Resale price maintenance and minimum requirements

Andrei Hagiu† and Julian Wright‡

February, 2017

Abstract

We study how much revenue a principal (e.g. a manufacturer) should extract from an agent (e.g. a retailer) and how much control it should grant the agent over decisions that it can monitor (e.g. the retailer’s price or the level of local advertising). This provides a new theory of resale price maintenance, explaining when price ceilings or price floors should be used. It also explains why the frequently observed practice of granting agents control over costly decisions, subject to minimum requirements, is oftentimes optimal. Our analysis applies to the contracting choices facing franchisors, manufacturers, shopping malls, online platforms, and movie studios, among other examples.

JEL classification: D4, L1, L5

Keywords: resale price maintenance, partial delegation, channel coordination, platform governance.

1 Introduction

With the shift away from per-se illegality of resale price maintenance (RPM) in the U.S. and elsewhere, business-format franchisors and manufacturers increasingly face the issue of whether and how to restrict the prices set by their respective agents (i.e. franchisees and retailers). For instance, franchisors and manufacturers can consider setting prices themselves, or imposing price caps or price floors on the prices set by franchisees and retailers. The issue of who controls prices is also highly relevant for the many digital platforms that have risen to prominence in recent years (e.g. Airbnb, Amazon, TaskRabbit, Uber). Similarly, when contracting with franchisees, a business-format franchisor such as a hotel chain, a fast-food restaurant or a car rental company has to decide how much to control

*We thank Juanjuan Zhang, Yossi Spiegel and participants in seminars at Boston University, University of Utah, and the Conference on the Economics of Intellectual Property, Software and the Internet at the Toulouse School of Economics for their helpful comments.

†MIT Sloan School of Management, Boston, MA 02142, E-mail: ahagiu@mit.edu
‡Department of Economics, National University of Singapore, Singapore 117570, E-mail: jwright@nus.edu.sg
local advertising choices. The franchisor could decide the level of local advertising itself and write it into the franchise contract, delegate the choice entirely to the franchisees, or let the franchisees decide subject to a minimum advertising expense. Similar choices arise in a wide range of other principal-agent settings (e.g. shopping malls and retailers with respect to their opening hours, movie studios and movie theaters with respect to the run length of the movie, and so on).

We build a theoretical model to evaluate the principal’s optimal levels of control over such transferable decisions (i.e. price or costly investment choices). We compare partial delegation—in which the principal retains some control—with both full control and full delegation. In particular, we determine when threshold delegation—in which the principal imposes a minimum or maximum threshold on the agent’s choice—does better than either full control or full delegation. We also provide sufficient conditions for threshold delegation to be the optimal form of partial delegation.

In our baseline model, demand is determined by (i) price, (ii) two ongoing and costly investment decisions, and (iii) a random demand shock privately observed by the agent. We first consider the case in which price can be controlled by either the principal or the agent (i.e. RPM is allowed), and both costly decisions are non-contractible and non-transferable, one always undertaken by the principal (e.g. the manufacturer’s ongoing investment in quality and marketing of the product) and one always undertaken by the agent (e.g. the retailer’s ongoing sales effort and store maintenance). We then consider the case in which price is non-contractible and always controlled by the agent (i.e. RPM is not allowed), one of the costly decisions remains non-contractible and under the principal’s control, while the other costly decision can now be chosen by either the principal or the agent, i.e. it is contractible and transferable (e.g. local advertising).

In both versions of the model, we assume the principal can make use of a two-part tariff—a fixed fee (or payment) and a wholesale fee—in its contract. We use the model to compare the principal’s expected profit from three types of contract: (i) the principal fixes the level of the transferable decision variable (i.e. the price or the costly decision) in its contract (which we call the $P$-mode), (ii) the agent is granted full control over the transferable decision variable (which we call the $A$-mode), and (iii) a hybrid mode in which the principal restricts the agent’s choice of the transferable decision variable according to a threshold rule that provides the agent with discretion subject to satisfying either a maximum or a minimum constraint (which we call the $H$-mode).

Thus, at a high level, we study the interaction between two channel coordination instruments: the wholesale price paid by the agent to the principal, and the allocation of control over transferable decision variables between them. Both are decided by the principal as part of her contract choice. To incentivize the principal to choose positive levels of her non-transferable investment, the wholesale price must always be positive. However, a positive wholesale price means that the agent’s choice of the transferable decision (price or costly decision) will be distorted. The wholesale price endogenously determines the magnitude of the agent’s bias (distortion) in choosing the transferable decision variable. In turn, the bias determines the extent to which the principal wishes to delegate control of this variable to the agent.

When the transferable decision is price (RPM is allowed), the direction of the distortion depends
on whether demand increases or decreases when price is higher. In the standard case that demand decreases in price, the agent’s pricing decision is biased upwards, reflecting the effect of double marginalization (the agent only receives a fraction of the full revenues from its costly choices to raise demand). This happens when the agent’s moral hazard is not too important. However, if the agent’s moral hazard is sufficiently important, demand will increase in price. This reflects that a higher price will induce the agent to choose a higher level of investment (or effort) to such an extent that it more than offsets the direct effect of the higher price on demand. In this case, the agent’s pricing decision will suffer from a downward bias.

The existence of an upward or a downward bias suggests that the principal may prefer the $P$-mode, in which it controls the choice of price by fixing it contractually, thereby avoiding the distortion when the agent alone sets the price. This provides the principal with a commitment benefit and reduces the need to set a low wholesale price so as to leave a large share of revenue with the agent. On the other hand, the agent may have better demand information, relevant to setting the price. This consideration suggests that the principal may prefer the $A$-mode, in which it delegates the setting of the price to the agent and sets a lower wholesale price.

Rather than fully controlling ($P$-mode) or fully delegating ($A$-mode), we show that the principal often does best using partial delegation ($H$-mode), i.e. restricting the agent’s choice of the price to be above some minimum threshold or below some maximum threshold. Threshold delegation is a way to get some of the advantages of each of the pure modes: the commitment benefit of the $P$-mode and the responsiveness to the agent’s information that is enabled by the $A$-mode, while avoiding the worst pricing biases that can arise when the agent controls the price but only keeps some of the associated variable revenue. In contrast to most existing theories of RPM, our theory explains the use of both minimum and maximum RPM, and when each would be used. For instance, we show that with two-part tariffs based on wholesale pricing, minimum (respectively, maximum) RPM is preferred when demand is increasing (respectively, decreasing) in price after taking into account the effect of price on the agent’s demand-increasing investments (e.g. local advertising, service etc). We show how this condition changes when instead the principal makes use of a revenue-sharing contract. Our theory also differs from existing theories of RPM in predicting a constraint on price that is not always binding, rather than specifying a specific level of price or a constraint that is always binding.

When the transferable decision is a costly action rather than price (i.e. RPM is not allowed), the agent’s choice of costly action is always biased downwards, reflecting that the agent only receives a fraction of the full revenues resulting from its choice of this costly action. In this case, we show that the principal often does best by restricting the agent’s choice of the costly action to be above some minimum threshold. Thus, our theory can explain the imposition of minimum requirements such as a franchisor imposing a minimum requirement on the level of franchisee advertising or investment, a manufacturer imposing a minimum requirement on an authorized dealer’s in-store promotion or investments, a shopping mall imposing a minimum requirement on a retailer’s opening hours, a movie studio imposing a minimum requirement on the number of weeks a movie theater must show a particular film, or platforms like Uber imposing minimum requirements on the quality of car a driver can
In the rest of the paper, we review the related literature (Section 2), discuss a number of examples that motivate our theory (Section 3), and set up our baseline model (Section 4). Section 5 obtains the main results, focusing on the case when contracting on price is legal (i.e. RPM is allowed). Section 6 focuses on the case when RPM is not allowed and delegation concerns a costly decision instead of price. Section 7 provides three types of extensions: general demand and cost specifications, contractual instruments other than wholesale prices, and different types of private information. Section 8 concludes.

2 Literature review

Our paper combines elements from the literatures on retail channel coordination, agency vs. wholesale pricing, partial delegation and double-sided moral hazard.

The choice between $P$-mode and $A$-mode in our model is reminiscent of the choice studied by Simon (1951) between contracting on a decision ex-ante (before uncertainty is resolved) vs. giving full authority to the employer (principal) or the employee (agent) to unilaterally choose the decision ex-post. In our model, giving full authority to the principal to unilaterally make the transferable decision ex-post is never optimal. This is because the principal never observes the realization of the agent’s private information and can always extract the entire surplus from the agent through its two-part tariffs—this means the principal can always do better by committing to the choice of the transferable action ex-ante.

A related and more recent strand of literature has emerged that studies conditions under which retailers/platforms take control over transferable decisions pertaining to the sale of products to end-consumers or allow their suppliers/complementors to keep control over these decisions. Bhardwaj (2001), Foros et al. (2013), Abhishek et al. (2015), and Johnson (2016) focus on price as the main decision that can be controlled by the retailers (wholesale model) or by the suppliers (agency model). Desiraju and Moorthy (1997), Jerath and Zhang (2010), and Hagiu and Wright (2015, 2016), study delegation of both price and costly investment (e.g. service) decisions.

The key novelty that we introduce relative to the articles mentioned in the two paragraphs above is that we allow for an intermediate option between fixing the transferable decision in the principal’s ex-ante contract and giving full authority to the agent: the agent can be given authority to choose the transferable action subject to restrictions imposed by the principal’s ex-ante contract. This is known as “partial delegation” following the seminal work by Holmstrom (1977, 1984). Several papers have proven that threshold delegation is optimal in similar settings—see for example, Melumad and Shibano (1991), Martimort and Semenov (2006), Alonso and Matouschek (2008), and Amador and Bagwell (2013). We directly show that threshold delegation is optimal in our benchmark setting under fairly general conditions on the distribution of private information. There are three key contributions in our model relative to the partial delegation literature: (1) we allow for monetary transfers between the principal and the agent in the form of two-part tariffs set by the principal in the contracting.
stage; (2) we introduce double-sided moral hazard; and (3) the bias of the agent’s objective function relative to the principal’s is endogenously determined by the two-part tariff, which in turn depends on the importance of the agent’s moral hazard relative to the principal’s, and on the importance of the agent’s private information. By contrast, the partial delegation literature to date assumes an exogenously given bias, no transfers between principal and agent and no moral hazard for either the principal or the agent.

In retail contexts, partial delegation can be viewed as an additional instrument that can help improve channel coordination. Our modelling approach is entirely consistent with the principal-agent view of channel coordination taken by the marketing and management literature to date (see Lal, 1990, Gal-Or, 1995, Ingen and Parry, 1995, Cachon and Lariviere, 2005, Raju and Zhang, 2005, and Foros et al., 2009). However, this literature has focused on improving channel coordination through various payment instruments (revenue sharing, wholesale pricing, slotting fees, quantity discounts, buy-backs) and/or through monitoring, which is modelled as enforcing a specific level of a non-contractible investment in service. We extend this work by showing that the addition of threshold delegation with respect to transferable decision variables can improve channel coordination. Threshold delegation is also commonly used in practice (see Section 3 below).

Since in our model revenues must be shared between the principal and the agent to incentivize both sides to make non-contractible investments, we also directly build upon principal-agent models with double-sided moral hazard (Lal, 1990, Romano, 1994, and Bhattacharyya and Lafontaine, 1995). The key difference relative to these papers is that our model can explain partial delegation of a transferable action, which is frequently observed in practice. Indeed, the transferable action is entirely absent from the models of Lal (1990) and Bhattacharyya and Lafontaine (1995), and is deterministic in Romano (1994), which means there is no scope for partial delegation. Furthermore, Romano (1994) focuses on price as the transferable action, whereas we cover both the case in which the transferable action is a costly investment and the case in which it is price.

In contrast to most existing theories of resale price maintenance (RPM) such as Shaffer (1991), Deneckere et al. (1996), Jullien and Rey (2007) and Asker and Bar-Isaac (2014), our model can explain when either minimum or maximum RPM should be used. Moreover, the use of price ceilings or floors in our theory of RPM is distinct from theories which assume that the manufacturer uses RPM to fix a particular level of price, e.g. Mathewson and Winter (1984), Shaffer (1991), Romano (1994). In previous theories, RPM is a binding constraint on the price, whereas in our theory the agent will sometimes set a price above the minimum or below the maximum price prescribed by RPM.

Given that our model is applicable to many industry settings featuring platforms, we also contribute to the emerging literature on platform governance rules, i.e. mechanisms employed by platform owners to regulate the access and behavior of platform participants (Boudreau and Haghlu, 2009). Specifically, the minimum requirements placed by the principal on the agent’s choice of investment can be viewed as a form of governance whenever the principal is a platform provider (e.g. Uber and Lyft as discussed in Section 3); likewise for the maximum price imposed by app stores on apps.
3 Examples

There are a wide variety of examples that our theory can be applied to. Table 1 summarizes a few key examples, listing decisions that are transferable and potentially subject to restrictions (minimum or maximum requirements), decisions that are non-transferable and subject to moral hazard, and the source of the agent’s private information.

Table 1: Examples

<table>
<thead>
<tr>
<th>source of agent’s private information</th>
<th>Non-transferable investment decisions made by the principal</th>
<th>transferable decisions (possibly subject to restrictions)</th>
<th>Non-transferable investment decisions made by the agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>local demand; effectiveness of local advertising; revenue from cross-selling other products; franchisee’s costs</td>
<td>national advertising of the brand</td>
<td>local advertising of the outlet; opening hours; price</td>
<td>outlet manager’s effort</td>
</tr>
<tr>
<td>local demand; revenue from cross-selling other products or services; dealer’s costs</td>
<td>quality and marketing of the product</td>
<td>investment in quality of outlet; local promotion and advertising; price</td>
<td>dealer’s effort</td>
</tr>
<tr>
<td>retailer’s demand, costs and outside revenues that originate from mall traffic</td>
<td>maintenance and advertising of mall</td>
<td>retailer’s opening hours</td>
<td>quality and advertising of retail store</td>
</tr>
<tr>
<td>revenues and cross-selling opportunities outside of app</td>
<td>technological upkeep (e.g. payment) and advertising of service</td>
<td>app licensing terms; price of app</td>
<td>advertising and upkeep of app</td>
</tr>
<tr>
<td>repeat business for the driver off the platform; cash tips</td>
<td>quality of car; price</td>
<td>customer service</td>
<td>technological upkeep (e.g. payment, dispatch) and advertising of service</td>
</tr>
<tr>
<td>local demand for movies; theater’s concession revenue and opportunity costs</td>
<td>run length</td>
<td>quality and maintenance of the theater’s facilities</td>
<td>advertising and promotion of the movie</td>
</tr>
<tr>
<td>local demand; revenue from cross-selling other products</td>
<td>price</td>
<td>sales effort and in-store promotion</td>
<td>quality and marketing of the product</td>
</tr>
</tbody>
</table>

In many of these examples, price is a key transferable decision. Our theory therefore applies to the case of resale price maintenance (RPM). In RPM contracts, the principal (e.g. a manufacturer, producer or franchisor) exercises control over the price set by the agent (e.g. a retailer, distributor or franchisee). A classic case is Albrecht v. Herald Co. (1968). The newspaper owner (the Herald) granted the distributor and delivery agent (Albrecht) an exclusive territory for selling the Herald’s newspaper, the St. Louis Globe-Democrat. Rather than Herald requiring Albrecht to set a particular price, or leaving Albrecht free to set the price, Herald imposed a price ceiling below which Albrecht could choose any price it wanted. The Herald only removed this maximum RPM contract following a 1968 U.S. Supreme Court decision, which ruled that fixing a maximum price was illegal. In other countries, where maximum RPM is explicitly allowed (e.g. India), retailers routinely set prices below the price ceilings imposed by manufacturers, consistent with our theory. With the shift away from per-se illegality of RPM in the U.S., business-format franchisors and manufacturers will increasingly face the issue of whether and how to restrict the prices set by their respective agents (i.e. franchisees

---

and retailers). The issue of who sets prices is also highly relevant for modern platforms, which make use of all three modes featured in our theory. Most often, developers or suppliers have complete freedom to set prices (e.g. sellers on eBay), but sometimes restrictions are imposed on them (e.g. Apple’s App Store or Google’s Play Store\(^3\)), or the platform may even set the prices itself (e.g. Uber and Lyft).

As shown in Table 1, there are also many costly (non-price) transferable decisions that can be subject to minimum threshold requirements. To illustrate, consider first “business format” franchising (e.g. hotel, fast-food, car rental, etc.). The franchisor is the principal in our setting, while the franchisee is the agent. One can often distinguish national from local advertising, with the latter being a decision that could be made either by the franchisor or the franchisee. For local advertising, contracts often specify a minimum spending requirement by the franchisee, consistent with our \(H\)-mode. Other decisions that are typically chosen by franchisees subject to minimum requirements imposed by franchisors include the number of staff that have to be on-site at various days/times, cleanliness, and opening hours.

Branded manufacturers that distribute their products through authorized dealers provide another large set of related examples. For instance, manufacturers oftentimes impose minimum standards for retail premises and minimum advertising or promotion levels by the retailers, but these same transferable decision variables can also sometimes be stipulated by the manufacturer or left unrestricted. Likewise, a manufacturer that uses sales representatives would face a similar situation—the extent to which it controls their transferable decisions (and therefore the extent to which sales representatives would be considered employees or independent contractors).

Our theory also applies to an increasing number of platforms, both offline and online. Consider three examples: shopping malls, digital app stores (e.g. Apple’s App Store for iPhone apps and Google’s Play Store for Android apps) and ride-hailing apps (e.g. Lyft and Uber). First, all three types of platforms place minimum requirements on important transferable decisions. In the case of shopping malls, the lease agreements often specify minimum opening hours for retailers (those hours could be set by the mall or by each respective retailer). Apple and Google place minimum requirements on the terms of the licensing agreement provided by app developers to their users.\(^4\) For instance, both Apple and Google require developers to assume sole responsibility for any defects or performance issues related to their apps and Google requires developers to respond to customer support inquiries within three business days. UberX and Lyft drivers have to use cars that satisfy a minimum age requirement (e.g. 2001 or newer in many cities for UberX, and 2004 or newer in many cities for Lyft). The two companies also impose minimum requirements on the cars’ functionality (e.g. 4 doors, at least 5 seat belts) and on their state of maintenance (e.g. fully functioning A/C and heating, no major cosmetic damage). In contrast, traditional taxi companies can be viewed as functioning in our \(P\)-mode, since they completely control and incur the costs corresponding to the choice of cars used by their drivers.

A further application of our theory is to the movie industry. A movie studio is the principal in our

\(^3\)App Store developers cannot charge more than $999.99, while Play Store app prices are capped at $400 in the U.S. The vast majority of apps sell below these price ceilings, but some do price at the maximum allowed level.

setting, with the movie theater the studio’s agent for the “distribution” of its movies to consumers. One important transferable variable is the run length of the movie, i.e. the number of weeks it will be shown. The studio could fix the number of weeks in its contract (P-mode), leave it unrestricted (A-mode), but most often it imposes a minimum run length on the theater (H-mode).

4 Model set-up

We assume the demand $D(p, q, Q)$ generated by a principal and an agent is determined by the choice of three decision variables: the price $p$ and two costly actions $q$ and $Q$. Throughout the paper, $Q$ is a costly, non-transferable and non-contractible action always chosen by the principal. It captures the on-going investments made by the principal that increase demand. Examples are given in the second-to-last column of Table 1.

We will consider two variants of the model. In the first, we assume the price $p$ is transferable and contractible, while the costly action $q$ is non-contractible and always chosen by the agent (this could be an effort decision—see some of the examples in the third column of Table 1). This case applies if RPM is feasible: the price can then be set by the principal or by the agent (possibly subject to a restriction). In the second variant of the model, we assume that the price $p$ is non-contractible and always chosen by the agent (this would apply if RPM is considered illegal), while the costly action $q$ is the transferable and contractible decision variable. Examples of such actions are given in the second column of Table 1.

For the baseline model we assume that demand is linear in these variables, and can be written as

$$D(p, q, Q) = \theta - \beta p + \phi q + \Phi Q,$$

where $\beta$, $\phi$ and $\Phi$ are positive constants, measuring the impact of $p$, $q$ and $Q$, respectively, on demand, and $\theta$ is an additive demand shock observed only by the agent. Formally, we assume $\theta$ is a random variable drawn from the distribution function $G$, with positive density $g(.)$ over $[\theta_L, \theta_H]$, with finite mean $E(\theta) = \bar{\theta} > 0$ and variance $V_\theta$. We assume $0 \leq \theta_L < \theta_H$, but do not require that $\theta_H$ is finite, so we allow for distributions with unbounded support on the right tail (such as the exponential or normal distributions). The distribution function $G(.)$ is twice continuously differentiable. We assume the fixed costs of the respective costly actions are $\frac{1}{2}q^2$ and $\frac{1}{2}Q^2$. The total revenue net of fixed costs generated by the principal and the agent is therefore

$$p(\theta - \beta p + \phi q + \Phi Q) - \frac{1}{2}q^2 - \frac{1}{2}Q^2.$$

We assume that the principal makes a take-it-or-leave-it offer to the agent, which is in the form

---


6Although we focus on a demand shock, in Section 7.3 we show that the logic of our analysis also extends to other sources of private information such as those noted in the last column of Table 1.

7In Section 7.1 we show that our main results hold for more general demand and cost functions.
of a two-part tariff involving a per-unit wholesale price \( w \) and a fixed participation fee \( F \). According to this contract, the principal receives \( wD(p, q, Q) + F \) from the agent, with the agent retaining \( (p - w)D(p, q, Q) - F \). The fixed fee \( F \) will always be set to leave the agent indifferent between participating or not. In practice, this could be negative (the agent may be paid a fixed salary to participate) but without loss of generality we normalize the agent’s outside option to zero, which means the optimal \( F \) will always be positive in our model. Such two-part tariffs are widespread and are used in some of the examples discussed in Section 3. They also lead to simpler expressions, and so we focus on them for expositional purposes. In Section 7.2 we explore the alternative formulation in which the principal extracts a fixed fee and a fixed share of revenues instead of a wholesale price per unit. This is the other commonly used form of two-part tariff.

In addition to the two-part tariff \((w, F)\), the contract offered by the principal also specifies the level of control over the transferable decision. If the contract directly specifies the choice of the transferable decision, we say that the principal has chosen the \( P \)-mode to reflect that only the principal determines it. If the contract leaves the agent free to set the transferable decision with no restrictions, we say that the principal has chosen the \( A \)-mode to reflect that only the agent determines it. Finally, if the contract partially restricts the agent’s choice of the transferable decision to be above or below a certain threshold, we say that the principal has chosen the \( H \)-mode, which is a hybrid of the two pure modes in that both the principal and agent determine the transferable decision.

The timing of the players’ moves is as follows: In the first stage, the principal offers its contract (which includes the choice of the transferable decision if in \( P \)-mode\(^8\)) and the agent decides whether or not to accept the contract. If the contract is accepted, in the second stage the agent observes the realization of \( \theta \) and decides on the transferable decision (if in \( A \)-mode or \( H \)-mode). In the second stage, the agent also chooses \( q \) (when \( p \) is the transferable decision) or \( p \) (when \( q \) is the transferable decision), while the principal always chooses \( Q \). Finally, payoffs are realized.

Two remarks are in order. First, when the transferable decision is the costly action \( q \), we assume the cost of \( q \) (i.e. \( \frac{1}{2}q^2 \)) is incurred by the agent in the second stage in \( A \)-mode and \( H \)-mode, and is incurred by the principal in the first stage in \( P \)-mode. Whether the cost of \( q \) is incurred by the principal or the agent in \( P \)-mode is actually immaterial to the outcome because \( q \) is set contractually and the principal can use a two-part tariff to extract the agent’s entire expected surplus in excess of a fixed outside option. Thus, our assumption that the principal incurs the cost of \( q \) in \( P \)-mode is made without any loss of generality.

Second, to keep the analysis tractable, we focused on a two-part tariff and did not consider more complicated contracts such as contracts that make the principal’s payment conditional on the level of the transferable decision (e.g. for the case of price, adding price-dependent royalties to the two-part tariffs we consider). Other papers on channel coordination have explored the benefits of using additional contractual instruments that depend on price such as fixed royalty payments in addition

---

\(^8\)In our setup, the fourth logically possible delegation option—in which the principal maintains control over the transferable decision but only chooses it in the second stage instead of fixing it in its contract—is always dominated. This is because the principal never observes any private information, so there is no benefit in waiting rather than committing to the choice of the transferable decision ex-ante.
to wholesale prices (Lal, 1990) and price-dependent royalties (Foros et al., 2009). We discuss the possibility of using such instruments in Section 7.2.

5 Resale price maintenance

In this section we analyze the variant of the benchmark model in which the price \( p \) is contractible and transferable, while the costly actions \( q \) and \( Q \) are non-contractible and non-transferable (and therefore create double-sided moral hazard). This captures the case in which resale price maintenance is allowed. In order to ensure all decision variables and profits are positive and second-order conditions hold, we add the assumptions

\[
(2\beta - \phi^2)(\Phi^2 + \phi^2) - \Phi^4 > 0
\]

and

\[
\frac{\theta_L}{\theta} > \frac{(\beta - \Phi^2)\Phi^2}{(2\beta - \phi^2)(\Phi^2 + \phi^2) - \Phi^4 + (\beta - \phi^2)^2}.
\]

The first assumption requires \( \beta \) is not too small. The second assumption always holds if \( \Phi^2 \geq \beta \), or, in case \( \beta > \Phi^2 \), if \( \theta_L \) is not too small. We first analyze whether the principal prefers to set the level of \( p \) in its contract (P-mode) or entirely delegate that choice to the agent (A-mode), before considering whether the principal can do better than both pure modes through threshold delegation (H-mode).

5.1 Full control vs. no control

Consider first the P-mode. The principal solves

\[
\max_{w,F,p} \left\{ \mathbb{E} [w (\theta - \beta p + \phi q + \Phi Q)] + F - \frac{1}{2} Q^2 \right\}
\]

subject to

\[
F \leq \mathbb{E} \left[ (p - w) (\theta - \beta p + \phi q + \Phi Q) - \frac{1}{2} q^2 \right]
\]

\[
q = \arg \max_q \left\{ (p - w) (\theta - \beta p + \phi q' + \Phi Q') - \frac{1}{2} q'^2 \right\}
\]

\[
Q = \arg \max_{Q'} \left\{ \mathbb{E} [w (\theta - \beta p + \phi q + \Phi Q')] - \frac{1}{2} Q'^2 \right\}.
\]

The first constraint is the agent’s participation constraint—recall the agent must make the participation decision prior to learning the realization of \( \theta \). The second and third constraints are the agent’s and principal’s respective stage 2 optimization problems. They reflect that the agent and the principal’s incentives to invest in \( q \) and \( Q \) respectively are driven by the margins that each extracts: \( p - w \) for the agent and \( w \) for the principal.

At the optimum, the first constraint is also binding (the principal uses the fixed fee \( F \) to extract the entire net expected payoff from the agent) and, after solving the optimization problems in the last
two constraints, the principal’s program can be re-written

\[ \max_{w,p} \mathbb{E} \left[ p(\theta - \beta p + \phi q + \Phi Q) - \frac{1}{2} q^2 - \frac{1}{2} Q^2 \right] \]

subject to

\[
\begin{align*}
q &= (p - w) \phi \\
Q &= w \Phi.
\end{align*}
\]

Note that, after substituting the two constraints in \((q,Q)\) into the principal’s objective function, whether demand is decreasing or increasing in price depends on whether \(\beta\) (the direct price effect) is higher or lower than \(\phi^2\) (the indirect price effect through the agent’s investment in \(q\)). Even if demand is increasing in price, the profit maximization problem remains well behaved. This is because a higher price increases \(q\) and so the marginal cost of the agent’s investment, which ensures profit eventually decreases with price provided \(\phi^2 < 2\beta\), a condition which follows from (2). Plugging in the two constraints, the principal’s profit as a function of \((w,p)\) is

\[
\Pi^P(w,p) \equiv \frac{1}{2} p \left( 2\bar{\theta} - (2\beta - \phi^2) p + 2w\Phi^2 \right) - \frac{1}{2} w^2 (\Phi^2 + \phi^2). \tag{4}
\]

Optimizing \(\Pi^P(w,p)\) over \(p\) we obtain that the price set by the principal for a given wholesale price \(w\) is \(p^P(w,\theta)\), where

\[
p^P(w,\theta) = \frac{\theta + w\Phi^2}{2\beta - \phi^2}. \tag{5}
\]

Thus, \(p^P(w,\theta)\) is the hypothetical price the principal would want to set given \(w\) if it could observe \(\theta\). This price is higher when \(w\) is higher, reflecting that increasing \(w\) induces the principal to invest more in stage 2, which in turn increases demand. Given that the principal sets \(p^P(w,\theta)\) in the first stage, the agent will set \(q^P(w,\theta)\) in the second stage, where

\[
q^P(w,\theta) = \left( \frac{\theta - w(2\beta - \phi^2 - \Phi^2)}{2\beta - \phi^2} \right) \phi. \tag{6}
\]

Substituting \(p^P(w,\theta)\) back into \(\Pi^P(w,p)\), we obtain

\[
\Pi^P(w) \equiv \frac{(\bar{\theta} + w\Phi^2)^2}{2(2\beta - \phi^2)} - \frac{1}{2} w^2 (\Phi^2 + \phi^2). \tag{7}
\]

Optimizing \(\Pi^P(w)\) over \(w\), we obtain

\[
w^{P*} = \frac{\Phi^2 \bar{\theta}}{(2\beta - \phi^2)(\Phi^2 + \phi^2)} - \Phi^4.
\]

It is easily verified that (2) implies \(w^{P*} > 0\), \(p^P(w^{P*},\bar{\theta}) > w^{P*}\) and \(q^P(w^{P*},\bar{\theta}) > 0\). The resulting
optimal profit for the principal in $P$-mode is

$$
\Pi^P = \frac{\theta^2 (\phi^2 + \Phi^2)}{2((2\beta - \phi^2)(\phi^2 + \Phi^2) - \Phi^4)}.
$$

Consider now the $A$-mode. The principal solves

$$
\max_{w,F} \left\{ \mathbb{E} \left[ w \left( \theta - \beta p + \phi q + \Phi Q \right) \right] + F - \frac{1}{2} Q^2 \right\}
$$
subject to

$$
F \leq \mathbb{E} \left[ (p - w) \left( \theta - \beta p + \phi q + \Phi Q \right) - \frac{1}{2} q^2 \right]
$$

$$
p = \arg \max_{p'} \left\{ (p' - w) \left( \theta - \beta p' + \phi q + \Phi Q \right) - \frac{1}{2} q^2 \right\}
$$

$$
q = \arg \max_{q'} \left\{ (p - w) \left( \theta - \beta p + \phi q' + \Phi Q \right) - \frac{1}{2} q'^2 \right\}
$$

$$
Q = \arg \max_{Q'} \left\{ \mathbb{E} \left[ w \left( \theta - \beta p + \phi q + \Phi Q' \right) \right] - \frac{1}{2} Q'^2 \right\}.
$$

The additional constraint facing the principal compared to its $P$-mode problem is that the agent sets the price in the second stage optimally given the principal’s wholesale price set in the first stage and the observed demand shock $\theta$. Again, the fixed fee (or salary) $F$ is set such that the principal extracts the entire expected payoff in excess of the agent’s outside option. Thus, after replacing the last three constraints with the corresponding second-stage first order conditions, the principal solves

$$
\max_w \left\{ \mathbb{E} \left[ p \left( \theta - \beta p + \phi q + \Phi Q \right) - \frac{1}{2} q^2 - \frac{1}{2} Q^2 \right] \right\}
$$
subject to

$$
p = \frac{w}{2} + \frac{1}{2\beta} (\theta + \phi Q)
$$

$$
q = (p - w) \phi
$$

$$
Q = w\Phi.
$$

The price is determined by the usual monopoly pricing formula based on linear demand. As in $P$-mode, whether demand is decreasing or increasing in price depends on whether $\beta$ (the direct price effect) is higher or lower than $\phi^2$ (the indirect price effect through the agent’s investment in $q$). Solving for $(p,q,Q)$ as functions of $w$, we obtain

$$
p^A(w,\theta) = \frac{\theta + (\Phi^2 + \beta - \phi^2) w}{2\beta - \phi^2}
$$

$$
q^A(w,\theta) = \frac{(\theta - w(\beta - \Phi^2)) \phi}{2\beta - \phi^2}
$$

$$
Q^A(w) = w\Phi.
$$
Comparing (5) and (9) reveals that for the same positive level of $w$, the agent has an upward bias in choosing the price (i.e. $p^A(w, \theta) > p^P(w, \theta)$) if $\beta > \phi^2$ and a downward bias in choosing the price if $\beta < \phi^2$. The existence of a bias (upward or downward) is due to the positive wholesale price, which means the agent only receives a fraction of the full revenues from its costly choice of $q$. This leads to choices of the agent that are distorted away from the levels preferred by the principal. Indeed, the absolute value of the bias is increasing in $w$. The direction of the bias is determined by whether demand is decreasing or increasing in price. If demand is decreasing in price ($\beta > \phi^2$), then the wholesale price $w$ leads the agent to set $p$ too high from the principal’s perspective—the normal double marginalization effect dominates. On the other hand, if demand is increasing in price ($\beta < \phi^2$), then the wholesale price $w$ leads the agent to set $p$ too low from the principal’s perspective. Indeed, in this case the effect that dominates is the fact that the agent does not fully internalize the benefit to the principal of a higher investment $q$ which would be induced by setting a higher price.

If there was no moral hazard by the principal (i.e. if $\Phi = 0$), then the principal could eliminate the agent’s bias by setting the wholesale price equal to its marginal cost (zero in this case), so that the agent retains the full margin associated with its decisions. However, with the principal facing a moral hazard problem ($\Phi > 0$), this would not be profitable. Note also that $q^A(w, \theta) < q^P(w, \theta)$ if and only if $\beta < \phi^2$, i.e. if and only if the agent has a downward bias in choosing price. Thus, when $\beta < \phi^2$, the strategic complementarity of $p$ and $q$ implies that the expected level of effort $q$ chosen by the agent will also be too low.

Substituting expressions (9)-(11) back into the expression of the principal’s profits, we obtain that the principal solves $\max_w \Pi^A(w)$, where

$$
\Pi^A(w) = \left(\frac{w \Phi^2 + \theta}{\Phi^2 + \phi^2} + V_\theta - w^2 (\beta - \phi^2)^2 - \frac{1}{2} w^2 (\Phi^2 + \phi^2)\right).
$$

The optimal wholesale price extracted by the principal in $A$-mode is then

$$
w^{A^*} = \frac{\Phi^2 \theta}{(2 \beta - \phi^2)(\Phi^2 + \phi^2) + (\beta - \phi^2)^2 - \Phi^4}.
$$

It is easily verified that (2) implies $w^{A^*} > 0$, while (3) ensures $p^A(w^{A^*}, \theta) > w^{A^*}$ and $q^A(w^{A^*}, \theta) > 0$ for all $\theta$. Note that $0 < w^{A^*} < w^{P^*}$. Plugging $w^{A^*}$ into (12), the principal’s optimal $A$-mode profit is

$$
\Pi^{A^*} = \frac{\theta^2 (2 \Phi^2 + \beta - \Phi^2 \phi^2)}{2(2 \beta - \phi^2) \left(\frac{\Phi^2 + \phi^2}{(2 \beta - \phi^2)(\Phi^2 + \phi^2) + (\beta - \phi^2)^2 - \Phi^4} + \frac{V_\theta}{2(2 \beta - \phi^2)}\right)}.
$$

Comparing $\Pi^{A^*}$ with $\Pi^{P^*}$, we obtain the following proposition.

---

9 Equation (10) implies that the agent’s investment $q^A(w, \theta)$ is increasing in the wholesale price $w$ whenever $\beta < \phi^2$. To understand this, note that, whenever $\Phi > 0$, increasing $w$ increases $p$ through two channels: (i) the direct double marginalization effect ($w$ is the agent’s marginal cost), and (ii) the indirect effect through a higher choice of $Q$ by the principal, which increases the level of demand. Thus, if $\Phi$ is large enough, the increase in $p$ can be larger than the increase in $w$. In this case, $q$ is increasing in $w$. 

13
Proposition 1 The principal’s profit is higher in A-mode compared to P-mode if and only if the variance of the agent’s private information is sufficiently large, i.e.

\[ \frac{V_\theta}{\theta^2} > \frac{\Phi^4 (\beta - \phi^2)^2}{(2\beta - \phi^2)(\phi^2 + \Phi^2) - \Phi^4} \left( (2\beta - \phi^2)(\Phi^2 + \phi^2) + (\beta - \phi^2)^2 - \Phi^4 \right). \] (14)

The inequality in (14) captures the key tradeoff between the two pure modes. On the one hand, the A-mode leverages the agent’s private information on demand as captured by \( V_\theta \). On the other hand, the P-mode removes the distortion created by the agent setting \( p \) in A-mode (this explains why the right-hand side of (14) is positive).

It is straightforward to verify that the right-hand side of (14) is increasing in \( \Phi \), so the tradeoff shifts in favor of the P-mode when the principal’s non-contractible investment becomes more important. This is easily understood: as noted above, the principal extracts a larger wholesale price \( w \) in P-mode than in A-mode, so the level of the principal’s investment is higher in P-mode. Thus, as this investment becomes more important, the associated gain from shifting to the P-mode is larger.

However, the effect of \( \phi \) on the tradeoff is ambiguous (the sign of the derivative of the right-hand side of (14) in \( \phi \) is indeterminate). To understand why this is the case, recall that the agent’s pricing bias is proportional to \( w (\beta - \phi^2) \). Thus, when \( \phi^2 \) is below \( \beta \) and increases, the bias is reduced, so the loss associated with distorted pricing by the agent in A-mode is also reduced. Furthermore, increasing \( \phi^2 \) reduces the effective price sensitivity of demand and therefore allows the agent to better respond to his private information in A-mode (this is captured by the term \( \frac{V_\theta}{\sqrt{2\beta - \phi^2}} \) in the optimal A-mode profit (13)). Thus, both effects contribute to making the A-mode relatively more attractive. By contrast, when \( \phi^2 \) is above \( \beta \) and increases, the bias is exacerbated. This effect makes the A-mode relatively less attractive and can outweigh the other, positive effect of increasing \( \phi^2 \) (operating through \( \frac{V_\theta}{\sqrt{2\beta - \phi^2}} \)).

These and other comparative static effects are illustrated in the figures in Section 5.3.

5.2 Threshold delegation

Now suppose the principal can monitor \( p \) and therefore restrict the agent’s choice of \( p \) according to some rule (i.e. H-mode). In particular, the principal could restrict the agent’s choice of \( p \) to a degenerate interval \( \{p_0\} \) that only contains one point—this effectively replicates the P-mode where the principal sets \( p = p_0 \) in its contract. At the other extreme, the principal’s restriction could be so lax that it places no effective constraint on the agent’s choice of \( p \)—this replicates the A-mode. For the sake of clarity, we will only use the label H-mode when the principal’s restriction is neither one of these two extremes, but instead places some partial restriction on the agent’s choices. Otherwise, we will refer to the contract choice as P-mode or A-mode given the equivalence noted above.

Throughout the paper we focus on partial restrictions that are threshold rules. A threshold rule is one in which the principal restricts the agent’s choice to be above or below a certain threshold. We first determine sufficient conditions for the H-mode with threshold delegation to dominate both the A-mode and the P-mode. Subsequently, we will provide a sufficient condition for threshold delegation
to be the optimal form of partial delegation. Even when threshold delegation is not the optimal form of partial delegation, it does have the advantage of being simple to write down in a contract and relatively easy to monitor (as opposed to, for example, delegation that involves multiple intervals). This explains why threshold delegation is often used in practice and justifies our focus on it here.

As pointed out above, the agent has an upward (respectively, downward) bias in A-mode relative to what the principal would set in P-mode if and only if \( \beta > \phi^2 \) (respectively, \( \beta < \phi^2 \)). Thus, the relevant form of threshold delegation is that with a maximum threshold if \( \beta > \phi \) and that with a minimum threshold if \( \beta < \phi^2 \).

Consider first the case \( \beta > \phi^2 \). In this case, given a wholesale price \( w \) and a maximum threshold \( x \), the agent chooses

\[
p(\theta, \beta) = \begin{cases} 
\frac{\theta + (\beta + \Phi^2 - \phi^2)w}{2\beta - \phi^2} & \text{if } \theta \leq (2\beta - \phi^2)x - w(\beta + \Phi^2 - \phi^2) \\
\Phi & \text{if } \theta \geq (2\beta - \phi^2)x - w(\beta + \Phi^2 - \phi^2).
\end{cases}
\]

As in the pure modes, the principal extracts the agent’s entire expected payoff through the fixed fee, so the principal’s profit is

\[
\max_{w, x} \left\{ \mathbb{E} \left[ p(\theta - \beta p + \phi q + \Phi Q) - \frac{1}{2}q^2 - \frac{1}{2}Q^2 \right] \right\}
\]

subject to

\[
p = \min \left\{ p^A (w, \theta), x \right\}
\]

\[
q = \arg \max_{q'} \left\{ (p - w)(\theta - \beta p + \phi q' + \Phi Q) - \frac{1}{2}q'^2 \right\} = (p - w) \phi
\]

\[
Q = \arg \max_{q'} \left\{ \mathbb{E} \left[ w(\theta - \beta p + \phi q + \Phi Q') \right] - \frac{1}{2}Q'^2 \right\} = w \Phi.
\]

Substituting in the three constraints, we obtain that the principal’s profit is \( \max_{w, x} \Pi^H (w, x) \), where

\[
\Pi^H (w, x) \equiv \int_{\theta_L}^{\theta_H} \left(2\beta - \phi^2\right)x - w(\beta + \Phi^2 - \phi^2) \left(\theta + w\Phi^2\right)^2 - w^2 (\beta - \phi^2)^2 \frac{dG(\theta)}{2 (2\beta - \phi^2)} + \frac{1}{2}x(2\theta - (2\beta - \phi^2)x + 2w\Phi^2) dG(\theta) - \frac{1}{2}w^2 (\Phi^2 + \phi^2).
\]

If \( (2\beta - \phi^2)x - w(\beta + \Phi^2 - \phi^2) \geq \theta_H \), then \( x \) places no effective constraint on the agent, who chooses \( p = p^A (w, \theta) \) for all \( \theta \). This replicates the A-mode, so the principal’s profits are:

\[
\Pi^H (w, x) = \frac{(\theta + w\Phi^2)^2 + V_\theta - w^2 (\beta - \phi^2)^2}{2 (2\beta - \phi^2)} - \frac{1}{2}w^2 (\Phi^2 + \phi^2) = \Pi^A (w).
\]

Similarly, if \( (2\beta - \phi^2)x - w(\beta + \Phi^2 - \phi^2) \leq \theta_L \), then the constraint on \( p \) is always binding, so \( p = x \) for all \( \theta \). This is equivalent to the principal choosing \( p = x \) contractually, i.e. the P-mode. The
principal’s resulting profits are

\[ \Pi^H (w, x) = \frac{1}{2} x (2 \theta - (2 \beta - \phi^2) x + 2 w \Phi^2) - \frac{1}{2} w^2 (\Phi^2 + \phi^2) = \Pi^P (w, x). \]  

(16)

As a result, the \( H \)-mode only refers to the case when \((w, x)\) are “interior”, i.e. such that

\[ \theta_L < (2 \beta - \phi^2) x - w (\beta + \Phi^2 - \phi^2) < \theta_H. \]  

(17)

The principal’s profit as a function of \((w, x)\) is then \( \Pi^H (w, x) \) given by expression (15) above.

Things are very similar when \( \beta < \phi^2 \) and the principal sets a minimum (rather than a maximum) threshold. The same calculations yield

\[ \Pi^H (w, x) = \int_{\theta_L}^{\theta_H} \frac{(\theta + w \Phi^2)^2 - w^2 (\beta - \phi^2)^2}{2 (2 \beta - \phi^2)} dG (\theta) \]

\[ + \int_{\theta_L}^{(2 \beta - \phi^2) x - w (\beta + \Phi^2 - \phi^2)} \frac{1}{2} x (2 \theta - (2 \beta - \phi^2) x + 2 w \Phi^2) dG (\theta) - \frac{1}{2} w^2 (\Phi^2 + \phi^2) \]

for interior \((w, x)\), which is defined in the same way:

\[ \theta_L < (2 \beta - \phi^2) x - w (\beta + \Phi^2 - \phi^2) < \theta_H. \]

In both cases, the advantage of delegating to the agent is that the agent will take into account the realized value of \( \theta \) when choosing \( p \), so will set \( p \) closer to the first-best level, and the principal can extract this additional expected payoff through its fixed fee \( F \). But the principal also needs to extract a positive margin (i.e. \( w > 0 \)) in order to maintain an incentive to invest in \( Q \). This in turn distorts the agent’s choice of \( p \)—upwards if \( \beta > \phi^2 \) and downwards if \( \beta < \phi^2 \). The principal therefore prefers to stipulate a maximum (respectively, minimum) level of \( p \) to help offset the upward (respectively, downward) bias, although at the cost of having \( p \) set too low (respectively, too high) whenever \( \theta \) turns out to be particularly high (respectively, particularly low).

Thus, in some sense, threshold delegation would seem like a way for the principal to combine some of the benefits of both delegation and control. The following proposition establishes the conditions under which threshold delegation dominates the two pure modes.\(^\text{10}\)

**Proposition 2** (Resale price maintenance)

**Maximum RPM:** If \( \beta > \phi^2 \), the \( H \)-mode with maximum RPM dominates the \( A \)-mode. If in addition \( \frac{\partial \theta}{\partial p} < 1 - \frac{\Phi^2 (\beta - \phi^2)}{(2 \beta - \phi^2) (\phi^2 + \Phi^2) - \phi^2} \), the \( H \)-mode with maximum RPM also dominates the \( P \)-mode.

**Minimum RPM:** If \( \beta < \phi^2 \), the \( H \)-mode with minimum RPM dominates the \( A \)-mode. If in addition \( \frac{\partial \theta}{\partial p} > 1 + \frac{\Phi^2 (\phi^2 - \beta)}{(2 \beta - \phi^2) (\phi^2 + \Phi^2) - \phi^2} \), the \( H \)-mode with minimum RPM also dominates the \( P \)-mode.

\(^\text{10}\)It is straightforward to verify that \( \frac{\partial \theta}{\partial p} \) is not incompatible with assumption (3).
Whether maximum or minimum RPM is optimal depends on whether the agent’s pricing bias is upwards or downwards. In turn, recall that this is determined by whether demand is decreasing or increasing in price taking into account the indirect positive effect through the agent’s choice of q. If the magnitude of the normal price sensitivity of demand parameter β is larger than the magnitude of the indirect positive effect of price on demand as measured by φ^2, then the agent’s pricing distortion due to the wholesale price leads to an upward pricing bias. And vice versa if β < φ^2. Put differently, the direction of the bias depends on the strength of the agent’s moral hazard problem. If it is not very important, then the usual double marginalization distortion dominates and maximum RPM is optimal. If the agent’s moral hazard problem is sufficiently important, then the agent will set prices too low from the principal’s perspective since higher prices actually expand demand on net, and minimum RPM is optimal. This explanation is aligned with the standard justification given by manufacturers for the imposition of minimum RPM requirements on retailers, namely that this preserves sufficient incentives by the retailers to invest in services so as to promote consumer demand.

In contrast to most existing theories of RPM, Proposition 2 can explain when either minimum or maximum RPM should be used. The use of a constraint on prices rather than imposing a specific price level reflects the realistic feature that a principal often wants to give the agent discretion to react to its private information about demand shocks, while mitigating the worst pricing biases that can arise when the agent controls the price but only keeps some of the associated variable revenue. In practice, as noted in Section 3, retailers sometimes set prices below price ceilings or above price floors, consistent with our theory but inconsistent with standard RPM theories.

While the proof of Proposition 2 (along with those of other propositions not proven directly in the text) is given in the appendix, here we sketch the idea behind the proof. The comparison with the A-mode implies that putting some restriction on the agent’s choice of p is always optimal. To see this, suppose β > φ^2 and consider the H-mode with w = w^A* and the maximum threshold 

\[ x = p^A(w^A*, \theta_H) = \theta_H + \frac{(\beta + \phi^2 - \phi^2)w^{A^*}}{2\beta - \phi^2}. \]

This replicates the outcome of the A-mode since the maximum restriction on price would never be binding. Now consider decreasing x by a small amount κ. The only change in the resulting profits occurs when the realized θ is in the interval \( [\theta_H - \kappa (2\beta - \phi^2), \theta_H] \). In this interval the maximum requirement on p binds. The logic is that for all realizations of θ sufficiently close to the highest possible value θ_H, the price that the agent wants to choose is higher than even the highest price the principal would ever want to choose if it could observe θ (due to the agent’s upward bias). This means that even without observing θ, the principal must be better off by preventing the agent from choosing its optimal response to the highest values of θ. Put differently, there is no value of θ for which the principal would be happy with the agent’s choice of p in this range. By the same logic when β < φ^2, the principal can always do strictly better than the A-mode by adding a minimum requirement that prevents the agent from choosing the lowest values of p.

When β > φ^2, the comparison with the P-mode implies that leaving the agent some discretion over the choice of p is optimal provided that the realized value of the demand shock θ is sometimes sufficiently low relative to its expected value. The simple explanation for this is that the principal can always do at least as well as in the P-mode by setting w = w^P* and allowing the agent to choose
$p$ below the level that the principal would choose in the $P$-mode, i.e. $p^P (w, \theta)$. Given there is an upward bias in the agent’s choice of $p$, when the agent prefers to set $p$ lower than $p^P (w, \theta)$, the principal must also be better off compared to the case when it chooses $p^P (w, \theta)$ (i.e. $P$-mode). The inequality $\theta < \left(1 - \frac{\Phi^2(\beta - \phi^2)}{(2\beta - \phi^2)(\phi^2 + \beta^2) - \phi^2} \right) \theta$ imposed in Proposition 2 can be written as $p^A (w, \theta_L) < p^P (w, \theta)$, which ensures that there are some realizations of $\theta$ for which the agent would indeed want to set $p$ below $p^P (w, \theta)$. If the condition $\theta < \left(1 - \frac{\Phi^2(\beta - \phi^2)}{(2\beta - \phi^2)(\phi^2 + \beta^2) - \phi^2} \right) \theta$ does not hold, then even for the lowest realization of $\theta$, the agent’s choice of $p$ would still be above what the principal would like to choose in $P$-mode, so the simple logic above no longer applies.

By the same logic, when $\beta < \phi^2$ and $\theta_H > \left(1 + \frac{\Phi^2(\phi^2 - \beta)}{(2\beta - \phi^2)(\phi^2 + \beta^2) - \phi^2} \right) \theta$, the principal can always do at least as well as in the $P$-mode by setting $w = w^P$ and allowing the agent to choose $p$ above the level that the principal would choose in the $P$-mode.

We analyze the cases not covered by Proposition 2 in Section 5.3, taking $G$ as the uniform distribution. We show there that the $P$-mode can dominate threshold delegation.

Proposition 2 says that the $A$-mode can never be optimal. However, this conclusion holds in the absence of any fixed costs that the principal might incur when operating in a particular mode. In reality, the $H$-mode and the $P$-mode are likely to incur higher fixed costs than the $A$-mode, due to the need to monitor the agent in order to ensure it respects the constraint imposed by the threshold $x$ in $H$-mode or by the principal’s contractual choice of $p$ in $P$-mode. Clearly, if this monitoring cost is sufficiently large and condition (14) holds, then the $A$-mode will be optimal. In Section 5.3 we examine how the optimal choice of mode depends on the magnitude of such monitoring costs.

Proposition 2 established conditions under which the principal prefers threshold delegation to the two pure modes. However, there are more complex forms of partial delegation that the principal could utilize (e.g. delegating subject to requirement that price be in one of two disjoint intervals). The next proposition provides a sufficient condition for the threshold delegation described in Proposition 2 to be the optimal form of delegation.

**Proposition 3** If $g' (\theta) \leq 0$ for all $\theta \in [\theta_L, \theta_H]$, the optimal contract in $H$-mode involves threshold delegation.

The condition in the proposition requires that the density is non-increasing on the support of $\theta$. It is obviously satisfied by the uniform distribution (which has constant density) and all distributions that have decreasing density on the positive domain (e.g. the normal or log-normal distributions). When this condition holds, Proposition 3 implies that we should never observe a price floor and price cap used together, other than in the trivial case where they coincide (i.e. the $P$-mode).

In what follows, we will equate the $H$-mode with the principal imposing a minimum or maximum threshold. Assuming that the first-order conditions of $\Pi^H (w, x)$ characterize a unique interior maximum $(w^H, x^*)$ (recall that $(w, x)$ is said to be interior if it satisfies condition 17), we can also provide some additional properties of the optimal solution in $H$-mode with threshold delegation, as summarized in the following proposition.
Proposition 4 The optimal wholesale price extracted by the principal is highest in $P$-mode, lowest in $A$-mode, and intermediate in $H$-mode, i.e. $0 < w^A_* < w^H_* < w^P_*$. Moreover, if the agent has a downward bias (i.e. if $\beta < \phi^2$), the optimal minimum requirement for $p$ in $H$-mode is below the fixed choice of $p$ in $P$-mode, i.e. $x^* < p^P (w^P*, \overline{\theta})$.

Compared to its optimal solution in $A$-mode, the principal extracts a higher wholesale price in $P$-mode. This reflects that in $P$-mode, the principal sets the price so there is no bias, and $w$ is determined by trading off the two moral hazards due to the agent’s and the principal’s non-contractible costly investments. By contrast, in $A$-mode, the agent sets the price and the magnitude of the resulting bias is increasing in $w$. This additional concern means the optimal $w$ is lower. The $H$-mode is intermediate.

To understand the second part of the proposition, recall from the earlier discussion that if the agent has a downward pricing bias, then the principal prefers to allow the agent to set the price above $p^P (w^P*, \overline{\theta})$ whenever the agent wants to do so, rather than forcing it to set the price equal to $p^P (w^P*, \overline{\theta})$. Once it allows the agent this pricing freedom, we know from the first part of Proposition 4 that the principal wants to lower $w$ below $w^P_*$. This in turn means the principal’s preferred price also decreases below $p^P (w^P*, \overline{\theta})$, so the principal now wants to allow the agent freedom to price above this new preferred price, i.e. even more freedom than before. And this further lowers the principal’s preferred $w$ and so on until we reach $w^H_* < w^P_*$. This means that the optimal minimum threshold must be strictly lower than $p^P (w^P*, \overline{\theta})$.

However, when the agent has an upward pricing bias ($\beta > \phi^2$), the symmetric result does not necessarily hold, i.e. we do not necessarily have $x^* > p^P (w^P*, \overline{\theta})$. Indeed, in this case the principal prefers to allow the agent to set price below $p^P (w^P*, \overline{\theta})$ whenever it wants to rather than forcing it to set price equal to $p^P (w^P*, \overline{\theta})$. Once the agent has this pricing freedom, the principal again wants to decrease $w$ below $w^P_*$. This decreases the principal’s preferred price below $p^P (w^P*, \overline{\theta})$, which now results in less freedom for the agent. As a result, we cannot conclude whether $x^* > p^P (w^P*, \overline{\theta})$ or not.

5.3 Comparative statics

Given our result in Proposition 2 that $H$-mode dominates $A$-mode, in order to look at interesting tradeoffs between all three modes, in this section we introduce a monitoring cost that must be incurred whenever the principal wants to exercise some control over the price. This monitoring cost arises because the principal wants to ensure either that the price is actually set at the particular contracted level ($P$-mode) or that the agent complies with the maximum or minimum restrictions placed on price ($H$-mode). To keep things as simple as possible, we assume that the monitoring technology requires a fixed ex-ante investment (e.g. hiring additional managers and staff) and no extra cost ex-post (see Gal-Or, 1995), thus abstracting away from any strategic monitoring game that the principal and the agent might engage in (as in Lal, 1990). We denote the fixed monitoring costs by $K$. When $K > 0$, any of the three modes can be optimal.

At the end of Section 5.1 we have seen that when the principal’s ongoing investment in quality
becomes more important in determining revenue (i.e. when $\Phi$ increases), the $P$-mode becomes more desirable relative to the $A$-mode. We can extend this result to the comparison among all three modes.

**Proposition 5** A larger $\Phi$ shifts the tradeoff between $A$-mode and $H$-mode in favor of $H$-mode, and shifts the tradeoff between $H$-mode and $P$-mode in favor of $P$-mode. I.e.

$$\frac{d\Pi^P_*}{d\Phi} > \frac{d\Pi^H_*}{d\Phi} > \frac{d\Pi^A_*}{d\Phi}.$$  
Moreover, the optimal wholesale prices $w^P_*$, $w^A_*$ and $w^H_*$ are all increasing in $\Phi$.

These results are easily understood. Since the principal’s investment in $Q$ is determined by the wholesale price, when this investment is more important, it is natural that the optimal wholesale prices increase in all three modes. Furthermore, because the wholesale price is highest in $P$-mode, lowest in $A$-mode and intermediate in $H$-mode (Proposition 4), the principal’s investment is also highest in $P$-mode, lowest in $A$-mode and intermediate in $H$-mode. Thus, when $Q$ becomes more important, the $P$-mode becomes more attractive relative to the $H$-mode, which in turn becomes more attractive relative to the $A$-mode.

Next, we wish to explore the effect of increasing the importance of the agent’s private information about $\theta$, which is captured by its variance $V_\theta$. To do so, we focus on the case when $\theta$ follows a uniform distribution. Specifically, let $\theta_L = -\sigma$ and $\theta_H = +\sigma$, so that the variance of $\theta$ is $V_\theta = \frac{\sigma^2}{3}$. The next proposition characterizes the interior solution for the $H$-mode and the effect of $\sigma$ on the optimal choice of mode. The proof, which contains lengthy calculations, is relegated to the online appendix.

**Proposition 6** The optimal solution in the $H$-mode is interior if and only if

$$\sigma > \frac{\Phi^2 \beta |\beta - \phi^2|}{(2\beta^2 - \Phi^4 + 2\beta\Phi^2 - 2\Phi^2\phi^2)}$$

or

$$\frac{4\Phi^2 \beta |\beta - \phi^2|^3}{(\beta^2 - \Phi^4 + 2\beta\Phi^2 - 2\Phi^2\phi^2)^2} \leq \sigma \leq \frac{\Phi^2 \beta |\beta - \phi^2|}{(2\beta^2 - \Phi^4 + 2\beta\Phi^2 - 2\Phi^2\phi^2)}$$

and

$$(2\beta - \phi^2)(\phi^2 + \Phi^2) - \Phi^4 < |\beta - \phi^2|^2.$$  
It is then characterized by

$$w^* = \frac{\sigma}{2|\beta - \phi^2|^3} \left( \beta^2 - \Phi^4 + 2\beta\Phi^2 - 2\Phi^2\phi^2 \right) - \sqrt{\left( \beta^2 - \Phi^4 + 2\beta\Phi^2 - 2\Phi^2\phi^2 \right)^2 - \frac{4\Phi^2 \beta |\beta - \phi^2|^3}{\sigma}},$$

and

$$x^* = \begin{cases} \bar{\theta} + \sigma + w^*(\Phi^2 - \beta^2 + \phi^2) & \text{(maximum threshold) if } \beta > \phi^2 \\ \bar{\theta} - \sigma + w^*(\Phi^2 + \phi^2 - \beta) & \text{(minimum threshold) if } \beta < \phi^2 \end{cases}.$$  
Furthermore, if the optimal solution in the $H$-mode is interior, then a larger $\sigma$ shifts the tradeoff between $A$-mode and $H$-mode in favor of $A$-mode, and shifts the tradeoff between $H$-mode and $P$-mode in favor of $H$-mode, i.e.

$$\frac{d\Pi^A_*}{d\sigma} > \frac{d\Pi^H_*}{d\sigma} > \frac{d\Pi^P_*}{d\sigma}.$$
The second part of the proposition says that the larger the variance of the demand shock (holding its expectation constant), the more attractive the $A$-mode becomes relatively to the $H$-mode, and the more attractive the $H$-mode becomes relatively to the $P$-mode. This reflects that the $A$-mode fully leverages the agent’s private information about demand, the $H$-mode leverages it to an extent limited by the threshold constraint placed on price, and the $P$-mode does not leverage the agent’s information at all. In the absence of monitoring costs, this result implies that there exists a cutoff $\hat{\sigma}$, such that the $H$-mode is optimal for $\sigma \geq \hat{\sigma}$ and the $P$-mode is optimal for $\sigma \leq \hat{\sigma}$ (recall from Proposition 2 that the $A$-mode is dominated by the $H$-mode in this case).

We illustrate the above results with three figures, which show the roles of monitoring costs $K$, the variance of the demand shock (captured by $\sigma$), the importance of the principal’s investment ($\Phi$) and the importance of the agent’s investment ($\phi$). In all three figures, black indicates the region for which $P$-mode is optimal, dark gray the region for which $H$-mode is optimal, and light gray indicates the region for which $A$-mode is optimal.

Figure 1 illustrates the optimal choice of mode as a function of $(\sigma, K) \in [0, 0.5] \times [0, 0.015]$, with the other parameter values set at $\bar{v} = 2$, $\beta = 1$, $\phi = 0.75$ and $\Phi = 0.5$. Fixing $K > 0$, as $\sigma$ increases, the optimal mode shifts from $P$-mode to $H$-mode and then to $A$-mode (consistent with the second part of Proposition 6), or directly from $P$-mode to $A$-mode when $K$ is high enough (consistent with Proposition 1).

Figure 2 illustrates the optimal choice of mode as a function of $(\Phi, K) \in [0, 0.5] \times [0, 0.015]$, with the other parameter values set at $\bar{v} = 2$, $\beta = 1$, $\phi = 0.5$ and $\sigma = 0.25$. Fixing $K > 0$, as $\Phi$ increases, the optimal mode shifts from $A$-mode to $H$-mode and then to $P$-mode (consistent with the second
part of Proposition 5), or directly from $A$-mode to $P$-mode when $K$ is high enough.

Figure 2: Optimal mode as a function of $\Phi$ and $K$

Figure 3: Optimal mode as a function of $\phi$ and $K$

Figure 3 illustrates the optimal choice of mode as a function of $(\phi, K) \in [0.25, 1.75] \times [0, 0.015]$, with the other parameter values set at $\theta = 2$, $\beta = 1$, $\Phi = 0.25$ and $\sigma = 0.25$. As anticipated by
the discussion following Proposition 1, this figure illustrates that the effect of $\phi$ on the tradeoffs is ambiguous. In particular, fixing $K > 0$, when $\phi$ is small and increases, the optimal mode shifts from $P$-mode to $H$-mode and then to $A$-mode (or directly from $P$-mode to $A$-mode), whereas when $\phi$ is large and increases, the optimal mode shifts from $A$-mode to $H$-mode and then to $P$-mode (or directly from $A$-mode to $P$-mode). The reason is that when $\phi$ is low, an increase in $\phi$ reduces the agent’s pricing bias (recall the bias is proportional to $\beta - \phi^2$), so the principal prefers to delegate more. By contrast, when $\phi$ is already large, an increase in $\phi$ increases the agent’s pricing bias, so the principal prefers to delegate less.

Obviously, when $K$ is large enough, the $A$-mode (which does not require a monitoring cost) dominates the $H$-mode and the $P$-mode. An interesting feature illustrated in all three graphs is that as $K$ increases from zero, initially only the margin between the $H$-mode and the $A$-mode is affected, with the optimal choice of mode shifting from $H$-mode to the $A$-mode. Only after the $H$-mode is no longer optimal for any parameter values because $K$ has become sufficiently high, does the optimal choice of mode start shifting from the $P$-mode to the $A$-mode as $K$ increases. The reason is that the $H$-mode is intermediate between the two pure modes, so when the monitoring cost incurred in $P$-mode and $H$-mode starts increasing, the shift can initially only be between the $H$-mode and the $A$-mode. Only once $K$ is high enough that the intermediate $H$-mode option is no longer viable, does $K$ start affecting the margin between the $P$-mode and the $A$-mode. This has a clear empirical implication: we expect to observe only pure modes when monitoring costs are high and hybrid modes only to appear when monitoring costs are low enough.

6 Minimum requirements for costly decisions

Suppose now RPM is illegal, so the price $p$ is always chosen by the agent. Instead, we assume $q$ is transferable and contractible. Recall that the total profit net of fixed costs is

$$p(\theta - \beta p + \phi q + \Phi Q) - \frac{1}{2}q^2 - \frac{1}{2}Q^2.$$ 

Given the additive nature of demand, $\theta$ has no direct impact on the agent’s choice of $q$, so one may wonder why the principal would ever want to delegate the choice of $q$ to the agent. The reason is that the agent chooses $p$, which is positively impacted by $\theta$ and is a strategic complement to $q$. Thus, delegation of $q$ to the agent allows the principal to leverage the agent’s private information on the demand shock, which indirectly affects the agent’s choice of $q$. However, this benefit must be traded off against the inefficiency due to the fact that the agent sets $q$ based on a lower margin ($p - w$ rather than $p$), and so tends to underinvest in $q$ all other things equal.

Following the same steps as in Section 5, we first analyze the choice between $P$-mode (the principal chooses $q$ in its contract) and $A$-mode (the choice of $q$ is entirely delegated to the agent), before considering whether the principal instead prefers threshold delegation ($H$-mode).

For this section, we replace (2) and (3) with the following two assumptions, which ensure all
second-order conditions hold and all decision variables and profits are positive at equilibrium values:

\[(2\beta - \phi^2)\left(\frac{\beta}{2} + \Phi^2\right) - \Phi^4 > 0\]  

(19)

and

\[\theta_L > \max \left\{ \Phi^2 \left( \frac{\beta - \Phi^2}{2} \right) - \phi^2 \left( \frac{\beta}{2} + \Phi^2 \right), \frac{\Phi^2 (\beta - \Phi^2)}{(2\beta - \phi^2) \left(\frac{\beta}{2} + \Phi^2\right) - \Phi^4}, \frac{\Phi^2 (\beta - \Phi^2)}{(2\beta - \phi^2) \left(\frac{\beta}{2} + \Phi^2\right) + \frac{\beta}{2} \phi^2 - \Phi^4} \right\}.\]  

(20)

Note that the second assumption is satisfied whenever \(\Phi^2 > \beta\).

6.1 Full control vs. no control

Consider first the \(P\)-mode in which the principal sets \(q\). As in the RPM case, the principal uses the fixed fee \(F\) to extract the agent’s expected surplus, so the principal solves

\[
\max_{w,q} \mathbb{E} \left[ p(\theta - \beta p + \phi q + \Phi Q) - \frac{1}{2} q^2 - \frac{1}{2} Q^2 \right]
\]

subject to

\[
p = \arg \max_{p'} \left\{ (p' - w) \left( \theta - \beta p' + \phi q + \Phi Q \right) - \frac{1}{2} q^2 \right\} = \frac{1}{2} w + \frac{1}{2\beta} \left( \theta + \phi q + w\Phi \right)
\]

\[
Q = \arg \max_{Q'} \left\{ \mathbb{E} \left[ w \left( \theta - \beta p + \phi q + \Phi Q' \right) \right] - \frac{1}{2} Q'^2 \right\} = w\Phi.
\]

Substituting \(p\) and \(Q\) back into the principal’s objective function and maximizing over \(q\) implies that, for a given \(w\), the principal’s optimal choice of \(q\) is \(q^P(w^*, \theta)\), where

\[q^P(w, \theta) = \frac{(\theta + w\Phi^2) \phi}{2\beta - \phi^2}.\]

Thus, \(q^P(w, \theta)\) is the hypothetical level of \(q\) that the principal would choose for a given \(w\) if it were able to observe \(\theta\). Plugging the expression of \(q^P(w^*, \theta)\) back into the principal’s objective function and optimizing over \(w\), we obtain that the optimal wholesale price \(w\) in \(P\)-mode is

\[w^P = \frac{\Phi^2 \theta}{(2\beta - \phi^2) \left(\frac{\beta}{2} + \Phi^2\right) - \Phi^4},\]

which is positive given (19). Furthermore, (19) ensures that the agent’s margin is always positive when the principal sets \(w = w^P\), i.e. \(P^P(w^P, \theta) > w^P\) for all \(\theta\). The principal’s resulting profits are

\[\Pi^P = \frac{V\theta}{4\beta} + \frac{\left(\frac{\beta}{2} + \Phi^2\right) \theta^2}{2 \left(2\beta - \phi^2\right) \left(\frac{\beta}{2} + \Phi^2\right) - \Phi^4}.\]  

(21)
Consider now the A-mode. The principal solves

$$\max_w \left\{ E \left[ p (\theta - \beta p + \phi q + \Phi Q) - \frac{1}{2} q^2 - \frac{1}{2} Q^2 \right] \right\}$$

subject to

$$(p, q) = \arg \max_{p', q'} \left\{ (p' - w) (\theta - \beta p' + \phi q' + \Phi Q) - \frac{1}{2} q'^2 \right\}$$

$$Q = \arg \max_{Q'} \left\{ E \left[ w (\theta - \beta p + \phi q + \Phi Q') \right] - \frac{1}{2} Q'^2 \right\}.$$

Solving the three constraints in $(p, q, Q)$ implies

$$p^A (w, \theta) = \frac{\theta + w (\beta + \Phi^2 - \phi^2)}{2\beta - \phi^2}$$

$$q^A (w, \theta) = \frac{(\theta + w (\Phi^2 - \beta)) \phi}{2\beta - \phi^2}$$

$$Q^A (w) = w \Phi.$$

Note $q^A (w, \theta) < q^P (w, \theta)$ for all $w$ and $\theta$, which means the agent always has a downward bias in setting $q$ (unlike the case with price as the transferable decision analyzed in Section 5). This reflects that (i) $q$ is a costly investment and (ii) the agent only gets a fraction of the revenue if $w > 0$ (when the principal sets $q$ in the contract, it takes into account the full revenue created by this investment).

Plugging the expressions of $p^A (w, \theta)$, $q^A (w, \theta)$ and $Q^A (w)$ back into the principal’s objective function, we obtain that the optimal wholesale price set by the principal in A-mode is

$$w^{A^*} = \frac{\Phi^4 \theta}{(2\beta - \phi^2) \left( \frac{\beta}{2} + \Phi^2 \right) + \frac{\beta}{2} \phi^2 - \Phi^4}.$$

It is easily verified that (19) implies $w^{A^*} > 0$, while (20) ensures $p^A (w^{A^*}, \theta) > w^{A^*}$ and $q^A (w^{A^*}, \theta) > 0$ for all $\theta$. Note also that $0 < w^{A^*} < w^{P^*}$.

The principal’s optimal A-mode profit is then

$$\Pi^{A^*} = \frac{V\theta}{2 (2\beta - \phi^2)} + \frac{\Phi^2 (2\beta - \phi^2) + \beta^2}{2 (2\beta - \phi^2) \left( (2\beta - \phi^2) \left( \frac{\beta}{2} + \Phi^2 \right) + \frac{\beta}{2} \phi^2 - \Phi^4 \right)}.$$

Comparing $\Pi^{A^*}$ with $\Pi^{P^*}$, we obtain the following proposition.

**Proposition 7** The principal’s profit is higher in A-mode compared to P-mode if and only if the variance of the agent’s private information on demand is sufficiently large, i.e.

$$\frac{V\theta}{\theta^2} > \frac{2\Phi^4 \beta^2}{((2\beta - \phi^2) (\beta + 2\Phi^2) - 2\Phi^4) (\Phi^2 (2\beta - \phi^2) + \beta^2 - \Phi^4)}.$$

The only meaningful difference relative to the case when price was the transferable action (Propo-
sition 1) is that now, the effect of an increase in $\phi$ on the tradeoff between the two pure modes is unambiguous. Indeed, the right hand side of (23) is increasing in $\phi$, so when the transferable investment $q$ becomes more important in determining demand, the tradeoff shifts in favor of the $P$-mode. This is because the agent’s choice of $q$ is always downward biased and the magnitude of the bias is increasing in $\phi$.

6.2 Threshold delegation

Following the same logic as in section 5, we now consider the $H$-mode in which, given the agent’s downward bias in setting $q$, the principal places a minimum threshold $x$ on the agent’s choice of $q$. Given $w$ and a minimum threshold $x$, the agent chooses

$$ q = \begin{cases} q^A(w, \theta) = \frac{(\theta + w(\Phi^2 - \beta)) \phi}{2(\beta - \phi^2)} - \frac{1}{2} q^2 - \frac{1}{2} Q^2 & \text{if } \theta \geq \frac{2\beta - \phi^2}{\phi} x + w(\beta - \Phi^2) \\ x & \text{if } \theta \leq \frac{2\beta - \phi^2}{\phi} x + w(\beta - \Phi^2). \end{cases} $$

The principal extracts the agent’s entire expected payoff, so the principal’s profit is

$$ \Pi^H = \max_{w, x} \left\{ \mathbb{E} \left[ p(\theta - \beta p + \phi q + \Phi Q) - \frac{1}{2} q^2 - \frac{1}{2} Q^2 \right] \right\} $$

subject to

$$ p = \arg\max_{p'} \left\{ \left( p' - w \right) \left( \theta - \beta p' + \phi q + \Phi Q \right) - \frac{1}{2} q^2 \right\} = \frac{1}{2} w + \frac{1}{2} \beta (\theta + \phi q + w \Phi^2) $$

$$ q = \max \left\{ q^A(w, \theta) , x \right\} $$

$$ Q = \arg\max_{Q'} \left\{ \mathbb{E} \left[ w \left( \theta - \beta p + \phi q + \Phi Q' \right) - \frac{1}{2} Q'^2 \right] \right\} = w \Phi. $$

We then obtain the corresponding version of Proposition 2 from the previous section. (The proof is very similar to that of Proposition 2, so we have relegated it to the online appendix.)

**Proposition 8** The $H$-mode with minimum requirement on $q$ dominates the $A$-mode. If in addition $\frac{\theta \mu}{\theta} > 1 + \frac{2\beta^2 \beta}{(\beta - \phi^2)(\beta + 2\Phi^2) - 2\Phi^2}$, the $H$-mode also dominates the $P$-mode.

The logic of this result is similar to that for the downward bias case when the transferable action was price. The use of a minimum constraint on the agent’s costly investment (rather than imposing a specific level) allows the principal to give the agent discretion to react to its private information about demand shocks, while eliminating the agent’s worst under-investment scenarios which occur when the demand shock is particularly low. By setting a minimum constraint above the lowest possible level of $q$ that the agent would choose, the principal prevents the agent from choosing investment levels that could never be optimal from the principal’s perspective even if it could observe $\theta$. This necessarily does better than giving the agent full discretion, i.e. the $A$-mode. And the principal can do strictly better than in $P$-mode by giving the agent freedom to choose any investment level above what the
principal would like to choose in $P$-mode (but not below), provided there are realizations of $\theta$ for which the agent would want to do so.

We can also obtain the corresponding results to Proposition 4, following an almost identical proof.

**Proposition 9** The optimal wholesale price extracted by the principal is highest in $P$-mode, lowest in $A$-mode, and intermediate in $H$-mode, i.e. $0 < w^A < w^H < w^P$. Moreover, the optimal minimum requirement for $q$ in $H$-mode is below the fixed choice of $q$ in $P$-mode, i.e. $x^* < q^P (w^P, \theta)$.

The logic of Proposition 9 is essentially the same as that in Proposition 4 for the the case of a downward bias.

7 Extensions

In this section we show that our main results are robust to several extensions: (i) general demand and cost functions, (ii) the principal uses revenue sharing instead of a per-unit wholesale price, and (iii) alternative sources of private information. The proofs of all results in this section are contained in an online appendix. For concision, we focus on the case in which RPM is allowed. The analysis for the case when RPM is not possible and $q$ is the relevant transferable action is similar.

7.1 General demand and cost functions

In this section we work with general demand $D (\theta, p, q, Q)$ and fixed costs $c (q)$ and $C (Q)$. We assume $D (\theta, p, q, Q)$ is increasing in $\theta$, $q$ and $Q$, and decreasing in $p$, while $c (q)$ and $C (Q)$ are increasing and convex. We can work out conditions under which Proposition 2 and Proposition 8 continue to hold.

If there is a demand interaction effect between $Q$ and either $p$ or $q$, then the principal’s choice of $Q$ in $H$-mode depends on the full distribution of the agent’s choice of the transferable decision ($p$ or $q$) implied by the distribution of $\theta$ and the principal’s threshold contract. This creates a technical problem in generalizing the results from Propositions 2 and Proposition 8. There are two alternative ways to avoid this problem: (i) assume demand is additively separable in $Q$ (this was the case in the previous sections), or (ii) assume that in stage two of the $A$-mode and the $H$-mode, the transferable action ($p$ or $q$) is chosen by the agent before the two non-transferable decisions are made.

Since in this section we want to show results hold with general demand and cost functions, in what follows, we focus on the second solution. Start with the case in which RPM is allowed, so $p$ is transferable and the non-contractible decisions $q$ and $Q$ are always chosen in the second stage after $p$ is determined. Denote by $(q (p, w, \theta), Q (p, w))$ the joint solutions to

$$
\left\{ \begin{array}{l}
q = \text{arg max}_{q'} \{ (p - w) D (\theta, p, q', Q) - c (q') \} \\
Q = \text{arg max}_{Q'} [\mathbb{E}_{\theta} [w D (\theta, p, q, Q') - C (Q')]]
\end{array} \right.
$$

Thus, $(q (p, w, \theta), Q (p, w))$ is the Nash equilibrium of the game in which the agent and the principal
simultaneously choose \( q \) and \( Q \), given \( p \) and \( w \). Let then

\[
p^P (w, \theta) \equiv \arg \max_{p'} \{p' D (\theta, p', q (p', w, \theta), Q (p', w)) - c (q (p', w, \theta)) - C (Q (p', w))\}
\]

\[
p^A (w, \theta) \equiv \arg \max_{p'} \{(p' - w) D (\theta, p', q (p', w, \theta), Q (p', w)) - c (q (p', w, \theta))\}.
\]

Thus, \( p^P (w, \theta) \) is the hypothetical price that the principal would like to choose for a given \( w \) if it could observe \( \theta \) in stage 1 when setting \( p \), but not in stage 2 when setting \( Q \). Meanwhile, \( p^A (w, \theta) \) is the price chosen by the agent in \( A \)-mode, given \( w \) and \( \theta \).

We assume that \( p^P (w, \theta) \) and \( p^A (w, \theta) \) are increasing in \( \theta \) (consistent with the random term \( \theta \) being a positive demand shock) and that the principal’s profit as a function of \( p \),

\[
pD (\theta, p, q (p, w, \theta), Q (p, w)) - c (q (p, w, \theta)) - C (Q (p, w)),
\]

is concave in \( p \) for any \((w, \theta)\). Denoting by \( p^P^* \) and \( w^P^* \) the principal’s optimal choices of price and wholesale price in \( P \)-mode, and by \( w^A^* \) the principal’s optimal choice of wholesale price in \( A \)-mode, we obtain the following result.

**Proposition 10** (Resale price maintenance)

**Maximum RPM:** If \( p^A (w^A^*, \theta) > p^P (w^A^*, \theta) \) for all \( \theta \) (i.e. the agent has an upward bias when the principal charges \( w^A^* \)), then the \( H \)-mode with maximum RPM dominates the \( A \)-mode. If \( p^A (w^P^*, \theta) > p^P (w^P^*, \theta) \) for all \( \theta \) (i.e. the agent has an upward bias when the principal charges \( w^P^* \)) and \( p^A (w^P^*, \theta_L) < p^P^* \), then the \( H \)-mode with maximum RPM dominates the \( P \)-mode.

**Minimum RPM:** If \( p^A (w^A^*, \theta) < p^P (w^A^*, \theta) \) for all \( \theta \) (i.e. the agent has a downward bias when the principal charges \( w^A^* \)), the \( H \)-mode with minimum RPM dominates the \( A \)-mode. If \( p^A (w^P^*, \theta) < p^P (w^P^*, \theta) \) for all \( \theta \) (i.e. the agent has a downward bias when the principal charges \( w^P^* \)) and \( p^A (w^P^*, \theta_H) > p^P^* \), the \( H \)-mode with minimum RPM dominates the \( P \)-mode.

Thus, for threshold delegation (minimum or maximum) to dominate both full control and full delegation, the agent’s pricing bias must be in the same direction (downward or upward) when \( w = w^A^* \) and \( w = w^P^* \). When demand is additively separable in \( Q \), the principal’s choice of \( Q \) does not depend on \( p \) and so the timing of the choice of \( p \) does not affect the outcome. Thus, Proposition 10 applies to the model with linear demand and quadratic costs used in Section 5, in which the price \( p \) was assumed to be chosen at the same time as the non-transferable actions \( q \) and \( Q \) in \( A \)-mode or \( H \)-mode. Furthermore, we have shown that in that model, the agent’s bias was always in the same direction for any \( w \).

### 7.2 Revenue sharing and other contractual instruments

Suppose that the principal’s two-part tariff involves a revenue share \( t \) instead of the wholesale price \( w \), in addition to the fixed fee \( F \). In this case, the variable revenue extracted by the principal becomes
tpD (θ, p, q, Q), while the agent receives \((1 - t) pD (θ, p, q, Q)\). We wish to explore whether threshold delegation continues to dominate full control and full delegation.

Assuming the general demand and cost functions, as well as the alternative timing from Section 7.1 (i.e. that in A-mode and H-mode, the agent chooses \(p\) before the principal and the agent simultaneously choose \(Q\) and \(q\)), it is easily shown that Proposition 10 continues to hold (after replacing \(w\) with \(t\)). The only difference arises in the specific conditions determining whether the bias is in the upward or downward direction and in the additional condition on the distribution of \(θ\) required for the H-mode to dominate the P-mode.

To illustrate, consider the model with linear demand and quadratic costs used in the previous sections, i.e. \(D (θ, p, q, Q) = θ - βp + φq + ΦQ\), \(c (q) = \frac{1}{2}q^2\) and \(C (Q) = \frac{1}{2}Q^2\). To ensure all second order conditions hold, assume \(2β > φ^2\).

We then obtain the following proposition.

**Proposition 11 (Resale price maintenance with revenue sharing)**

**Maximum RPM:** If \(Φ^2 > φ^2\), the H-mode with maximum RPM dominates the A-mode. If in addition \(\frac{q_L}{β} < \frac{2β(Φ^2 + φ^2) - φ^2 - 2Φ^2}{2β(Φ^2 + φ^2) - φ^2 - 2Φ^2}\), then the H-mode with maximum RPM also dominates the P-mode.

**Minimum RPM:** If \(Φ^2 < φ^2\), the H-mode with minimum RPM dominates the A-mode. If in addition \(\frac{q_H}{β} > \frac{2β(Φ^2 + φ^2) - φ^2 - 2Φ^2}{2β(Φ^2 + φ^2) - φ^2 - 2Φ^2}\), the H-mode with minimum RPM also dominates the P-mode.

Whether threshold delegation should involve minimum or maximum RPM is determined by whether \(φ^2 < Φ^2\) or \(Φ^2 > φ^2\). In the model with wholesale prices, the corresponding conditions were \(β < φ^2\) and \(β > φ^2\). There are two main differences relative to the case with wholesale prices. First, here the revenue share \(t\) does not directly (i.e. for given \(q\) and \(Q\)) distort the agent’s choice of price \(p\), which is why the result no longer depends on the level of \(β\). Put differently, there is no longer a double marginalization problem when the agent sets price. Second, here the choice of price influences the level of \(Q\) chosen by the principal in the second stage, which is proportional to \(p\). By contrast, with wholesale pricing, \(p\) had no impact on \(Q\).

As a consequence, the agent’s bias in setting price induced by revenue sharing is determined by two effects, which work in opposite directions. On the one hand, the agent tends to set price too low from the principal’s perspective, reflecting that the agent only receives revenue \((1 - t) φpq\) from his investment \(q\) instead of the full amount \(φpq\). This leads the agent to under-invest in \(q\), and since \(q\) and \(p\) are strategic complements, it will also set \(p\) too low. This effect is the same as under wholesale pricing. On the other hand, when the agent sets price, it does not take into account the costs incurred by the principal’s investment \(Q\). Given that the choice of \(Q\) is proportional to \(p\), this leads the agent to set price too high compared to what the principal prefers.

The net result of these two opposing effects turns out to depend on which type of investment is more important. If the agent’s investment \(q\) is more important (\(Φ^2 < φ^2\)), then, without any contract
restriction, the agent will tend to set prices too low and so minimum RPM is required. Once again, this effect is consistent with the traditional argument for minimum RPM to promote the agent’s investment in service. On the other hand, we obtain here the entirely new prediction that if the principal’s investment is more important \( (\Phi^2 > \phi^2) \), the agent will tend to set prices too high and so maximum RPM is required.

More generally, threshold delegation would continue to have a role to play even if the principal could use other, frequently observed contractual instruments (e.g. slotting fees, quantity discounts, buy-backs, or price-dependent royalties) instead of—or in addition to—wholesale prices and revenue shares. This includes the possibility that the principal can make the payment conditional on price (e.g. price-dependent royalties). The difference would be in the conditions determining whether the agent has a bias and the sign of that bias at the optimal values of the contractual instruments for the A-mode and the P-mode, as well as in the conditions on the distribution of the demand shock required for the H-mode to dominate the P-mode. Thus, although other contractual instruments might reduce the likelihood of existence and the magnitude of a bias, as long as such instruments cannot achieve complete revelation of the agent’s private information about demand, there can still be scope for threshold delegation to improve over 100% delegation and 100% control. Finally, a virtue of threshold delegation relative to some other more sophisticated contractual instruments is its simplicity—the principal only has to confirm that the level of the transferable decision variable satisfies a single threshold.

7.3 **Alternative sources of private information**

We have assumed throughout that the agent’s private information pertained to a demand shock \( \theta \). In practice, there can be other sources for the agent’s private information, e.g. private benefits or costs, the effect of the agent’s investment on demand, and so on. In this section we briefly explain how our model can be applied to the case of private benefits or costs. (The analysis in which the source of private information is the effect of the agent’s investment on demand was analyzed in an earlier version of the paper, available from the authors upon request.)

Suppose demand is deterministic and given by \( D(p, q, Q) \). Costs are once again \( c(q) \) and \( C(Q) \). However, suppose the agent derives a private benefit \( b \) per unit of demand, where \( b \) is distributed with mean \( \bar{b} \) and variance \( V_b > 0 \). We allow \( \bar{b} \) to be positive or negative, so \( b \) can also be a private cost. Thus, given a two-part tariff \((w, F)\), the principal extracts

\[
wD(p, q, Q) - C(Q) + F,
\]

whereas the agent receives

\[
(p + b - w)D(p, q, Q) - c(Q) - F.
\]

All other assumptions are unchanged. In particular, \( b \) is only observed by the agent in the second stage.

A private benefit reflects that when the agent has a higher level of sales through its contract
with the principal, the agent may also have increased opportunities to sell complementary services or products to the same customers, but these opportunities fall outside the scope of the contractual relationship with the principal. For example, movie theaters can sell more popcorn and soda when a movie attracts more viewers, but do not share the resulting revenues with the movie producer. Alternatively, the agent may incur private costs proportional to demand that are not accounted for in the contract with the principal. Examples of such private costs include the opportunity costs of franchisees, dealers and retailers, Uber drivers, etc.

It is then easily seen that, assuming either additive separability of \( D(p, q, Q) \) in \( Q \) or the alternative timing in stage two discussed in Section 7.1, the set-up and analysis from Section 7.1 goes through almost unchanged. In particular, if \( D(p, q, Q) = \theta - \beta p + \phi q + \Phi Q \), where \( \theta > 0 \) is now a constant, \( c(q) = \frac{1}{2} q^2 \) and \( C(Q) = \frac{1}{2} Q^2 \), then the condition determining whether the principal should employ maximum or minimum RPM is once again \( \beta > \phi^2 \). Furthermore, the entire discussion in Section 7.2 regarding the robustness to additional pricing instruments applies unchanged to this model.

8 Concluding remarks

The emergence of digital monitoring and data analytics technologies has created more opportunities for firms to enforce different degrees of delegation in a cost effective way. As a result, partial delegation is likely to become a contractual instrument that a greater number of firms that act as principals (e.g. franchisors, platforms, manufacturers, movie studios) can consider using when setting the terms for their agents (e.g. franchisees, third-party suppliers, retailers, movie theaters). The recent shift towards the legality of resale price maintenance (RPM) also means that placing restrictions on price setting is becoming increasing relevant. Our theory provides several lessons for managers in this regard.

At a high level, we have shown that delegation subject to minimum requirements strikes a middle ground between complete delegation and full control and oftentimes does better than both. It is a way to get the best of both worlds, by leveraging the agent’s private information, while also eliminating the more extreme biases that arise when an agent only keeps some of the revenues it produces. For platforms, one can view the use of minimum requirements as a governance rule designed to achieve strategic positions that are intermediate along the spectrum between pure platform (e.g. relying on independent contractors) and pure vertical integration (e.g. relying on employees). In manufacturer-retailer contexts, threshold delegation is an additional contracting instrument that can improve channel coordination beyond what can be achieved with typical pricing instruments (wholesale pricing, revenue shares, quantity discounts, etc).

More specifically, when the principal charges wholesale prices, it should use RPM with price ceilings (respectively, price floors) whenever the double marginalization concern is more (respectively, less) important relative to the concern that the agent under-invests in non-transferable actions. When the principal extracts a revenue share, it should use RPM with price ceilings (respectively, price floors) whenever the principal’s moral hazard is more (respectively, less) important than the agent’s.

In settings where there is more uncertainty regarding the agent’s private information (i.e. higher
variance of private shocks), the principal should give the agent more autonomy, i.e. switch from full control to partial delegation, reduce minimum or increase maximum requirements if they are already in place, or even switch from partial delegation to full delegation to the agent. When the principal’s moral hazard hazard becomes more important, the principal should delegate less (switch to partial delegation from full delegation, increase minimum or decrease maximum requirements if they are already in place, or even switch from partial delegation to full control). Finally, when the agent’s moral hazard is already large relative to the price sensitivity of demand, an increase in the agent’s moral hazard increases the agent’s upward bias in setting price, so the principal prefers to delegate less. The opposite is the case when the agent’s moral hazard is small relative to the price sensitivity of demand—an increase in the agent’s moral hazard decreases the agent’s downward bias in setting price, so the principal prefers to delegate more. For delegating costly actions, the principal should follow the same advice as for delegating price in the case it faces a downward bias.

There are several promising directions in which our analysis can be extended. Using the existing framework it would be quite straightforward to introduce multiple agents and (positive or negative) spillovers from each agent’s choice of the transferable action on the revenues derived by other agents. The interesting question would then be to determine whether spillovers exacerbate or offset the agent’s bias, and therefore how spillovers affect the principal’s delegation decision. Another direction worth exploring would be to introduce risk aversion or wealth constraints for agents, so that they cannot pay large fixed fees upfront. This should increase the principal’s wholesale price in all three modes, and so the degree of bias, thereby shifting the tradeoff in favor of less delegation. Finally, it would also be interesting (but challenging) to extend our model to allow for multiple competing principals. This could possibly generate equilibria in which principals compete with different delegation models.

References


9 Appendix

9.1 Proof of Proposition 2

We start with the case $\beta > \phi^2$. The principal’s $H$-mode profit is

$$\Pi^H(w, x) = \int_{\theta_L}^{\theta_H} (2\beta-\phi^2) x - w(\Phi^2-\phi^2+\beta) \left( p^A (w, \theta) (\theta - \beta p^A (w, \theta) + w\Phi^2) + \frac{\phi^2}{2} p^A (w, \theta)^2 \right) dG(\theta)$$

$$+ \int_{(2\beta-\phi^2) x - w(\Phi^2-\phi^2+\beta)}^{\theta_H} x (\theta - \beta x + w\Phi^2) + \frac{\phi^2}{2} x^2 dG(\theta) - \frac{\phi^2 + \phi^2}{2} w^2.$$

We first compare the $A$-mode to the $H$-mode with $w = w^A*$ and

$$x = p^A (w^A*, \theta_H - \kappa) = \frac{\theta_H - \kappa + w^A* (\Phi^2 - \phi^2 + \beta)}{2\beta - \phi^2},$$

where $\kappa > 0$ is small. We can then write

$$\Pi^H (w^A*, x = p^A (w^A*, \theta_H - \kappa)) - \Pi^{A*} = \Pi^H (w^A*, x = p^A (w^A*, \theta_H - \kappa)) - \Pi^H (w^A*, x = p^A (w^A*, \theta_H)).$$
Since the principal fixes these assumptions, we just have to show threshold delegation is optimal for any given \( w \).

Next, we compare the \( P \)-mode to the \( H \)-mode with \( w = w^{p*} = \frac{\overline{\beta} \Phi^2}{(2 \beta - \phi^2)(\phi^2 + \Phi^2) - \phi^2} \) and \( x = p^{p*} = \frac{\Phi + \beta \Phi^4}{2 \beta - \phi^2} \).

Note that this \((w^H, x)\) is strictly interior because

\[
\theta_L < (2 \beta - \phi^2) p^{p*} - w^{p*} (\Phi^2 - \phi^2 + \beta) = \overline{\theta} - w^{p*} (\beta - \phi^2) < \theta_H,
\]

where the first inequality follows from the assumption \( \theta_L < \overline{\theta} \left( 1 - \frac{\Phi^2 (\beta - \phi^2)}{(2 \beta - \phi^2)(\phi^2 + \Phi^2) - \phi^2} \right) \) and the second inequality from \( \beta > \phi^2 \). Then, using (4) we obtain

\[
\Pi^H (w^{p*}, x = p^{p*}) - \Pi^{P*} = \Pi^H (w^{p*}, p^{p*}) - \Pi^P (w^{p*}, p^{p*})
\]

\[
= \int_{\theta_L}^{(2 \beta - \phi^2)p^{p*} - w^{p*}(\Phi^2 - \phi^2 + \beta)} \left( p^A (w^{p*}, \theta) \left( \theta - \beta p^A (w^{p*}, \theta) + w^{p*} \Phi^2 \right) + \frac{\phi^2}{2} p^A (w^{p*}, \theta)^2 \right) dG (\theta)
\]

\[
= \int_{\theta_L}^{(2 \beta - \phi^2)p^{p*} - w^{p*}(\Phi^2 - \phi^2 + \beta)} \left( p^A (w^{p*}, \theta) - \beta p^A (w^{p*}, \theta) + w^{p*} \Phi^2 \right) - \left( p^{p*} (\theta - \beta p^{p*} + w^{p*} \Phi^2) + \frac{\phi^2}{2} (p^{p*})^2 \right) \right) dG (\theta)
\]

\[
= \int_{\theta_L}^{2 \beta - \phi^2} \left( \frac{\phi^2 (\beta - \phi^2)}{(2 \beta - \phi^2)(\phi^2 + \Phi^2) - \phi^2} \right) dG (\theta) > 0,
\]

because \( \beta > \phi^2 \) in this case. This implies \( H \)-mode strictly dominates \( P \)-mode.

Turning now to the case \( \beta < \phi^2 \), we can use the same steps as above to show that \( \Pi^H (w^{A*}, x = p^{A*} (w^{A*}, \theta_L + \kappa)) > \Pi^{A*} \) for \( \kappa \) sufficiently small, which implies that \( H \)-mode dominates \( A \)-mode. Similarly, we can show that \( \Pi^H (w^{p*}, x = p^{p*}) > \Pi^{p*} \). In this case, \((w^H = w^{p*}, x = p^{p*})\) is strictly interior because

\[
\theta_L < (2 \beta - \phi^2) p^{p*} - w^{p*} (\Phi^2 - \phi^2 + \beta) = \overline{\theta} + w^{p*} (\phi^2 - \beta) < \theta_H
\]

due to the assumptions \( \phi^2 > \beta \) and \( \theta_H > \overline{\theta} \left( 1 + \frac{\phi^2 (\phi^2 - \beta)}{(2 \beta - \phi^2)(\phi^2 + \Phi^2) - \phi^2} \right) \). Thus, \( H \)-mode dominates \( P \)-mode under these assumptions.

9.2 Proof of Proposition 3

Since the principal fixes \( w \) in its contract at the same time as deciding on the type and nature of any delegation, we just have to show threshold delegation is optimal for any given \( w \). To do so, we will show that any contract
that differs from threshold delegation can be improved upon by a contract with the same \( w \) and threshold delegation.

The principal’s delegation problem for a fixed choice of \( w \) is

\[
\max_D \left\{ \mathbb{E} \left[ p \left( \theta - \beta p + \phi q + \Phi Q \right) - \frac{1}{2} q^2 - \frac{1}{2} Q^2 \right] \right\}
\]

subject to

\[
(p, q) = \arg \max_{p' \in D, q} \left\{ (p' - w) \left( \theta - \beta p' + \phi q + \Phi Q \right) - \frac{1}{2} q^2 \right\}
\]

\[
Q = w \Phi,
\]

where \( D \) denotes the delegation set to which the principal restricts the agent’s choice of \( p \). This can be rewritten as

\[
\max_D \left\{ \mathbb{E} \left[ p \left( \theta - \beta p + (p - w) \phi^2 + w \Phi^2 \right) - \frac{1}{2} (p - w)^2 \phi^2 - \frac{1}{2} w^2 \Phi^2 \right] \right\}
\]

subject to

\[
p = \arg \max_{p' \in D} \left\{ (p' - w) \left( \theta - \beta p' + (p' - w) \phi^2 + w \Phi^2 \right) - \frac{1}{2} (p' - w)^2 \phi^2 \right\}.
\]

Ignoring terms that do not depend on \( p \), the program that defines the principal’s optimal delegation set \( D(w) \) can be re-written more simply as

\[
\max_D \mathbb{E} \left[ -\alpha_0 p^2 + (\theta + w \Phi^2) p \right]
\]

subject to

\[
p = \arg \max_{p' \in D} \left\{ -\alpha_0 (p')^2 + (\theta + w \Phi^2 - \alpha_1) p' \right\},
\]

where \( \alpha_0 = \beta - \frac{2^2}{T} > 0 \) due to our assumption that \( \left( 2 \beta - \phi^2 \right) (\Phi^2 + \phi^2) - \Phi^4 > 0 \) and \( \alpha_1 = w \left( \phi^2 - \beta \right) \).

In this model, for a given \( \theta \), the ideal choice of \( p \) for the principal is \( p = \frac{\theta + w \Phi^2}{2 \alpha_0} \), while for the agent it is \( p = \frac{\theta + w \Phi^2 - \alpha_1}{2 \alpha_0} \). Note that \( \alpha_1 \) is positive or negative depending on whether the agent has a downward or upward bias in setting price. We will first consider the case in which \( \alpha_1 \geq 0 \), so that without any restrictions, the agent prefers a lower \( p \) than the principal. The principal is therefore interested in restricting the agent from setting \( p \) too low. The question remains whether the principal can do better by requiring the agent to choose from some specific values of \( p \) or some disjoint intervals that exclude some high values of \( p \). Formally, we want to show that the delegation set \( D(w) \) is a threshold interval, i.e. \( D(w) = \{ p \geq x(w) \} \) for some \( x(w) \).

Suppose first the principal restricts the agent to choose \( p \) from some subset of \( p \leq p_0 \equiv \frac{\theta_0 + w \Phi^2 - \alpha_1}{2 \alpha_0} \) which includes \( p = p_0 \), where \( \theta_0 < \theta_H \). This covers the possibility that the agent can only choose \( p = p_0 \) or can choose any \( p \leq p_0 \). In this case, when \( \theta \in [\theta_0, \theta_H] \), it is easily seen that the agent chooses \( p = p_0 \) because the agent’s objective function is increasing in \( p \) for all \( p \leq p_0 \). But the principal could strictly improve expected profits by adding the range \( p \geq p_0 \) to the set of permissible choices of \( p \) by the agent. To see this, note that the only change comes from the different choices of \( p \) by the agent when \( \theta \in [\theta_0, \theta_H] \). The change in expected profits is

\[
\frac{1}{4 \alpha_0} \int_{\theta_0}^{\theta_H} \left( \theta - \theta_0 \right) \left( (\theta - \theta_0) + 2 \alpha_1 \right) dG(\theta) > 0.
\]
Suppose now that the agent is allowed to choose from some set that does not include \( p \in (p_0, p_1) \), where \( p_0 \equiv \frac{\theta_0 + w \Phi^2 - \alpha_1}{2 \alpha_0} \) and \( p_1 \equiv \frac{\theta_1 + w \Phi^2 - \alpha_1}{2 \alpha_0} \) for some \( \theta_L \leq \theta_0 < \theta_1 \leq \theta_H \). In this case, since the agent’s objective function is quadratic in \( p \), if the agent’s draw of \( \theta \) is in the range \([\theta_0, \theta_1]\), then the agent chooses \( p = p_0 \) when \( \theta \leq \frac{\theta_0 + \theta_1}{2} \) and \( p = p_1 \) when \( \theta > \frac{\theta_0 + \theta_1}{2} \). The principal can profitably deviate by adding the range \([p_0, p_1]\) to the set of permissible choices of \( p \). The change in profits is

\[
\frac{1}{4 \alpha_0} \left( \int_{\theta_0}^{\theta_1} (\theta - \theta_0) (\theta - \theta_0 + 2 \alpha_1) dG(\theta) + \int_{\theta_0}^{\theta_1} (\theta - \theta_1) (\theta - \theta_1 + 2 \alpha_1) dG(\theta) \right) = \frac{1}{4 \alpha_0} \left( \int_{\theta_0}^{\theta_1} (2 \alpha_1 (\theta - \theta_0) + (\theta - \theta_0)^2) g(\theta) d\theta + \int_{\theta_0}^{\theta_1} (2 \alpha_1 (\theta - \theta_1) + (\theta - \theta_1)^2) g(\theta) d\theta \right).
\]

Using integration by parts, we have

\[
\int_{\theta_0}^{\theta_1} (2 \alpha_1 (\theta - \theta_0) + (\theta - \theta_0)^2) g(\theta) d\theta = \alpha_1 (\theta_1 - \theta_0)^2 g \left( \frac{\theta_0 + \theta_1}{2} \right) + \int_{\theta_0}^{\theta_1} (g(\theta) - \alpha_1 g'(\theta)) d\theta
\]

\[
\int_{\theta_0}^{\theta_1} (2 \alpha_1 (\theta - \theta_1) + (\theta - \theta_1)^2) g(\theta) d\theta = -\alpha_1 (\theta_1 - \theta_0)^2 g \left( \frac{\theta_0 + \theta_1}{2} \right) + \int_{\theta_0}^{\theta_1} (g(\theta) - \alpha_1 g'(\theta)) d\theta.
\]

With these expressions, the profit change above becomes equal to

\[
\frac{1}{4 \alpha_0} \left( \int_{\theta_0}^{\theta_1} (\theta - \theta_0)^2 (g(\theta) - \alpha_1 g'(\theta)) d\theta + \int_{\theta_0}^{\theta_1} (\theta - \theta_1)^2 (g(\theta) - \alpha_1 g'(\theta)) d\theta \right).
\]

This expression is clearly positive under the assumption \( \alpha_1 g'(\theta) \leq g(\theta) \) for all \( \theta \).

Thus, we can conclude that the optimal range of admissible \( p \) for the agent must take the form of a threshold interval \( p \geq x \). Since \( \alpha_1 = w (\phi^2 - \beta) > 0 \), threshold delegation is optimal provided

\[
w (\phi^2 - \beta) g'(\theta) \leq g(\theta).
\]

(24)

Since \( w > 0 \) and \( g(\theta) > 0 \) for all \( \theta \in [\theta_L, \theta_H] \), a sufficient condition for (24) to hold is \( g'(\theta) \leq 0 \).

If instead \( \alpha_1 < 0 \), then the same proof applies, except that now the optimal range of admissible \( p \) for the agent must take the form of a threshold interval with a maximum requirement \( p \leq x \).

### 9.3 Proof of Proposition 4

Taking the respective first-order conditions in \( w \) of profit expressions (7) and (12), we obtain that \( w^{P^*} \) is the solution to \( h^P(w) = 0 \), where

\[
h^P(w) = \frac{(\bar{\theta} + \Phi^2 w) \Phi^2}{2 \beta - \phi^2} - (\Phi^2 + \phi^2) w,
\]

and \( w^{A^*} \) is the solution to \( h^A(w) = 0 \), where

\[
h^A(w) = \frac{(\bar{\theta} + \Phi^2 w) \Phi^2}{2 \beta - \phi^2} - w (\beta - \phi^2)^2 - (\Phi^2 + \phi^2) w.
\]
If $\beta > \phi^2$, then, from (15), the first-order conditions of $\Pi^H (w, x)$ in $x$ and $w$ are

$$
\int_{(2\beta - \phi^2)x - w(\beta + \Phi^2 - \phi^2)}^{\theta_L} \Phi^2 (\theta + w\Phi^2) - w (\beta - \phi^2)^2 dG (\theta) + \int_{(2\beta - \phi^2)x - w(\beta + \Phi^2 - \phi^2)}^{\theta_H} x\Phi^2 dG (\theta) - (\Phi^2 + \phi^2) w = 0.
$$

Combining these two equations, we obtain the $w^H$ is the solution to $h_{max}^H (w) = 0$, where

$$
h_{max}^H (w) \equiv \frac{(\bar{\theta} + w\Phi^2) \Phi^2}{2\beta - \phi^2} - \frac{w (\beta - \phi^2)^2}{2\beta - \phi^2} G ((2\beta - \phi^2) x - w (\beta + \Phi^2 - \phi^2)) - (\Phi^2 + \phi^2) w.
$$

Suppose now $\beta < \phi^2$. From (15), the first-order conditions of $\Pi^H (w, x)$ in $x$ and $w$ are

$$
\int_{(2\beta - \phi^2)x - w(\beta + \Phi^2 - \phi^2)}^{\theta_L} \Phi^2 (\theta + w\Phi^2) - w (\beta - \phi^2)^2 dG (\theta) + \int_{(2\beta - \phi^2)x - w(\beta + \Phi^2 - \phi^2)}^{\theta_H} x\Phi^2 dG (\theta) - (\Phi^2 + \phi^2) w = 0.
$$

Combining these two equations, we obtain that $w^H$ is the solution to $h_{min}^H (w) = 0$, where

$$
h_{min}^H (w) \equiv \frac{(\bar{\theta} + w\Phi^2) \Phi^2}{2\beta - \phi^2} - \frac{w (\beta - \phi^2)^2}{2\beta - \phi^2} (1 - G ((2\beta - \phi^2) x - w (\beta + \Phi^2 - \phi^2))) - (\Phi^2 + \phi^2) w.
$$

Comparing the expressions of $h^P (w)$, $h^A (w)$, $h_{max}^H (w)$ and $h_{min}^H (w)$ above, it is clear that

$$
h^P (w) > h_{max}^H (w) > h^A (w) \text{ and } h^P (w) > h_{min}^H (w) > h^A (w)
$$

for all $w$, which implies $w^{A*} < w^{H*} < w^{P*}$ both when $\beta > \phi^2$ and when $\beta < \phi^2$.

Suppose $\beta < \phi^2$. Then

$$
\frac{\partial \Pi^H (w^{H*}, x = p^P (w^{H*}, \bar{\theta}))}{\partial x} = \int_{\max \{\theta_L, \min \{\theta_H - (2\beta - \phi^2) p^P (w^{H*}, \bar{\theta}) - w^{H*} (\beta + \Phi^2 - \phi^2)\}\}}^{\theta_H} \left( \frac{\theta + w^{H*} \Phi^2}{- (2\beta - \phi^2) p^P (w^{H*}, \bar{\theta})} \right) dG (\theta)
$$

$$
= \int_{\max \{\theta_L, \min \{\theta_H - w^{H*} (\Phi^2 - \phi^2)\}\}}^{\theta_H} (\theta - \bar{\theta}) dG (\theta) \leq 0 = \frac{\partial \Pi^H (w^{H*}, x = x^*)}{\partial x}.
$$

Combined with $w^{H*} < w^{P*}$, this implies $x^* \leq p^P (w^{H*}, \bar{\theta}) < p^P (w^{P*}, \bar{\theta})$.

### 9.4 Proof of Proposition 5

Using expressions (7) and (12) and applying the envelope theorem, we obtain

$$
\frac{d\Pi^P}{d\Phi^2} = \frac{w^{P*} (\bar{\theta} + w^{P*} \Phi^2)}{2\beta - \phi^2} - \frac{(w^{P*})^2}{2}
$$

$$
\frac{d\Pi^A}{d\Phi^2} = \frac{w^{A*} (\bar{\theta} + w^{A*} \Phi^2)}{2\beta - \phi^2} - \frac{(w^{A*})^2}{2}
$$
Suppose first $\beta > \phi^2$. Then, using expression (15) and applying the envelope theorem, we have

$$\frac{d\Pi^{H*}}{d\Phi} = \int_{\theta_L}^{\theta_H} \left( (2\beta - \phi^2) x^* - w^{H*}(\Phi^2 - \phi^2 + \beta) \right) \frac{w^{H*} (\theta + w^{H*} \Phi^2)}{2\beta - \phi^2} dG(\theta) + \int_{(2\beta - \phi^2)x^* - w^{H*}(\Phi^2 - \phi^2 + \beta)}^{\theta} x^* w^{H*} dG(\theta) - \frac{(w^{H*})^2}{2}.$$

At the same time, the first-order condition of $\Pi^H(w, x)$ in $x$ (also obtained from (15)) implies

$$\int_{(2\beta - \phi^2)x^* - w^{H*}(\Phi^2 - \phi^2 + \beta)}^{\theta} (\theta + w^{H*} \Phi^2 - (2\beta - \phi^2) x^*) dG(\theta) = 0.$$

Using this equation to replace $x^*$ in the expression of $\frac{d\Pi^{H*}}{d\Phi}$ above, we obtain

$$\frac{d\Pi^{H*}}{d\Phi} = \frac{w^{H*} (\bar{\theta} + w^{H*} \Phi^2)}{2\beta - \phi^2} - \frac{(w^{H*})^2}{2}.$$

If $\beta < \phi^2$, a very similar calculation leads to the same expression

$$\frac{d\Pi^{H*}}{d\Phi} = \frac{w^{H*} (\bar{\theta} + w^{H*} \Phi^2)}{2\beta - \phi^2} - \frac{(w^{H*})^2}{2}.$$

Consider now the function

$$f(w) \equiv \frac{w (\bar{\theta} + w \Phi^2)}{2\beta - \phi^2} - \frac{w^2}{2}.$$

We have

$$f'(w) = \frac{\bar{\theta} + w (2\Phi^2 - 2\beta + \phi^2)}{2\beta - \phi^2}.$$

If $2\Phi^2 - 2\beta + \phi^2 \geq 0$, then $f(w)$ is increasing. Thus, since $w^{A*} < w^{H*} < w^{P*}$, we have

$$f(w^{A*}) < f(w^{H*}) < f(w^{P*}).$$

Suppose $2\Phi^2 - 2\beta + \phi^2 < 0$. Then $f'(w) \geq f'(w^{P*})$ for all $w \leq w^{P*}$ and

$$f'(w^{P*}) = \frac{\bar{\theta} ((2\beta - \phi^2) \phi^2 + \Phi^4)}{(2\beta - \phi^2)((2\beta - \phi^2)(\Phi^2 + \phi^2) - \Phi^4)} > 0,$$

so we still have

$$f(w^{A*}) < f(w^{H*}) < f(w^{P*}).$$

Thus, we can conclude that

$$\frac{d\Pi^{A*}}{d\Phi^2} < \frac{d\Pi^{H*}}{d\Phi^2} < \frac{d\Pi^{P*}}{d\Phi^2}.$$