CENTER FOR ECONOMIC STUDIES

OPTIMAL REGULATORY LAG UNDER PRICE CAP REGULATION

Mark Armstrong, Ray Rees
and John Vickers

Working Paper No. 4

UNIVERSITY OF MUNICH

CES

Working Paper Series

CES Working Paper No. 4
June 1991

OPTIMAL REGULATORY LAG UNDER PRICE CAP REGULATION

Abstract

We present a model of monopoly regulation in which the probability of cost being low rather than high follows a Markov process that depends on the firm's efforts. The regulator chooses price and regulatory lag (i.e. the length of time until the next price review) to maximize welfare subject to an expected break-even constraint for the firm. Between reviews the firm has a finite horizon cost minimization problem. We characterize its solution and the resulting dilemmas for the regulator. Starting from a high cost state, longer lag postpones the date at which price might be reduced, but lowers price in the interim and improves the chance that cost will be low at the next review. These pros and cons are reversed starting from a low cost state. With inelastic demand infinite lags are optimal. If costs are unresponsive to efforts, then minimal lags are best. Numerical simulations indicate the importance of the demand elasticity and effort responsiveness more generally. Finally, the desirability of non-constant prices between reviews is analyzed.

Ray Rees
Department of Economics
University of Guelph
Ontario, Canada
visiting CORE
Louvain-la-Neuve, Belgium

Mark Armstrong
John Vickers
Institute of Economics and Statistics
Oxford OX1 3UL
United Kingdom

1. Introduction

The use of price caps as a method of monopoly regulation is becoming increasingly common — see RAND Journal (1989). In Britain, the form of price cap regulation known as "RPI–X" was introduced for British Telecom when it was privatized in 1984, and has since been used for the regulation of the prices of gas, airport use, water supply, and electricity transmission and distribution.[1] In the United States, price caps have been recently introduced on long–distance telephone rates.[2]

There are at least three broad dimensions to the analysis of the possible merits of price cap regulation. There are multiproduct issues, concerning the relative balance of prices when an index of prices is kept below some ceiling.[3] There is the cost pass–through question concerning which of the firm's costs should be automatically reflected (and to what degree) in the prices faced by consumers. Finally, there is the question of the appropriate length of time between price reviews. This dynamic issue of optimal regulatory lag is the focus of this paper. Regulatory lag is often claimed to be the source of an important advantage of price cap regulation over traditional forms of rate–of–return regulation, namely that it gives regulated firms good incentives for cost reduction — see Littlechild (1983).[4] This advantage, if it exists, derives from the fact that regulatory lag is typically longer with price cap regulation than with rate–of–return regulation. Indeed, if

---

[1] RPI–X regulation requires an index of the prices of the regulated products to fall in real terms by X per cent per annum, i.e. to increase by no more than the rate of inflation of the retail price index (RPI) minus X per cent. (In cases where real price increases are allowed, the system is usually described as RPI+X regulation.)

[2] *Policy and Rules Concerning Rates for Dominant Carriers*, FCC 89–91, CC Docket No. 87–313, Report and Order and Second Further Notice of Proposed Rule Making (released April 17, 1989).

[3] In this case the precise form of the index is of crucial importance. See Bradley and Price (1988), Armstrong and Vickers (1990) and Bös (1990) for an analysis of this aspect of price cap regulation.

[4] The Littlechild Report emphasized not only the likely superiority of price caps in terms of allocative efficiency and incentives for cost reduction but also the reduced costs of operating the regulatory mechanism itself. The prospect of "regulation with a light hand" was very attractive to policy makers at the time of privatisation.

rate of return remains the basis for price review, then price cap regulation can be very similar to rate–of–return regulation with longer lags. With this perspective the central issue is: what is the optimal length of time between regulatory price reviews?

The idea that regulatory lag can stimulate cost efficiency is far from new. Baumol and Klevorick (1970, page 184) suggested that the period before a regulatory review can be regarded as

> "the time when the firm has the opportunity of earning a profit rate exceeding that specified by the [price] constraint. When the regulatory review occurs, this excess is eliminated by the regulator's adjustment of the prices the firm can charge".

Bailey (1974) drew the analogy between regulatory lag and patent protection to provide some formal analysis of the idea.[5] In both cases there is a tradeoff between allocative efficiency — price in line with cost — which is favoured by shorter lags, and cost efficiency, which is stimulated by longer lags.

In the model, which we present in section 2, there is a monopoly with known demand that is constant through time and with unit cost that follows a Markov process where the transition probabilities between high and low cost states depend on the firm's cost–reducing expenditures.[6] These expenditures are not observed by the regulator. At the time of each regulatory review there are two instruments available to the regulator — price and regulatory lag before the next review. The regulator is assumed not to be able to commit to policies lasting beyond the end of the review period, nor able to make lump–sum transfers to the firm. Thus we follow the spirit of Schmalensee (1989) on "good regulatory regimes", rather than the literature on the use of transfers and prices to implement

[5] See also the further discussion by Lesourne (1976) and Bailey (1976).

[6] It is usual in the principal–agent literature to take the agent's effort as determining the parameters of the probability density over a fixed support, the set of possible values of an output variable. The idea of the Markov process formulation here is to extend this to the dynamic case: if the firm did nothing, its costs would change in a stochastic way. By incurring costs the firm is able to influence the transition probabilities of this stochastic process, which is defined relative to a fixed set of possible states, the firm's unit cost levels.

(second–best) optimal behaviour in models with asymmetric information. We take as given the essential features of price cap mechanisms found in practice and consider the question of the optimal choice of their main parameters.

Between reviews the regulator commits not to intervene, and the firm chooses cost–reducing efforts to maximize the value of its profits before the next review. At that point — just as in the quotation from Baumol and Klevorick above — the regulatory parameters are reset so that any excess (expected) profits are eliminated, and the process repeats itself.

For modelling purposes a decision must be made about whether to measure time discretely or continuously. The continuous formulation has the advantages that results need fewer assumptions, and also that differential calculus may be used. However, the discrete case perhaps uses more familiar techniques and is more intuitive. The numerical simulations described later also require discrete time. We have therefore used this formulation in the body of the text and provide rather more briefly for the continuous case in Appendix I.

In sections 3 and 4 respectively we consider the dynamic optimization problems of the firm and the regulator, which are neatly nested. We show, among other results, that the firm's efforts decrease as a review approaches, and that the minimum regulated prices consistent with the firm breaking even decrease as the lag extends for review periods that begin in a high cost state, but increase with the lag for periods beginning in a low cost state. These facts present the regulator with a dilemma: for a period beginning in a high cost state, a longer lag lowers price in the immediate future and improves the chances that cost will be low at the next review time, but it delays the coming of that time. For a period beginning in a low cost state, a longer lag raises price in the immediate future and reduces the chance of remaining in the low cost state, but it does delay the possibility of switching to a high cost state with its corresponding high price. In the special case of totally inelastic demand, infinite lags are optimal, because allocative inefficiency is of no

consequence. And if transition probabilities are insensitive to efforts, then no lag is optimal, because effort incentives are irrelevant. Analytical results for the general case are hard to obtain, and so (in section 5) we use numerical methods. In the simulations that we have performed, less elastic demand and more sensitivity of costs to effort favour longer lags.

In the basic version of the model, the regulator commits to a uniform price for a review period, but an interesting question, which we take up in section 6, is whether commitment to a non–constant price path might be superior if it is feasible, even though the model has a stationary structure. We show that it will always be better to have a non–constant price path, but there appear to be few general results as to its shape. However, a natural Ramsey result is derived. Finally, extensions and conclusions are discussed in section 7.

2. The model

There is a regulated profit–maximizing monopolist supplying a single product. Time is discrete and there is an infinite horizon. The demand function $q(p)$ — where q and p respectively denote output and price — is decreasing, invariant over time and known to the regulator.[7] There are two possible levels of unit production cost, denoted $\underline{c}$ and $\bar{c}$, with $\Delta c = \bar{c} - \underline{c} > 0$. The probability that unit cost in a given period is $\underline{c}$ rather than $\bar{c}$ depends on the cost–reducing efforts chosen by the firm. In particular, if the preceding period's cost level was $\underline{c}$, and if the firm spends x on cost reduction in period n, then the probability that cost will be $\underline{c}$ in period n is given by $\underline{a}(x)$. And if $\bar{c}$ was the previous period's cost level, then $\bar{a}(x)$ is the probability that cost will be $\underline{c}$ in period n. We assume that (i) $a'(x) > 0$ and $a'(0) = \infty$; (ii) $a'' < 0$; and (iii) $\underline{a}(x) > \bar{a}(x)$ for $x \geq 0$.

[7] See Lewis and Sappington (1989) for an analysis of regulation when the demand curve is not known with certainty.

Price, or equivalently output, is determined by the regulator. The regulator's objective is the discounted sum (possibly weighted) of consumer and producer surplus — see below. The regulator and the firm both have discount factor $\delta \in (0, 1)$ and are risk–neutral. If the current period is one of regulatory review, the regulator (accurately) observes the cost level inherited from the past and specifies:

— the number of periods r before the next review, and

— the price p to prevail until that review.

The regulator cannot affect price between reviews — the commitment not to intervene then is assumed to be total, but there is full discretion at review times. However, when setting price and the number of periods before the next review, the regulator must ensure that the firm at least breaks even in expected terms over the forthcoming period.[8] This participation constraint will always bind, and it follows that the regulated price p is therefore implied by the choice of regulatory lag r. Thus the regulator's problem reduces to choosing r (contingent on the cost level at review time).

Moreover, the problem to be analyzed neatly decomposes into two nested parts. Suppose that a review has just occurred. The firm will get an expected payoff of zero in the subgame following the next price review, irrespective of the cost level that prevails at the time, and so the firm now faces a *finite* horizon optimization problem, even though the game as a whole has an infinite horizon. In particular, the firm has an r–period cost minimization problem, since its revenues have been determined by regulation. It is convenient to number the periods before the next review in reverse order: $n = r,\ldots,1$. Given q, the regulated level of output, let $\bar{x}_n$ and $\underline{x}_n$ denote the expenditure on cost reduction in period n (i.e. n periods before the next review) when the cost level

[8] An alternative formulation, which we do not examine here, would be to require that the firm should never incur a loss. For a well financed firm, our assumption appears to be the more appropriate. The firm should be willing to accept the regulated monopoly franchise if the prospect has a non–negative expected present value. We also assume that, should the firm midway through a review period expect to incur losses in the future (because of an unlucky switch in cost level, say) it is not then allowed to withdraw from production.

inherited from the previous period is high and low respectively. We can think of the firm as choosing $2r$ expenditure levels, namely $\bar{x}_n$ and $\underline{x}_n$ for $n = r,...,1$, to minimize expected discounted costs over the review period.

Turning now to the regulator's problem, it follows from the fact that the firm has a zero expected payoff at a review time that maximizing the regulator's objective is the same as maximizing (expected discounted) consumer surplus. We assume that there is no cost incurred in auditing the firm at review times, so any benefits which result from lags are due solely to their incentive effects, rather than from reducing the cost burden of regulation. The regulator cares about the cost–reducing efforts induced by the choice of r (and hence of p), because they affect the probability that the cost level will be low rather than high at the next review. Because of the stationary structure of the model, the regulator's optimal strategy is simply a pair $\{\underline{r}, \bar{r}\}$ of regulatory lags — one for each initial cost state.[9]

To summarize:

(i) between regulatory reviews the firm chooses efforts to solve a finite horizon cost minimization problem, and

(ii) at regulatory reviews the regulator chooses regulatory lags to maximize expected discounted welfare, given the solution to (i).

We now examine each optimization problem in turn.

3. The firm's cost–minimization problem between reviews

Suppose the regulatory parameters have been fixed, so that the firm is faced with a given price p, hence output level $q(p)$, and review period r. The value of its revenues over the next r periods is $\frac{1-\delta^r}{1-\delta} R(p)$, where $R(p) = pq(p)$.[10] Given q, define $\bar{C}_n$ as the minimum expected discounted value of costs — production costs plus effort costs — when n periods remain before the next review and when the initial cost state is $\bar{c}$. Similarly define $\underline{C}_n$ for when the initial cost state is $\underline{c}$. More formally, $\bar{C}_n$ and $\underline{C}_n$ are defined recursively as follows: $\bar{C}_0 = \underline{C}_0 = 0$, and for $n \geq 1$

$$\text{(1a)} \quad \bar{C}_n = \min_{x \geq 0}: \; q\{\bar{a}(x)\underline{c} + [1-\bar{a}(x)]\bar{c}\} + x + \delta\{\bar{a}(x)\underline{C}_{n-1} + [1-\bar{a}(x)]\bar{C}_{n-1}\}$$

$$\text{(1b)} \quad \underline{C}_n = \min_{x \geq 0}: \; q\{\underline{a}(x)\underline{c} + [1-\underline{a}(x)]\bar{c}\} + x + \delta\{\underline{a}(x)\underline{C}_{n-1} + [1-\underline{a}(x)]\bar{C}_{n-1}\} \; .$$

We use this pair of dynamic programming equations to characterize the firm's effort choices. From equations (1) the first–order conditions for the optimal x_n are

$$\text{(2)} \qquad \bar{a}'(\bar{x}_n) = \underline{a}'(\underline{x}_n) = 1/(q\Delta c + \delta\Delta C_{n-1})$$

where $\Delta C_{n-1} \equiv \bar{C}_{n-1} - \underline{C}_{n-1}$. The second order conditions are necessarily satisfied given the assumptions on $a(\cdot)$. We now make a key assumption:

$$\text{(A1)} \qquad \bar{a}'(\bar{x}) = \underline{a}'(\underline{x}) \text{ implies } \bar{a}(\bar{x}) \leq \underline{a}(\underline{x}).$$

We regard this as a mild assumption. Its effect is to ensure that with optimal efforts, the probability of being in the low cost state in the next period is higher if the current cost

[9] The two lags $\underline{r}$ and $\bar{r}$ may differ. We could further constrain the problem so that the regulator must choose a common r, but then we would not be able to use dynamic programming methods so straightforwardly. In any event, it is not clear why the regulator should be constrained in that way. British Telecom provides an example of a different lag being chosen at review time. In 1989, at the end of its first review period, which had been five years in length, four years was chosen as the time before the next review.

[10] Since a price cap is a ceiling, it is conceivable that, given a low cost realization, the firm's profit–maximizing price could be lower than the regulated cap. In that case the price cap constraint would not bind. To avoid the resulting complications, and because we doubt the practical relevance of the point in industries that are regulated, we make the mild assumption that the price cap constraint is always binding. Essentially this involves the cost difference Δc and/or the elasticity of demand being not too large.

state is low than if it is high: $\underline{a}(\underline{x}_n) > \bar{a}(\bar{x}_n)$ for all n. Whether or not $\underline{x}_n > \bar{x}_n$ is ambiguous in this model.[11]

The first question of interest is how efforts depend on the length of time until the next review. In keeping with the intuition that effort incentives weaken as regulatory review approaches, we have:

RESULT 1: The effort levels $\underline{x}_n$ and $\bar{x}_n$ are increasing in n.

PROOF: The result is true if ΔC_n is increasing in n. This follows from the first–order conditions (2), and the concavity of the $a(\cdot)$ functions. It is easy to establish that $\Delta C_1 > \Delta C_0 = 0$. Now define the function

$$f(z) = [\underline{a}(\underline{x}(z)) - \bar{a}(\bar{x}(z))](q\Delta c + \delta z) + [\bar{x}(z) - \underline{x}(z)]. \tag{3}$$

Here we define $\underline{x}(z)$ to be given by $\underline{a}'(\underline{x}(z)) = 1/z$, and similarly for $\bar{x}(z)$. From (1a) and (1b) we see that $\Delta C_n = f(\Delta C_{n-1})$. If $f(z)$ is strictly increasing, then $\Delta C_{n-1} > \Delta C_{n-2}$ implies $\Delta C_n > \Delta C_{n-1}$, and an inductive argument works. Differentiating and cancelling terms, we have

$$f'(z) = \delta[\underline{a}(\underline{x}(z)) - \bar{a}(\bar{x}(z))] > 0,$$

where the inequality follows from (A1). □

[11] If we made the stronger assumption that $\bar{a}'(x) \leq \underline{a}'(x)$, then we would have the result that $\bar{x}_n \leq \underline{x}_n$. This assumption — that expenditure on effort is always more productive at the margin when the firm was previously in the low cost state — is not a natural one, and we do not need it for our results.

Thus the firm invests less on cost reduction as the review date nears.[12] It can also be shown that the firm spends a suboptimal amount on cost reduction, given q, because the firm does not take account of the benefit to consumers of increasing the probability that the cost state at the time of the next review is low rather than high (see below on probabilities). The next question is how efforts vary with output.

RESULT 2: Both $\underline{x}_n$ and $\bar{x}_n$ are increasing in q, the regulated level of output.

PROOF: From (3) we see that $f(z)$ is increasing in q, and therefore ΔC_n is increasing in q. Equation (2) then implies the stated result. □

Thus lower price caps stimulate more cost–reducing expenditure. This happens because savings in expected unit cost will be multiplied by a greater output.

We now establish some results concerning the probabilities with which $\underline{c}$ and $\bar{c}$ will be realised at the time of the next review. This is important because these help determine the ensuing optimal price cap regime. For given q, let $\bar{\pi}_r$ [resp. $\underline{\pi}_r$] denote the probability of realising cost level $\underline{c}$ in the last year of a review period of length r, given that the cost observation in the year before the review period started was $\bar{c}$ [resp. $\underline{c}$]. Then we have

RESULT 3: $\bar{\pi}_r < \underline{\pi}_r$. Moreover, $\bar{\pi}_r$ is increasing in r, and $\underline{\pi}_r$ is decreasing in r.

PROOF: By induction. It is true for $r = 1$ by (A1), so suppose it is true for $r-1 \geq 1$. Then we have

[12] More precisely, if the cost level does not switch, effort decreases over time. The sign of a term such as $(\bar{x}_n - \underline{x}_{n-1})$ is ambiguous.

(4a) $$\bar{\pi}_r = (1 - \bar{a}(\bar{x}_r))\bar{\pi}_{r-1} + \bar{a}(\bar{x}_r)\underline{\pi}_{r-1}$$

(4b) $$\underline{\pi}_r = (1 - \underline{a}(\underline{x}_r))\bar{\pi}_{r-1} + \underline{a}(\underline{x}_r)\underline{\pi}_{r-1}.$$

Therefore

$$\underline{\pi}_r - \bar{\pi}_r = (\underline{a}(\underline{x}_r) - \bar{a}(\bar{x}_r))(\underline{\pi}_{r-1} - \bar{\pi}_{r-1})$$

$$> 0,$$

where the inequality follows from (A1) and the inductive assumption. Tha $\bar{\pi}_r$ is increasing now follows from (4a), and $\underline{\pi}_r$ is decreasing from (4b). □

It is natural that $\bar{\pi}_r$ is increasing in r, because efforts increase with r and there is mor time for luck to improve the state as r increases. A reverse argument holds for the cas for $\underline{\pi}_r$. Result 3 was for a given output level. We next show how these probabilities var with output.

RESULT 4: $\bar{\pi}_r$ and $\underline{\pi}_r$ are both increasing in q.

PROOF: By induction on r. It is true for r = 1, so assume it is true for r−1 ≥ 1 Then, as above,

(4a) $$\bar{\pi}_r = (1 - \bar{a}(\bar{x}_r))\bar{\pi}_{r-1} + \bar{a}(\bar{x}_r)\underline{\pi}_{r-1}.$$

Result 2 tells us that $\bar{x}_r$ and $\underline{x}_r$ — and hence $\bar{a}(\bar{x}_r)$ and $\underline{a}(\underline{x}_r)$ — a increasing in q. By assumption, $\bar{\pi}_{r-1}$ and $\underline{\pi}_{r-1}$ are both increasing in q

and so (4a) tells us that $\bar{\pi}_r$ is then also increasing in q. A similar reasoning holds for $\underline{\pi}_r$. □

It is natural for these probabilities to increase with output, because the firm's choices of efforts do.

Results 3 and 4 stated how the probabilities vary as r and q change independently. But of course r and q are linked by the break–even constraint for the firm, and it is time to consider this explicitly. The regulator will want this constraint to hold for three reasons — lower prices are closer to expected marginal costs, they induce greater cost–reducing efforts, and they are distributionally beneficial if consumer interests carry more weight than profit in the welfare criterion.

Let the optimal price to set when the lag is r and the initial state is $\bar{c}$ [resp. $\underline{c}$] be denoted $\bar{p}(r)$ [resp. $\underline{p}(r)$]. Then $\bar{p}(r)$ is the smallest price p such that

(5) $$\frac{1-\delta^r}{1-\delta} R(p) = \bar{C}_r(q(p)) \ ,$$

and similarly for $\underline{p}(r)$. The left–hand side of (5) is the discounted sum of revenues over the period. Since $\Delta C_r > 0$, we see that $\bar{p}(r) \geq \underline{p}(r)$.

RESULT 5: $\bar{p}(r)$ is decreasing in r, and $\underline{p}(r)$ is increasing in r.

PROOF: See Appendix II.

The intuition behind this result is that, starting from the high cost state, as r increases the *annualized* expected cost given by $\frac{1-\delta}{1-\delta^r}\bar{C}_r$ falls. This is because the probability of ending up in the low cast state rises with r — the longer r means that the expected

proportion of time spent in a low cost state, given costs were high to start with, is greater
Exactly opposing arguments hold when costs are initially low, so here the annualized cost
rises with r.

In addition, the fact that $\bar{p}(r) > \underline{p}(r)$ combined with Result 5 tells us that
$\bar{p}(\bar{r}) > \underline{p}(\underline{r})$ for any pair of lags, $\bar{r}$ and $\underline{r}$. Thus prices will always be lower in low–cost
regimes than in high–cost regimes.

We can now say how the probability of the cost level being low at the next review
date varies with the lag when the dependence of price on the lag is taken into account. Let
$\bar{\Pi}(r) = \pi(r, q(\bar{p}(r)))$ be this probability, and similarly for $\underline{\Pi}(r)$. We obtain a result that
mirrors Result 3 (that it is true follows from Results 3, 4 and 5).

RESULT 6: $\bar{\Pi}(r) < \underline{\Pi}(r)$. Moreover, $\bar{\Pi}(r)$ is increasing in r, and $\underline{\Pi}(r)$ is decreasing in r.

Moreover, we have the important corollary that $\bar{\Pi}(\bar{r}) < \underline{\Pi}(\underline{r})$ for all pairs of lags.

To summarize, a number of facts about the firm's cost minimization problem have
been established. For a given cost state, efforts decrease as regulatory review approaches.
For a given output level, as regulatory lag extends, the probability that cost will be low at
the next review date increases if the current period began with a high cost level, but
decreases if the current period began with a low cost level. Efforts and the probabilities
that cost will be low at the next review are both increasing in the output level. The
regulated price is given by the condition that the firm expects to break even over the
review period. As regulatory lag extends, the regulated price decreases if the cost level at
the review time is high, but increases if it is low. This implies that the result above about
the probability of cost being low at the next review date is true — indeed is strengthened
— when the dependence of price on the lag is taken into account. The solution to the
firm's problem having been characterized, we now turn to the problem facing the regulator.

...tory lags

...ulator has a stationary welfare maximization problem with an infinite
...expected future welfare, as a function of the chosen lags $\bar{r}$ and $\underline{r}$, be
...the initial cost state is $\bar{c}$, and $\underline{W}(\bar{r}, \underline{r})$ when the initial cost state is $\underline{c}$. Let
...indirect utility function, which measures consumer surplus per period. We

$$\bar{W}(\bar{r}, \underline{r}) = \frac{1-\delta^{\bar{r}}}{1-\delta} v(\bar{p}(\bar{r})) + \delta^{\bar{r}}\{\bar{\Pi}(\bar{r})\underline{W}(\bar{r}, \underline{r}) + (1-\bar{\Pi}(\bar{r}))\bar{W}(\bar{r}, \underline{r})\}$$

$$\underline{W}(\bar{r}, \underline{r}) = \frac{1-\delta^{\underline{r}}}{1-\delta} v(\underline{p}(\underline{r})) + \delta^{\underline{r}}\{\underline{\Pi}(\underline{r})\underline{W}(\bar{r}, \underline{r}) + (1-\underline{\Pi}(\underline{r}))\bar{W}(\bar{r}, \underline{r})\}.$$

...these two simultaneous equations to give welfare explicitly in terms of $\bar{r}$ and $\underline{r}$,
...is

$$(1-\delta)\bar{W} = \bar{\gamma}v(\underline{p}) + (1-\bar{\gamma})v(\bar{p}) \quad \text{, and}$$

$$(1-\delta)\underline{W} = \underline{\gamma}v(\underline{p}) + (1-\underline{\gamma})v(\bar{p}) \quad \text{, where}$$

$$\bar{\gamma}(\bar{r}, \underline{r}) = \frac{\delta^{\bar{r}}(1-\delta^{\underline{r}})\bar{\Pi}(\bar{r})}{(1-\delta^{\bar{r}})(1-\delta^{\underline{r}}\underline{\Pi}(\underline{r})) + \delta^{\bar{r}}(1-\delta^{\underline{r}})\bar{\Pi}(\bar{r})} \quad \text{, and}$$

$$\underline{\gamma}(\bar{r}, \underline{r}) = \frac{(1-\delta^{\underline{r}})(1-\delta^{\bar{r}}(1-\bar{\Pi}(\bar{r})))}{\delta^{\underline{r}}(1-\delta^{\bar{r}})(1-\underline{\Pi}(\underline{r})) + (1-\delta^{\underline{r}})(1-\delta^{\bar{r}}(1-\bar{\Pi}(\bar{r})))} .$$

...e equations state that annualized welfare is equal to a weighted average of high– and
...cost regime surpluses. One may think of γ as the "proportion" of time that the

low–cost regime is expected to prevail, appropriately discounted (clearly $0 < \gamma < 1$).[13]

One way to solve for the optimal $\{\bar{r}, \underline{r}\}$ is to use (7a) and (7b) directly.[14] However, it is perhaps more illuminating to use the principles of dynamic programming, and express *optimal* welfare (denoted by W^* and W_*) implicitly in the following pair of simultaneous equations:

$$(8a) \quad W^* = \max_{r \geq 1}: \frac{1-\delta^r}{1-\delta} v(\underline{p}(r)) + \delta^r\{\Pi(r)W_* + (1-\Pi(r))W^*\}$$

$$(8b) \quad W_* = \max_{r \geq 1}: \frac{1-\delta^r}{1-\delta} v(\underline{p}(r)) + \delta^r\{\underline{\Pi}(r)W_* + (1-\underline{\Pi}(r))W^*\}.$$

These relations hold necessarily because regulatory behaviour that is optimal overall involves optimal behaviour at the current decision stage followed by optimal behaviour thereafter.[15] Put another way, if a one–shot deviation from the $\{\bar{r}, \underline{r}\}$ policy increased welfare, then that policy could not be optimal. (Equations (8) are the Bellman equations for the regulator, just as equations (1) are the Bellman equations for the firm.)

The lags which solve the right hand sides of (8a) and (8b) are clearly the optimal

---

[13] It will always be the case that $\bar{\gamma} \leq \underline{\gamma}$. Discounting is why different γ's are required for each initial state. If there were no discounting — i.e. if $\delta = 1$ — then

$$\bar{\gamma} = \underline{\gamma} = \frac{\underline{r}\underline{\Pi}}{\bar{r}(1-\underline{\Pi}) + \underline{r}\underline{\Pi}} .$$

This fraction is now precisely the proportion of time spent in low–cost regimes.

[14] While it is clear from the stationary structure of the problem that the same pair of lags will maximize both equations (7a) and (7b), it is not transparent in terms of mathematics why this is so. But, necessarily, equation (7b) adds no information to the problem given equation (7a).

[15] Equations (8) are also *sufficient* for optimality since solutions to (8) are unique. To see this, we can write (8a) as $W_* = g(W^*)$ where $g(0) < 0$ and $g'(.) > 1$, and similarly for (8b). Such simultaneous equations will always have exactly one solution.

e the right hand side of (8a) as follows:

$$\frac{v(\underline{p}(r))}{1-\delta} + \delta^r\{\Pi(r)W_* + (1-\Pi(r))W^* - \frac{v(\underline{p}(r))}{1-\delta}\}.$$

tradeoffs facing the regulator starting from high costs apparent. A higher

duce $\underline{p}$ (from Result 5) and so increase $v(\underline{p})$,

crease Π (from Result 6) and so increase $[\Pi(r)W_* + (1-\Pi(r))W^*]$, but

crease the coefficient of $\{\Pi(r)W_* + (1-\Pi(r))W^* - \frac{v(\underline{p}(r))}{1-\delta}\}$ in the above

pression.

equation (7) and the fact that $\bar{\gamma} < \underline{\gamma}$ imply that W_* is greater than W^*.

cketed term in (iii) is always positive because both W_* and W^* are greater

(r))

In short, the effects of a longer lag on price in the short term and on the the

ity that cost — and hence price — will be low at the next next review are both

, but that review — which will always increase expected welfare — is postponed.

Conversely, if the regulator faces a low present cost, the relevant tradeoffs are the

mirror–image of the above: increasing r will raise the short–run price and reduce

obability that cost will still be low by the next review; it will, however, put off the

which, when it comes, will necessarily reduce expected welfare.

Analytical results about the optimal resolution of these tradeoffs are hard to obtain

eral. However, there are two extreme cases in which it is straightforward to

terise optimal regulatory lag.

RESULT 7: If demand is completely inelastic, $\underline{r} = \bar{r} = \infty$ is optimal.

PROOF: Suppose demand is fixed at $\hat{q}$. Then maximizing consumer surplus is equivalent to minimizing the loss function $p\hat{q}$, which we denote by L. If lags are infinite then $(1-\delta)\bar{L} = \bar{p}(\infty)\hat{q}$ and $(1-\delta)\underline{L} = \underline{p}(\infty)\hat{q}$. In addition, (5) tells us that

$$\frac{1-\delta^r}{1-\delta}\, p(r)\hat{q} = \bar{C}_r(\hat{q}).$$

Then, analogously to (8a) and (8b), optimal lags are infinite if and only if the following pair of simultaneous equations are satisfied:

$$\bar{C}_\infty = \min_{r \geq 1}:\ \bar{C}_r + \delta^r\{\Pi(r)\underline{C}_\infty + (1-\Pi(r))\bar{C}_\infty\}$$

$$\underline{C}_\infty = \min_{r \geq 1}:\ \underline{C}_r + \delta^r\{\underline{\Pi}(r)\underline{C}_\infty + (1-\underline{\Pi}(r))\bar{C}_\infty\}.$$

One possible option for the firm to minimise expected costs over the infinite horizon is to expend effort in the way that is optimal for the r–period subproblem, and then return to the optimal strategy for the infinite game. The resulting payoff is precisely what is written on the right hand sides of each of the above, and this can clearly be no better than the optimal cost–minimizing strategy. Therefore, the above pair of equations do hold, and so for the case of inelastic demand it is best never to have regulatory reviews. □

The reason is that infinite lag gives the firm perfect incentives for cost reduction. Price is

…e cost, but this causes no welfare loss when demand is inelastic. In fact the … attained.

If the transition probabilities $\bar{a}$ and $\underline{a}$ are insensitive to efforts, then the policy of setting $\underline{r} = \bar{r} = 1$ is optimal.

…tion here is that, since promoting effort is not an issue, short lags are …ously desirable, because prices are kept more often in line with costs, thereby … allocative efficiency, and there are no deleterious effects on cost efficiency. We …ve Result 8 for the case of discrete time, preferring instead to prove it for the … continuous time (see Appendix I), where the analysis is somewhat simpler, and … a more general result (Result 10) can be obtained. Result 10 relates the effort …iveness of the transition probabilities to a term that is linked to the demand …city. A corollary of the result is that if effort responsiveness is zero, then zero lag is …al. If, on the other hand, effort responsiveness is infinite at the origin, then positive … optimal — in this case a small lag is better than no lag because it achieves a … order gain in cost efficiency at the expense of only a second–order loss in allocative …ency.

Numerical results

Further analytical results being difficult to obtain, we use numerical methods to … further insights.[16] In order to compute numerical solutions, we must set explicit forms … both the responsiveness functions $\bar{a}(\cdot)$ and $\underline{a}(\cdot)$, as well well as the demand function

[16] A listing of the simple Fortran program is available on request. A subroutine calculates … function $\bar{C}_n(q)$, and an iterative procedure uses this to find the function $\bar{p}(r)$. …lfare, $\bar{W}(\bar{r}, \underline{r})$, is then obtained from (7).

$q(\cdot)$. We suppose that

$$\bar{a}(x) = \bar{A} + bx^{\alpha}$$
$$\underline{a}(x) = \underline{A} + bx^{\alpha}$$
$$q(p) = 2 - \frac{p}{d}$$

over the relevant ranges.[17] Thus $\underline{a}(.)$ is simply a constant plus $\bar{a}(.)$. Throughout the following we will take:

$$\bar{A} = 0\ ;\ \underline{A} = 0.8\ ;\ \alpha = 0.2\ ;\ \delta = 0.9\ ;\ \bar{c} = 1\ ;\ \underline{c} = 0.$$

We considered the effect of changing the two remaining parameters d and b.

Table 1 below shows the effect on welfare of changing the lags for a particular specification, namely $d = 0.9$, $b = 0.18$.[18] Clearly $\bar{r} = 3$, $\underline{r} = 4$ is optimal. Table 2 shows the effect of changing the demand parameter d. Reducing d is similar to increasing the demand elasticity, with the extreme case of $d = \infty$ resulting in inelastic demand. Result 6 perhaps suggests that there is a tendency for optimal lag to increase with this demand parameter. The table displays welfare for various values of d and for lags up to 9 (b is set at $b = 0.18$). It is in agreement with intuition and with earlier findings: a higher value of d increases the optimal lags.

Table 3 shows the effect of changing b, keeping d fixed at $d = 0.9$. The parameter b represents the responsiveness of expected cost reduction to effort, a higher b

[17] We enter this qualification because $a(\cdot)$ would otherwise exceed one for large enough x. In our example effort incentives are always such that this complication never arises.

[18] To be precise, the table shows how $\bar{W}$ varies with lags. While $\underline{W}$ is always higher than $\bar{W}$, it behaves in exactly the same way. All three tables show welfare starting from the high cost state.

ing this responsiveness. Result 7 perhaps suggests that increasing b will again e optimal regulatory lag. The table lends support to this general principle.[19]

n–constant prices between reviews

So far it has been assumed that the regulator commits to a uniform price at review n. If, however, the regulator has the power to commit to a *path* of prices over the next ew period, some new questions arise. In particular, we are interested in whether the might be increasing or decreasing. If so, then there could be reasons for having ve or negative X terms in RPI–X price cap regulation that have nothing to do rends in technology or demand (recall that our model has a stationary structure). To tigate this issue, we focus on the subproblem of choosing the optimal price/output given a regulatory lag r. We do not analyze the choice of r when a non–constant e path can be chosen (although many earlier results will carry through to this case).[20]

The problem is (using natural notation):

$$\max_{q_r \ldots q_1} : \sum_{n=1}^{r} \delta^{r-n} u(q_n) - \bar{C}_r(q_r,\ldots,q_1) \quad \text{s.t.} \quad \sum_{n=1}^{r} \delta^{r-n} R(q_n) \geq \bar{C}_r(q_r,\ldots,q_1)$$

when the initial cost level is high, and analogously when costs were low. Here, u(q) is s consumer utility (i.e. $u(q) \equiv v(p(q)) + R(q)$).

It is convenient at this point to introduce one final piece of notation. Given the put path and the length of the review period r, define the functions $\bar{h}(n)$ and $\underline{h}(n)$ to e probability of being in the low cost state n periods from the end of the review,

[19] In addition, the table shows that it is sometimes the case that $\bar{r}$ is longer than $\underline{r}$. It is curious that $\bar{r}$ sticks at 3 for high b.

[20] For instance, $\bar{\pi}$ will now be increasing in all components of $(q_r,\ldots,q_1)$, and will be asing in r. However, it will not necessarily be the case that ΔC is increasing in n, so efforts may not be falling over time.

given that initial costs were high or low respectively. Then, from the identity $\bar{\pi}(r) = \bar{h}(n)\underline{\pi}(n) + (1-\bar{h}(n))\bar{\pi}(n)$, and similarly for $\underline{h}$, we see that

$$\bar{h}(n) = \frac{\bar{\pi}(r) - \bar{\pi}(n)}{\underline{\pi}(n) - \bar{\pi}(n)}, \qquad \underline{h}(n) = \frac{\underline{\pi}(r) - \bar{\pi}(n)}{\underline{\pi}(n) - \bar{\pi}(n)}.$$

Using these functions, we can obtain the following partial result.

RESULT 9: The optimal time path of p obeys the following relation:

$$p(n)[1 - \frac{\alpha}{\eta(n)}] = h(n)\underline{c} + (1-h(n))\bar{c}.$$

where h is either $\bar{h}$ or $\underline{h}$ according to the initial cost state, $\eta(n)$ is the (modulus of the) elasticity of demand, and α is a constant satisfying $0 < \alpha < 1$. If (as seems reasonable) the left–hand side of the above expression is (weakly) decreasing with output, this means that the output path follows the path of $h(\cdot)$.

Result 9 is a form of Ramsey pricing. We prove the result, which is simply an application of the maximum lemma, in continuous time in Appendix I. Unfortunately, it is not possible to predict the behaviour of the $h(\cdot)$ functions in general. Since $\bar{h}(r) = 0$, we must have $\bar{h}(r-1) > \bar{h}(r)$, and so if initial costs were high, the optimal output path is initially increasing. Since $\underline{h}(r) = 1$, we similarly have the optimal output path initially decreasing if costs were low to start with. However, in many cases, $\bar{h}$ will be decreasing near $n = 0$, and so it will be optimal for output to fall towards the end of the review period if costs were initially high. Therefore, a general monotonicity result cannot be obtained.

remarks

has never been any doubt that regulatory lag increases the incentives for cost the regulated firm. In this paper we have been concerned with the question of trade off this benefit against the welfare losses associated with prices remaining with costs for longer periods of time. Though stylized, our two–state model several aspects of the tradeoff.[21] Analysis of the firm's cost–minimization vealed that, starting from a low cost state, longer lag postpones the next review ch price might be increased, but worsens the chance that cost will still be low the next review, and price in the meantime has to be raised to ensure that is even. Starting from a high cost state, longer lag has the disadvantage of the next review, but advantages with respect to the probability of cost being review and price in the interim. Analysis (for special cases) and numerical (more generally) indicate that longer lags are favoured by more responsiveness to effort and less elastic demand, which is in accordance with intuition. Minimal optimal only under restrictive circumstances.

me dynamic aspects of the long–running debate about the relative merits of price e–of–return regulation can be set in the context of the model. For the case of a od monopolist, there are perhaps two natural ways to model rate–of–return on:

The regulator constantly monitors the costs and revenues of the firm to keep its rate return at the regulated level. In our framework, this would mean having the gulator's policy being to set very short lags, and ensuring that prices are set so

---

model could be extended straightforwardly to n cost states. Exactly the same would apply, with the modification that the simple transition functions become ion functions over the various states. We believe that, while many of the results wo–state model would apply to the highest and lowest cost states, results for states would be less clear–cut regarding such variables as price and end–of–period ities. If an infinite number of states were allowed, however, new complications e. For example, if costs could get arbitrarily high, it might be desirable for the r to stop production.

that the firm does not make a loss during these lags.

— Alternatively, the regulator might review the firm's rate of return only when the firm files a request for a price increase. This would correspond to a policy of having a very long lag if the cost level was high initially (in our simple two–state model if the firm started off with high costs, costs would never rise subsequently), and a very short lag when costs were low to start with (in order to catch the cost increase quickly).

In our model neither of these policies is likely to be optimal (though we gave conditions under which the first one would be). It might be said that the model therefore supports price caps (at least to some degree) over rate–of–return regulation,[22] but we do not favour this interpretation. We regard the "price cap versus rate–of–return" question as being too restrictive: it is more fruitful to explore the underlying issues such as the question of regulatory lag. Another such issue is the question of how much "pass–through" of costs the firm should be allowed, which Schmalensee (1989) investigated using a static principal–agent model. Our model, which focuses on dynamics and excludes cost pass–through, is complementary to that approach.

One dynamic aspect of the problem that we have not addressed is *adverse selection* (see Baron and Besanko (1984)). In our model both the regulator and the firm are equally well informed at the start of each review period. If, for example, the regulator (but not the firm) was uncertain about the potential for cost reduction, then the analysis would be considerably more complex. The firm would then have an incentive to behave strategically to influence the regulator's beliefs about the scope for cost reduction in order to receive a more lenient price for the next review period.[23] In the resulting signaling equilibrium, the

[22] A third variant on the rate of return theme has been analyzed by Bawa and Sibley (1980), who have a model where reviews are stochastic, and the probability of a review increases with the firm's excess rate of return. This would correspond to a review policy of random lags — a complication we have not covered.

[23] That is to say there would be a *ratchet effect* in the sense of Weitzman (1980), an idea further explored in Freixas, Guesnerie and Tirole (1985).

yoff in the continuation game after a review period will generally not be pendent, and so the firm will care about the end–of–review probabilities for the various cost states. As the analysis in this paper was greatly facilitated by le to ignore this consideration, with the result that the firm faced a *finite* horizon c programming problem, we do not imagine the extension to include adverse on would be a simple one.

Be that as it may, we hope that the present paper might provide a suitable ork in which to explore further issues concerning dynamic issues in price regulation.

Appendix I: The problem in continuous time

In this Appendix we set out the natural continuous time version of the discrete time model analyzed in the text. We follow the same structure and, as far as possible, the same notation as there. Equations corresponding to equations in the discrete case are marked with a prime.

The main difference between the continuous and discrete cases concerns the nature of the probabilities of transition between cost states. We assume that costs follow a continuous Markov process with instantaneous transition probabilities given by the functions $\bar{m}(.)$ and $\underline{m}(.)$:

$$\text{Prob}\{c(s-\delta s) = \underline{c} \mid c(s) = \bar{c}\} = \bar{m}(\bar{x})\delta s + [\text{terms of second order in } \delta s],$$

where $\bar{m}(\cdot)$ is non–negative, increasing and strictly concave. (It corresponds to the $\bar{a}(.)$ function in the main text.) And

$$\text{Prob}\{c(s-\delta s) = \bar{c} \mid c(s) = \underline{c}\} = \underline{m}(\underline{x})\delta s + [\text{terms of second order in } \delta s],$$

where $\underline{m}(.)$ is non–negative, decreasing and strictly convex. (It corresponds to the function $1-\underline{a}(.)$ in the main text.)

Thus in the first case, increasing effort increases the instantaneous probability of costs switching from the high to the low state, while in the second case, increasing effort reduces the probability of switching to the high cost state. There is no longer any restriction on either of these functions being less than one, and neither is a condition analogous to (A1) needed. Both the firm and the regulator face a common interest rate of $\rho > 0$ (so ρ is related to $1-\delta$ in the main text).

The firm's cost–minimization problem between reviews

The firm faces a cost–minimization problem over a continuous time interval of length t ($t \geq s \geq 0$). Denote expenditure on cost–reducing effort at time s by $\bar{x}(s)$ if the cost level at time s is $\bar{c}$, and by $\underline{x}(s)$ if it is $\underline{c}$.

Minimum expected value of discounted future costs is denoted by $\bar{C}(s)$ if the cost level at time s is $\bar{c}$, and by $\underline{C}(s)$ if it is $\underline{c}$. The Bellman equations are:

$$(1a') \qquad \bar{C}'(s) = q\bar{c} - \rho\bar{C}(s) + \bar{\psi}(\Delta C(s)),$$

where $\bar{\psi}(z) \equiv \min\{x - \bar{m}(x)z \mid x \geq 0\}$ is the dual function of $\bar{m}$, and

$$(1b') \qquad \underline{C}'(s) = q\underline{c} - \rho\underline{C}(s) + \underline{\psi}(\Delta C(s)),$$

where $\underline{\psi}(z) \equiv \min\{x + \underline{m}(x)z \mid x \geq 0\}$ is the dual function of $\underline{m}$. The first–order

...s for effort levels are

$$\bar{m}'(\bar{x}(s)) = -\underline{m}'(\underline{x}(s)) = 1/\Delta C(s).$$

...1': The effort levels $\bar{x}$ and $\underline{x}$ are increasing in s.

...: Again, it is sufficient to show that ΔC increases with s. Equations (1a') and ...imply that

$$\Delta C'(s) = q\Delta c - \rho\Delta C(s) + \bar{\psi}(\Delta C(s)) - \underline{\psi}(\Delta C(s)).$$

...

$$f(z) = \rho z + \underline{\psi}(z) - \bar{\psi}(z).$$

...since $\underline{\psi}$ is increasing and $\bar{\psi}$ is decreasing, f is a strictly increasing function. ...z^* to solve $f(z) = q\Delta c$. Then by construction, $\Delta C(s)$ is increasing [resp. ...ing] if $\Delta C(s) < z^*$ [resp. $> z^*$]. Since $z^* > 0$ and $\Delta C(0) = 0$, we have shown that ...increasing for all s (in fact, we have shown that it tends monotonically to z^* over ... □

...worth noting that we did not need to use any assumption analogous to (A1) above in ...to prove this result.

...T 2: Both $\bar{x}$ and $\underline{x}$ are strictly increasing in q, the regulated level of output.

...T 3: $\bar{\pi}(t) \leq \underline{\pi}(t)$. Moreover, $\bar{\pi}(t)$ is increasing in t, and $\underline{\pi}(t)$ is decreasing in t.

...: In this case, equations (4a) and (b) become

$$\bar{\pi}'(t) = \bar{m}(\bar{x}(t))\Delta\pi(t),$$

$$\underline{\pi}'(t) = -\underline{m}(\underline{x}(t))\Delta\pi(t) \quad \text{where } \Delta\pi \equiv \underline{\pi} - \bar{\pi}.$$

...ore,

$$\Delta\pi'(t) = -(\bar{m}(\bar{x}) + \underline{m}(\underline{x}))\Delta\pi(t).$$

... $\Delta\pi(0) = 1$, we see that $\Delta\pi(t)$ is non–negative (and decreasing in t). ...the remaining part of the result is true follows upon examination of (4'). □

...T 4: $\bar{\pi}(t)$ and $\underline{\pi}(t)$ are both increasing in output.

...mit proofs of analogous results from now on.) As before, $\bar{p}(t)$ is the smallest price

such that

$$(5') \qquad \frac{(1-e^{-\rho t})}{\rho} R(p) \geq \bar{C}(t, q(p))$$

and similarly for $\underline{p}(t)$.

RESULT 5: $\bar{p}$ [resp. $\underline{p}$] is decreasing [resp. increasing] in t.

RESULT 6: $\bar{\Pi}(t)$ [resp. $\underline{\Pi}(t)$] is increasing [resp. decreasing] in t.

In addition, the chain

$$\bar{\Pi}(t) = \bar{\pi}(t, \bar{p}(t)) < \underline{\pi}(t, \bar{p}(t)) \leq \underline{\pi}(t, \underline{p}(t)) = \underline{\Pi}(t)$$

shows that, even when the effect on prices is taken into account, we still have $\bar{\Pi} < \underline{\Pi}$.

Optimal regulatory lag

Denote the lag starting from a high cost state by $\bar{t}$, and from a low cost state by $\underline{t}$. Then, as before, W^* and W_* must satisfy the following pair of simultaneous equations:

$$(8a') \qquad W^* = \max_{t \geq 0} \{\frac{1-e^{-\rho t}}{\rho} v(\bar{p}(t)) + e^{-\rho t}[\bar{\Pi}(t)W_* + (1-\bar{\Pi}(t))W^*]\},$$

$$(8b') \qquad W_* = \max_{t \geq 0} \{\frac{1-e^{-\rho t}}{\rho} v(\underline{p}(t)) + e^{-\rho t}[\underline{\Pi}(t)W_* + (1-\underline{\Pi}(t))W^*]\}.$$

Clearly, the times that solve the right–hand sides of the above equations are the optimal lags. If there is no solution, then an infinite lag is optimal (in which case the "max" should strictly be read as a "sup"). Result 10 is concerned with whether positive lags are optimal.

RESULT 10: Optimal $\bar{t}$ is positive iff $\dfrac{-\bar{m}'(0)}{\bar{m}(0)\,\bar{m}'(0)} > \dfrac{\Delta v}{\bar{q}\Delta c} - 1 \quad (\geq 0)$,

and similarly for $\underline{t}$.

PROOF: Suppose that zero lags are optimal. Then it may be calculated that

$$(9) \qquad W^* = \frac{(\rho+\underline{m}_0)v(\bar{c}) + \bar{m}_0 v(\underline{c})}{\rho\lambda}$$

$$W_* = \frac{\underline{m}_0 v(\bar{c}) + (\rho+\bar{m}_0)v(\underline{c})}{\rho\lambda},$$

$\underline{m}_0 = \underline{m}(0)$ and $\lambda \equiv \rho + \bar{m}_0 + \underline{m}_0$. If no lags were indeed optimal sides of (8a′) and (8b′) with the values of W^* and W_* given in (9) at $t = 0$. The derivative of the right–hand sides of (8a′) and (8b′) at y,

$+ \bar{m}_0\Delta W$, and

$+ \underline{m}_0\Delta W$, (where $\Delta W \equiv W_* - W^*$).

to the above makes both derivatives vanish, and so this provides no on whether lags are good or bad. Therefore look for second–order the fact that $v'(p) = -q(p)$, we get the second derivatives of the equations (8) at zero to be, respectively,

$p'(\bar{q})\bar{q}'(0) + \rho^2 W^* - 2\rho\bar{m}_0\Delta W + \bar{m}'(0)\bar{x}'(0)\Delta W - \bar{m}_0(\bar{m}_0+\underline{m}_0)\Delta W,$

$p'(\underline{q})\underline{q}'(0) + \rho^2 W_* + 2\rho\underline{m}_0\Delta W - \underline{m}'(0)\underline{x}'(0)\Delta W + \underline{m}_0(\bar{m}_0+\underline{m}_0)\Delta W,$

$(\bar{c})$ and $\underline{q} = q(\underline{c})$. It may be calculated that

$$(0) = \frac{\bar{m}_0\Delta c}{-2p'(\bar{q})}, \quad \underline{q}'(0) = \frac{\underline{m}_0\Delta c}{2p'(\underline{q})},$$

$$(0) = \frac{\bar{q}\Delta c}{-\bar{m}'(0)}, \quad \underline{x}'(0) = \frac{\underline{q}\Delta c}{\underline{m}'(0)}.$$

ting these into (10) give the second derivatives at zero to be

$$\left\{\frac{\bar{m}_0\Delta v}{\lambda}\left[\frac{\bar{q}\Delta c}{\Delta v} - 1\right]\right\} + \left\{\frac{-\Delta v\bar{m}'(0)\bar{q}\Delta c}{-\lambda\bar{m}'(0)}\right\}, \text{ and}$$

$$\left\{\frac{\underline{m}_0\Delta c}{\lambda}\left[1 - \frac{\underline{q}\Delta c}{\Delta v}\right]\right\} + \left\{\frac{-\Delta v\underline{m}'(0)\underline{q}\Delta c}{\lambda\underline{m}'(0)}\right\},$$

$\Delta v \equiv v(\underline{c}) - v(\bar{c})$. Now, the convexity of $v(.)$ tells us that

$$\bar{q}\Delta c \leq \Delta v \leq \underline{q}\Delta c,$$

we see that the left–hand brace in each of the above is always non–positive, whilst

right–hand brace is always non–negative. Therefore, optimal $\bar{t}$ is positive if and only if

$$\frac{-\bar{m}'(0)}{\bar{m}(0)\,\bar{m}'(0)} > \frac{\Delta v}{\bar{q}\,\Delta c} - 1. \quad \square$$

The left–hand side of the above is measure of the responsiveness of $\bar{m}(.)$ to effort at zero, the right hand side is larger the more elastic is demand (it is zero for inelastic demand). It also tends to zero as Δc tends to zero. A very similar condition determines whether $\underline{t}$ is positive. Therefore, we can sum up this result by saying that optimal lags are positive if:

(i) if $\bar{m}'(0) = +\infty$ and $\underline{m}'(0) = -\infty$
(ii) if demand is sufficiently inelastic compared to the index of responsiveness given above, or
(iii) if the cost differential Δc is small relative to this index.

COROLLARY: If the functions $m(.)$ are insensitive to efforts, then $\bar{t} = \underline{t} = 0$ is optimal.

PROOF: In this case the right–hand braces in (11) above are not well–defined. However, in this case these become, respectively (from (10)),

$$\bar{m}'(0)\bar{x}'(0)\Delta W, \text{ and } \underline{m}'(0)\underline{x}'(0)\Delta W.$$

Both of these are clearly zero, and thus the second derivative at zero is negative in each case and zero lags are optimal. $\square$

Note that we have shown that zero lags to be *locally* optimal, but directly calculating welfare as a function of the two lags establishes that they are in fact globally optimal.

Non–constant prices between reviews

Proof of Result 9: For this we shall use different techniques to those in the rest of the paper — instead of Bellman's dynamic programming method, we will use Pontryagin's maximum principle. In this section we are only concerned with the firm's behaviour between reviews, and so we will take the length of the review as fixed at t.

We shall look for a different representation of the firm's costs. For a given output path, $q(s)$, define the functions $\bar{h}(s)$ and $\underline{h}(s)$ to be the probability of being in the low cost state at time s from the end of the review, given that initial costs were either high or low respectively. Then

$$(12) \qquad \bar{h}(s) = \frac{\bar{\pi}(t) - \bar{\pi}(s)}{\underline{\pi}(s) - \bar{\pi}(s)}, \quad \underline{h}(s) = \frac{\underline{\pi}(t) - \bar{\pi}(s)}{\underline{\pi}(s) - \bar{\pi}(s)}.$$

Using this notation we can express the firm's total discounted costs, given the output path

...m's effort outlay schedules, $\bar{x}(\cdot)$ and $\underline{x}(\cdot)$, as

$$\ldots) = \int_0^t e^{-\rho(t-s)}[\bar{h}(s)(q(s)\underline{c} + \underline{x}(s)) + (1-\bar{h}(s))(q(s)\bar{c} + \bar{x}(s))]\,ds,$$

...for $\underline{C}(q(.))$. Clearly $\bar{h}(t) = 0$ and $\underline{h}(t) = 1$. In addition, $\bar{h}(.)$ obeys the ...uation (from equations (4′)):

$$\frac{d\bar{h}}{ds}(s) = \bar{h}(s)[\bar{m}(\bar{x}(s)) + \underline{m}(\underline{x}(s))] - \bar{m}(\bar{x}(s)),$$

...for $\underline{h}$. Therefore, in this context the firm's problem is to minimize (13) ...(14) and $\bar{h}(t) = 0$ [resp. $\underline{h}(t) = 1$]. This is a Hamiltonian–type problem with ... control variables, and h as the state variable. This will (almost) always be ...lved using Hamiltonian methods.[24] Therefore, if $\mu(s)$ is the correctly chosen ... variable, the firm will choose $\underline{x}$, $\bar{x}$ and h at each moment in time in order ...

$$\int_0^t e^{-\rho(t-s)}[g(q\underline{c} + \underline{x}) + (1-g)(q\bar{c} + \bar{x})] + \mu\{g[\bar{m}(\bar{x}) + \underline{m}(\underline{x})] - \bar{m}(\bar{x})) + \mu' g\}\,ds.$$

... problems (R) from section 6 in continuous time, the regulator will wish to:

...x: $\int_0^t e^{-\rho(t-s)}u(q(s))\,ds - \bar{C}(q(.))$
...(.)

... to: $\int_0^t e^{-\rho(t-s)}R(q(s))\,ds \geq \bar{C}(q(.))$,

... maximization occurs over all output paths, $q(.)$, and $\bar{C}(q(.))$ is the solution to ...iltonian problem (15) above. If $\mu > 0$ is the lagrange multiplier associated with ...e problem, then the first–order conditions for the optimal output path are:

$$p(s)[1 - \frac{\alpha}{\eta(s)}] = h(s)\underline{c} + (1-h(s))\bar{c},$$

... $h(\cdot)$ is either $\bar{h}(\cdot)$ or $\underline{h}(\cdot)$ depending on initial costs, $\eta(s)$ is the (modulus of) ...y of demand at time s, and $\alpha = \frac{\mu}{1+\mu}$ $(\alpha > 0)$. Since the right hand side of the ...irst–order condition is (expected) marginal cost of producing at time s from the

... an extensive exposition of the methods of solving such problems, see Kamien and ...ts (1981).

end of the period, (16) tells us that the price–cost markup is inversely proportional to the elasticity at time s — a familiar result. (However, it is interesting to see that "cost" in the above condition does not include effort.) If, as seems reasonable, elasticity falls (weakly) with output, then the function:

$$q \longmapsto p(q)\left[1 - \frac{\alpha}{\eta(q)}\right]$$

will be decreasing. The right hand side of (16) is decreasing in h, and therefore, the time path of $q(\cdot)$ follows that of $h(\cdot)$.[25] □

APPENDIX II

Proof of Result 5:

(i) Substituting (2) into (1a) gives

$$\bar{D}_r \equiv \bar{C}_r - \delta\bar{C}_{r-1} = q\underline{c} + (\bar{x}_r - \bar{a}/\bar{a}'). \tag{17}$$

But the function $x \mapsto x - \bar{a}/\bar{a}'$ is necessarily decreasing since it has derivative equal to $\bar{a}\bar{a}''/(\bar{a}')^2$, which is negative since $\bar{a}$ is both positive and concave. Result 1 tells us that $\bar{x}_r$ is increasing in r, and so we have shown that $\bar{D}_r$ is decreasing in r (for $r \geq 1$). By construction

$$\bar{C}_r = \sum_{i=1}^{r} \delta^{r-i}\bar{D}_i \quad (r \geq 1).$$

We claim that $\frac{\bar{C}_r}{1-\delta^r}$ is decreasing in r (for $r \geq 1$). But we have that

$$\frac{\bar{C}_r}{1-\delta^r} - \frac{\bar{C}_{r-1}}{1-\delta^{r-1}} \overset{\text{sgn}}{=} \frac{1-\delta^r}{1-\delta}\bar{D}_r - \sum_{i=1}^{r-1}\delta^{r-1-i}\bar{D}_i \quad (\text{for } r \geq 2)$$

$$= \sum_{i=1}^{r-1}\delta^{r-1-i}(\bar{D}_r - \bar{D}_i)$$

$$< 0 \quad (\text{since } \bar{D}_i \text{ is decreasing}).$$

Therefore, since $\bar{p}(r-1)$ is the lowest price such that

$$\frac{1-\delta^r}{1-\delta} R(p) \geq \bar{C}_{r-1}(p),$$

[25] This observation allows us to deduce that, since $h(.)$ is never constant, it is never optimal to keep output constant within the review period.

own that

$$\frac{1-\delta^r}{1-\delta} R(p) \geq \bar{C}_r(\bar{p}(r-1)),$$

$$\bar{p}(r) \leq \bar{p}(r-1) \quad (\text{for } r \geq 2).$$

demonstration that $\underline{p}(r)$ is increasing in r is analogous to part (i):
ting (2) into (1b) gives

$$\underline{D}_r \equiv \underline{C}_r - \delta\underline{C}_{r-1} = q\underline{c} + (\underline{x}_r + (1-\underline{a})/\underline{a}').$$

same method as above tells us that in this case $\underline{D}_r$ is *increasing*, and thus that
creasing in r. □

References

ARMSTRONG, C.M. and VICKERS, J.S. "Welfare Effects of Price Discrimination by a Regulated Monopolist". Mimeo, Oxford University, April 1990.

BAILEY, E.E. "Innovation and Regulation." *Journal of Public Economics*, Vol.3 (1974), pp.285–295.

BAILEY, E.E. "Innovation and Regulation: A Reply." *Journal of Public Economics*, Vol.5 (1976), pp.393–394.

BARON, D.P. and BESANKO, D. "Regulation, asymmetric information, and auditing." *RAND Journal of Economics*, Vol.15 (1984), pp.447–470.

BAUMOL, W.J. and KLEVORICK, A. "Input choices and rate of return regulation." *Bell Journal of Economics* Vol.1 (1970), pp.162–190.

BAWA, V.S. and SIBLEY, D.S. "Dynamic behaviour of a firm subject to stochastic regulatory review." *International Economic Review*, Vol.21 (1980), pp.627–642.

BÖS, D. "Notes on Price–Cap Regulation." Discussion Paper A–278, Rheinische Friedrich–Wilhelms–Universität, Bonn, 1990.

BRADLEY, I. and PRICE, C. "The Economic Regulation of Private Industries by Price Constraints." *Journal of Industrial Economics*, Vol.37 (1988), pp.99–106.

FREIXAS, X., GUESNERIE, R. and TIROLE, J. "Planning under Incomplete Information and the Ratchet Effect." *Review of Economic Studies* Vol.52 (1985), pp.173–191.

KAMIEN, M.I. and SCHWARTZ, N.L., *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management.* Amsterdam: North Holland, 1981.

LEWIS, T.R. and SAPPINGTON, D.E.M. "Regulating a Monopolist with Unknown Demand and Cost Functions." *RAND Journal of Economics*, Vol.19 (1988), pp.438–457.

LESOURNE, J. "Innovation and Regulation: A Comment." *Journal of Public Economics*, Vol.5 (1976), pp.389–392.

LITTLECHILD, S.C. "Regulation of British Telecom's Profitability." Report to the Secretary of State, Department of Industry, London, February 1983.

RAND JOURNAL, Symposium on Price–cap Regulation. *RAND Journal of Economics*, Vol.20 (1989), pp.369–472.

SCHMALENSEE, R.L. "Good Regulatory Regimes." *RAND Journal of Economics*, Vol.20 (1989), pp.417–436.

VICKERS, J.S. and YARROW, G.K., *Privatization: An Economic Analysis.* Cambridge: MIT Press, 1988.

WEITZMAN, M.L. "The Ratchet Principle and Performance Incentives." *Bell Journal of Economics* Vol.11 (1980), pp.302–308.

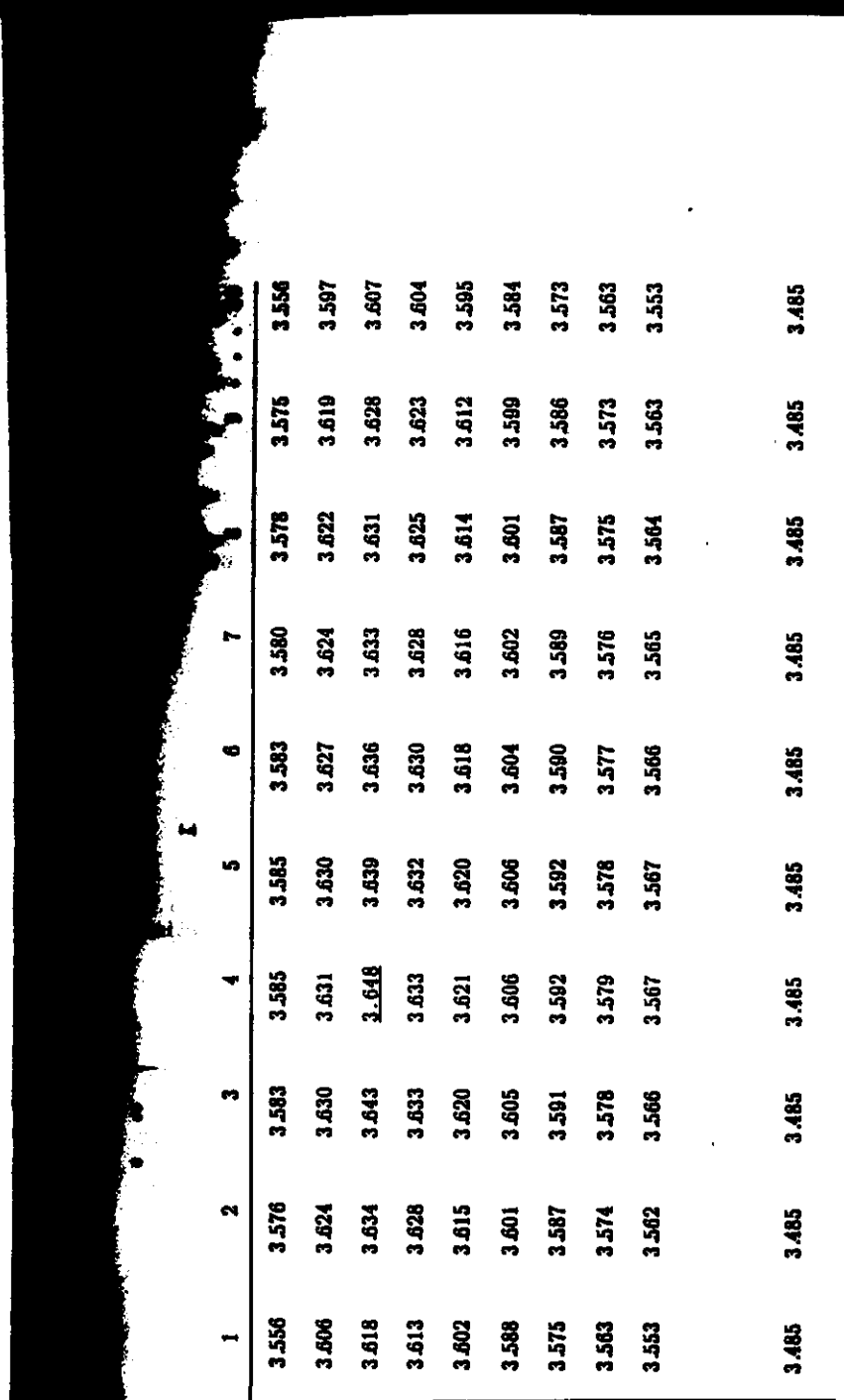

TABLE 2: The effect of changing demand on optimal lags

d =	.6	.9	1.0	1.1	5
F	1	3	4	6	∞
I	1	4	6	9	∞

TABLE 3: The effect of changing the responsiveness to effort on optimal lags

b =	.05	.1	.15	.17	.19	.25
F	1	2	3	3	3	3
I	1	1	1	3	7	∞

CES Working Paper Series

01 Richard A. Musgrave, Social Contract, Taxation and the Standing of Deadweight Loss, May 1991

02 David E. Wildasin, Income Redistribution and Migration, June 1991

03 Henning Bohn, On Testing the Sustainability of Government Deficits in a Stochastic Environment, June 1991

04 Mark Armstrong, Ray Rees and John Vickers, Optimal Regulatory Lag under Price Cap Regulation, June 1991

05 Dominique Demougin and Aloysius Siow, Careers in Ongoing Hierarchies, June 1991

06 Peter Birch Sørensen, Human Capital Investment, Government and Endogenous Growth, July 1991