimal value of $(P|\hat{a})$, by strong duality and since $w_{\lambda',\delta'}(\cdot|\hat{a})$ is an optimal solution of the inner maximization of (3.2) given λ' and δ'. On the other hand, we have

$$\begin{aligned}\mathcal{L}(w_{\lambda,\delta}(\cdot|\hat{a}),\lambda',\delta'|\hat{a}) &= V(w_{\lambda,\delta}(\cdot|\hat{a}),a^*) + \lambda'[U(w_{\lambda,\delta}(\cdot|\hat{a}),a^*) - \underline{U}] \\ &\qquad + \delta'[U(w_{\lambda,\delta}(\cdot|\hat{a}),a^*) - U(w_{\lambda,\delta}(\cdot|\hat{a}),\hat{a})] \\ &\geq V^*\end{aligned}$$

where the equality comes from writing out (3.1) and the inequality follows since $w_{\lambda,\delta}(\cdot|\hat{a})$ is an optimal solution to $(P|\hat{a})$ and $\lambda' \geq 0$. Observe that this inequality cannot be strict as it will violate (A.5). Hence, the inequality is tight, implying that $w_{\lambda,\delta}(\cdot|\hat{a})$ is also a maximizer of the inner maximization of (3.2) given λ' and δ'. Hence, $w_{\lambda,\delta}(\cdot|\hat{a}) = w_{\lambda',\delta'}(\cdot|\hat{a})$ for almost all x by the uniqueness of solutions to the inner maximization.

Now we show that $w_{\lambda,\delta}(\cdot|\hat{a}) = w_{\lambda',\delta'}(\cdot|\hat{a})$ implies $(\lambda,\delta) = (\lambda',\delta')$. We discuss two cases. These cases refer to the set $\mathcal{X}_{\underline{w}}$ defined in (3.6).

Case 1: $\Pr(X \in \mathcal{X}_{\underline{w}}|a^*) = 0$ where $\Pr(\cdot|a^*)$ is the measure for output associated with action a^* given by the pdf $f(x,a^*)$. Since $w_{\lambda,\delta}(\cdot|\hat{a}) = w_{\lambda',\delta'}(\cdot|\hat{a})$ then for almost all x we have via (3.5):

$$\lambda + \delta\left(1 - \tfrac{f(x,\hat{a})}{f(x,a^*)}\right) = \lambda' + \delta'\left(1 - \tfrac{f(x,\hat{a})}{f(x,a^*)}\right). \tag{A.6}$$

Taking the expectation of both sides of (A.6) over the domain $\mathcal{X}$ yields $\lambda = \lambda'$, since

$$\int \left(1 - \tfrac{f(x,\hat{a})}{f(x,a^*)}\right) f(x,a^*)dx = 0.$$

Thus, $\delta\left(1 - \tfrac{f(x,\hat{a})}{f(x,a^*)}\right) = \delta'\left(1 - \tfrac{f(x,\hat{a})}{f(x,a^*)}\right)$ for almost all x. Then via Assumption (A1.3) we can conclude $\delta = \delta'$.

Case 2: $\Pr(X \in \mathcal{X}_{\underline{w}}|a^*) > 0$.

We discuss two subcases, depending on whether $1 - \frac{f(x,\hat{a})}{f(x,a)}$ is a constant in the region $x \in \overline{\mathcal{X}}_{\underline{w}}$.

Subcase 2.1: $\left(1 - \tfrac{f(x,\hat{a})}{f(x,a^*)}\right)$ is not constant for $\overline{\mathcal{X}}_{\underline{w}}$. There exist two values x_1 and x_2 in $\overline{\mathcal{X}}_{\underline{w}}$ such that $\left(1 - \tfrac{f(x_1,\hat{a})}{f(x_1,a^*)}\right) \neq \left(1 - \tfrac{f(x_2,\hat{a})}{f(x_2,a^*)}\right)$. Then since $w_{\lambda,\delta}(\cdot|\hat{a}) = w_{\lambda',\delta'}(\cdot|\hat{a})$, by (3.5) we have for $i = 1, 2$:

$$\lambda + \delta(1 - \tfrac{f(x_i,\hat{a})}{f(x_i,a^*)}) = \lambda' + \delta'(1 - \tfrac{f(x_i,\hat{a})}{f(x_i,a^*)}).$$

Taking the difference of these two equations yields:

$$\delta\left[(1 - \tfrac{f(x_1,\hat{a})}{f(x_1,a^*)}) - (1 - \tfrac{f(x_2,\hat{a})}{f(x_2,a^*)})\right] = \delta'\left[(1 - \tfrac{f(x_1,\hat{a})}{f(x_1,a^*)}) - (1 - \tfrac{f(x_2,\hat{a})}{f(x_2,a^*)})\right],$$

which implies $\delta = \delta'$, and thus $\lambda = \lambda'$.

Subcase 2.2: $1 - \frac{f(x,\hat{a})}{f(x,a)}$ is a constant.

We show by contradiction that this subcase will not occur. Suppose $1 - \frac{f(x,\hat{a})}{f(x,a)} = C$ for all $x \in \overline{\mathcal{X}}_{\underline{w}}$. We show first under this supposition that the contract is a constant. Then we show that when $1 - \frac{f(x,\hat{a})}{f(x,a)} = C$, the optimal optimal contract is the first best contract, which implies $\frac{v'(\pi-w)}{u'(w)}$ will be also a constant. These two facts are in contradiction, by the continuity of $\frac{v'(\pi-w)}{u'(w)}$ if $\Pr(X \in \mathcal{X}_{\underline{w}}) > 0$.

Now we show $w_{\lambda,\delta}$ is constant for $x \in \overline{\mathcal{X}}_{\underline{w}}$. This result comes from the contrary statement that that the Lagrangian multipliers are not unique. Note that the Lagrangian multipliers are the solution to the following equation system

$$\begin{aligned} \lambda[U(w_{\lambda,\delta},a) - \underline{U}] &= 0 \\ U(w_{\lambda,\delta},a) - U(w_{\lambda,\delta},\hat{a}) &= 0. \end{aligned} \tag{A.7}$$

When the IR constraint is slack and $\lambda = 0$, then since $w_{\lambda,\delta}$ is monotone in δ (and thus U is monotone in δ) then if δ is not unique, we have

$$\begin{aligned} 0 &= \tfrac{\partial}{\partial\delta}[U(w_{\lambda,\delta},a) - U(w_{\lambda,\delta},\hat{a})] \\ &= \tfrac{\partial}{\partial\delta}\int u(w_{\lambda,\delta})(1 - \tfrac{f(x,\hat{a})}{f(x,a)})f(x,a)dx, \end{aligned}$$

which implies $(1-\frac{f(x,\hat{a})}{f(x,a)}) = C = 0$ since u is an increasing function. Therefore, we obtain a contradiction of (3.5) for $w_{\lambda,\delta}$ that

$$0 < \frac{v'(\pi - w_{\lambda,\delta})}{u'(w_{\lambda,\delta})} = \lambda + \delta C = 0.$$

Therefore, we only consider the case where $\lambda > 0$. Let $\lambda(\delta)$ be the unique λ solving $U(w_{\lambda,\delta}, a) = \underline{U}$ (this is unique since $w_{\lambda,\delta}$ is strictly increasing in λ) and hence so is $U(w_{\lambda,\delta}, a)$, and plugging $\lambda(\delta)$ into the no-jump equality (A.7) above and take the derivative with respect to δ, we obtain

$$\begin{aligned}
&\frac{\partial}{\partial\delta}[U(w_{\lambda(\delta),\delta}, a) - U(w_{\lambda(\delta),\delta}, \hat{a})] \\
= \; &\frac{\partial\lambda(\delta)}{\partial\delta}\int u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\lambda}(1-\frac{f(x,\hat{a})}{f(x,a)})f(x,a)dx \\
&+\int u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\delta}(1-\frac{f(x,\hat{a})}{f(x,a)})f(x,a)dx \\
= \; &\int u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\delta}(1-\frac{f(x,\hat{a})}{f(x,a)})f(x,a)dx \\
&-\frac{\int u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\delta}f(x,a)dx}{\int u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\lambda}f(x,a)dx}\int u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\lambda}(1-\frac{f(x,\hat{a})}{f(x,a)})f(x,a)dx. \qquad \text{(A.8)}
\end{aligned}$$

By the characterization of $w_{\lambda,\delta}$, we have

$$\frac{\partial w_{\lambda,\delta}}{\partial\delta} = \frac{\partial w_{\lambda,\delta}}{\partial\lambda}(1-\frac{f(x,\hat{a})}{f(x,a)}) = u'(w_{\lambda,\delta})\frac{1}{\frac{\partial}{\partial w}(\frac{v'(\pi-w)}{u'(w)})\Big|_{w=w_{\lambda,\delta}}}(1-\frac{f(x,\hat{a})}{f(x,a)}).$$

Therefore, if we write $\sqrt{u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\lambda}} = Z_1$ and $\sqrt{u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\lambda}}(1-\frac{f(x,\hat{a})}{f(x,a)}) = Z_2$ as two random variables, we can rewrite the derivative over δ as

$$\frac{\partial}{\partial\delta}[U(w_{\lambda(\delta),\delta}, a) - U(w_{\lambda(\delta),\delta}, \hat{a})] = \mathbb{E}[Z_2^2 | X \in \overline{\mathcal{X}}_{\underline{w}}] - \frac{\mathbb{E}[Z_1Z_2|X\in\overline{\mathcal{X}}_{\underline{w}}]^2}{\mathbb{E}[Z_1^2|X\in\overline{\mathcal{X}}_{\underline{w}}]}.$$

using (A.8). By the Cauchy-Schwartz inequality,

$$\mathbb{E}[Z_2^2 | X \in \overline{\mathcal{X}}_{\underline{w}}] - \frac{\mathbb{E}[Z_1Z_2|X\in\overline{\mathcal{X}}_{\underline{w}}]^2}{\mathbb{E}[Z_1^2|X\in\overline{\mathcal{X}}_{\underline{w}}]} = 0$$

only occurs if Z_1 and Z_2 are perfectly linearly correlated (that is, $Z_1 = \alpha_1 + \alpha_2 Z_2$), which implies

$$\alpha_1 + \frac{\alpha_2}{\sqrt{u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\lambda}}} = 1 - \frac{f(x,\hat{a})}{f(x,a)}$$

for all $x \in \overline{\mathcal{X}}_{\underline{w}}$. When $1-\frac{f(x,\hat{a})}{f(x,a)} = C$ is a constant, it follows that $u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\lambda} = \frac{\partial}{\partial\lambda}u(w_{\lambda,\delta})$ is a constant over all $x \in \overline{\mathcal{X}}_{\underline{w}}$, which implies $w_{\lambda,\delta}$ is a constant over $\overline{\mathcal{X}}_{\underline{w}}$. Therefore, we have a step contract

$$w_{\lambda,\delta} = \begin{cases} \underline{w} \text{ for } x \in \mathcal{X}_{\underline{w}} \\ w^c \text{ for } x \in \overline{\mathcal{X}}_{\underline{w}} \end{cases}$$

where w^c solves $\alpha_1 + \frac{\alpha_2}{\sqrt{u'(w_{\lambda,\delta})\frac{\partial w_{\lambda,\delta}}{\partial\lambda}}} = 1 - \frac{f(x,\hat{a})}{f(x,a)} = C$.

Now we show the second result that $\frac{v'(\pi - w_{\lambda,\delta})}{u'(w_{\lambda,\delta})}$ is also a constant and $w_{\lambda,\delta} = w^{fb}$. From the first order condition we have

$$\frac{v'(\pi - \underline{w})}{u'(\underline{w})} \leq \frac{v'(\pi - w_{\lambda^*,\delta^*})}{u'(w_{\lambda^*,\delta^*})} = \lambda^* + \delta^* C$$

is a constant, where (λ^*, δ^*) is any Lagrangian multiplier associated with the optimal solution. The constraint

$$\int u(w_{\lambda^*,\delta^*}) C f(x,a) dx = c(a) - c(\hat{a})$$

is satisfied. Now we replace (λ^*, δ^*) by $(\lambda', \delta') = (\lambda^* + \delta^* C, 0)$, the solutions $w_{\lambda',\delta'}$ and $w_{\lambda,\delta}$ are the same. Therefore $\delta = 0$ is an alternative Lagrangian multiplier of the problem. If so, by the strong duality

$$\lambda' = \arg\min_{\lambda \geq 0} \max_{w \geq \underline{w}} \mathcal{L}(w, \lambda, 0),$$

we have $\lambda' = \lambda^{fb}$. It follows that $\frac{v'(\pi - w_{\lambda^*,\delta^*})}{u'(w_{\lambda^*,\delta^*})} = \lambda^{fb}$. As we have argued, for $x \in \overline{\mathcal{X}}_{\underline{w}}$, $w_{\lambda^*,\delta^*} = w^c = w^{fb}$ is a constant, then $v'(\pi - w_{\lambda^*,\delta^*})$ must be a constant over all $x \in \overline{\mathcal{X}}_{\underline{w}}$. Then we derive a contradiction by the continuity of $\frac{v'(\pi - w^{fb})}{u'(w^{fb})}$. As we know when $\Pr(X \in \mathcal{X}_{\underline{w}}) > 0$, there must exist a cut-off x^c such that

$$\frac{v'(\pi(x^c) - \underline{w})}{u'(\underline{w})} = \lambda^* + \delta^*(1 - \frac{f(x^c,\hat{a})}{f(x^c,a)}) = \frac{v'(\pi(x^c) - w^c)}{u'(w^c)},$$

which however contradicts the fact that

$$\frac{v'(\pi(x^c) - w^c)}{u'(w^c)} = \frac{v'(\pi - w^{fb})}{u'(w^{fb})} = \lambda^{fb} > \frac{v'(\pi - \underline{w})}{u'(\underline{w})}$$

since $\frac{v'(\pi - w^{fb})}{u'(w^{fb})}$ is a constant over $x \in \overline{\mathcal{X}}_{\underline{w}}$. Therefore, we show that non-uniqueness of Lagrangian multiplier will not occur where $1 - \frac{f(x,\hat{a})}{f(x,a)}$ is a constant and $\Pr(X \in \mathcal{X}_{\underline{w}}) > 0$.

This completes the proof.

B Appendix: Proofs for the penalty function approach in Section 4.1

B.1 Proof of Proposition 4.1

We prove in two directions. The first is "$\geq$" and its proof is straightforward since

$$\begin{aligned}
B^k(z^k, \hat{a}^k | \hat{a}^*) &\geq B^k(0, \hat{a}^k | \hat{a}^*) \\
&= B(0, a^*) - \tfrac{k}{2} \min\{0, b(0, a^*) - \underline{U}\}^2 + \tfrac{k^{3/4}}{2}(\hat{a}^k - \hat{a}^*)^2 \\
&\quad - \tfrac{k}{2} \min\{0, b(0, a^*) - b(0, \hat{a}^k)\}^2 \\
&\geq B(0, a^*) \\
&= V(w^{a^*}, a^*),
\end{aligned}$$

where we note that $\min\{0, b(0, a^*) - \underline{U}\} = 0$ and $\min\{0, b(0, a^*) - b(0, \hat{a})\} = 0$ since $b(0, a^*) \geq b(0, \hat{a})$ for any $\hat{a}$.

It remains to show the other direction "≤". We start by observing

$$
\begin{aligned}
& \lim_{k\to\infty} B^k(z^k, \hat{a}^k \,|\hat{a}^*) \\
\leq\ & \lim_{k\to\infty} B^k(z^k, \hat{a}^* \,|\hat{a}^*) \\
=\ & \lim_{k\to\infty} \left[\begin{array}{c} B(z^k, a^*) - \frac{k}{2}\min\{0, b(z^k, a^*) - \underline{U}\}^2 - \frac{\alpha}{2}\left|z^k\right|^2 \\ -\frac{k}{2}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^*)\}^2 - \frac{\sqrt{k}}{2}\min\{0, -z^k\}^2 \end{array} \right] \\
\leq\ & \lim_{k\to\infty} B(z^k, a^*), \qquad \text{(B.1)}
\end{aligned}
$$

where the first inequality follows from the definition of z^k and that $\hat{a}^*$ may not be in $\zeta^k(z^k)$. The equality comes from writing out $B^k(z^k, \hat{a}^* \,|\hat{a}^*)$ and noting the penalty term (iii) equals zero since we take $\hat{a} = \hat{a}^*$. The second inequality follows from dropping negative terms.

To work with equation (B.1) note that z^k is bounded sequence so it has a convergent subsequence. We take any such subsequence and denote its limit as z_∞. By the continuity of B, we continue from (B.1) to write:

$$\lim_{k\to\infty} B^k(z^k, \hat{a}^k \,|\hat{a}^*) \leq B(z_\infty, a^*). \qquad \text{(B.2)}$$

The result follows if we establish the following claim:

Claim 5. z_∞ is a feasible solution to (4.6).

Indeed, if the claim holds then from (B.2), then the "≤" direction holds:

$$\lim_{k\to\infty} B^k(z^k, \hat{a}^k \,|a^*, \hat{a}^*) \leq B(z_\infty, a^*) \leq B(0, a^*) = V(w^{a^*}, a^*), \qquad \text{(B.3)}$$

which follows by the fact $z^* = 0$ is an optimal solution to (4.6).

It remains to show that Claim 5 holds. This is achieved by showing z_∞ satisfies the two constraints of (4.6):

$$
\begin{aligned}
b(z_\infty, a^*) - \underline{U} \geq 0, \text{ and} \qquad \text{(B.4)} \\
b(z_\infty, a^*) \geq b(z_\infty, \hat{a}), \ \forall \hat{a} \in [\underline{a}, \overline{a}]. \qquad \text{(B.5)}
\end{aligned}
$$

To show (B.4) holds we leverage the fact that we have already shown the "≥" direction of exactness. Indeed, suppose (B.4) does not hold and $b(z_\infty, a^*) - \underline{U} < 0$. Then term (i) in the penalty function diverges to $-\infty$ at a linear rate in k. This implies $\lim_{k\to\infty} B^k(z^k, \hat{a}^k|\hat{a}^*) \to -\infty$ since all terms in the penalty function except term (iii) are negative and term (iii) goes to $+\infty$ in at most rate $k^{3/4}$ since $(\hat{a}^k - \hat{a}^*)^2$ is a bounded sequence, (since $\mathbb{A}$ is a bounded set). Then the "≥" direction of exactness implies $V(w^{a^*}, a^*) = -\infty$, but this is a contradiction since we have assumed (P) has an optimal solution and so $V(w^{a^*}, a^*) > -\infty$. Hence we may conclude (B.4) holds.

To establish (B.5) we again proceed by contradiction. Suppose

$$\exists \hat{a}' \in [\underline{a}, \overline{a}] \text{ such that } b(z_\infty, a^*) - b(z_\infty, \hat{a}') < 0. \qquad \text{(B.6)}$$

We again use the "≥" direction of exactness to derive a contradiction. Let $\tilde{a}^k \in \arg\max_{\hat{a}} b(z^k, \hat{a})$ and let $\tilde{a}_\infty$ be the limit of a convergence subsequence of the $\tilde{a}^k$ (such a limit exists since $[\underline{a}, \overline{a}]$ is compact). We redefine the k sequence to that subsequence, and abuse notation by keeping the index k the same. Now, write

$$\begin{aligned}
&\lim_{k\to\infty} B^k(z^k, \hat{a}^k \,|\hat{a}^*) \\
&= \lim_{k\to\infty} \min_{\hat{a}} B^k(z^k, \hat{a} \,|\hat{a}^*) \\
&\leq \lim_{k\to\infty} B^k(z^k, \tilde{a}^k \,|\hat{a}^*) \\
&\leq \lim_{k\to\infty} \left[B(z^k, a^*) + \tfrac{k^{3/4}}{2}(\tilde{a}^k - \hat{a}^*)^2 - \tfrac{k}{2}\min\{0, b(z^k, a^*) - b(z^k, \tilde{a}^k)\}^2\right], \qquad \text{(B.7)}
\end{aligned}$$

where the first equality is by the definition of $\hat{a}^k$ and the first inequality comes from the definition of the "min". The second inequality writes out the definition of $B^k(z^k, \tilde{a}^k \,|\hat{a}^*)$ dropping negative terms of our choosing.

For the subsequence $\tilde{a}^k \to \tilde{a}_\infty$, we have $\tilde{a}_\infty \in \arg\max b(z_\infty, \hat{a})$ by the upper hemicontinuity of the argmax set. Moreover, by (B.6)

$$b(z_\infty, a^*) < b(z_\infty, \hat{a}') \leq b(z_\infty, \tilde{a}_\infty)$$

and this drives $\min\{0, b(z^k, a^*) - b(z^k, \tilde{a}^k)\}$ to be a strictly negative number in the limit. Therefore, right-hand side of (B.7) diverges to $-\infty$ at rate k. This dominates the only positive term (iii) that diverges to $+\infty$ at rate $k^{3/4}$. Hence, "$\geq$" contradicts the feasibility of the moral hazard problem. This establishes (B.5) and hence Claim 5. This establishes the result.

B.2 Proof of Corollary 4.2

Consider the following chain of inequalities:

$$\begin{aligned}
\lim_{k\to\infty} B^k(z^k, \hat{a}^*|\hat{a}^*) &= \lim_{k\to\infty} B(z^k, a^*) - \tfrac{k}{2}\min\left\{0, b(z^k, a^*) - \underline{U}\right\}^2 \\
&\quad - \tfrac{\alpha}{2}|z^k|^2 - \tfrac{k}{2}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^*)\}^2 - \tfrac{\sqrt{k}}{2}\min\{0, -z^k\}^2 \qquad \text{(B.8)} \\
&\leq B(z_\infty, a^*) \\
&\leq B(z^*, a^*)
\end{aligned}$$

where the first equality holds since term penalty term (iii) is zero, the first inequality holds by dropping negative terms and noting $z^k \to z_\infty$ and B is continuous in z and the last inequality holds since z_∞ is a feasible solution to (4.6) and z^* is an optimal solution.

By exactness we know $\lim_{k\to\infty} B^k(z^k, \hat{a}^k|\hat{a}^*) = B(z^*, a^*)$ and so all of the above inequalities are equalities. In particular, all the negative terms in (B.8) are equal to 0. This shows that $z^k \to z^*$.

B.3 Proof of Corollary 4.3

For now we assume that each $\hat{a}^k$ satisfies the first-order condition for sufficiently large k (we discuss corner solutions below):

$$B^k_{\hat{a}}(z^k, \hat{a}^k \,|\hat{a}^*) = k^{3/4}(\hat{a}^k - \hat{a}^*) + k\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\}b_a(z^k, \hat{a}^k) = 0.$$

Dividing the above equality by $k^{3/4}$, we get for all interior points $\hat{a}^k$:

$$(\hat{a}^k - \hat{a}^*) + k^{1/4}\sqrt{k}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\}b_a(z^k, \hat{a}^k) = 0. \qquad \text{(B.9)}$$

If we can show that the second term in (B.9) converges to 0 as $k \to \infty$ then we may conclude

$$\lim_{k\to\infty} (\hat{a}^k - \hat{a}^*) = 0,$$

as desired, since $b_a(z^k, \hat{a}^k)$ is uniformly bounded. That is, it suffices to show

$$k^{1/4}\sqrt{k}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\}b_a(z^k, \hat{a}^k) \to 0 \tag{B.10}$$

First, observe that $\lim_{k\to\infty} \sqrt{k}(\min\{0, -z^k\})^2 = 0$ from the exactness of penalty function (term v). Next we claim $\sqrt{k}\min\{0, z^k\} \to 0$. From term (i) of the penalty function we take the Taylor expansion of $b(z, a^*)$ in z around $z = 0$. By Assumption 3 we know $b(0, a^*) = U(w^{a^*}, a^*) = \underline{U}$ and from (4.9) we have $b_z(0, a^*) > 0$. Hence, term (i) of the penalty function diverges to $-\infty$ (violating exactness) unless $k\min\{0, z^k\} \to 0$. We have thus shown

$$\sqrt{k}(z^k)^2 = \max\{\sqrt{k}(\min\{0, z^k\})^2, \sqrt{k}(\min\{0, -z^k\})^2\} \to 0. \tag{B.11}$$

We now return to establishing (B.10). Note that by the Taylor expansion around $z = 0$,

$$b(z^k, a^*) - b(z^k, \hat{a}^k) = b(0, a^*) - b(0, \hat{a}^k) + z^k(b_z(0, a^*) - b_z(0, \hat{a}^k)) + o(z^k).$$

Since $b(0, a^*) - b(0, \hat{a}^k) \geq 0$ by the definition of a^*, we have

$$\begin{aligned}
0 &\geq k^{\frac{1}{4}}(\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\}) \\
&= k^{\frac{1}{4}}(\min\{0, b(0, a^*) - b(0, \hat{a}^k) + z^k(b_z(0, a^*) - b_z(0, \hat{a}^k)) + o(z^k)\}) \\
&\geq k^{\frac{1}{4}}(\min\{0, z^k(b_z(0, a^*) - b_z(0, \hat{a}^k)) + o(z^k)\}) \\
&= \min\{0, k^{\frac{1}{4}}z^k(b_z(0, a^*) - b_z(0, \hat{a}^k)) + o(k^{\frac{1}{4}}z^k)\} \\
&\to 0,
\end{aligned}$$

where the last step is by $\sqrt{k}(z^k)^2 \to 0$ from (B.11) and the fact that $b_z(0, a^*) - b_z(0, \hat{a}^k)$ is uniformly bounded for all k.

It only remains to consider corner solutions. We may assume that $\hat{a}^k$ are lower corner solutions $\hat{a}^k = \bar{a}$ for sufficiently large k, upper corner solutions are analogous. Note that it suffices to consider the case where $\hat{a}^k$ is a corner for sufficiently large k since if the current sequence of $\hat{a}^k$, for instance, alternated between interior and corner solutions for sufficiently large k we could simply restrict to the subsequence that converged to interior solutions and use the above argument.

Since $\hat{a}^k$ is an lower corner solution we know

$$B^k_{\hat{a}}(z^k, \hat{a}^k\,|\hat{a}^*) = k^{3/4}(\hat{a}^k - \hat{a}^*) + k\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\}b_a(z^k, \hat{a}^k) \geq 0.$$

Again dividing through by $k^{3/4}$ and using (B.11), this boils down to

$$\hat{a}^k - \hat{a}^* \to \delta$$

where $\delta \geq 0$. Since $\hat{a}^k = \underline{a}$ for sufficiently large k this implies that $\hat{a}^* \leq \underline{a}$. If $\hat{a}^* \neq \underline{a}$ then $\hat{a}^* < \underline{a}$, a contradiction of feasibility. Hence we conclude that $\hat{a}^k \to \hat{a}^*$, as desired.

B.4 Proof of Lemma 4.4

We treat two separate cases, whose proofs are quite different.

Case 1: $\hat{a}^*$ is a corner solution and $b_a(0, \hat{a}^*) \neq 0$. Suppose that $\hat{a}^*$ is the upper boundary $\overline{a}$ (the proof for $\underline{a}$ is analogous). For k sufficiently large

$$B^k_{\hat{a}}(z^k, \hat{a}^k\,|a^*, \hat{a}^*) = k^{3/4}(\hat{a}^k - \hat{a}^*) + k\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\}b_a(z^k, \hat{a}^k) < 0$$

since $\hat{a}^k \leq \hat{a}^* = \bar{a}$ and $b_a(z^k, \hat{a}^*) > 0$ since b_a is an increasing function in $\hat{a}$ when a approaches $\hat{a}^*$. This implies that all interior points cannot be optimal and thus $\zeta^k(z^k)$, a singleton.

Case 2: $\hat{a}^*$ is an interior point solution or $b_a(0, \hat{a}^*) = 0$. Suppose by way of contradiction that there exist at least two distinct solutions $\hat{a}^k_1$ and $\hat{a}^k_2$ in $\zeta^k(z^k)$. Consider the first-order condition satisfied by $\hat{a}^k_i$ (for $i = 1, 2$):

$$B^k_{\hat{a}}(z^k, \hat{a}^k_i\,|\hat{a}^*) = k^{\frac{3}{4}}(\hat{a}^k_i - \hat{a}^*) + k\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k_i)\}b_z(z^k, \hat{a}^k_i) = 0. \tag{B.12}$$

(since we can take $\hat{a}^k_i$ sufficiently close to $\hat{a}^*$ we may assume they are interior point solutions). Dividing the above equality by $k^{5/8}$ we get for $\hat{a}^k$:

$$k^{1/8}(\hat{a}^k - \hat{a}^*) + k^{-1/8}\sqrt{k}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\}b_a(z^k, \hat{a}^k) = 0. \tag{B.13}$$

Denote the second term in (B.13) by

$$e^k(\hat{a}) := k^{-1/8}\tfrac{\sqrt{k}}{2}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\}b_a(z^k, \hat{a}^k).$$

Our contradiction is to show

$$e^k(\hat{a}^k_1) - e^k(\hat{a}^k_2) = O(\hat{a}^k_1 - \hat{a}^k_2). \tag{B.14}$$

This is indeed a contradiction since (B.13) implies that

$$e^k(\hat{a}^k_1) - e^k(\hat{a}^k_2) = k^{1/8}(\hat{a}^k_1 - \hat{a}^k_2),$$

which contradicts (B.14) since $k^{\frac{1}{8}} \to \infty$. To show (B.14), we consider two subcases.

Subcase 2.1: $\hat{a}^k_1, \hat{a}^k_2 \neq \hat{a}^*$. The significance of $\hat{a}^k_i \neq \hat{a}^*$ is the following. If $\hat{a}^k_i \neq \hat{a}^*$ then the second term in (B.13) *cannot* be zero. This implies

$$\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k_i)\} = b(z^k, a^*) - b(z^k, \hat{a}^k_i) \tag{B.15}$$

holds for $i = 1, 2$. On our way to (B.14) we write:

$$\begin{aligned}
&e^k(\hat{a}^k_1) - e^k(\hat{a}^k_2)\\
=&k^{-1/8}\tfrac{\sqrt{k}}{2}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k_1)\}b_a(z^k, \hat{a}^k_1)\\
&-\tfrac{\sqrt{k}}{2}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k_2)\}b_a(z^k, \hat{a}^k_2)\\
=&k^{-1/8}\tfrac{\sqrt{k}}{2}[\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k_1)\} - \min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k_2)\}]b_a(z^k, \hat{a}^k_1)\\
&+k^{-1/8}\tfrac{\sqrt{k}}{2}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k_2)\}\{b_a(z^k, \hat{a}^k_1) - b_a(z^k, \hat{a}^k_2)\}
\end{aligned} \tag{B.16}$$

where (B.16) holds by adding and subtracting

$$k^{-1/8}\tfrac{\sqrt{k}}{2}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}_2^k)\}b_a(z^k, \hat{a}_1^k).$$

Observe that the second term in (B.16) is $\frac{\sqrt{k}}{2}\min\{0, b(z^k, a^*)-b(z^k, \hat{a}_2^k)\}\{b_a(z^k, \hat{a}_1^k)-b_a(z^k, \hat{a}_2^k)\} = o(b_a(z^k, \hat{a}_1^k) - b_a(z^k, \hat{a}_2^k)) = o(\hat{a}_1^k - \hat{a}_2^k)$ by the differentiability of $b_a(z^k, a)$ in a for any z^k and $\frac{\sqrt{k}}{2}\min\{0, b(z^k, a^*) - b(z^k, \hat{a}_2^k)\} \to 0$.

It remains to consider the growth of the first term in (B.16). Note that (for $i = 1, 2$)

$$\begin{aligned}
&\min\{0, b(z^k, a^*) - b(z^k, \hat{a}_i^k)\} \\
= \;&\min\{0, b(0, a^*) - b(0, \hat{a}_i^k) + z^k(b_z(0, a^*) - b_z(0, \hat{a}_i^k)) + h.o.t\}
\end{aligned}$$

by taking Taylor expansions around $z^k = 0$. Also, by (B.15) we may write the latter as

$$\min\{0, b(z^k, a^*) - b(z^k, \hat{a}_i^k)\} = b(0, a^*) - b(0, \hat{a}_i^k) + z^k(b_z(0, a^*) - b_z(0, \hat{a}_i^k)) + h.o.t. \tag{B.17}$$

Continuing from (B.16) we can now rewrite its first term using (B.17) as:

$$\begin{aligned}
&k^{-1/8}\tfrac{\sqrt{k}}{2}[(b(0, a^*) - b(0, \hat{a}_1^k) + z^k(b_z(0, a^*) - b_z(0, \hat{a}_1^k))) \\
&- (b(0, a^*) - b(0, \hat{a}_2^k) + z^k(b_z(0, a^*) - b_z(0, \hat{a}_2^k)))]b_a(z^k, \hat{a}_1^k)
\end{aligned}$$

taking k sufficiently large so that the $h.o.t$'s disappear. Collecting terms on z^k we may rewrite the above as (with canceling terms):

$$k^{-1/8}\tfrac{\sqrt{k}}{2}b_a(z^k, \hat{a}_1^k) \cdot \big(b(0, \hat{a}_2^k) - b(0, \hat{a}_1^k) + z^k(b_z(0, \hat{a}_2^k) - b_z(0, \hat{a}_1^k))\big). \tag{B.18}$$

We now attempt to bound the first term $b(0, \hat{a}_2^k) - b(0, \hat{a}_1^k)$ in the parenthesis above. We do so by taking the Taylor expansion of $b(0, \hat{a}_2^k)$ in $\hat{a}$ around $\hat{a}_1^k$ to rewrite that first term as:

$$b(0, \hat{a}_2^k) - b(0, \hat{a}_1^k) = b_a(0, \hat{a}_1^k)(\hat{a}_2^k - \hat{a}_1^k). \tag{B.19}$$

Moreover, since $b_a(0, \hat{a}_1^k) - b_a(0, \hat{a}^*) = O(\hat{a}_1^k - \hat{a}^*)$ by the second order differentiability of b with respect to a, we may write

$$b(0, \hat{a}_2^k) - b(0, \hat{a}_1^k) = O(\hat{a}_1^k - \hat{a}^*)\Theta(\hat{a}_2^k - \hat{a}_1^k). \tag{B.20}$$

We require the following intermediate claim.

Claim 6. Term (iii) in the penalty function converges to 0 in k; that is, $k^{3/4}(\hat{a}^k - \hat{a}^*)^2 \to 0$.

Proof. By the exactness of the penalty function and the Proof of Corollary 4.2, we have

$$\tfrac{1}{2}k^{3/4}(\hat{a}^k - \hat{a}^*)^2 - k(\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\})^2 \to 0 \tag{B.21}$$

which are terms (iii) and (iv) of the penalty function. To show $k^{3/4}(\hat{a}^k - \hat{a}^*)^2 \to 0$, it suffices to show $k(\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\})^2 \to 0$.

Note that

$$\begin{aligned}
&k(\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\})^2 \\
= \;&k(\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^*) + b(z^k, \hat{a}^*) - b(z^k, \hat{a}^k)\})^2 \\
\leq \;&2k(\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^*)\})^2 + 2k(\min\{0, b(z^k, \hat{a}^*) - b(z^k, \hat{a}^k)\})^2.
\end{aligned} \tag{B.22}$$

The first term $k(\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^*)\})^2 \to 0$ by (B.1), it remains to show $k(\min\{0, b(z^k, \hat{a}^*) - b(z^k, \hat{a}^k)\})^2 \to 0$.

By Taylor expansion around $z^k = 0$,

$$b(z^k, \hat{a}^*) - b(z^k, \hat{a}^k) = b(0, \hat{a}^*) - b(0, \hat{a}^k) + z^k(b_z(0, \hat{a}^*) - b_z(0, \hat{a}^k)) + O((z^k)^2)$$

By Taylor expansion around $\hat{a}^k = 0$, $b_z(0, \hat{a}^*) - b_z(0, \hat{a}^k) = b_{za}(0, \hat{a}^*)(\hat{a}^* - \hat{a}^k) + o(\hat{a}^* - \hat{a}^k)$, which, by Claim 1 below, implies

$$z^k(b_z(0, \hat{a}^*) - b_z(0, \hat{a}^k)) = O(z^k(\hat{a}^* - \hat{a}^k)) = o((z^k)^2).$$

Therefore, we have

$$\begin{aligned} & k\left(\min\{0, b(z^k, \hat{a}^*) - b(z^k, \hat{a}^k)\}\right)^2 \\ = \; & k\left(\min\{0, b(0, \hat{a}^*) - b(0, \hat{a}^k) + z^k(b_z(0, \hat{a}^*) - b_z(0, \hat{a}^k)) + O((z^k)^2)\}\right)^2 \\ \leq \; & k\left(\min\{0, z^k(b_z(0, \hat{a}^*) - b_z(0, \hat{a}^k)) + O((z^k)^2)\}\right)^2 \\ = \; & k\left(\min\{0, O((z^k)^2)\}\right)^2 \\ \to \; & 0. \end{aligned}$$

It follows the second term $k(\min\{0, b(z^k, \hat{a}^*) - b(z^k, \hat{a}^k)\})^2$ in (B.21) attains zero as $k \to \infty$. Therefore, $k^{3/4}(\hat{a}^k - \hat{a}^*)^2 \to 0$ follows. □

With the claim in hand, note that the first term

$$k^{-1/8}\frac{\sqrt{k}}{2}b_a(z^k, \hat{a}_1^k) \cdot (b(0, \hat{a}_2^k) - b(0, \hat{a}_1^k))$$

in (B.18) is $o(\hat{a}_1^k - \hat{a}_2^k)$ since $b_a(z^k, \hat{a}_1^k)$ is bounded.

Now, for the second term in (B.18) involving z^k. Observe that exactness of the penalty function tells us that $\sqrt{k}\min\left\{0, z^k\right\} \to 0$, given that $b_z(z^k, a^*) > 0$ by the assumptions of the moral hazard problem and $\sqrt{k}\min\{0, b(z^k, a^*) - \underline{U}\} = O(\sqrt{k}\left|\min\{0, z^k\}\right|)b_z(z^k, a^*) \to 0$ from (4.9). If $z^k < 0$ then this implies $\sqrt{k}z^k \to 0$ and so

$$k^{-1/8}\frac{\sqrt{k}}{2}b_a(z^k, \hat{a}_1^k)z^k \to 0.$$

Hence the second term in (B.18) involving z^k is $o(\hat{a}_1^k - \hat{a}_2^k)$, as required.

It remains to argue that $z^k < 0$ for sufficiently large k. Observe from (B.17) that since $b(0, a^*) - b(0, \hat{a}_i^k) \geq 0$ by the optimality of a^* and $b_z(0, a^*) - b_z(0, \hat{a}_i^k) > 0$ by (4.10), if $z^k > 0$ then this contradicts the definition of the minimum in (B.17).

Taken together we have shown that both terms in (B.18) are $o(\hat{a}_1^k - \hat{a}_2^k)$. This, in turn shows that first term in (B.16) is also then $o(\hat{a}_1^k - \hat{a}_2^k)$. We have already shown the second term is $o(\hat{a}_1^k - \hat{a}_2^k)$ and so we have shown (B.14). This concludes the proof of Subcase 2.1.

Subcase 2.2: One of $\hat{a}_i^k = \hat{a}^*$. For Subcase 2.2 a similar contradiction follows by showing $e^k(\hat{a}_1^k) - e^k(\hat{a}^*) = O(\hat{a}_1^k - \hat{a}^*)$. This concludes Case 2 and the proof.

B.5 Proof of Proposition 4.6

Fix a k sufficiently large so that $\zeta^k(\bar{z})$ is a singleton for every $\bar{z} \in \mathcal{N}_{1/k}(z^k)$ (such a k is guaranteed by Lemma 4.4). Then ζ^k is a real-valued function (no longer set-valued) on the set $\mathcal{N}_{1/k}(z^k)$. Moreover, by the Theorem of Maximum it is continuous on that set.

Since $\mathcal{N}_{1/k}(z^k)$ is a full-dimensional open ball, $(\bar{z}+\epsilon)$ remains in $\mathcal{N}_{1/k}(z^k)$ for $\epsilon > 0$ sufficiently small. Then for any such ϵ, $\zeta^k(\bar{z}+\epsilon)$ is a real number and we can write:

$$
\begin{aligned}
&\tfrac{B^k(\bar{z}+\epsilon,\zeta^k(\bar{z}+\epsilon)|\hat{a}^*)-B^k(\bar{z},\zeta^k(\bar{z}+\epsilon)|\hat{a}^*)}{\epsilon} \\
&\leq \tfrac{B^k(\bar{z}+\epsilon,\zeta^k(\bar{z}+\epsilon)|\hat{a}^*)-B^k(\bar{z},\zeta^k(\bar{z})|\hat{a}^*)}{\epsilon} \\
&\leq \tfrac{B^k(\bar{z}+\epsilon,\zeta^k(\bar{z})|\hat{a}^*)-B^k(\bar{z},\zeta^k(\bar{z})|\hat{a}^*)}{\epsilon}
\end{aligned}
$$

where both inequalities come from the definition of minimum. We can write the right derivative as

$$
\begin{aligned}
\lim_{\epsilon\to 0^+} \tfrac{\varphi^k(\bar{z}+\epsilon)-\varphi^k(\bar{z})}{\epsilon} &= \lim_{\epsilon\to 0^+} \tfrac{B^k(\bar{z}+\epsilon,\zeta^k(\bar{z}+\epsilon)|\hat{a}^*)-B^k(\bar{z},\zeta^k(\bar{z})|\hat{a}^*)}{\epsilon} \\
&= \lim_{\epsilon\to 0^+} \tfrac{B^k(\bar{z}+\epsilon,\zeta^k(\bar{z})|\hat{a}^*)-B^k(\bar{z},\zeta^k(\bar{z})|\hat{a}^*)}{\epsilon} \\
&= B_z^k(\bar{z},\zeta^k(\bar{z})|\hat{a}^*)
\end{aligned}
$$

where the second equality follows from the continuity of ζ^k.

A similar argument establishes that the left limit exists (taking $\epsilon \to 0^-$) and is also equal to $B_z^k(\bar{z},\zeta^k(\bar{z})|\hat{a}^*)$ so φ^k is differentiable in z for all $\bar{z} \in \mathcal{N}_{1/k}(z^k)$ with k sufficiently large.

B.6 Proof of Claim 1

There are two cases to establish, depending on whether $\hat{a}^*$ is an interior solution or not.

Case 1: $\hat{a}^*$ is an interior solution. In this case $\hat{a}^k$ is an interior solution when k is large, which satisfies the first-order condition

$$\frac{\partial}{\partial \hat{a}} B^k(z^k,\hat{a}^k|\hat{a}^*) = k(\min\{0, b(z^k,a^*) - b(z^k,\hat{a}^k)\})b_a(z^k,\hat{a}^k) + k^{3/4}(\hat{a}^k - \hat{a}^*) = 0.$$

Dividing both sides by $k^{3/4}$, we have

$$k^{1/4}\min\{0, b(z^k,a^*) - b(z^k,\hat{a}^k)\}b_a(z^k,\hat{a}^k) + (\hat{a}^k - \hat{a}^*) = 0. \tag{B.23}$$

Taking the Taylor's expansion with respect to z^k around 0 yields:

$$b_a(z^k,\hat{a}^k) = b_a(0,\hat{a}^k) + z^k b_{az}(0,\hat{a}^k) + o(z^k).$$

Taking again the Taylor's expansion with respect to $\hat{a}^k$ around $\hat{a}^*$ then yields:

$$
\begin{aligned}
b_a(z^k,\hat{a}^k) &= b_a(0,\hat{a}^*) + (\hat{a}^k - \hat{a}^*)b_{aa}(0,\hat{a}^*) + z^k b_{az}(0,\hat{a}^*) + O(\hat{a}^k - \hat{a}^*) + O(z^k) \\
&= O(\hat{a}^k - \hat{a}^*) + O(z^k),
\end{aligned} \tag{B.24}
$$

since the derivatives $b_{az}(0,a^*)$ and $b_{aa}(0,\hat{a}^*)$ are bounded (due to them arising as integrals involving pdf functions).

Now, putting (B.10) and (B.24) into (B.23) we see that $\hat{a}^k - \hat{a}^*$ is $o(z^k)$.

Case 2. $\hat{a}^*$ is a corner solution. In this case, if $\frac{\partial}{\partial \hat{a}} B^k(z^k, \hat{a}^k|\hat{a}^*) \neq 0$, ($\hat{a}^k = \underline{a}$ if $\frac{\partial B^k(z^k,\hat{a}^k|\hat{a}^*)}{\partial \hat{a}} > 0$ and $\hat{a}^k = \bar{a}$ if $\frac{\partial B^k(z^k,\hat{a}^k|\hat{a}^*)}{\partial \hat{a}} < 0$), then $\hat{a}^k = \hat{a}^*$, we have $\frac{\hat{a}^k - \hat{a}^*}{z^k} = 0$, which is even of smaller order than $o(1)$. If $\frac{\partial}{\partial \hat{a}} B^k(z^k, \hat{a}^k|\hat{a}^*) = 0$ still holds, and $b_a(0, \hat{a}^*) = U_a(w^{a^*}, \hat{a}^*) = 0$ in particular, we can apply the same analysis in Case 1. It remains to consider $\frac{\partial}{\partial \hat{a}} B^k(z^k, \hat{a}^k|\hat{a}^*) = 0$ but $b_a(0, \hat{a}^*) \neq 0$. In this situation, if $\hat{a}^*$ is the lower corner, then $b_a(0, \hat{a}^*) < 0$, by the continuity of $b_a(\cdot, \cdot)$ we have that $b_a(z^k, \hat{a}^k) \leq 0$ when k is large, then

$$\frac{\partial}{\partial \hat{a}} B^k(z^k, \hat{a}^k|\hat{a}^*) = k(\min\{0, b(z^k, a^*) - b(z^k, \hat{a}^k)\}) b_a(z^k, \hat{a}^k) + k^{3/4}(\hat{a}^k - \hat{a}^*) > 0,$$

a contradiction of the supposition that $\frac{\partial}{\partial \hat{a}} B^k(z^k, \hat{a}^k|\hat{a}^*) = 0$. Similarly, if $\hat{a}^*$ is the upper corner, then $b_a(0, \hat{a}^*) > 0$, and so $b_a(z^k, \hat{a}^k) \geq 0$ when k is large. Hence, $\frac{\partial}{\partial \hat{a}} B^k(z^k, \hat{a}^k|\hat{a}^*) < 0$, and we obtain another contradiction. In either case, $\hat{a}^k - \hat{a}^*$ converges to 0 faster than z^k converges to 0. This establishes Claim 1.

C Appendix: Technical details of Lemmas 4.10 and 4.14

C.1 Proof of Lemma 4.10

We first establish Claim 2 under the conditions in (4.47). The proof for (4.48) is analogous. This requires the following claim.

Lemma C.1. λ_h and δ_h are invariant under any linear transformation of h.

Proof. Recall the first-order condition (4.20)

$$\int (-T(x) + \lambda_h + \delta_h R(x)) h(x) f(x, a^*) dx = 0. \tag{C.1}$$

Let h_0 satisfy restrictions (4.9) and (4.10). Then αh_0 (for any $\alpha \in (0, 1]$) also satisfies these conditions. Hence,

$$\begin{aligned} 0 &= \int (-T(x) + \lambda_{\alpha h_0} + \delta_{\alpha h_0} R(x)) h_0(x) f(x, a^*) dx \\ &= \int (-T(x) + \frac{\alpha \theta_{\alpha h_0}}{\theta_{h_0}} (\lambda_{h_0} + \delta_{h_0} R(x)) h_0(x) f(x, a^*) dx \\ &= \int (-T(x) + \theta_{h_0} (\lambda_{h_0} + \delta_{h_0} R(x)) h_0(x) f(x, a^*) dx, \end{aligned}$$

which implies $\alpha \theta_{\alpha h_0} = \theta_{h_0}$, $\lambda_{\alpha h_0} = \lambda_{h_0}$ and $\delta_{\alpha h_0} = \delta_{h_0}$. That is, the linear transformation of h_0 does not change the value of λ_{h_0} and δ_{h_0}. □

We now return to the proof of Claim 2.

Proof of Claim 2. We first prove (i). Since (4.47) holds we know that $T(x)$ crosses C_{h_0} within the subset $(L_1^- \cup L_2^-)$. As we will show, since $\Pr(L_i^-) > 0$ for $i = 1, 2$ we have the flexibility to construct a new variation $h_1(x)$ for $x \in (L_1^- \cup L_2^-)$ to satisfy our properties.

Let $g_i(x) \in \mathcal{H}$ and $\alpha_i > 0$ $(i = 1, 2)$. Define

$$h_1(x) = \begin{cases} \alpha_1 g_1(x) & \text{if } x \in L_1^- \\ \alpha_2 g_2(x) & \text{if } x \in L_2^- \\ 0 & \text{otherwise} \end{cases}. \tag{C.2}$$

We give conditions on α_1, α_2, g_1 and g_2 so that (4.40) (4.42) hold. By linear algebra, provided

$$\int_{L_1^-} g_1(x) f(x, \hat{a}^*) dx \int_{L_2^-} g_2(x) f(x, a^*) dx - \int_{L_1^-} g_1(x) f(x, a^*) dx \int_{L_2^-} g_2(x) f(x, \hat{a}^*) dx \neq 0, \tag{C.3}$$

(a determinant condition), then the linear system (4.40)–(4.41) has a solution

$$\begin{aligned} \alpha_1 &= \tfrac{t_0 - t_2}{t_1 - t_2} \tfrac{\int h_0 f(x, a^*) dx}{\int_{L_1^-} g_1(x) f(x, a^*) dx}, \\ \alpha_2 &= \tfrac{t_1 - t_0}{t_1 - t_2} \tfrac{\int h_0 f(x, a^*) dx}{\int_{L_2^-} g_2(x) f(x, a^*) dx}, \end{aligned} \tag{C.4}$$

where

$$t_0 = \tfrac{\int h_0(x) f(x, \hat{a}^*) dx}{\int h_0(x) f(x, a^*) dx}, \quad t_1 = \tfrac{\int_{L_1^-} g_1(x) f(x, \hat{a}^*) dx}{\int_{L_1^-} g_1(x) f(x, a^*) dx}, \quad t_2 = \tfrac{\int_{L_2^-} g_2(x) f(x, \hat{a}^*) dx}{\int_{L_2^-} g_2(x) f(x, a^*) dx},$$

provided the denominators defining t_1, t_1 and t_2 are nonzero; that is,

$$\int h_0(x) f(x, a^*) dx \neq 0, \tag{C.5}$$

$$\int_{L_1^-} g_1(x) f(x, a^*) dx \neq 0, \tag{C.6}$$

$$\int_{L_2^-} g_2(x) f(x, a^*) dx \neq 0, \tag{C.7}$$

and the denominator

$$t_1 - t_2 = \frac{\int_{L_1^-} g_1(x) f(x, \hat{a}^*) dx}{\int_{L_1^-} g_1(x) f(x, a^*) dx} - \frac{\int_{L_2^-} g_2(x) f(x, \hat{a}^*) dx}{\int_{L_2^-} g_2(x) f(x, a^*) dx} \neq 0 \tag{C.8}$$

in the definition of the α_i.

Observe that (C.5) follows by (4.9). We have a lot of choice on how to define g_1 and g_2, here we choose a specific functional form and verify that (C.3), (C.6)–(C.8) hold for this choice.

Let $g_1(x) = [\beta_1 + \gamma_1(x)]$ and $g_2(x) = \gamma_2(x)\beta_2(x)$ where γ_2 is a positive (almost everywhere) function and β_2 is an indicator function of a positive subset of L_2^-, and β_1 is a scalar. By the definition of β_2, (C.7) immediately holds. It remains to establish (C.6). We establish these below, and continue instead to work from (C.4).

Note that the α_i defined in (C.4) are chosen to satisfy (4.40)–(4.41). We now show how to choose β_1 to guarantee that (4.42) also holds. It suffices to solve the following equality for β_1:

$$\begin{aligned}
&(t_0 - t_2)\int_{L_1^-} T(x)[\beta_1 + \gamma_1(x)]\beta_0(x))f(x,a^*)dx \\
&+\Big(\int_{L_1^-}[\beta_1+\gamma_1(x)]f(x,\hat{a}^*)dx \\
&\quad -t_0\int_{L_1^-}[\beta_1+\gamma_1(x)]\beta_0(x)f(x,a^*)dx\Big)\frac{\int_{L_2^-}T(x)g_2(x)f(x,a^*)dx}{\int_{L_2^-}g_2(x)f(x,a^*)dx} \\
= \;& T(x^1)\left(\int_{L_1^-}[\beta_1+\gamma_1(x)]\beta_0(x)f(x,\hat{a}^*)dx - t_2\int_{L_1^-}[\beta_1+\gamma_1(x)]f(x,a^*)dx\right).
\end{aligned}$$

Then it is straightforward to solve for

$$\beta_1 = \frac{T(x^1)\frac{\int_{L_1^-}\gamma_1(x)f(x,\hat{a}^*)dx - t_2\int_{L_1^-}\gamma_1(x)f(x,a^*)dx}{\int_{L_1^-}f(x,a^*)dx} - \frac{(t_0-t_2)\int_{L_1^-}T(x)\gamma_1(x))f(x,a^*)dx}{\int_{L_1^-}f(x,a^*)dx} - \frac{\int_{L_1^-}\gamma_1(x)f(x,\hat{a}^*)dx - t_0\int_{L_1^-}\gamma_1(x)f(x,a^*)dx}{\int_{L_1^-}f(x,a^*)dx}\frac{\int_{L_2^-}T(x)g_2(x)f(x,a^*)dx}{\int_{L_2^-}g_2(x)f(x,a^*)dx}}{(t_0-t_2)\frac{\int_{L_1^-}T(x)f(x,a^*)dx}{\int_{L_1^-}f(x,a^*)dx} + \left(\frac{\int_{L_1^-}f(x,\hat{a}^*)dx}{\int_{L_1^-}\beta_0(x)f(x,a^*)dx} - t_0\right)\frac{\int_{L_2^-}T(x)g_2(x)f(x,a^*)dx}{\int_{L_2^-}g_2(x)f(x,a^*)dx} - T(x^1)\left(\frac{\int_{L_1^-}f(x,\hat{a}^*)dx}{\int_{L_1^-}\beta_0(x)f(x,a^*)dx} - t_2\right)}. \tag{C.9}$$

with the additional solvability condition that we have not divided by zero; that is,

$$\begin{aligned}
D \equiv (t_0 - t_2)\frac{\int_{L_1^-}T(x))f(x,a^*)dx}{\int_{L_1^-}f(x,a^*)dx} + \left(\frac{\int_{L_1^-}f(x,\hat{a}^*)dx}{\int_{L_1^-}\beta_0(x)f(x,a^*)dx} - t_0\right)\frac{\int_{L_2^-}T(x)g_2(x)f(x,a^*)dx}{\int_{L_2^-}g_2(x)f(x,a^*)dx} \\
- T(x^1)\left(\frac{\int_{L_1^-}f(x,\hat{a}^*)dx}{\int_{L_1^-}f(x,a^*)dx} - t_2\right) \neq 0.
\end{aligned} \tag{C.10}$$

This is also established below. For now, we assume β_1 can be defined this way and thus, we have a family of g_1 and g_2 such that h_1 satisfies (4.40)–(4.42). Recall that this immediately implies that (4.43) holds, however we now argue that β_2 can be chosen as indicators of sufficiently small subsets so that (4.44) also holds, our contradiction.

To see this, observe that

$$
\begin{aligned}
&\int [T(x) - R_{h_0}(x)] h_1 f(x,a^*) dx \\
= \;& \tfrac{t_0 - t_2}{t_1 - t_2} \tfrac{\int h_0 f(x,a^*)dx}{\int_{L_1^-} g_1(x) f(x,a^*)dx} \int_{L_1^-} [T(x) - R_{h_0}(x)] g_1(x) f(x,a^*) dx \\
& + \tfrac{t_1 - t_0}{t_1 - t_2} \tfrac{\int h_0 f(x,a^*)dx}{\int_{L_2^-} g_2(x) f(x,a^*)dx} \int_{L_2^-} [T(x) - R_{h_0}(x)] g_2(x) f(x,a^*) dx \\
= \;& \frac{\int h_0 f(x,a^*)dx}{\frac{\int_{L_2^-} R(x) g_2(x) f(x,a^*)dx}{\int_{L_2^-} g_2(x) f(x,a^*)dx} - \frac{\int_{L_1^-} R(x) g_1(x) f(x,a^*)dx}{\int_{L_1^-} g_1(x) f(x,a^*)dx}} \Bigg[\Bigg(\frac{\int_{L_2^-} R(x) g_2(x) f(x,a^*)dx}{\int_{L_2^-} g_2(x) f(x,a^*)dx} - R(x^2) \Bigg) \\
& \times \frac{\int_{L_1^-} [T(x) - R_{h_0}(x)] g_1(x) f(x,a^*)dx}{\int_{L_1^-} g_1(x) f(x,a^*)dx} + \Bigg(R(x^2) - \frac{\int_{L_1^-} R(x) g_1(x) f(x,a^*)dx}{\int_{L_1^-} g_1(x) f(x,a^*)dx} \Bigg) \\
& \times \frac{\int_{L_2^-} [T(x) - R_{h_0}(x)] g_2(x) f(x,a^*)dx}{\int_{L_2^-} g_2(x) f(x,a^*)dx} \Bigg] \\
\to \;& \frac{\int h_0 f(x,a^*)dx}{R(x^2) - \frac{\int_{L_1^-} R(x) g_1(x) f(x,a^*)dx}{\int_{L_1^-} g_1(x) f(x,a^*)dx}} \Bigg[\Bigg(R(x^2) - \frac{\int_{L_1^-} R(x) g_1(x) f(x,a^*)dx}{\int_{L_1^-} g_1(x) f(x,a^*)dx} \Bigg) (T(x^2) - R_{h_0}(x^2)) \Bigg] \\
= \;& (T(x^2) - T(x^1)) \int h_0 f(x,a^*) dx \\
> \;& 0,
\end{aligned}
$$

where the convergence is by letting β_2 indicate a subset of $[x^2, x^2 + \epsilon_2]$ where $\epsilon_2 \to 0$. The above uses the fact that when (4.47) holds we have $x^0 < x^1 < x^2$.

To establish Claim 2(i) it only remains to check that (C.6), (C.8), and (C.10) hold. To establish (C.10) observe that:

$$
\begin{aligned}
D = (t_0 - t_2) & \frac{\int_{L_1^-} T(x)) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} + \Bigg(\frac{\int_{L_1^-} f(x,\hat{a}^*)dx}{\int_{L_1^-} \beta_0(x) f(x,a^*)dx} - t_0 \Bigg) \frac{\int_{L_2^-} T(x) g_2(x) f(x,a^*)dx}{\int_{L_2^-} g_2(x) f(x,a^*)dx} \\
& - T(x^1) \Bigg(\frac{\int_{L_1^-} f(x,\hat{a}^*)dx}{\int_{L_1^-} f(x,a^*)dx} - t_2 \Bigg) \\
& = \Bigg(\frac{\int_{L_2^-} R(x) g_2(x) f(x,a^*)dx}{\int_{L_2^-} g_2(x) f(x,a^*)dx} - R(x^2) \Bigg) \frac{\int_{L_1^-} T(x)) f(x,a^*)dx}{\int_{L_1^-} \beta_0(x) f(x,a^*)dx} \\
& + \Bigg(R(x^2) - \frac{\int_{L_1^-} R(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \Bigg) \frac{\int_{L_2^-} T(x) g_2(x) f(x,a^*)dx}{\int_{L_2^-} g_2(x) f(x,a^*)dx} \\
& - T(x^1) \Bigg(\frac{\int_{L_2^-} R(x) g_2(x) f(x,a^*)dx}{\int_{L_2^-} g_2(x) f(x,a^*)dx} - \frac{\int_{L_1^-} \beta_0(x) R(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \Bigg) \\
& \to [R(x^2) - \frac{\int_{L_1^-} \beta_0(x) R(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}][T(x^2) - T(x^1)] \\
& > 0.
\end{aligned}
$$

again with convergence as defined above.

Next to establish (C.6), we will show that $\int_{L_1^-} g_1(x)f(x,a^*)dx \neq 0$, even when $g_1(x)$ could be negative for some $x \in L_1^-$. By the definition of g_1 is suffices to show

$$\beta_1 + \frac{\int_{L_1^-} \gamma_1(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \neq 0.$$

Recall the elaborate expression for β_1 in (C.9) and write simply $\beta_1 = \frac{N}{D}$ where N is the numerator of (C.9) and D is the denominator of (C.9). Multiplying the above displayed equation through by D, it suffices to show

$$N + D\frac{\int_{L_1^-} \gamma_1(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \neq 0.$$

some careful manipulation (suppressed for brevity) yields:

$$\begin{aligned}
&N + D\frac{\int_{L_1^-} \gamma_1(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \\
&= \left(\frac{\int_{L_2^-} T(x)g_2(x)f(x,a^*)dx}{\int_{L_2^-} g_2(x)f(x,a^*)dx} - T(x^1)\right)\left[\frac{\int_{L_1^-} \gamma_1(x)R(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} - \frac{\int_{L_1^-} \beta_0(x)R(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}\frac{\int_{L_1^-} \gamma_1(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}\right] \\
&+ \left(\frac{\int_{L_2^-} R(x)g_2(x)f(x,a^*)dx}{\int_{L_2^-} g_2(x)f(x,a^*)dx} - R(x^2)\right)\left(\frac{\int_{L_1^-} T(x))f(x,a^*)dx}{\int_{L_1^-} \beta_0(x)f(x,a^*)dx}\frac{\int_{L_1^-} \gamma_1(x)\beta_0(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} - \frac{\int_{L_1^-} T(x)\gamma_1(x))f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}\right)
\end{aligned}$$

Since the support of g_2 shrinks to $[x^2, x^2 + \epsilon_2]$, the second term above

$$\left(\frac{\int_{L_2^-} R(x)g_2(x)f(x,a^*)dx}{\int_{L_2^-} g_2(x)f(x,a^*)dx} - R(x^2)\right)\left(\frac{\int_{L_1^-} T(x))f(x,a^*)dx}{\int_{L_1^-} \beta_0(x)f(x,a^*)dx}\frac{\int_{L_1^-} \gamma_1(x)\beta_0(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} - \frac{\int_{L_1^-} T(x)\gamma_1(x))f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}\right)$$

is of smaller asymptotic order. However, we choose $\gamma'(x) > 0$ to ensure

$$\frac{\int_{L_1^-} \gamma_1(x)R(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} > \frac{\int_{L_1^-} \beta_0(x)R(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}\frac{\int_{L_1^-} \gamma_1(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}$$

since both $\gamma_1(x)$ and $R(x)$ are increasing. Therefore, the first term in the above expression has

$$\begin{aligned}
&\left(\frac{\int_{L_2^-} T(x)g_2(x)f(x,a^*)dx}{\int_{L_2^-} g_2(x)f(x,a^*)dx} - T(x^1)\right)\Bigg[\frac{\int_{L_1^-} \gamma_1(x)R(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \\
&\qquad - \frac{\int_{L_1^-} R(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}\frac{\int_{L_1^-} \gamma_1(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}\Bigg] > 0.
\end{aligned}$$

Since the second term is of smaller order as $\epsilon_2 \to 0$ this implies $N + D\frac{\int_{L_1^-} \gamma_1(x)\beta_0(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \neq 0$.

Finally, we check (C.8) that $t_1 - t_2 \neq 0$. Recall that $g_1(x) = \frac{N}{D} + \gamma_1(x)$ then it suffices to show that

$$\begin{aligned}
E &\equiv \left(N\frac{\int_{L_1^-} R(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} + D\frac{\int_{L_1^-} R(x)\gamma_1(x)f(x,a^*)dx}{\int_{L_1^-} \beta_0(x)f(x,a^*)dx}\right) \\
&\quad - R(x^2)\left(N + D\frac{\int_{L_1^-} \gamma_1(x)\beta_0(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}\right) \\
&= N\left(\frac{\int_{L_1^-} R(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} - R(x^2)\right) \\
&\quad + D\left(\frac{\int_{L_1^-} R(x)\gamma_1(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} - R(x^2)\frac{\int_{L_1^-} \gamma_1(x)f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx}\right) \neq 0
\end{aligned}$$

By some algebra, we have

$$E = \left(\frac{\int_{L_2^-} R(x) g_2(x) f(x,a^*) dx}{\int_{L_2^-} g_2(x) f(x,a^*) dx} - R(x^2) \right) \mathbb{F}[\gamma_1]$$

where $\mathbb{F}[\gamma_1]$ the linear functional of γ_1 defined as

$$\begin{aligned}
\mathbb{F}[\gamma_1] \equiv & \Big(\frac{\int_{L_1^-} R(x)\gamma_1(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \frac{\int_{L_1^-} T(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \\
& - \frac{\int_{L_1^-} T(x)\gamma_1(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \frac{\int_{L_1^-} R(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \Big) \\
& - T(x^1) \left(\frac{\int_{L_1^-} R(x)\gamma_1(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} - \frac{\int_{L_1^-} R(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \frac{\int_{L_1^-} \gamma_1(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \right) \\
& - R(x^2) \left(\frac{\int_{L_1^-} T(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \frac{\int_{L_1^-} \gamma_1(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} - \frac{\int_{L_1^-} \gamma_1(x) T(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \right) \\
= & \int_{L_1^-} \Big[C_1 R(x) + C_2 T(x) + C_3 \Big] \gamma_1(x) f(x,a^*) dx,
\end{aligned}$$

where

$$C_1 = \left(\frac{\int_{L_1^-} T(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} - T(x^1) \right) \frac{1}{\int_{L_1^-} f(x,a^*)dx},$$

$$C_2 = \left(R(x^2) - \frac{\int_{L_1^-} R(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \right) \frac{1}{\int_{L_1^-} f(x,a^*)dx} \neq 0 \quad \text{(since } R(x) \text{ is strictly monotone)},$$

and

$$C_3 = \left(T(x^1) \frac{\int_{L_1^-} R(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} - R(x^2) \frac{\int_{L_1^-} T(x) f(x,a^*)dx}{\int_{L_1^-} f(x,a^*)dx} \right) \frac{1}{\int_{L_1^-} f(x,a^*)dx}.$$

Since $\frac{\int_{L_2^-} R(x) g_2(x) f(x,a^*)dx}{\int_{L_2^-} g_2(x) f(x,a^*)dx} - R(x^2) > 0$ by the increasing of $R(x)$, it suffices to show $\mathbb{F}[\gamma] \neq 0$.

Now if $T(x)$ is linear function of $R(x)$ for $x \in \overline{\mathcal{X}}_{\underline{w}}$ we are done, which is exactly what we want to show. It remains to consider the case where $T(x)$ and $R(x)$ are not linearly dependent. In this case, if $T(x)$ and $R(x)$ are linearly independent in a domain $x \in \mathcal{X}_L$, where we can find the infimum or supremum of $T(x)$ since $T(\cdot)$ is monotone. Let x^1 be an extremum of $T(x)$ for $x \in \mathcal{X}_L$ that is not boundary of $\overline{\mathcal{X}}_{\underline{w}}$. Then we can choose h_0 such that

$$C_{h_0} = T(x^1),$$

which is doable by adjusting the weight function $\frac{h_0(x)}{\int h_0(x) f(x,a^*)dx}$. Therefore, $T(x)$ and $R(x)$ cannot be linearly dependent in L_1^-, and thus there must exist some strictly increasing function $\gamma_1(x)$ such that $\mathbb{F}[\gamma_1] \neq 0$ (based on the following lemma). This completes the proof of Claim 2(i).

Lemma C.2. There exists a strictly increasing function $\gamma_1(x)$ such that $\mathbb{F}[\gamma_1] \neq 0$, otherwise

$$C_1 R(x) + C_2 T(x) + C_3 = 0, \quad \forall\, x \in L_1^-.$$

Proof. Suppose that $\mathbb{F}[\gamma_1] = 0$ for all strictly increasing function $\gamma_1(x)$. In particular, we can take $\gamma_1'(x) \geq 1$ in L_1^-. Let $\phi(x)$ be any C^1 function on L_1^-, then it is easy to check $\gamma_1(x) + \epsilon\phi(x)$ is a strictly increasing function on L_1^-. Hence $\mathbb{F}[\gamma_1(x) + \epsilon\phi(x)] = 0$. By the linearity of $\mathbb{F}[\cdot]$, we have $\mathbb{F}[\phi] = 0$. That is,

$$\mathbb{F}[\phi] = \int_{L_1^-} \Big[C_1 R(x) + C_2 T(x) + C_3\Big]\phi(x) f(x, a^*) dx = 0,\ \forall \phi(x) \in C^1(L_1^-),$$

which implies that

$$C_1 R(x) + C_2 T(x) + C_3 = 0, \quad \forall\, x \in L_1^-.$$

□

To prove Claim 2(ii) we can verify that, in fact, the $\alpha_i > 0$ by showing that $t_2 < t_0 < t_1$. This follows from the definition of L_i. Indeed, for $x \in L_1$ we have $R_{h_0} < C_{h_0}$. Writing out the definition of R_{h_0} and C_{h_0} implies that $x \in L_1$ when $\frac{f(x,\hat{a}^*)}{f(x,a^*)} > \frac{\int h_0(x) f(x,\hat{a}^*)dx}{\int h_0(x) f(x,\hat{a})dx}$. Then integrating by g_1 yields $\frac{\int_{L_1^-} g_1(x) f(x,\hat{a}^*)dx}{\int_{L_1^-} g_1(x) f(x,\hat{a})dx} > \frac{\int h_0(x) f(x,\hat{a}^*)dx}{\int h_0(x) f(x,\hat{a})dx}$ since g_1 is a nonnegative function. Similarly, $\frac{\int_{L_2^-} g_2(x) f(x,\hat{a}^*)dx}{\int_{L_2^-} g_2(x) f(x,\hat{a})dx} > \frac{\int h_0(x) f(x,\hat{a}^*)dx}{\int h_0(x) f(x,\hat{a})dx}$ since g_2 is a nonnegative function. Thus $t_2 < t_0 < t_1$ and (4.40) and (4.41) hold.

To have $h_1 \in \mathcal{H}$ it remains to argue that $h_1(x) \leq \overline{h}$ almost everywhere. Lemma C.1 says that if we construct a variation h_1, as long as it is positive we may assume it is essentially bounded by $\overline{h}$, as required in $\mathcal{H}$. Indeed, if $h_1(x)$ is not essentially bounded by $\bar{h}$, i.e., $\Pr(h_1(X) > \bar{h}) > 0$ we choose $\tilde{h}_0 = \frac{\bar{h}}{\max_x h_1(x)} h_0$, where a maximum of $h_1(x)$ exists because $g_i(x)$ and α_i are bounded. Recall that by Lemma C.1 , a linear transformation does not change λ_{h_0}, δ_{h_0}, and $\frac{\int h_0(x) f(x,\hat{a}^*)dx}{\int h_0(x) f(x,a^*)dx}$, so the areas $\mathcal{X}^-$, $\mathcal{X}^+$, $\mathcal{X}^{h_0-}$ and $\mathcal{X}^{h_0+}$ are the same under $\tilde{h}_0$. Repeating the above reasoning based on $\tilde{h}_0$ and $\tilde{\alpha}_i = \frac{\bar{h}}{\max_x h_1(x)} \alpha_i$ and keeping the same $g_i(x)$, we have that $\tilde{h}_1(x) \leq \bar{h}$. So, without loss of generality, we may assume $h_1(x) \leq \bar{h}$. Similarly, (4.40) and (4.41) are preserved. This establishes Claim 2(ii). □

The rest of proof continues from the main body of the text, with Claim 2 in hand. We use the fact, already established in the proof in the main text that the sets in (4.45) all have positive measure. We discuss three possible cases, which enumerate the possible crossing patterns of T, R_{h_0} and C_{h_0}.

Case 1. $T(x)$ crosses $R_{h_0}(x)$ at some $x^c \subset \mathcal{X}^{h_0-}$. Observe that $T(x)$ crosses R_{h_0} at x^c means that the sign of $R_{h_0}(x) - T(x)$ is not constant in a neighborhood $\{x : \|x - x^c\| \leq \epsilon\}$ for some $\epsilon > 0$. In this case, $T(x)$ crosses $R_{h_0}(x)$ while $T(x)$ is below the constant C_{h_0}. By the continuity of $T(x)$ and $R_{h_0}(x)$, there is positive measure of x's such that $C_{h_0} > T(x) > R_{h_0}(x)$, which, by the definition of C_{h_0}, implies

$$\Pr(L_1^- \cap \mathcal{X}^{h_0-}) > 0. \tag{C.11}$$

Meanwhile, there also is positive measure of x's such that $T(x) < R_{h_0}(x) < C_{h_0}$, which implies

$$\Pr(L_1^+ \cap \mathcal{X}^{h_0-}) > 0. \tag{C.12}$$

We now discuss three subcases (they are mutually exclusive).

Subcase 1. The set of $x \in L_2$ such that $T(x) > R_{h_0}(x)$ has positive measure. This subcase implies the existence of a positive measure of x such that $T(x) > R_{h_0}(x) > C_{h_0}$, which, by definition of C_{h_0} and $\mathcal{X}^{h_0+}$, implies $\Pr(L_2^- \cap \mathcal{X}^{h_0+}) > 0$. Together with (C.11), we confirm that (4.48) is satisfied.

Subcase 2. For almost all $x \in L_2$, $T(x) < R_{h_0}(x)$. Both $T(x)$ and $R_{h_0}(x)$ are increasing, there must be some positive measure of x's such that $T(x) > C_{h_0}$ and $R_{h_0}(x) > C_{h_0}$. Since for all $x \in L_2$, $T(x) < R_{h_0}(x)$, we have that there is positive measure x satisfying $R_{h_0}(x) > T(x) > C_{h_0}$, which is equivalent to say $\Pr(L_2^+ \cap \mathcal{X}^{h_0+}) > 0$. Together with (C.12), this yields (4.47).

Subcase 3. For almost all $x \in L_2$, $T(x) = R_{h_0}(x)$. This is an unstable subcase that can be converted to Subcase 1 or 2. Consider the situation $\Pr(L_1^-) > \Pr(L_1^+)$. We construct

$$h_1(x) = \begin{cases} \alpha_1 g_1(x) \text{ if } x \in L_1 \\ \alpha_2 g_2(x) \text{ if } x \in L_2 \end{cases} \tag{C.13}$$

where $g_i(x) \in \mathcal{H}$. By Claim 2(ii), we can find $\alpha_i > 0$ satisfying (4.40) and (4.41), for any $g_i(x) \in \mathcal{H}$. Now since for $x \in L_1^+$, $T(x) < R_{h_1}(x)$ and $x \in L_1^-$, $T(x) > R_{h_1}(x)$, where $R_{h_1}(x) := \lambda_{h_1} + \delta_{h_1}(1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)})$, we can adjust $g_1(x)$ in L_1^- or L_1^+ to obtain $\int T(x)h_1(x)f(x,a^*)dx > \int T(x)h_0(x)f(x,a^*)dx$. Therefore, we have

$$\begin{aligned} 0 &= \int (-T(x) + R_{h_0}(x))h_1(x)f(x,a^*)dx \\ &< \int -T(x)h_0(x)f(x,a^*)dx + \lambda_{h_1} + \delta_{h_1}\int R(x)h_1(x)f(x,a^*)dx \\ &= \int -T(x)h_0(x)f(x,a^*)dx + \frac{\theta_{h_1}}{\theta_{h_0}}\left(\lambda_{h_0} + \delta_{h_0}\int R(x)h_0(x)f(x,a^*)dx\right) \end{aligned}$$

where the second equality is by equalities (4.40) and (4.41). From the first-order condition for h_0 again we can conclude $\theta_{h_1}/\theta_{h_0} > 1$. Then, the new $R_{h_1}(x) = \theta_{h_1}/\theta_{h_0}R_{h_0}(x)$ moves up. The curve $T(x)$ crosses $R_{h_1}(x)$ while $T(x)$ is below the new constant C_{h_1}. Recall that we are in the situation that for all $x \in L_2$, $T(x) = R_{h_0}(x) < R_{h_1}(x)$. We essentially return to Subcase 2, when replacing h_0 with h_1 and taking h_1 as the initial variation to begin with. If $\Pr(L_1^-) \le \Pr(L_1^+)$ adjust $g_1(x)$ in L_1^- or L_1^+ to obtain $\int T(x)h_1(x)f(x,a^*)dx < \int T(x)h_0(x)f(x,a^*)dx$, which results in $\theta_{h_1}/\theta_{h_0} < 1$. Similar reasoning to Subcase 1 now applies.

Case 2. $T(x)$ crosses $R_{h_0}(x)$ at some $x^c \subset \mathcal{X}^{h_0+}$. This case is analogous to Case 1, so we omit the details.

Case 3. $T(x)$ crosses $R_{h_0}(x)$ only at x^c where $T(x^c) = C_{h_0}$. In this case, there is no cross in the sets L_1 or L_2, nor in the sets $\mathcal{X}^{h_0-}$ or $\mathcal{X}^{h_0+}$. We want to show that this case is unstable by choosing some h_1, and it will eventually return to either Case 1 or Case 2. Recall that neither $T(x)$ nor $R(x)$ are constants by Proposition 4.9. Therefore we can move $R_h(x) := \lambda_h + \delta_h(1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)})$ by choosing some suitable h using Claim 2(ii) until there is the desired cross. Then we can return to one of the earlier two cases.

Finally we deal with one assumption we made at the outset of the proof.

<u>Unbounded $\frac{1}{u'(w^{a^*})}$.</u> Finally, we discuss the case $\frac{1}{u'(w^{a^*})} \to \infty$, which occurs only if $u'^{-1}(\cdot)$ is unbounded. By the monotonicity of $u'^{-1}(\cdot)$ and the Chebyshev inequality, we have that $\Pr(\frac{1}{u'(w^{a^*}(X))} > n) = \Pr(w^{a^*}(X) > u'^{-1}(\frac{1}{n})) \le \frac{\int w^{a^*}(x)f(x,a^*)dx}{u'^{-1}(\frac{1}{n})}$, which implies that as $n \to \infty$, $\Pr(\frac{1}{u'(w^{a^*}(X))} > n) \to 0$ since $\int w^{a^*}(x)f(x,a^*)dx$ is bounded. Therefore, we choose a sequence of $h_0^n(x) = h_0(x)\mathbf{1}[w^{a^*}(x) \le$

$(u')^{-1}(1/n)]$ where $0 \leq h_0(x) \leq \frac{\epsilon}{u'^{-1}(\frac{1}{n})}$. For every n and $h_0^n(x)$, repeat the same reasoning as in Cases 1-3. This yields $T(x) = R_{h_0^n}(x)$ for almost every $x \in \{x : w^{a^*}(x) \leq u'^{-1}(\frac{1}{n})\}$. As $n \to \infty$, $\Pr(\frac{1}{u'(w^{a^*}(X))} > n) \to 0$, then for $h_0^\infty(x)$ such that $h_0^\infty(x) = 0$ for $x \in \{x : u'(w^{a^*}(x)) = 0\}$, gives the same conclusion $T(x) = R_{h_0^\infty}(x)$, a.e. This suffices to establish the result.

C.2 Proof of Lemma 4.14

We break up the proof into two stages. The first is to show that if the MLRP holds then $T(x)$ and $R(x)$ are comonotone on a set of positive measure. The second stage is to establish how this comonotonicity can be extended to all of $\mathcal{X}$. In the main body of the paper we provide details for the first stage. Details of the second stage are in Appendix C.

C.2.1 Stage 1

By Assumption (A3.1) and the definition of $(P|\hat{a})$ we have

$$U(w^*_{\hat{a}^*}, a^*) - U(w^*_{\hat{a}^*}, \hat{a}^*) = 0 = U(w^{a^*}, a^*) - U(w^{a^*}, \hat{a}^*).$$

Therefore,

$$\int u(w^{a^*}(x))R(x)f(x, a^*)dx = \int u(w^*_{\hat{a}^*}(x))R(x)f(x, a^*)dx. \tag{C.14}$$

which implies

$$0 = \int [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))]R(x)f(x, a^*)dx.$$

We want to show a contradiction of the above equality if $T(x)$ and $R(x)$ are not comonotone on any subset of the domain.

We only show the case $a^* > \hat{a}^*$, the other case $a^* < \hat{a}^*$ follows analogous logic. From $\int u(w^{a^*}(x))f(x, a^*)dx = \int u(w^*_{\hat{a}^*}(x))f(x, a^*)dx$ and the fact u is a increasing and continuous function, $w^*_{\hat{a}^*}$ is continuous from Proposition 3.1, and w^{a^*} is continuous from Remark 4.11, we know $w^{a^*}(x)$ must cross $w^*_{\hat{a}^*}(x)$ at some point. Note that by *cross* we mean the sign of the difference of the functions is not constant on a small neighborhood of the point of intersection. This implies that the crossing point x must lie in the domain $\overline{\mathcal{X}^*_{\underline{w}}}$ and where $w^*_{\hat{a}^*}(x) \neq \underline{w}$. This, in turn, implies $T(x)$ crosses $\hat{T}(x)$ at some point $x \in \overline{\mathcal{X}^*_{\underline{w}}}$ where $\hat{T}(x) = C$ for some constant C, where $\hat{T}(x)$ is as defined in (4.49). We should have $C > \frac{v'(\pi(x) - \underline{w})}{u'(\underline{w})}$ because $T(x) \geq \frac{v'(\pi(x) - \underline{w})}{u'(\underline{w})}$. Given $\hat{T}(x) > C$ by the definition of a GMH contract in (3.5), we have

$$\hat{T}(x) = \lambda^*(\hat{a}^*) + \delta^*(\hat{a}^*)R(x),$$

which means that $T(x)$ will cross $\lambda^*(\hat{a}^*) + \delta^*(\hat{a}^*)R(x)$ at least once.

Suppose by contradiction, $T(x)$ and $R(x)$ are not comonotone almost everywhere on $\mathcal{X}^*_{\underline{w}}$, then $T(x)$ crosses $\lambda^*(\hat{a}^*) + \delta^*(\hat{a}^*)R(x)$ only once, given that $R(x)$ is nondecreasing by Proposition 4.9(iv). For convenience, let

$$\mathcal{X}^c \equiv \{x \in \overline{\mathcal{X}^*_{\underline{w}}} : T(x) \leq C\}.$$

We consider two cases.

Case 1. $\delta^*(\hat{a}^*) = 0$.

In this case $\hat{T}(x) = \lambda^*$, implying $w^*_{\hat{a}^*}(x)$ is increasing in x since π is increasing. Note that $\frac{v'(\pi(x)-y)}{u'(y)}$ is increasing in y, so $w^{a^*}(x) - w^*_{\hat{a}^*}(x) \geq 0$ or $u(w^{a^*}(x)) \geq u(w^*_{\hat{a}^*}(x))$, if and only if $T(x) \geq \hat{T}(x)$. It follows that there is some x^0 such that $u(w^{a^*}(x)) \geq u(w^*_{\hat{a}^*}(x))$ if $x > x^0$ and vice versa. Therefore, we have

$$
\begin{aligned}
0 &= \int [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))]R(x)f(x,a^*)dx \\
&= \int_{\mathcal{X}^c} [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))](1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)})f(x,a^*)dx \\
&\quad + \int_{\overline{\mathcal{X}^c}} [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))](1 - \tfrac{f(x,\hat{a}^*)}{f(x,a^*)})f(x,a^*)dx \\
&< (1 - \tfrac{f(x^0,\hat{a}^*)}{f(x^0,a^*)}) \int_{\mathcal{X}^c} [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))]f(x,a^*)dx \\
&\quad + (1 - \tfrac{f(x^0,\hat{a}^*)}{f(x^0,a^*)}) \int_{\overline{\mathcal{X}^c}} [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))]f(x,a^*)dx \\
&= (1 - \tfrac{f(x^0,\hat{a}^*)}{f(x^0,a^*)}) \int [u(w^{a^*}(x|\hat{a}^*)) - u(w^*_{\hat{a}^*}(x))]f(x,a^*)dx = 0
\end{aligned}
$$

which is a contradiction.

Case 2. $\delta^*(\hat{a}^*) \neq 0$.

We further break this case into two subcases.

Subcase 2.1. The (IR) constraint in $(P|\hat{a})$ is binding; that is, $U(w^*_{\hat{a}^*}, a^*) = \underline{U}$. Note that this implies

$$
\int u(w^{a^*}(x))f(x,a^*)dx = \int u(w^*_{\hat{a}^*}(x))f(x,a^*)dx. \tag{C.15}
$$

since $U(w^*_{\hat{a}^*}, a^*) = \underline{U} = U(w^{a^*}, a^*)$ by Assumption ((A3.2)).

Suppose that (i) $\delta^* > 0$, $\hat{T}(x)$ is nondecreasing, we have

$$
\delta^* R(x) \begin{cases} \geq C - \lambda^* \text{ for all } x \in \mathcal{X}^c \\ < C - \lambda^* \text{ for all } x \notin \mathcal{X}^c \end{cases} .
$$

where $C = \hat{T}(x)$ is the point where $T(x)$ crosses $\hat{T}(x)$. (ii) When $\delta^* < 0$, $\hat{T}(x)$ is nonincreasing, we also have

$$
\delta^* R(x) \begin{cases} \geq C - \lambda^* \text{ for all } x \in \mathcal{X}^c \\ < C - \lambda^* \text{ for all } x \notin \mathcal{X}^c \end{cases} .
$$

Therefore, it follows

$$
\begin{aligned}
0 &= \delta^* \int [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))] R(x) f(x,a^*) dx \\
&= \int_{\mathcal{X}^c} [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))] \delta^* R(x) f(x,a^*) dx \\
&\qquad + \int_{\overline{\mathcal{X}^c}} [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))] \delta^* R(x) f(x,a^*) dx \\
&< (C-\lambda^*) \int_{\mathcal{X}^c} [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))] f(x,a^*) dx \\
&\qquad + (C-\lambda^*) \int_{\overline{\mathcal{X}^c}} [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))] f(x,a^*) dx \\
&= (C-\lambda^*) \int [u(w^{a^*}(x)) - u(w^*_{\hat{a}^*}(x))] f(x,a^*) dx = 0.
\end{aligned}
$$

which is a contradiction.

Subcase 2.2. The (IR) constraint in $(P|\hat{a})$ is not binding. This implies $\lambda^*(\hat{a}^*) = 0$.

Suppose by contradiction that $T(x)$ and $\hat{T}(x)$ single cross is at some x^0. We know $x^0 \in \overline{\mathcal{X}^*_{\underline{w}}}$. Note that when $T(x)$ crosses $\hat{T}(x)$, w^{a^*} also crosses $w^*_{\hat{a}^*}$ at point x^0. Note that $w^*_{\hat{a}^*}$ is nondecreasing by that MLRP when $\delta^* > 0$ and nonincreasing when $\delta^* < 0$. (i) We consider the case $\delta^* > 0$ first. If w^{a^*} crosses $w^*_{\hat{a}^*}$ from below, then $T(x)$ also crosses $\hat{T}(x)$ from below, since $\hat{T}(x)$ is increasing, then $T(x)$ must be increasing with positive measure around a neighborhood around x^0. We are done. If w^{a^*} crosses $w^*_{\hat{a}^*}$ from above, then $T(x)$ crosses $\hat{T}(x)$ from above, which implies that when $w^*_{\hat{a}^*} = \underline{w}$, $w^{a^*} > \underline{w}$. Then, we have

$$
\begin{aligned}
0 &= \int [u(w^{a^*}) - u(w^*_{\hat{a}^*})] R(x) f(x,a^*) dx \\
&= \int_{R(x)\geq 0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})] R(x) f(x,a^*) dx + \int_{R(x)<0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})] R(x) f(x,a^*) dx \\
&< R(x^0) \int_{R(x)\geq 0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})] f(x,a^*) dx \\
&< 0,
\end{aligned}
$$

where the first inequality follows since $R(x)$ is increasing and $R(x) < 0$ for $x \in \mathcal{X}^*_{\underline{w}}$. The last inequality is implied by the slackness of the (IR) constraint in (P$|\hat{a}^*$):

$$
\int_{R(x)\geq 0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})] f(x,a^*) dx < -\int_{R(x)<0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})] f(x,a^*) dx < 0.
$$

Therefore we have a contradiction. (ii) Now we consider the case $\delta^* < 0$, then $\hat{T}(x)$ is decreasing. If $T(x)$ crosses $\hat{T}(x)$ from above, then they must comonotone with positive measure, we are done.

Suppose that $T(x)$ crosses $\hat{T}(x)$ from below, which means w^{a^*} crosses $w^*_{\hat{a}^*}$ from below. We have

$$\begin{aligned}
0 &= \delta^* \int [u(w^{a^*}) - u(w^*_{\hat{a}^*})]R(x)f(x,a^*)dx \\
&= \int_{R(x)\geq 0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})]\delta^* R(x)f(x,a^*)dx + \int_{R(x)<0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})]\delta^* R(x)f(x,a^*)dx \\
&= \int_{R(x)\geq 0} [u(w^{a^*}) - u(\underline{w})]\delta^* R(x)f(x,a^*)dx + \int_{R(x)<0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})]\delta^* R(x)f(x,a^*)dx \\
&\leq \int_{R(x)<0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})]\delta^* R(x)f(x,a^*)dx \\
&< \hat{T}(x^0) \int_{R(x)<0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})]f(x,a^*)dx \\
&< 0,
\end{aligned}$$

where the second to last inequality follows from the definition of $\hat{T}(x)$ and the fact $\hat{T}(x)$ is decreasing. The last inequality is implied by the slackness of the (IR) constraint in (P$|\hat{a}^*$).

$$\int_{R(x)<0} [u(w^{a^*}) - u(w^*_{\hat{a}^*})]f(x,a^*)dx < -\int_{R(x)\geq 0} [u(w^{a^*}) - u(\underline{w})]f(x,a^*)dx < 0.$$

Again, we obtain a contradiction.

Putting all the cases together, we conclude that $T(x)$ must cross $\lambda^*(\hat{a}^*) + \delta^*(\hat{a}^*)R(x)$ at least twice. So both subsets where $T(x)$ is increasing or decreasing is of positive measure (by Proposition 4.9(iii) we know $T(x)$ is not a constant). Since $R(x)$ is nondecreasing, $T(x)$ must be comonotone with $R(x)$ at least for a positive measure subset of $\overline{\mathcal{X}^*_{\underline{w}}}$.

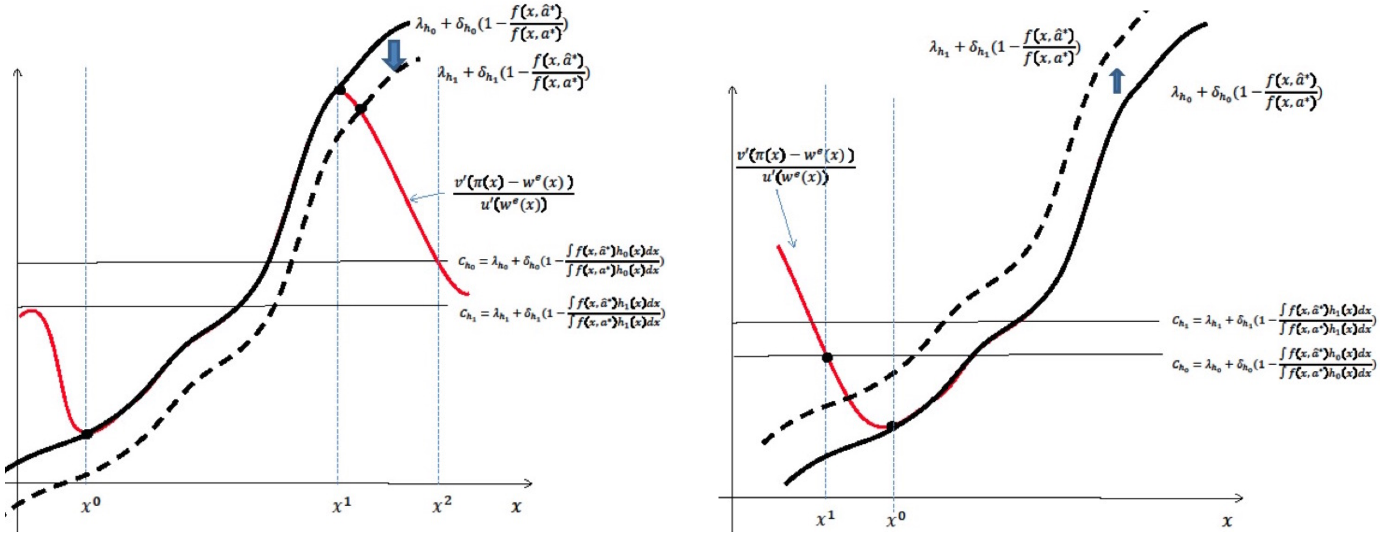
C.2.2 Stage 2

It remains to show that if $T(x)$ and $R(x)$ are comonotone on a subset of positive measure in $\overline{\mathcal{X}^*_{\underline{w}}}$ then they are comonotone on all of $\overline{\mathcal{X}^*_{\underline{w}}}$. Recall we are assuming that $\hat{a}^* < a^*$ and so $R(x)$ is nondecreasing on all of $\mathcal{X}$ (via Proposition 4.9(iv)). For simplicity of discussion below we will always be referring to the domain $\overline{\mathcal{X}^*_{\underline{w}}}$, so when we say "for all x" we mean for all $x \in \overline{\mathcal{X}^*_{\underline{w}}}$. Note that since w^{a^*} is continuous (via Remark 4.11), the set $\overline{\mathcal{X}^*_{\underline{w}}}$ has positive measure and consists of intervals in $\mathcal{X}$. The intervals explored in the proof below lie within $\overline{\mathcal{X}^*_{\underline{w}}}$.

Step 1. In this step, we show that once $T(x)$ starts to increase at point x^0, then $T(x)$ will not decrease for any $x > x^0$. To see this, by contradiction, suppose $T(x)$ turns to strictly decrease at point x^1. See Figure 3(a) for an illustration.

By Lemma 4.10, since for $x \in [x^0, x^1]$, $T(x)$ is increasing, choosing some h_0 that has support $[x^0, x^1]$, we obtain $T(x) = \lambda_{h_0} + \delta_{h_0} R(x)$, for $x \in [x^0, x^1]$. That is to say, within the interval $x \in [x^0, x^1]$, $T(x)$ should coincide with $\lambda_{h_0} + \delta_{h_0} R(x)$ for some constant λ_{h_0} and δ_{h_0}. Also we know that $T(x)$ crosses the constant C_{h_0} (defined in (4.46)) from below, by the fact that $T(x)$ increases in x. If $T(x)$ crosses the C_{h_0} again, let x^2 be the intersection otherwise, we let $x^2 = \bar{x}$. Now we construct h_1 to move $\lambda_h + \delta_h R(x)$ down. Let $h_1(x)$ have support on $[x^0, x^2]$, which is specified as:

$$h_1(x) = \begin{cases} \alpha_1 g_1(x) & \text{if } x \in L_1 \cap [x^0, x^2] \\ \alpha_2 g_2(x) & \text{if } x \in L_2 \cap [x^0, x^2] \\ 0 & \text{otherwise} \end{cases} . \tag{C.16}$$



(a) Step 1

(b) Step 2

Figure 3: Proof of Lemma 4.14

By the same argument as in Claim 2(ii), we can find $\alpha_i > 0$ satisfying (4.40) and (4.41), for any $g_i(x) \in \mathcal{H}$ and $x \in [x^0, x^2]$, given that $\Pr(L_1 \cap [x^0, x^2]) > 0$ and $\Pr(L_2 \cap [x^0, x^2]) > 0$. Moreover, $\Pr(L_1 \cap [x^0, x^2] \cap \mathcal{X}^{h_0-}) > 0$ and $\Pr(L_2 \cap [x^0, x^2] \cap \mathcal{X}^{h_0+}) > 0$ imply that we can choose some $g_i(x)$ to obtain $\int T(x)h_1(x)f(x,a^*)dx = \sum_{i=1}^{2} \int_{x^0}^{x^2} T(x)\alpha_i g_i(x) f(x,a^*)dx < \int T(x)h_0(x)f(x,a^*)dx$.

Therefore, we have $\theta_1/\theta_0 < 1$, $C_{h_1} < C_{h_0}$, and $\lambda_{h_1} + \delta_{h_1}R(x) = (\theta_1/\theta_0)[\lambda_{h_0} + \delta_{h_0}(1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)})] < \lambda_{h_0} + \delta_{h_0}R(x)$. within the interval $[x^0, x^2]$, $\lambda_{h_1} + \delta_{h_1}R(x)$ crosses $T(x)$ from below where $T(x) > C_{h_1}$. Recall the method in Case 2 of Lemma 4.10, taking h_1 as the initial variation, we can show that for some variation $h_2 \in \mathcal{H}$, $T(x) = \lambda_{h_2} + \delta_{h_2}R(x)$ for $x \in [x^0, x^2]$, which implies that $T(x)$ is increasing in $[x^1, x^2]$, a contradiction. Then, we conclude that once $T(x)$ starts to strictly increase, it will never turn to strictly decrease.

Step 2. By Step 1, if $T(x)$ starts to increase at $x = \underline{x}$, we are done. Otherwise, $T(x)$ is U-shaped. See Figure 3(b). That is, $T(x)$ is decreasing up to $x = x^0$ and starts to increase. In this case, for $x \in [x^0, \bar{x}]$, $T(x)$ is increasing and as we have shown in Step 1, it holds that $T(x) = \lambda_{h_0} + \delta_{h_0}(1 - \frac{f(x,\hat{a}^*)}{f(x,a^*)})$, for $x \in [x^0, \bar{x}]$, for some h_0 that has support only on $[x^0, \bar{x}]$.

We now construct a variation h_1 with support $[x^1, \bar{x}]$, where $x^1 < \bar{x}$ is the point where $T(x)$ crosses the constant C_{h_0} or $x^1 = \underline{x}$ if $T(x)$ does not cross C_{h_0} at $[\underline{x}, x^0]$. We can move the curve $\lambda_h + \delta_h R(x)$ up. Let $h_1(x)$ have support on $[x^1, \bar{x}]$, which is specified as follows:

$$h_1(x) = \begin{cases} \alpha_1 g_1(x) & \text{if } x \in L_1 \cap [x^1, \bar{x}] \\ \alpha_2 g_2(x) & \text{if } x \in L_2 \cap [x^1, \bar{x}] \\ 0 & \text{otherwise} \end{cases} . \tag{C.17}$$

By the same argument is in Claim 2(ii), $\alpha_i > 0$ is determined to satisfy (4.40) and (4.41), for any $g_i(x) \in \mathcal{H}$ and $x \in [x^1, \bar{x}]$, given that $\Pr(L_1 \cap [x^1, \bar{x}]) > 0$ and $\Pr(L_2 \cap [x^1, \bar{x}]) > 0$. Moreover, $\Pr(L_1 \cap [x^1, \bar{x}] \cap \mathcal{X}^{h_0-}) > 0$ and $\Pr(L_2 \cap [x^1, \bar{x}] \cap \mathcal{X}^{h_0+}) > 0$ imply that we can choose some $g_i(x)$ to obtain $\int T(x)h_1(x)f(x,a^*)dx = \sum_{i=1}^{2} \int T(x)\alpha_i g_i(x)f(x,a^*)dx > \int T(x)h_0(x)f(x,a^*)dx$.

Therefore, we have $\theta_1/\theta_0 > 1$, $C_{h_1} > C_{h_0}$, and $\lambda_{h_1} + \delta_{h_1} R(x) = (\theta_1/\theta_0)[\lambda_{h_0} + \delta_{h_0} R(x)] > \lambda_{h_0} + \delta_{h_0} R(x)$. So within the interval $[x^1, \bar{x}]$, $\lambda_{h_1} + \delta_{h_1} R(x)$ crosses $T(x)$ from below where $T(x) < C_{h_1}$. Recall Case 1 of Lemma 4.10, taking h_1 as the initial variation, we can show that there exists a variation h_2 such that

$$T(x) = \lambda_{h_2} + \delta_{h_2} R(x) \text{ for } x \in [x^1, \bar{x}],$$

which implies that $T(x)$ is increasing in $[x^1, x^0]$, a contradiction. Then, we conclude that $T(x)$ cannot be U-shaped and must be nondecreasing.

D Proof of Lemma 5.1

Proof by contradiction. Suppose $\hat{a}^* > a^*$ then $R(x)$ is nonincreasing in x, by the assumption of MLRP and Lemma 3.2(ii). By Theorem 4.14 this implies $T(x)$ is also a nonincreasing function of x on $\overline{\mathcal{X}^*_{\underline{w}}}$. Also, note that $T(x)$ is always a decreasing function of x on $\mathcal{X}^*_{\underline{w}}$ since in that region $T(x) = \frac{v'(\pi(x)-\underline{w})}{u'(\underline{w})}$ and v' is decreasing and π is an increasing (by Assumption 1). Hence, $\frac{dT}{dx} \leq 0$ for all x. Writing this out (given the definition of $T(x)$ in (4.32)) and isolating for $\frac{dw^{a^*}}{dx}$ yields:

$$\frac{dw^{a^*}}{dx} \leq \frac{-\frac{v''(\pi(x)-w^{a^*}(x))}{u'(w^{a^*}(x))}}{-\frac{v''(\pi(x)-w^{a^*}(x))}{u'(w^{a^*}(x))} - \frac{v'(\pi(x)-w^{a^*}(x))}{u'(w^{a^*}(x))} u''(w^{a^*}(x))} \frac{d\pi}{dx} \tag{D.1}$$

when the derivative exists (which is almost everywhere since w^{a^*} is almost everywhere differentiable by Proposition 3.4 and Theorem 4.15). Observe that the fractional coefficient on $\frac{d\pi}{dx}$ in (D.1) is less than 1 since $\frac{v'(\pi(x)-w^{a^*}(x))}{u'(w^{a^*}(x))} u''(w^{a^*}(x)) < 0$ by Assumption 1. This implies

$$\frac{dw^{a^*}}{dx} < \frac{d\pi}{dx}$$

whenever the derivative exists. Thus $v(\pi(x) - w^{a^*}(x))$ is increasing and the MLRP implies that $\int v(\pi(x) - w^{a^*}(x)) f(x, \hat{a}^*) dx > \int v(\pi(x) - w^{a^*}(x)) f(x, a^*) dx = V(w^{a^*}, a^*)$, contradicting the optimality of the optimal contract, since $\hat{a}^*$ is also implementable under w^{a^*} and yields a higher objective value for the principal.