Hotelling Under Pressure*

Soren T. Anderson[†] Ryan Kellogg[‡] Stephen W. Salant[§]

January 3, 2014

Abstract

We show that crude oil production from existing wells in Texas does not respond to current or expected future oil prices, contradicting a basic prediction of Hotelling's (1931) canonical model of exhaustible resource extraction. In contrast, the drilling of new wells exhibits a strong price response, as does the rental rate on drilling rigs. To explain these observations, we reformulate Hotelling's model as a drilling problem, in which firms choose when to drill new wells, but flow from existing wells is limited by a capacity constraint that decays toward zero as reservoir pressure declines. This drilling problem implies a modified Hotelling rule for discounted revenue flows net of drilling costs. Our model rationalizes the empirical findings from Texas and can replicate several other well-known features of the oil industry: local production peaks, backwardated price expectations following unanticipated positive demand shocks, and expectations that prices will rise faster than the interest rate following large, unanticipated negative demand shocks.

JEL classification numbers: Q3, Q4
Key words: crude oil prices; oil extraction; decline curve; oil drilling; rig rental rates; exhaustible resource; Darcy's Law

*For helpful comments and suggestions, we are grateful to Lutz Kilian, Jinhua Zhao, and seminar and conference audiences at AERE, the Energy Institute at Haas Energy Camp, the MSU/UM Energy and Environmental Economics Summer Workshop, Michigan State, the Occasional California Workshop in Environmental and Resource Economics, and the University of Michigan. Excellent research assistance was provided by Dana Beuschel.

[†]Michigan State University and NBER. Email: sta@msu.edu
[‡]University of Michigan and NBER. Email: kelloggr@umich.edu
[§]University of Michigan and Resources for the Future. Email: ssalant@umich.edu

1 Introduction

Ever since Hotelling's (1931) seminal article, economists have modeled the optimal extraction of an exhaustible resource as a cake-eating problem, in which extraction today is traded off directly against the opportunity of extracting tomorrow, leading—in the canonical model—to a production path in which price less per-barrel marginal extraction cost rises at the rate of interest. A common application of the Hotelling logic is to production from an oil reserve. We argue, however, that observed patterns of oil production and prices are not compatible with Hotelling (1931) nor any of its subsequent modifications in the literature. Instead, we show that to replicate structurally the dynamics of oil supply, economists must recognize the physical constraints on oil extraction rates that are imposed by fluid flow dynamics and declines in underground pressure and recast the Hotelling model as a well-drilling problem (or a keg-tapping problem, if one wishes to maintain an analogy to food and drink) rather than a cake-eating problem.

We begin by showing empirically, using data from Texas over 1990–2007, that oil production from drilled wells declines asymptotically toward zero and is not affected by shocks to either spot or expected future oil prices. We also show empirically that this behavior, which is inconsistent with most extraction models in the literature, is not simply driven by institutional factors, such as common-pool problems or oil lease contract provisions. Instead, we argue that the observed decline in production is rationalized by the cost structure of the oil industry and by the flow constraints imposed by underground pressure dynamics.

When a well is first drilled, the pressure in the underground oil reservoir is high. Thus, production may initially be rapid, since the maximum rate of fluid flow is roughly proportional to the pressure difference between the reservoir and the surface (or between the reservoir and the bottom of the well if the well has a pump). Over time, however, the reservoir pressure abates as less and less of the original oil remains, and the well's flow must gradually decay toward zero. Thus, while extractors can choose when to drill their wells, the maximum rate of oil flowing from these wells at any time is constrained by pressure. This constraint is roughly proportional to the volume of recoverable oil remaining, since it is this oil that determines the pressure underground. Because the marginal cost of producing oil from a drilled well is very low relative to oil prices and drilling costs, and because the flow constraint greatly dampens the incentive to defer production in anticipation of higher prices in the future, the flow constraint typically binds in equilibrium, yielding the asymptotic decline in production that we observe in the data.

While production from drilled wells is insensitive to oil prices, our Texas data also show that the drilling of new oil wells and the rental price of drilling rigs both respond strongly to oil price shocks. These findings imply an upward-sloping supply curve of drilling rigs for rent and crews to operate them. They also imply that if extractors are to arbitrage prices over time, such arbitrage would be based mainly on the timing of drilling.

These empirical results motivate us to recast Hotelling's (1931) canonical model as a drilling problem rather than a production problem. Consistent with our empirical findings, extractors in our model choose when to drill their wells, but the maximum flow from these wells is geologically constrained due to pressure, such that this maximum flow decays asymptotically toward zero as more oil is extracted. Thus, there are two state variables in our model: (1) the maximum flow of oil from existing wells, which rises as more wells are drilled and declines as more oil is produced, and (2) the number of new wells remaining to be drilled, which obviously declines as more wells are drilled.

Since there are millions of operating wells in the world, we focus on the realistic case of many infinitesimally small wells rather than a finite number of large wells. In this case, the planner's problem can be decentralized as a competitive equilibrium. We show that a modified Hotelling rule holds in our model: whenever drilling occurs, the discounted revenue stream that flows from each well, less the marginal cost of drilling it, rises at the rate of interest. We also show that the flow constraint will always bind in equilibrium under fairly weak sufficient conditions, so that the only relevant margin on which extractors control production is the rate at which they drill new wells, consistent with our empirical findings.

We show that the traditional Hotelling path can, via judicious control of the drilling rate, still emerge as optimal—but only under the assumptions that marginal drilling costs are constant and that the marginal utility of consuming oil is unbounded as consumption declines to zero. The first assumption is contradicted by our data on rental rates for drilling rigs, and the second is dubious given the existence of viable alternative fuel sources. Nevertheless, it is not the constraint on oil flow by itself that undermines Hotelling's path, but rather the interaction of this constraint with the nature of drilling supply and oil demand.

When we relax these two assumptions, our model easily and naturally replicates key qualitative features of the crude oil extraction industry. We first consider the behavior of private oil developers in a local region facing an exogenous price of oil and an upward-sloping local supply of drilling rigs for rent. This simple model generates the classic "peak oil" production profile observed in the world at every level of aggregation, from individual oil fields to large oil-producing regions (Hamilton 2013). Production initially rises as developers in a region rapidly drill wells but then inevitably declines as drilling slows and the flow from existing wells decays. Consistent with our data from Texas, an unanticipated, sudden increase in the oil price leads to an immediate increase in drilling activity and rig rental rates, shifting the peak in production earlier in time and making it more pronounced.

We next consider equilibrium with endogenous oil prices, beginning with a particularly tractable case in which there is an unlimited number of wells to be drilled—an approximation to a world in which the stock of oil is vast relative to current demand. We show that in this case our model bears a strong resemblance to a standard macroeconomic q model of investment in the presence of convex adjustment costs, with the decaying flow from existing wells playing the role of an industry's depreciating capital stock and the drilling of new wells playing the role of investment. This model naturally leads to an equilibrium that has a steady state with non-zero rates of drilling and extraction, with extraction always capacity-constrained. When we then impose resource scarcity onto this model, we return to the result that, starting from zero capacity, oil production will initially rise and then decline over time, ultimately to zero in the limit, with drilling also declining to zero. In this case, oil prices will initially fall as production builds but must eventually increase as production declines.

Using our endogenous price models, we also study how the equilibrium paths for extraction, drilling, and the oil price respond to unanticipated shocks to oil demand, motivated by the existence of numerous such shocks in the recent past (see, for example, Kilian (2009) and Kilian and Hicks (2013)). We show that positive demand shocks lead to an immediate increase in oil prices, drilling, and rig rental prices, and that oil prices may then subsequently fall if the increase in the drilling rate is sufficient to cause production to increase. These results are reversed for negative demand shocks, which can—if large enough—cause oil prices to temporarily rise faster than the rate of interest following the initial drop in prices, even though production and drilling need not fall all the way to zero. Thus, our model generates the same qualitative responses to demand shocks that we observe in real-world production data and futures markets.

Within the Hotelling literature, our paper is most closely related to models in which firms face convex costs to expand reserves or production capacity (Pindyck 1978; Switzer and Salant 1986; Livernois and Uhler 1987; Holland 2008).[1] Such models can also generate an initial period of falling prices as capacity builds, with prices eventually rising as production inevitably declines. But this literature ignores the geological fact that as the underground oil stock disappears, so too does the pressure that determines the maximal rate of extraction. It is this pressure—not the oil itself—that is scarce. When we take this geological relationship into account, there is never a final phase during which extraction occurs at an unconstrained

[1] The economics literature following Hotelling (1931) has also modeled, among other things: extraction costs that increase with the rate of production from a reserve or over time as a reserve is depleted; heterogeneity in extraction costs across different reserves; cost-reducing technological change; demand growth; extraction by firms with market power; exploration and development of new reserves; and uncertainty of various kinds (e.g., in demand or resource discovery). For a review of the theoretical literature through the 1970s, see Devarajan and Fisher (1981). For more recent reviews, see Krautkraemer (1998) and Gaudet (2007). For an accessible theoretical primer, see Salant (1995). In addition, a multiple demand curve model has been developed to take account of users in different countries who must pay different transport costs or taxes for the same resource or must pay to convert one resource before using it as a substitute for another type of resource. For a synthesis of this recent literature on "multiple demand curves" see Gaudet and Salant (2014).

rate, as in Switzer and Salant (1986). In our model, the constraint decays over time, shifting relentlessly toward zero whenever production occurs due to declining reservoir pressure. As a result, the constraint will typically bind, even when production is falling. Thus, unlike these other models, ours can explain why production from drilled wells declines steadily over time yet simultaneously does not respond to price shocks,[2] and why futures markets sometimes forecast prices to rise faster than the rate of interest. By identifying drilling as the means by which extraction capacity expands, we link capacity expansion to drilling activity and the marginal cost of capacity expansion to the rental rates on drilling rigs, each of which can be observed empirically.

A number of papers in the economics literature are, like ours, premised on the idea of an oil production constraint that decays over time due to declining pressure. To the best of our knowledge, however, only Mason and van't Veld (2013) attempts to derive equilibrium outcomes—for oil prices, not drilling rig rental prices, as they assume constant marginal drilling costs—but only in a simplified, two-period version of their model. The remaining papers all treat oil prices and drilling costs as exogenous, focusing on the valuation of individual reserves that have already been developed or on the optimal timing and form of development for individual new reserves (Nystad 1987; Adelman 1990; Davis and Cairns 1998; Cairns and Davis 2001; Thompson 2001; Smith 2012; Cairns forthcoming). Our paper differs in several crucial ways. First, we present an extensive set of empirical results to motivate every assumption in our model. Our key empirical finding—that oil production from drilled wells is unresponsive to price shocks, consistent with a binding capacity constraint—has not, to the best of our knowledge, been documented in prior work. Second, motivated by our empirical results, we develop a new model of resource extraction that includes convex drilling costs and emphasizes the rate of drilling as the central choice variable, rather than the rate of oil production. Third, we use our model to explain why the production

[2]Without a declining capacity constraint, an increasing marginal cost is necessary to explain a decline in production over time (with production necessarily below the constraint). However, in this case production will respond to oil price shocks.

constraint is observed to bind empirically, proving that this must be the case under weak sufficient conditions. Fourth, we extensively explore the short-run and long-run dynamics of drilling, production, and prices implied by this model, showing that they can match those of the real-world oil extraction industry: local oil-producing fields and regions exhibit production peaks, while expected future oil prices can be backwardated after positive demand shocks and can rise faster than the interest rate following negative demand shocks. Finally, we clarify that Hotelling's logic still applies in our model, with the discounted revenue stream minus marginal cost of drilling rising at the rate of interest. None of the other papers in this literature contains the depth or breadth of results presented here.

We also contribute to an extensive empirical literature testing whether the Hotelling Rule (that the resource price minus the marginal extraction cost must rise at the rate of interest) holds in practice, generally finding that it does not.[3] Of course, there are countless reasons why this rule may fail to hold, given the original model's many simplifying assumptions and the difficulty of specifying the parameters of a more realistic model. Thus, rather than test the precise numerical implications of a model that we know is unrealistic, we simply ask whether crude oil production responds *at all* to current and expected future prices in the way that such a model would predict. As we show, new drilling activity increases with spot prices, but production from existing wells is almost entirely price inelastic. These results motivate our rethinking of the canonical model to focus on a capacity constraint that decays with production and new drilling to expand capacity.

Finally, we contribute to a broader empirical literature estimating the response of crude oil production to prices. This work has generally found that production in aggregate is price inelastic, at least in the short run (Griffin 1985; Hogan 1989; Jones 1990; Dahl and Yucel 1991; Ramcharran 2002), though Rao (2010) finds evidence using well-level data that firms

[3]Smith (1979), Slade (1982), and Berck and Roberts (1996) find limited evidence for an upward trend in exhaustible resource prices, but tests based on price alone are not correctly specified unless extraction costs are negligible. Structural econometric papers estimating *in-situ* values, including Miller and Upton (1985), Halvorsen and Smith (1984), Black and LaFrance (1998), and Thompson (2001), find mixed results. See Krautkraemer (1998) and Slade and Thille (2009) for recent reviews.

can shift production across wells in response to well-specific taxes. The empirical response of crude oil supply to prices has important implications for the macroeconomic effects of oil supply and demand shocks, since inelastic supply and demand lead to volatile prices (see Hamilton (2009) and Kilian (2009) for insightful discussions). Our empirical findings provide a micro-foundation for the result that aggregate crude oil production is highly inelastic in the short run, with somewhat more elastic supply over the medium run as new wells come online.[4]

The remainder of the paper proceeds as follows. Section 2 presents our empirical evidence on crude oil production and drilling in Texas. Section 3 recasts Hotelling's model as a drilling problem, deriving our fundamental necessary conditions and showing that Hotelling's path can only emerge as optimal under implausible assumptions, including constant marginal drilling costs. Section 4 relaxes this assumption and discusses the short-run and long-run dynamics of oil prices, production, and drilling, first considering a local region facing an exogenous oil price path and then considering equilibrium with endogenous prices. Section 5 then briefly considers implications of above-ground storage for our results. Finally, section 6 concludes.

2 Empirical evidence from Texas

In this section we study how oil production and drilling in Texas respond to incentives generated by changes in current and expected future oil prices, as revealed in futures markets. We show that production exhibits nearly zero response to current and expected future prices, whereas drilling activity—along with the cost of renting drilling rigs—responds strongly. We then discuss how these results derive from the fundamental cost structure and technology of the crude oil extraction industry, and we rule out several alternative explanations.

[4]Kline (2008) shows that wages and employment in the oil and gas field services industry also increase with crude oil prices—though these labor market effects emerge more slowly.

2.1 Data sources and data cleaning

Our crude oil drilling and production data come from the Texas Railroad Commission (TRRC), covering the period 1990–2007. The drilling data come from the TRRC's "Drilling Permit Master" dataset, which provides the date, county, and lease name for every well drilled in Texas. A lease is a plot of land upon which an oil production company has obtained from the (usually private) mineral rights owner the right to drill for and produce oil and gas. Over 1990–2007, a total of 157,271 new wells were drilled, along with 42,893 "re-entries" of existing wells.[5]

The production data come from the TRRC's "Oil and Gas Annuals" dataset, which records monthly crude oil production at the lease-level.[6] The TRRC production data are at the lease-level rather than the well-level because individual wells are not flow-metered.[7] Thus, we generally cannot observe well-level production, though for some analyses we will take advantage of a set of leases that have a single well.

Our analysis of the production data focuses on whether firms respond to oil price shocks month-to-month by adjusting the flow rates of their existing wells (by speeding up or slowing down the pumping unit),[8] or shutting in (turning off) or restarting their wells. To distinguish these actions from investments in new production capital, such as drilling a new well, we

[5] A re-entry occurs when a rig is used to deepen the well, drill a "sidetrack" well off of the existing well bore, or attempt to stimulate production through perforating or fracturing the oil reservoir—possibly at a different depth than that from which the well was previously producing. These interventions are all similar to (though somewhat cheaper than) drilling a new well in that they require a substantial up-front capital investment and allow access to a previously untapped section of oil-bearing rock.

[6] Due to false zeros in the raw dataset provided by the TRRC for some leases in 1996 and December 2004-2007, we augmented these data with production information scraped from the TRRC's online production query tool. We verified that the online data matched the raw dataset for leases and months not affected by the data error.

[7] Direct production from a well is a mix of oil, gas, and often water. These products must first be separated before metering of oil flow can take place; this separation typically takes place at a single facility that serves all of the wells on the lease. Oil flows from the separation facility into storage tanks, from which production is metered when it is delivered to either a pipeline or truck for sale. Although firms often assess well-level productivity monthly by diverting each well's flow into a small "test separator," these data are not available from TRRC nor would they be particularly reliable for accurately measuring a well's flow over the course of an entire month.

[8] The vast majority of the wells in the dataset are pumped and do not flow naturally. The average lease-month in the data has 2.02 pumped wells and 0.06 naturally flowing wells.

discard from our production data leases in which any rig work took place.[9] In the remaining data, there exist 16,148 leases for which production data are not missing for any month from 1990–2007 and production is non-zero for at least one month, so that our dataset consists of 3,487,968 lease-months of production. The typical oil lease in Texas has a fairly low rate of production, reflecting the fact that oil fields in Texas are mature and have been heavily produced in the past. The average daily lease production in the data is 3.6 barrels of oil per day (bbl/d), with a standard deviation of 18.2 bbl/d. A total of 1,070,632 (31%) of the observed lease-months have zero production, and the maximum observed production is 9,510 bbl/d.

Our oil price data come from the New York Mercantile Exchange (NYMEX) and measure prices for West Texas Intermediate (WTI) crude oil delivered in Cushing, Oklahoma—the most common benchmark for crude oil prices in North America—from 1990–2007. We use the Bureau of Labor Statistics's All Urban, All Goods Consumer Price Index (CPI) to convert all prices to December 2007 dollars. We use the front-month (upcoming month) futures price as our measure of the spot price of crude oil and use prices for longer-term futures contracts to measure firms' price expectations.

The use of futures prices to measure price expectations in not without controversy. Alquist and Kilian (2010), for example, shows that futures markets do not out-perform a simple no-change forecast in out-of-sample forecasting over 1991-2007. We use futures markets here for several reasons. First, NYMEX futures are liquidly traded at the horizons we consider here, and with many deep-pocketed, risk-neutral traders, the futures price should equal the expected future spot price. Second, a majority of oil producers in Texas claim to use futures prices in making their own price projections (Society of Petroleum Evaluation Engineers 1995). Third, Kellogg (forthcoming) shows that the drilling activity of Texas oil

[9]Discarding these leases requires matching the drilling dataset to the production dataset. Accomplishing this match at the lease level is difficult because the datasets must be matched on lease name, which is not consistent across the two datasets. Rather than risk having our production dataset contaminated by leases in which rig activity occurred, we instead conservatively identify all county-firm pairs in which rig work took place and then discard *all* leases corresponding to those county-firm pairs (counties and firms are identified with numeric codes in both datasets and therefore do not suffer match problems).

Figure 1: Crude oil spot prices and futures curves

Note: This figure shows crude oil front month ("spot") and futures prices, all in real $2007. The solid black line represents the spot price of oil. The dashed colored lines represent the futures curves as of December in each year. See text for details.

production firms is more consistent with price expectations based on the futures market than on a no-change forecast. Finally, as we discuss below, we find that firms' on-lease oil stockpiles increase when futures prices are high relative to spot prices, as one would expect if firms' expectations aligned with the market.

Figure 1 shows the time series of crude oil spot prices (solid black line) as well as the futures curves as of December in each year (dashed colored lines).[10] For example, the left-most dashed line shows prices in December 1990 for futures contracts with delivery dates from January 1991 through December 1992. As is clear from the figure, the futures market

[10] We have converted all of the price data so that the slopes of the futures curves in figure 1 and the expected rates of price change used in our analysis reflect real rather than nominal changes. To convert the futures curves from nominal to real, we adjust for both the trade date's CPI and for expected annualized inflation of 2.50% between the trade date and delivery date. The average annual inflation rate from January 1990 to December 2007 is 2.50%, and inflation varies little over the sample. Thus, for example, we convert the nominal prices for futures contracts traded in December 1990 to real price expectations by multiplying by the December 2007 CPI, dividing by the December 1990 CPI, and then dividing each contract price by $1.025^{t/12}$, where t is the number of months between the trade date and the delivery date.

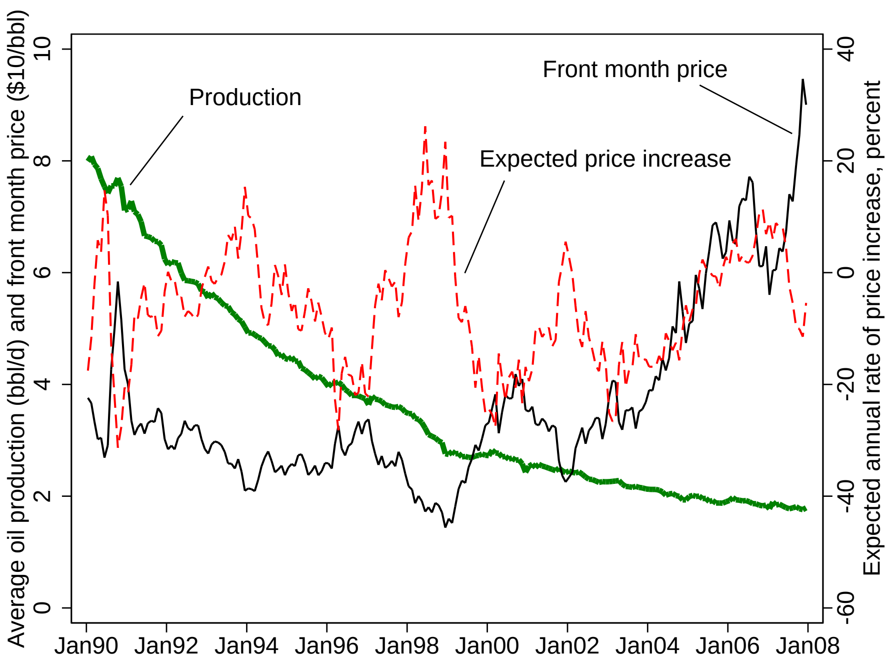
for crude oil is often backwardated (meaning that the future price is lower than the spot price), and was particularly strongly backwardated during the mid-2000s when the spot price was rapidly increasing. Kilian (2009) and Kilian and Hicks (2013) attribute the increase in spot prices during this period to a series of large, positive, and unanticipated shocks to the demand for oil, primarily from emerging Asian markets. However, figure 1 also reveals several periods of contango (meaning that the future price is higher than the spot price) during the sample, particulary during the 1998–1999 low price period, in which the futures curve is upward sloping. Kilian (2009) attributes the low oil prices during this period to a negative demand shock arising from the Asian financial crisis.

Finally, we have also obtained information on rental prices ("dayrates") for drilling rigs from RigData, a firm that publishes reports on the U.S. onshore oil and gas industry. As discussed in Kellogg (2011), the oil production companies that make drilling and production decisions do not drill their own wells but rather contract drilling out to independent service companies that own rigs. Paying to rent a rig and its crew typically constitutes the largest line-item in the overall cost of a well. The data provided by RigData are quarterly, covering Q4 1990 through Q4 2007, and are broken out by region and rig depth rating. We use dayrates for rigs with depth ratings between 6,000 and 9,999 feet (the average well depth in our drilling data is 7,425 feet) for the Gulf Coast / South Texas region. Observed dayrates range from \$6,315 to \$15,327 per day, with an average of \$8,008 (all real December 2007 dollars).

2.2 Production from existing wells does not respond to prices

Our main empirical results focus on the production data for leases on which there was no rig activity from 1990–2007, so that all production in the data come from pre-existing wells. Figure 2 presents daily average production (in bbl/d) for these leases, along with crude oil spot prices and the expected percent change in spot prices over one year. Average production is dominated by a long-run downward trend, with little apparent response either to the spot

Figure 2: Crude oil prices and production from existing wells in Texas



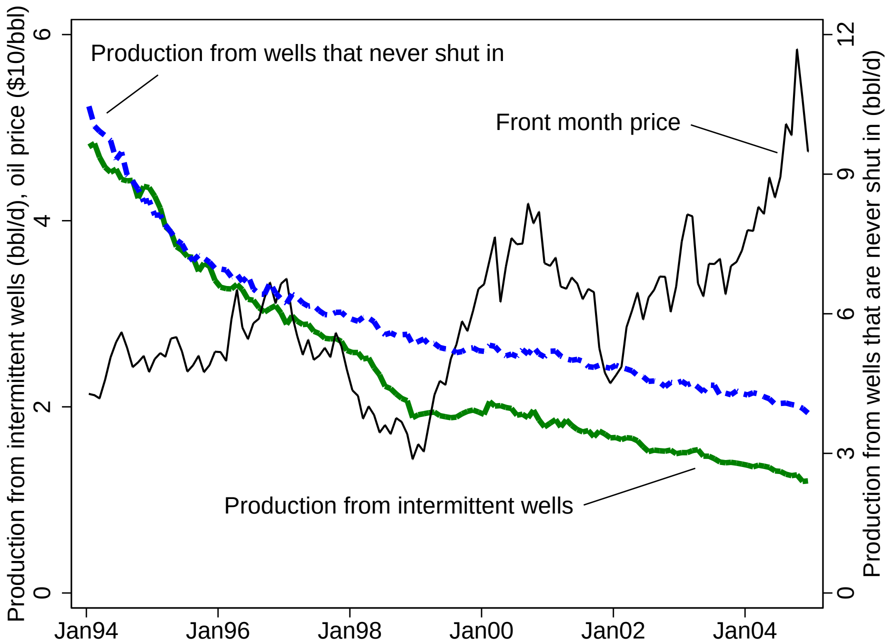
Note: This figure presents crude oil front month ("spot") prices and the expected percent change in prices over one year, as well as daily average lease-level production from leases on which there was no rig activity (so that all production comes from pre-existing wells). All prices are real $2007, and expected price changes are net of inflation. See text for details.

price of oil or to the expected percent change in prices over one year. In addition, we show in appendix A that the pattern shown in figure 2 holds for subsamples of leases that have relatively high production volumes and for production from wells that are drilled during the sample period. Thus, our results are not specific only to the low-volume wells that are the norm in Texas.[11]

These results contrast sharply with predictions from Hotelling models that are standard in the literature, which would predict a complete shutdown of production during periods, such as 1998–1999, when prices are expected to rise more quickly than the rate of interest. Moreover, under a commonly used assumption of increasing marginal extraction costs, Hotelling models will predict that production should increase with the spot price and de-

[11]We have also confirmed the overall lack of response to price incentives using a regression analysis. These results are available upon request.

Figure 3: Intermittent wells versus wells never shut in



Note: This figure shows production from wells that shut in at least once during 1994–2004 versus production from wells that never shut in during 1994–2004. Production data come from leases that had no more than one productive well and never experienced a rig intervention over 1994–2004. Oil prices are real \$2007. See text for details.

crease with the expected future change in price. None of these predictions appears strongly in the time series of production from pre-existing wells. Figure 2 suggests that production did deviate slightly from the long-run trend during the 1998–1999 period during which the spot price fell below \$20/bbl and the expected rate of price change exceeded 10% (and sometimes 20%). In particular, it appears that production accelerated its decline rate in 1998 while prices were falling, leveled off in 1999 while prices were rising, and then resumed its usual decline in 2000.

To assess whether this deviation is real and what mechanism lay behind it, we study whether it arose from wells being shut in or changes in production from active wells. We first isolate the sample to leases that had no more than one flowing well over 1994–2004, so that observed lease-level production during this time can be interpreted as well-level production.

We then split this sample into two groups: wells that are never shut in over 1994–2004 and "intermittent" wells that are shut in at least once. Figure 3 plots the time series of production from these two samples. This figure makes clear that the 1998 deviation from trend was driven entirely by marginal wells that sometimes have zero production. For wells that always produce, there is no adjustment on the intensive margin. It appears that when prices fell in 1998, an unusually large number of wells were shut in, temporarily accelerating the decline. Then, when prices recovered during 1999, many of these wells were returned to production, temporarily slowing the decline. Apart from these deviations, a large response of production to price signals does not appear anywhere in the data.

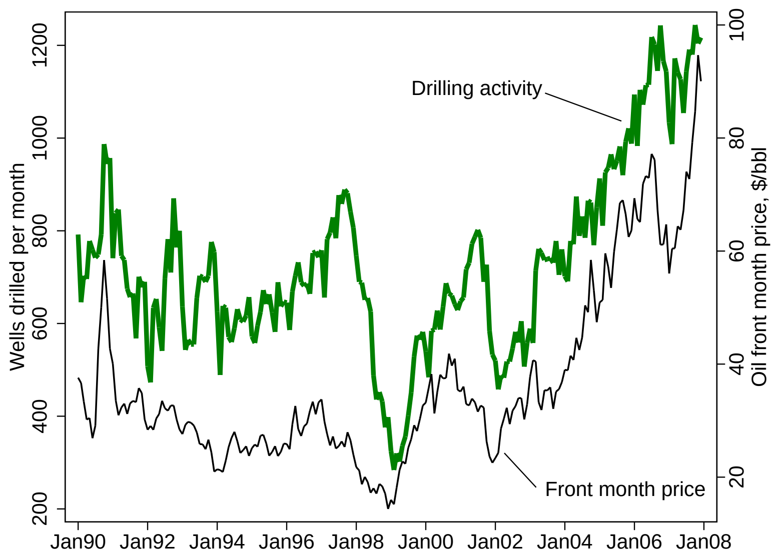
2.3 Rig activity does respond to price incentives

These no-response results based on existing wells stand in stark contrast to new drilling activity in Texas. Figure 4(a) shows the total number of new wells drilled across all leases in our dataset, along with the spot price for crude oil. The figure shows a pronounced monthly correlation between oil prices and new drilling activity. In addition to the graphical evidence presented here, we have found using regression analysis that, for every $1 increase in crude oil prices, the monthly rate at which new wells are drilled increases by a strongly statistically significant 1%–1.5%, while the relationship between drilling and the expected percent change in prices is negative but insignificant. We have also found that the use of rigs to re-enter old wells correlates with oil prices, though not as strongly as the drilling of new wells.

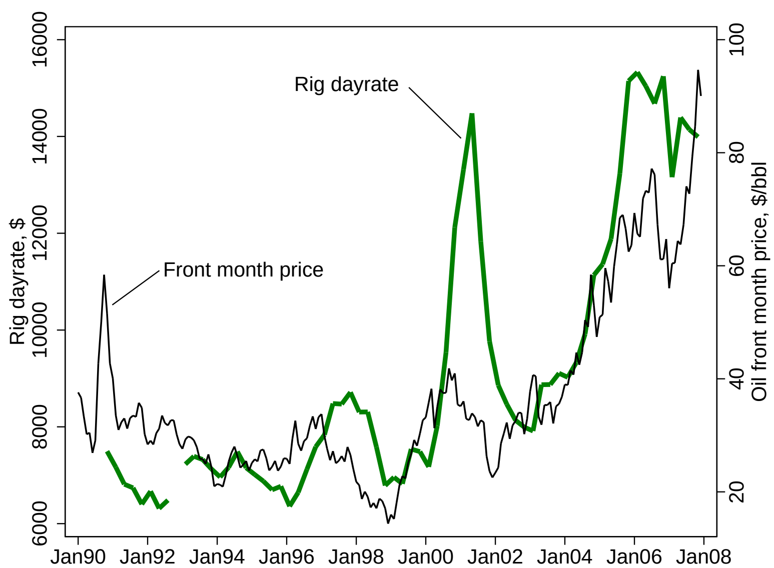
When oil production companies drill more wells in response to an increase in oil prices, more rigs (and crews) must be put into service in order to drill them. Figure 4(b) shows that these fluctuations in rig demand are reflected in a positive covariance between rig dayrates and oil prices. In addition to the graphical evidence presented here, we have found using regression analysis that this relationship is statistically significant and has a magnitude such that a $1 increase in the crude oil spot price is associated with a roughly 2% increase in the average rig dayrate. Thus, as the industry collectively wishes to drill more wells within a

Figure 4: Texas rig activity versus crude oil spot prices

(a) Drilling of new wells



(b) Rig dayrates



Note: Panel (a) shows the total number of new wells drilled across all leases in our dataset. Panel (b) shows dayrates for the Gulf Coast / South Texas region, for rigs with depth ratings between 6,000 and 9,999 feet. The dayrate data are quarterly rather than monthly. Data are available beginning in Q4 1990, and data for Q4 1992 are missing. See text for details.

given time frame, the marginal cost of drilling those wells increases.

2.4 Industry cost structure explains these price responses

The analysis above documents two empirical facts about oil production and drilling in Texas from 1990–2007. First, production from drilled wells is almost completely unresponsive to changes in spot or expected future oil prices, with an exception being an increased rate of shut-ins during the 1998 oil price crash. Second, drilling of new wells responds strongly to oil price changes, and rig dayrates respond commensurately. Here, we argue that these empirical results reflect an industry cost structure with the following characteristics:

1. The maximum rate of production from a well is physically constrained, and this constraint declines asymptotically toward zero as a function of cumulative production; this function is known as a well's production decline curve.

2. The marginal cost of production below a given well's capacity constraint is very small.

3. The fixed costs of operating a producing well are non-trivial. There may also be costs for restarting a shut-in well, but they are often small enough to be overcome.

4. Drilling rigs and crews are a relatively fixed resource, at least in the short run, leading to an upward-sloping supply curve of drilling rigs for rent.

The capacity constraint and low marginal production cost relate to the observation that production from existing wells does not respond to oil price shocks on the intensive margin. Because oil production firms in Texas are price-takers,[12] production will increase with oil prices only if the marginal cost of producing is upward-sloping and of similar magnitude to price. Because we do not observe such a response in the data, it must be that the oil price intersects marginal cost at a vertical, capacity-constrained section of the curve. While the

[12] The market for crude oil is global, and Texas as a whole (let alone a single firm) constitutes only 1.3% of world oil production (Texas and world oil production data for 2007 from the U.S. Energy Information Administration); thus, the exercise of market power by Texas oil producers is implausible.

marginal cost of production below the capacity constraint is not necessarily zero, it must be well below the range of oil prices observed in the data.

The existence of a capacity constraint for well-level production is consistent not only with the data presented above, but also with standard petroleum geology and engineering.[13] As noted recently in the economics literature by Mason and van't Veld (2013), the flow of fluid through reservoir rock and up the well bore is governed by Darcy's Law (Darcy 1856), which stipulates that the rate of flow is proportional to the pressure differential between the reservoir and the well.[14] In the simplest model of reservoir flow, the reservoir pressure is proportional to the volume of fluid in the reservoir. In this case, the maximum flow rate is proportional to the remaining reserves, which results in an exponential production decline curve. More complex cases, which might involve the presence of gas, water, or fractures in the reservoir, may yield a more general hyperbolic decline. Regardless, the physical laws governing fluid flow place a limit on the rate at which oil can be extracted from a reservoir, and this limit declines with the volume of oil remaining. In fact, this issue is sufficiently important that entire sub-fields of petroleum engineering—reservoir engineering and decline curve analysis—are devoted to understanding it.

The existence of a capacity constraint also explains why most producers do not shut in their wells during periods of severe contango, such as 1998, when oil prices are expected to rise more quickly than the rate of interest. A standard Hotelling analysis would suggest that producers should shut in at such times and then recover the lost production once price expectations are backwardated or at least in less severe contango. In the presence of a capacity constraint, however, when a well is restarted after a shut-in, it cannot instantly produce all of the oil that would have been produced were it not for the shut-in. Instead, production returns to its location on the decline curve from the time before the shut-in,

[13] We give the geologic and engineering basis for well-level capacity constraints only a brief treatment here. For a fuller discussion of fluid flow and production decline curves, Hyne (2001) is an excellent source that does not require a geology or engineering background.

[14] Installing a pump on a well effectively eliminates the need for the oil to overcome gravity as it rises up the well. Darcy's Law still binds on the flow of oil through the reservoir into the bottom of the well.

so that the shut-in effectively pushes the entire production profile back in time. Thus, for shutting-in a well to be optimal, prices must be expected to rise more quickly than the rate of interest for a long period of time, spanning most of a well's remaining life. Such long-lived contango does not occur in the futures data, as can be seen in figure 1.[15]

Per figure 3, some relatively low-volume wells were shut in during 1998. These shut-ins are consistent with the existence of fixed production costs, which intuitively arise from the need to monitor and maintain surface facilities such as pumps, flowlines, and separators so long as production is nonzero. When the oil price fell in 1998, production from these wells was no longer sufficient to cover their fixed costs, explaining the decision to shut in. When oil prices subsequently recovered, many of these wells restarted, suggesting that start-up costs are relatively minor.

In appendix A, we consider and rule out alternative explanations for the lack of response of oil production to oil prices. We show that the overall lack of price response cannot be explained by (1) well-specific production quotas (because production quotas are not binding); (2) by leasing agreements that require non-zero production (because multiple-well leases show the same results); (3) by races-to-oil induced by open-access externalities within oil fields (because fields controlled by a single operator show the same results); or (4) by producer myopia or price expectations that are not aligned with the futures market (because producers respond to high futures prices by stockpiling oil above ground).

[15]For instance, in December 1998, the price for delivery in December 1999 is 21.4% greater than the front month price, indicating severe short-run contango. However, the price for delivery in December 2000 is only 10.3% greater than that for December 1999, and the price for delivery in December 2001 is only 4.2% greater than that for December 2000. Thus, shutting in a well in December 1998, even for a brief period, causes production to be delayed over time periods when prices are not expected to rise more quickly than the rate of interest. In principle, deferring production can still be optimal if, as in this example, the price is expected to increase sharply and then remain persistently high. Our calculations suggest, however, that such a price increase would need to be several times greater than even this extreme episode.

3 Recasting Hotelling as a drilling problem

In this section, we develop a theory of optimal drilling that closely follows the industry cost structure described above. After setting up the problem, we derive conditions that necessarily hold at any optimum. This leads to our fundamental result for how drilling revenues and costs should evolve at the optimum or in the associated competitive equilibrium.

3.1 Planner's problem and definitions

The planner's problem is given by:

$$\max_{F(t),a(t)} \int_{t=0}^{\infty} e^{-rt} \left[U(F(t)) - D(a(t))\right] dt \tag{1}$$

subject to

$$0 \leq F(t) \leq K(t) \tag{2}$$

$$a(t) \geq 0 \tag{3}$$

$$\dot{R}(t) = -a(t),\ R_0 \text{ given} \tag{4}$$

$$\dot{K}(t) = a(t)X - \lambda F(t),\ K_0 \text{ given,} \tag{5}$$

where $F(t)$ is the rate at which oil is flowing to market at time t (a choice variable), $a(t)$ is the rate at which new wells are drilled (a choice variable), $K(t)$ is the capacity constraint on oil flow (a state variable), and $R(t)$ is the amount of wells that remain untapped (a state variable). The instantaneous utility derived from oil flow is given by $U(F(t))$, where $U(\cdot)$ is increasing and weakly concave; we assume $U(0) = 0$. The total instantaneous cost of drilling wells at rate $a(t)$ is given by $D(a(t))$, where $D(\cdot)$ is increasing and convex. We denote the derivative of the total cost function as $d(a(t))$ and assume that $d(0) \geq 0$. Consistent with our empirical results from Texas, we assume a trivially low (i.e., zero) marginal cost of extraction up to the constraint. We ignore any fixed costs for operating, shutting in, or restarting wells since such costs are only relevant for marginally productive wells or when oil prices are low—and since accounting for these costs would complicate the analysis substantially.[16] Utility and drilling costs are discounted continuously at rate r. If wealth-maximizing agents are involved, they also discount profit flows at rate r.

Condition (4) describes how the stock of untapped wells $R(t)$ evolves over time. The planning period begins with a continuum of untapped wells of measure R_0, and the stock of untapped wells thereafter declines one-for-one with the rate of drilling. Condition (23) describes how the oil flow capacity constraint $K(t)$ evolves over time. The planning period begins with capacity constraint K_0 inherited from previously tapped wells. As discussed above, the maximum rate of oil flow from a tapped well depends on pressure in the well, which is approximately proportional to the oil that remains underground. Thus, oil flow $F(t)$ erodes capacity at rate λ (the second term on the right). The planner can, however, rebuild capacity by drilling new wells. Thus, the rate of drilling $a(t)$ relaxes the capacity constraint at rate X (the first term on the right), where we can interpret X as the maximum flow from a newly drilled well—or to be more precise, a unit mass of newly drilled wells.[17] Thus, if production is set to the constraint ($F(t) = K(t)$) and there are no new wells being drilled ($a(t) = 0$), oil flow decays exponentially toward zero at rate λ with the capacity constraint. The total amount of oil in untapped wells is given by $R(t)X/\lambda$, so that the total amount of oil underground at the outset of the planning period is given by $S = (K_0 + R_0X)/\lambda$.[18]

[16] Among other complications, allowing for such fixed costs would require us to model, at each t, how the quantity of oil reserves remaining in tapped wells is distributed across the continuum of tapped wells, along with the shadow opportunity cost associated with extracting more oil from every point in this distribution.

[17] If the drilling cost is strictly convex, the planner would never find it optimal to set up a mass of wells instantaneously at $t = 0$—or at any other time—and the stock of untapped wells and oil flow capacity constraint would both evolve continuously over time. When the drilling cost is linear, however, such "pulsing" behavior may be optimal, leading to discontinuous changes in these state variables.

[18] A mathematically equivalent formulation of our problem would involve imposing resource scarcity directly on cumulative extraction by replacing condition (4) with $\dot{S}(t) = -F(t)$, where $S(t) = (K(t)+R(t)X)/\lambda$ is the total amount of oil remaining underground at time t. We find that our current formulation leads to necessary conditions that are easier to interpret and manipulate.

3.2 Necessary conditions and their implications

Following Léonard and Long (1992), the current-value Hamiltonian-Lagrangean of this maximization problem is given by:

$$H = U(F(t)) - D(a(t)) + \theta(t)[a(t)X - \lambda F(t)] + \gamma(t)[-a(t)] + \phi(t)[K(t) - F(t)], \tag{6}$$

where $\theta(t)$ and $\gamma(t)$ are the co-state variables on the two state variables $K(t)$ and $R(t)$, and $\phi(t)$ is the shadow cost of the oil flow capacity constraint.

Necessary conditions are given by:

$$F(t) \geq 0,\ U'(F(t)) - \lambda\theta(t) - \phi(t) \leq 0,\ \text{comp. slackness } (c.s.) \tag{7}$$

$$F(t) \leq K(t),\ \phi(t) \geq 0,\ c.s. \tag{8}$$

$$a(t) \geq 0,\ \theta(t)X - d(a(t)) - \gamma(t) \leq 0, c.s. \tag{9}$$

$$\dot{R}(t) = -a(t),\ R_0 \text{ given} \tag{10}$$

$$\dot{\gamma}(t) = r\gamma(t) \tag{11}$$

$$\dot{K}(t) = a(t)X - \lambda F(t),\ K_0 \text{ given} \tag{12}$$

$$\dot{\theta}(t) = -\phi(t) + r\theta(t) \tag{13}$$

$$K(t)\theta(t)e^{-rt} \to 0 \text{ and } R(t)\gamma(t)e^{-rt} \to 0 \text{ as } t \to \infty \tag{14}$$

We begin by interpreting condition (9), which characterizes drilling incentives. The $\theta(t)X$ term is the stream of marginal utilities (or revenues) that a well drilled at time t will generate over its lifetime, discounted back to time t, while the $d(a(t))$ term is the marginal cost (or rental rate) of drilling a well at time t when $a(t)$ other wells are being drilled. Thus, the $\theta(t)X - d(a(t))$ term is the marginal profitability of drilling a well at time t. The term $\gamma(t)$, which is the co-state variable on the stock of remaining wells, can be interpreted as the shadow cost (in current value terms) of having marginally fewer than $R(t)$ wells remaining at time t. Note that condition (11) implies that $\gamma(t) = \gamma_0 e^{rt}$, where $\gamma_0 \geq 0$ is a constant.

Thus, when drilling occurs ($a(t) > 0$), conditions (9) and (11) together imply that:

$$\theta(t)X - d(a(t)) = \gamma_0 e^{rt}. \tag{15}$$

Intuitively, whenever drilling occurs ($a(t) > 0$), the net marginal value of drilling *wells* must rise at the rate of interest so that every well, no matter when it is drilled, yields the same net payoff in present-value terms. This makes perfect sense since the total number of wells which can be drilled is fixed. If a planner could earn strictly higher net payoff by reducing drilling when, in present value terms, it was marginally less valuable and expanding drilling when it was more valuable, then she would surely do so. Condition (9) is analogous to the standard Hotelling rule, which states that, constrained to a fixed volume of oil, the planner should extract so that the net marginal value of extracting *barrels* rises at the rate of interest. Thus, every barrel extracted yields the same net payoff in present-value terms. Since this is a drilling problem, however, the Hotelling-like intuition applies to wells, not barrels.

Equation (15) holds whenever drilling occurs. But when would the planner choose not to drill? There are several such scenarios. For example, if the inherited constraint on oil flow were very high, implying a large stock of oil in previously tapped wells, then the planner might postpone drilling the first well. Alternatively, if prices needed to rise quickly over an interval—perhaps following a large, negative demand shock or across a highly convex region of the demand curve—then production might be set below the constraint, implying a likely pause in drilling. Or finally, if marginal utility were bounded, then the revenue stream from drilling would also be bounded, implying that the last well would need to be drilled in finite time.

We now interpret condition (7), which characterizes oil extraction incentives. The $U'(F(t))$ term is the marginal utility of oil flow (or price of oil) at time t. The $\lambda\theta(t)$ term captures the opportunity cost of this flow in terms of forgone future utility: extraction erodes the constraint on future oil flow at rate λ, due to the decline in pressure underground, while

θ is the marginal value of extraction capacity. When the constraint on oil flow is binding ($F(t) = K(t), \phi(t) > 0$), the marginal utility of oil flow is strictly greater than this opportunity cost ($U'(F(t)) > \lambda\theta(t)$), so that there is no incentive to defer extraction.

Note that conditions (7) and (13) admit the possibility that, when production is constrained and capacity is declining, the oil price may rise at a rate strictly greater than r over an interval without inducing extraction below the constraint, provided that this interval of rapidly rising prices does not last too long.[19] This possibility relates to the empirical observations from the 1998–1999 period, when the futures market indicated an expectation that the oil price would rise very quickly for several months, yet oil production still appeared to decline along the capacity constraint. For a theoretical example of when such a situation may occur, suppose that the inverse demand for oil $P(F) = U'(F)$ is linear and that all wells have been drilled. In this case, when production is constrained, $\dot{P}(t)/P(t)$ must strictly decrease over time as the oil price approaches its maximum value of $P(0)$. Thus, if the initial production capacity is sufficiently high and inverse demand is sufficiently steep (the local inverse demand elasticity $\eta(F) \equiv -F\frac{P'(F)}{P(F)}$ must be sufficiently large that $\lambda\eta(F) > r$)), the oil price will initially increase at a rate strictly greater than r, but the rate of increase must eventually fall to a value strictly less than r. In this case, deferring production during the initial phase when price is rising more quickly than r may not be optimal, since the capacity constraint implies that any deferred production can only be recovered over the course of the entire remaining lifetime of the stock of drilled wells.[20] Thus, some non-zero fraction

[19] To see this possibility formally, note that in this situation, $\theta(t)$ rises at a rate strictly less than r, so that $\phi(t)$ must be rising at a rate strictly greater than r (otherwise, condition (7) will not hold with equality). These dynamics can persist over a brief interval without violating any necessary conditions. However, they cannot persist forever, since $\theta(t)$ would eventually be driven below zero (a contradiction). So if production is to remain constrained, the rate of oil price increase must at some point decrease to a value less than r (this in turn requires the inverse demand elasticity to strictly decrease as production decreases).

[20] To see this more formally, suppose that production is reduced below the constraint by an amount ϵ for some time interval of length δ. Then, the total amount of oil production deferred equal $\epsilon\delta$, and the available production capacity after this time interval will be $\lambda\epsilon\delta$ greater than what it otherwise would have been. This additional capacity is not infinite, so the entire deferred volume cannot be extracted immediately. Moreover, the additional capacity declines with production. Thus, the fastest way to extract the deferred production is to produce at the capacity constraint, in which case the rate of production declines exponentially at rate λ, and the deferred production is only completely recovered in the limit as $t \to \infty$.

of the deferred production will not be recovered until the phase when price is rising more slowly than r, providing a disincentive to produce below the capacity constraint in the initial period.

When the constraint on oil flow is slack ($F(t) < K(t), \phi(t) = 0$), then the marginal utility of oil flow exactly equals the marginal opportunity cost of extraction in terms of forgone future utility ($U'(F(t)) = \lambda\theta(t)$). Moreover, condition (13) then implies that $\dot{\theta}(t) = r\theta(t)$, such that $\theta(t)$ rising at the rate of interest. Thus, the marginal utility of oil flow (or oil price) rises at the interest rate as in the standard Hotelling model, although the price increase in our model is driven by the scarcity rent on capacity to be utilized in the future.

If we continue to assume that the constraint on oil flow is slack and further assume that drilling is strictly positive ($F(t) < K(t), a(t) > 0$), then conditions (7), (9), (11), and (13) imply that marginal drilling costs ($d(a(t))$) must also be rising at the rate of interest. The contrapositive is that whenever drilling is strictly positive ($a(t) > 0$) over some interval and marginal drilling costs rise more slowly than the rate of interest, oil flow must be constrained. Intuitively, if the planner is willing to drill earlier despite the higher discounted marginal cost, it is because she is using all of the additional capacity immediately.

Thus, the necessary conditions alone do not guarantee that production is always constrained: it may be unconstrained when no drilling is occurring, and it may even be unconstrained while wells are being drilled so long as the marginal drilling cost is increasing at the rate of interest. Accordingly, it is possible to devise specific examples in which production will indeed be unconstrained in the optimal program.[21] In section 4 below, however, we provide realistic sufficient conditions ruling out such scenarios, the most stringent of which is that $U'(F)$ is such that the inverse elasticity of oil demand must be weakly rising with F. This condition is satisfied by most, if not all, functional forms for demand used in applied

[21] In the simplest example, $K_0 > 0$, drilling costs are sufficiently high that no wells are ever drilled, and demand is constant elasticity, with an elasticity sufficiently low that if production is constrained, price will rise more quickly than r. In this case, production must always be unconstrained. This example can be modified to allow for drilling, with production unconstrained before drilling begins: let $d(0)$ be sufficiently low that drilling is profitable when capacity is far below K_0 (i.e., when the oil price is high) but still sufficiently high that drilling is unprofitable when capacity is near K_0.

work. Overall, we will conclude that production is always optimally set to the constraint under realistic conditions. This result is consistent with the data from Texas and oil futures markets, in which anticipated prices are never observed to rise faster than the interest rate for a long enough period of time to make deferring production optimal.

3.3 Comparison with the canonical Hotelling model

In this section, we study the conditions under which the solution to our drilling problem can replicate the extraction and price dynamics of the canonical Hotelling model in which there is no capacity constraint but rather a constant marginal drilling cost. We assume that the cost of drilling is sufficiently low that drilling occurs in the optimal program, and we provisionally assume that production is constrained. Then, combining conditions (7) and (13) allows us to eliminate $\phi(t)$ from the system and obtain:

$$U'(F(t)) - (r+\lambda)\theta(t) = \dot{\theta}(t). \tag{16}$$

Meanwhile, differentiating condition (9) with respect to time and solving for $\dot{\theta}(t)$ yields:

$$\dot{\theta}(t) = \frac{r\gamma(t) + d'(a(t))\dot{a}(t)}{X}. \tag{17}$$

Finally, substitute for $\theta(t)$ and $\dot{\theta}(t)$ in condition (16) using condition (9) and condition (17). Remembering that we have assumed that $a(t) > 0$ and $F(t) = K(t)$, this yields equation 18, which can be thought of as a generalized Hotelling rule:

$$U'(F(t)) - \frac{(r+\lambda)d(a(t))}{X} + \frac{d'(a(t))\dot{a}(t)}{X} = \frac{\lambda\gamma_0}{X}e^{rt}. \tag{18}$$

On the right-hand side of equation (18) we have the shadow value of wells ($\gamma_0 e^{rt}$) divided by the total amount of oil stored in a unit mass of untapped wells (X/λ), which we can interpret as the per-barrel shadow value of oil in untapped wells. On the left-hand side, the first term is the marginal utility (or price) of oil. The second term is the amortized, per-barrel marginal cost of drilling a well at time t.[22] The third term captures the opportunity cost of drilling now versus waiting, which arises due to the convexity in the drilling cost function. When $\dot{a}(t) < 0$, drilling activity and marginal drilling costs are falling over time. Thus, drilling immediately incurs an additional opportunity cost relative to delaying. One implication is that, in contrast to standard models in which production and marginal cost decline together, oil flow in our model can *increase* over intervals during which the marginal cost of drilling is falling. The cases examined in section 4 illustrate this phenomenon (see, for example, the phase of rising production in figure 5(b)).

Equation (18) makes clear that the canonical Hotelling rule, which we derive for reference in appendix D, can in fact arise as a special case of our drilling problem under the assumption that the marginal cost of drilling is constant: $d(a) = \bar{d}$ for all $a \geq 0$. In this case, the $d'(a(t))\dot{a}(t)$ term disappears and the imputed per-barrel cost of drilling is given by $c = \bar{d}(r + \lambda)/X$, which we assume equals the marginal extraction cost of Hotelling's planner. Thus, condition (18) implies that, whenever drilling occurs, the marginal utility of oil flow minus the per-barrel marginal cost of extraction rises at the rate of interest:

$$U'(F(t)) - c = \frac{\lambda\gamma_0}{X}e^{rt}. \tag{19}$$

So our planner, though facing a binding production constraint among existing wells, would drill new wells in such a way that the marginal utility of consumption, net of per-barrel extraction costs, rises over time at the rate of interest. Since cumulative usage of oil is the same along this path and Hotelling's path, the two paths must coincide!

But there are several important caveats. First, for our planner to replicate Hotelling's

[22] To clarify, consider the special case of constant marginal drilling costs: $d(a(t)) = \bar{d}$. If the planner drills one well (or rather, she marginally increases the rate of drilling) she gets a marginal increase in oil flow h instants later of $Xe^{-\lambda h}$ assuming oil flow is set to the maximum. If each barrel of flow has imputed cost of c at that time, then at the time the well is drilled, the flow at h would have imputed cost $cXe^{-(r+\lambda)h}$. Since such flows continue indefinitely, we want to find c such that $\bar{d} = \int_{h=0}^{\infty} cXe^{-(r+\lambda)h}dh$. Integrating, we get $\bar{d} = cX/(r + \lambda)$. Solving for c we conclude: $c = \bar{d}(r + \lambda)/X$.

path exactly, we must also assume that marginal utility is unbounded: $U'(0) = \infty$. Why? Recall that oil flow in our model is constrained so that, even if the planner drilled every well immediately and produced at the maximum rate, oil would flow forever. Thus, for price to rise at the rate of interest whenever oil is flowing, as Hotelling's path requires, the price of oil must also grow without bound. Second, we must assume that the initial capacity constraint on oil flow is not too high, for otherwise the planning period would begin with drilling set to zero and production at the constraint, with prices rising exogenously—or in the extreme, production set below the constraint, and price rising at the rate of interest.[23] Lastly, the flow from producing wells must decay sufficiently quickly so that the planner can achieve the Hotelling path via judicious control of the drilling rate while producing at her constraint (otherwise, the planner will drill all of the wells in a pulse at $t = 0$, and price will subsequently rise at a rate strictly less than r). The weaker the concavity of the utility function, or the higher the interest rate, the more rapid the decay in oil flow must be. If any of these assumptions fail, then the planner is constrained to an inferior path, despite having constant marginal drilling costs. We illustrate these assumptions formally in appendix D.

While the latter two assumptions seem reasonable, given that new wells are constantly being drilled in the real world, the assumptions of constant marginal drilling costs and an unbounded oil price are not tenable. Our analysis of the Texas data shows clearly that marginal costs rise with the rate of drilling, while the viability of alternative fuels at *current* oil prices argues against an unbounded oil price. Nevertheless, this exercise makes clear that it is not the geological constraint on oil flow *per se* that undermines Hotelling's path. Rather, it is the way this constraint interacts with the nature of drilling supply and oil demand that relegates the planner to an inferior extraction path.

[23] In the latter case, the period of production below the constraint would be followed by a period of production at the constraint and drilling set to zero, followed by an extended period of non-zero drilling and production at the constraint.

4 Optimal drilling with increasing marginal costs

In this section, we explore the implications of positive and increasing marginal drilling costs, assuming throughout that the marginal utility of oil is bounded. In a market context, marginal utility equals price, and the marginal cost of drilling equals the rental rate on drilling equipment; we therefore interpret most of our results as market outcomes. We begin by considering the case of an individual oilfield that is small relative to the global market, so that the path of oil prices is exogenous but the local market for renting drilling rigs clears at each instant. We then endogenize the oil price path, first focusing on a case in which scarcity rents are negligible, and then studying a model in which only a finite measure of wells may be drilled.

4.1 Exogenous oil prices

As a preliminary step to analyzing the full competitive equilibrium, we examine industry supply in isolation. This case is particularly relevant for interpreting drilling and extraction behavior in a small, local region, such as Texas, and for describing how that behavior responds to both anticipated and unanticipated changes in the path of world oil prices. We assume that identical extractors face given time paths for oil prices $P(t)$ and drilling rig rental rates $\rho(t)$. We close the model on the input side by assuming a competitive local market for drilling rigs and that this market clears. These assumptions are analogous to the "small country" assumptions in international trade models of an exogenous world output price and endogenous local input prices that clear domestic factor markets.

Intuitively, well owners must decide when to drill based on their forecast of future oil prices $P(t)$ and drilling rig rental rates $\rho(t)$. If a well drilled at some date t earns strictly positive discounted profits net of the cost of renting the drilling rig, then every well will be drilled at some point; no well will remain untapped indefinitely. Since well owners all face the same oil price path and rig rental rate path, they agree in their rankings of the optimal time to rent drilling equipment to drill their wells. Rental rates must therefore adjust so that the owners of the wells are indifferent about when they drill. Otherwise, all well owners would attempt to rent drilling rigs at the same time and the rental market would not clear. The same intuition applies when we consider the case of endogenous oil prices below.

Industry supply of oil will maximize the discounted profits of the well owners. Thus, we can deduce the path of drilling ($a(t)$) that is most profitable by defining $U(F(t)) = P(t)F(t)$ and $D(a(t)) = \rho(t)a(t)$ and by reconsidering the necessary conditions in (9)-(11).[24] The $\theta(t)X$ term in necessary condition (9) equals the value of the gross revenue stream that results from drilling a well at t, discounted back to that date. If wells always produce at a strictly positive rate, then equation (7) can be used to eliminate $\phi(t)$ in equation (13), and the resulting linear first-order differential equation, in conjunction with the endpoint condition (14) can be solved to obtain:[25]

$$\theta(t)X = \int_{y=t}^{\infty} P(y)Xe^{-(r+\lambda)(y-t)}dy, \tag{20}$$

The $d(a(t))$ term in (9) is the rental price for drilling rigs at time t, where we interpret $\rho(t) = d(a(t))$ as the upward-sloping rig rental supply curve. In equilibrium, well owners are indifferent about when they drill their wells and each new well earns the same discounted profit regardless of when it is drilled. Aggregate profits earned on wells drilled after $t = 0$ are given by the shadow value of undrilled wells: $\gamma_0 R_0$. As we will see, exogenous changes that reduce γ_0 without altering R_0 reduce extractor wealth from new wells proportionately. Since the price of oil is exogenous, all that matters in determining the path of drilling activity is the measure of wells available to be drilled at the beginning of the program (R_0).

Suppose that well owners expect the oil price to remain forever at its current level ($\bar{P}$). Then equation (20) implies that $\theta(t)X = \frac{\bar{P}X}{r+\lambda}$, which we assume is sufficiently high that

[24] Note in this formulation that utility depends directly on time: $U(F(t), t)$. This could also be the case more generally. For example, incomes may be rising or population may be growing exogenously over time. To avoid complicating the notation, we suppress any implicit time dependence.

[25] To check this solution, note that differentiating (20) and replacing $P(t)$ using equation (7), under the assumption that $F(t) > 0$, yields (13).

drilling is profitable $\left(\frac{\bar{P}X}{r+\lambda} > d(0)\right)$. Since $\dot{\theta}(t)/\theta(t) < r$, conditions (13) and (8) imply that $F(t) = K(t)$ for $t \geq 0$: production is always at the constraint.

As long as drilling continues, condition (15) then implies that the following equation must hold:

$$\frac{\bar{P}X}{r+\lambda} - \gamma_0 e^{rt} = d(a(t)). \tag{21}$$

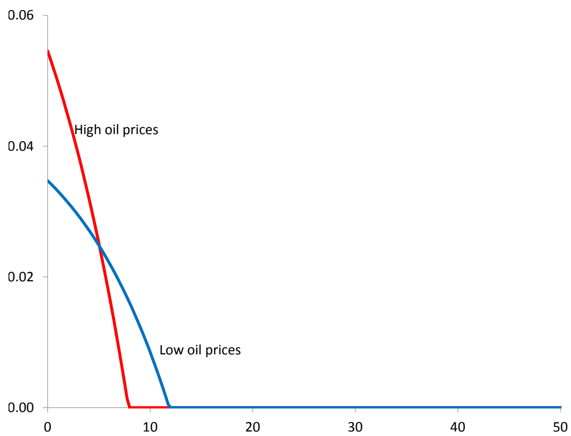
A given γ_0 determines the path of rental prices for drilling equipment and hence the path of rigs that would be offered for rent. If $\gamma_0 = 0$, the path of rental prices—the left-hand side of (21)—is horizontal and would result in an infinite cumulative supply of rentals, assuming an unbounded time horizon. For any $\gamma_0 > 0$, the left-hand side of (21) decreases over time, reaching $d(0)$ at some date, at which time drilling will cease forever. Thus, $\gamma_0 > 0$ results in a finite offering of rig rentals and finite number of wells drilled. As γ_0 increases, the number of wells drilled in equilibrium decreases, and there is a unique $\gamma_0 \in \left(0, \frac{\bar{P}X}{r+\lambda} - d(0)\right)$ that will induce a path of rental prices such that the full measure R_0 of wells will be drilled. Denote as $\hat{T}$ the date when drilling ceases.

Having determined that in response to constant oil prices production would continue at capacity forever and that drilling would decline monotonically until $\hat{T}$, we can use equation (12) to deduce the aggregate supply of oil that results. Drilling activity is intense at first but declines until it ceases altogether at $\hat{T}$. Differentiating equation (12), we conclude that $\ddot{F}(t) = \dot{a}(t)X - \lambda\dot{F}(t)$. Since $\dot{a}(t) < 0, F(t)$ is strictly concave whenever $\dot{F}(t) \geq 0$ and even when $\dot{F}(t)$ is slightly negative. However, for $t > \hat{T}, \dot{a}(t) = 0$ and $\dot{F}(t) < 0$ which implies $\ddot{F}(t) > 0$. Thus, the function $F(t)$ passes through the origin, increases at a decreasing rate from its initial rate of $a(0)X$, reaches a maximum, decreases at an increasing rate but then inflects and ultimately decays exponentially at rate λ.

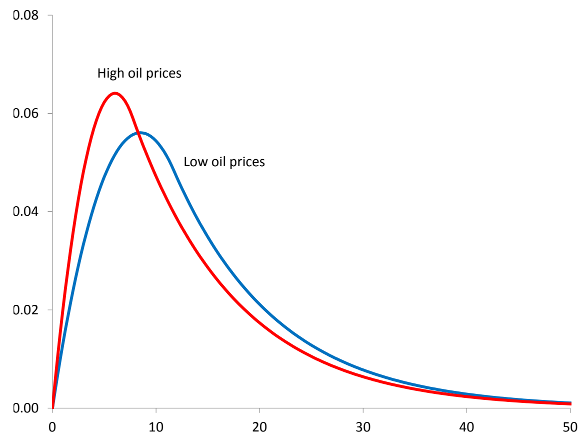
Figure 5 illustrates these results. The hump-shaped "peak oil" production profile in panel (b) of this figure is a well-known feature of production on oil fields all across the world. As we have shown, this production profile emerges from a simple model of a competitive oil extraction industry whose only features include a horizontal price path, an upward-sloping

Figure 5: Optimal drilling in a local region

(a) Rate of drilling ($a(t)$)



(b) Oil production ($F(t)$)



Note: This figure illustrates the equilibrium time paths of drilling rates (panel a) and oil production (panel b) for the case of a linear drilling rig supply function; rig rental rates move in tandem with drilling rates. The figure assumes initial extraction capacity of $K_0 = 0$ (which is equivalent to focusing exclusively on the drilling of, and on the production from, wells that have yet to be drilled as of time $t = 0$), remaining wells of $R_0 = 1$, decline rate of $\lambda = 0.1$, and initial oil flow of $X = 0.1$. Thus, the initial stock of oil underground is $S_0 = (K_0 + R_0X)/\lambda = 1$. The figure further assumes a discount rate of $r = 0.1$, a linear inverse supply function for drilling rigs with intercept of 0 and slope of 10, and low and high oil prices of 1 and 2. See text for details.

local supply function for drilling rigs, and constrained production from existing wells that decays asymptotically toward zero. We show below that a similar "peak oil" result also emerges in the case of endogenous oil prices.

How does the drilling time path in our local region vary with the exogenous oil price $\bar{P}$? Suppose the oil price expected to persist in the market were higher than that for the case described above. If the γ_0 corresponding to the remaining stock of wells (R_0) did not change at all, then rig rentals would be uniformly higher at every instant and would decline to zero later than before. But then $\int_{t=0}^{\infty} a(t) > R_0$, which cannot occur in equilibrium. Suppose instead that γ_0 increased so much that the initial rental price (the left-hand side of equation (21)) was unchanged at $t = 0$. Then the remainder of the rental price path would be uniformly lower and rigs would be supplied for a shorter period of time. But then

$\int_{t=0}^{\infty} a(t) < R_0$, which also cannot occur in equilibrium. Consequently the increase in the oil price must increase *both* γ_0 and the initial drilling rate. The rental price of rigs begins higher but eventually crosses the old equilibrium path and reaches $d(0)$ earlier. As a result, drilling activity starts higher than before but declines more quickly and ceases earlier in response to the higher oil price. Since γ_0 is higher, the increase in the oil price raises the aggregate wealth derived from the drilling of new wells; previously drilled wells are also more profitable. Figure 5(a) illustrates these results by comparing drilling rates (and implicitly, rig rental rates) for low and high oil prices.

What are the implications for the flow of oil from these newly drilled wells? First, with a relatively high oil price, the flow of oil from these wells will increase at a steeper rate initially, since $a(0)X$ is larger. Thus, oil flow must initially be strictly greater with a relatively high price. Eventually, however, oil flow must be strictly lower with the high price, since the total amount of oil underground is fixed. Thus, the overall effect of a higher oil price is to shift oil production earlier in time. Figure 5(b) illustrates these results.[26]

Note that this analysis comparing drilling paths under low and high price scenarios can be reinterpreted as an analysis of the response of drilling to an unanticipated price shock. Suppose that the price of oil is expected to remain low forever but then, at $t = 0$, suddenly and unexpectedly increases to a higher level, where it is then expected to persist. Drilling and rig rental rates immediately jump up at $t = 0$, as we shift from anticipating a low oil price forever to anticipating a high price forever and switch over to the drilling path corresponding to high oil prices. Thus, the model replicates the covariance of oil prices, drilling, and rig rental rates that we see in the Texas data. Moreover, since a price shock will lead to a surge in drilling, it could potentially reverse an overall decline in production in the region as a whole.

Finally, note that our analysis here is applicable if a price ceiling is binding or a backstop producing a perfect substitute at constant marginal cost is active. In either case, drilling

[26]In the case of a linear supply curve for drilling, it can be shown that peak production occurs earlier with a relatively high oil price.

activity would decline monotonically and would cease altogether in finite time while oil production would continue indefinitely. These predictions differ sharply from those of most Hotelling models.[27]

4.2 Equilibrium dynamics with an unlimited number of wells

We now close the model by endogenizing the path of oil prices and requiring that at every instant demand equals supply. This permits us to examine the equilibrium dynamics of oil prices, drilling rig rates, oil production, and drilling activity. We begin in this section by studying a tractable special case in which we assume that the number of wells remaining to be drilled is unlimited ($R_0 = \infty$), so that $\gamma(t) = 0$ in condition (9) above. This condition approximates a world in which scarcity rents are negligible, perhaps because the resource stock is perceived to be vast relative to demand. We will later impose scarcity rents in section 4.3 below. We assume that inverse demand ($U'(F)$) is strictly downward-sloping, that drilling supply ($d(a)$) is upward-sloping, and that both functions are continuously differentiable.

Note that with an unlimited stock of wells, our model bears a strong resemblance to a standard macroeconomic q model of optimal investment in an industry with convex adjustment costs. In that model, per-capita production (F) is determined by the stock of productive capital (K), which decays over time (at rate λ) but can be augmented via investment ($a(t)X$). The industry faces a downward-sloping inverse demand curve for its product (with price $P(t) = U'(F(t))$) and a convex investment cost function (with marginal cost $\rho(t) = d(a(t))$). Both revenues and costs are discounted over time at rate r. The differences between the models are that: (1) the production function in the macroeconomic model typically features diminishing returns to capital, whereas in our model $F = K$ (there is no labor input); (2) in the macroeconomic model there is never a reason to produce less than what the available inputs permit; and (3) in our model capital depreciates with use rather than

[27] For example, in Lee (1978), when the oil price reaches the price ceiling, extraction ceases because the extractors have exhausted their stocks. In Solow (1974), when price reaches the backstop marginal cost, extraction again ceases for the same reason.

deterministically over time. This last difference vanishes when production in our model is constrained.

The dynamic of macroeconomic investment models are often best understood using a phase diagram; we adopt this approach here. Our phase diagram is in (K, a) space. Given $a(t) > 0$ and $F(t) > 0$, we can deduce from the necessary conditions differential equations governing the rate of change of $a(t)$ and $K(t)$[28]

$$\dot{a}(t) = \frac{X[(r+\lambda)d(a)/X - U'(F)]}{d'(a)} \tag{22}$$

$$\dot{K}(t) = aX - \lambda F. \tag{23}$$

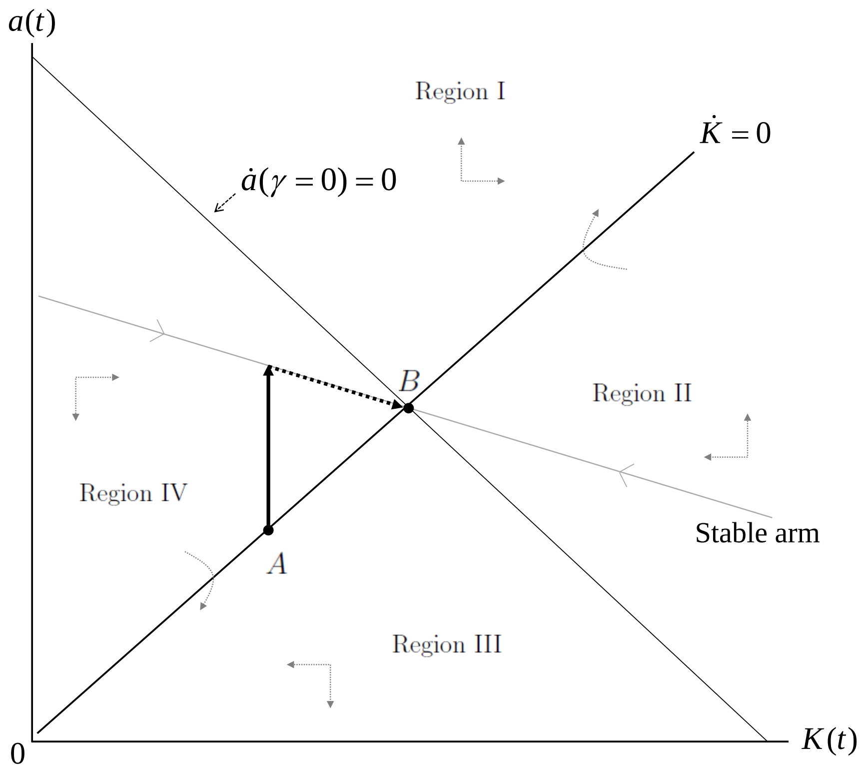
Figure 6 presents the phase diagram, including the two loci of (K, a) combinations such that $\dot{a}(t) = 0$ and $\dot{K}(t) = 0$, under the provisional assumption that $F(t) = K(t)$. The $\dot{a}(t) = 0$ locus will be downward sloping under the maintained assumptions that $d'(a) \geq 0$ and $U''(F) \leq 0$.[29] To the right of this locus, $\dot{a} > 0$ and to its left $\dot{a} < 0$. The ray through the origin with positive slope X/λ is the locus of (K, a) combinations such that $\dot{K} = 0$. Above this locus, $\dot{K} > 0$, and below it $\dot{K} < 0$. The two loci therefore divide the phase diagram into four regions, and in figure 6 the directions of motion for drilling and capacity in each region are indicated with gray arrows.

Note that at all points on and below the $\dot{a}(t) = 0$ locus such that $a(t) > 0$ (regions III and IV of the phase diagram), the marginal drilling cost is weakly decreasing over time, implying that production must be constrained; i.e., $F(t) = K(t)$. It can also be shown that the equilibrium (or equivalently, socially optimal) path for drilling and capacity can never enter region I: once in this region, both drilling and capacity explode without bound, violating the necessary conditions (a formal proof is given in appendix B, lemmas 8 through

[28]The equation for $\dot{a}$ comes from rearranging equation 18 and setting $\gamma_0 = 0$. The equation for $\dot{K}$ is simply the necessary condition 12.

[29]The $\dot{a}(t) = 0$ locus will, moreover, be a line if $d(a)$ and $U'(F)$ are both linear. If $d(a)$ is linear but $U'(F)$ is convex, the locus will be convex; if $U'(F)$ is instead concave, the locus will be concave. Our assumption that $d(a)$ and $U'(F)$ are continuously differentiable implies that the loci and the equilibrium paths are also continuously differentiable.

Figure 6: Phase diagram: infinite stock of oil (special case)



Note: Figure shows the phase diagram corresponding to conditions (12) and (18) with $\gamma(t) = 0$. See text for details.

11). It is only in region II (in which the marginal drilling cost is increasing) and along the $a(t) = 0$ axis that production may be unconstrained.[30]

The $\dot{a}(t) = 0$ locus and the $\dot{K}(t) = 0$ locus intersect at point B, which is the unique steady state. Since $\dot{a}(t) = 0$ at this point, $F(t) = K(t)$ at the steady state. Solving for the steady-state rate of drilling using (12) yields $a^* = \lambda F^*/X$. Thus, new drilling exactly offsets the declining in flow from existing wells. Likewise, solving for the steady-state rig rental rate using (18) yields $d(a) = U'(F^*)X/(r + \lambda)$. The right-hand side is equal to the present discounted stream of revenues generated by a well if the oil flowing from it is sold

[30]If production is unconstrained in parts of region II, the $\dot{K} = 0$ locus may be strictly concave above the $\dot{a} = 0$ locus, though it must still always be strictly upward sloping.

over time at the constant price of $U'(F^*)$. Since every well owner also pays this amount to rent a rig to drill his well, such well owners earn just enough revenue to cover their costs; the inframarginal rents go to the owners of the drilling rigs that are cheapest to mobilize and maintain. The market is in a long-run equilibrium.[31]

Figure 6 also depicts the stable arm (saddle path) running northwest to southeast through the steady state B. In equilibrium, from any initial K_0, the paths for drilling and capacity will follow this stable arm to the steady state.[32] If K_0 is less than capacity at point B, then production will always be constrained. If, on the other hand, K_0 is greater than capacity at B, the path will be in region II. Here, production will be set at capacity as long as K_0 does not grossly exceed the steady state capacity.[33] Thus, the phase diagram as drawn will accurately reflect the equilibrium dynamics in realistic cases.

The phase diagram of figure 6 is particularly useful for studying the effects of unanticipated demand shocks on drilling, extraction, and the oil price. Suppose that oil demand $U'(F)$ is initially relatively low, so that we are initially in a steady state located at point A. Consider an unanticipated outward shift in oil demand that shifts the $\dot{a} = 0$ locus up and right, yielding a new steady state (B) to the northeast of the old one. The optimal transition dynamics imply jumping up immediately to the new stable arm and then following it down gradually toward the new steady state, as depicted in the figure (thick black arrows). Thus, following a positive demand shock, oil prices, drilling activity, and rig rental rates all spike up on impact, and oil flow begins to increase. Prices then fall gradually over time as production builds up, as do drilling activity and rig rental rates, until we arrive at a new steady state with higher oil prices, flow, drilling, and rig rental rates. Notably,

[31] Note that we have implicitly assumed that a steady state with strictly positive drilling exists: $U'(0)X/(r+\lambda) > d(0))$.

[32] A path starting above the stable arm will ultimately lead drilling and capacity into region I, while a path starting below the stable arm will ultimately lead to a cessation of drilling. In either case, the necessary conditions will be violated (in the latter case, the transversality condition (14) is violated.

[33] Formally, note that $\theta^* = P^*/(r+\lambda)$ acts as an upper bound on the marginal value of capacity when initial capacity exceeds the steady state, where P^* is the steady-state price, for production must fall and price must rise monotonically toward the steady state. It can be shown that $P(K_0) \geq \lambda P^*/(r+\lambda)$ or equivalently $P^*/P(K_0) \leq (r+\lambda)/\lambda \approx 2$ is a sufficient condition for production to remain at the constraint, where K_0 is the initial capacity.

these dynamics imply that, following a positive demand shock, oil price expectations must be backwardated.[34] This theoretical result helps to explain the backwardation of oil futures markets during the mid-2000s when the demand for oil was repeatedly affected by positive demand shocks from Asian markets, per Kilian (2009) and Kilian and Hicks (2013).

For a negative demand shock, the story is reversed. Oil prices, drilling, and rig rental rates all fall sharply following an unanticipated inward shift in oil demand and then rise along the new stable arm to the new steady state.[35] If the demand shock is sufficiently large, the rate of drilling will fall to a point at which the capacity dynamics are dominated by the exponential production decline from previously drilled wells. In this event, the rise in oil price after the shock can actually proceed at a rate greater than r, at least initially, per the arguments given in section 3.2. This theoretical result helps to explain the severe contango of oil futures markets in 1998–1999 when the demand for oil was negatively affected by the Asian financial crisis, per Kilian (2009).

We view the ability of our model to predict how expectations of future oil prices react to both positive and negative unanticipated demand shocks as an important contribution. Canonical Hotelling models that do not include a capacity constraint will predict neither the backwardation result nor the result that prices may be expected to rise faster than r following a sufficiently large negative shock. Models such as Pindyck (1978) that include a capacity constraint but do not allow the constraint to decline can predict the backwardation result but not the severe contango result (price can never rise more quickly than r in such models).

[34] We emphasize that, following the demand shock, price expectations must be backwardated, but the actual path of future spot prices need not be, since subsequent demand shocks may occur. Thus, during the mid-2000s, spot prices increased steadily in response to unexpected increases in demand even though price expectations were backwardated.

[35] In the case of a sufficiently large negative demand shock, production may initially fall below the constraint, with the oil price and marginal cost of drilling both rising at the rate of interest. (If the demand shock were sufficiently large so as to knock drilling to zero initially, then this interval would be preceded by an interval of zero drilling and price rising at the rate of interest.)

4.3 Equilibrium dynamics when wells are scarce

Finally, we consider the case in which oil prices are endogenous and oil is scarce. Throughout this section, we assume that the marginal cost of drilling $d(a)$ is strictly upward sloping, that inverse demand $U'(F)$ is strictly downward sloping, and that both functions are continuously differentiable, so that equilibrium paths are continuous and differentiable. Although this general case does not permit analytical tractability, it is possible to draw qualitative conclusions about the equilibrium paths. In particular, we will show that imposing a relatively weak assumption on the shape of the demand curve is sufficient for production to be capacity constrained starting from an initial condition of $K_0 = 0$.

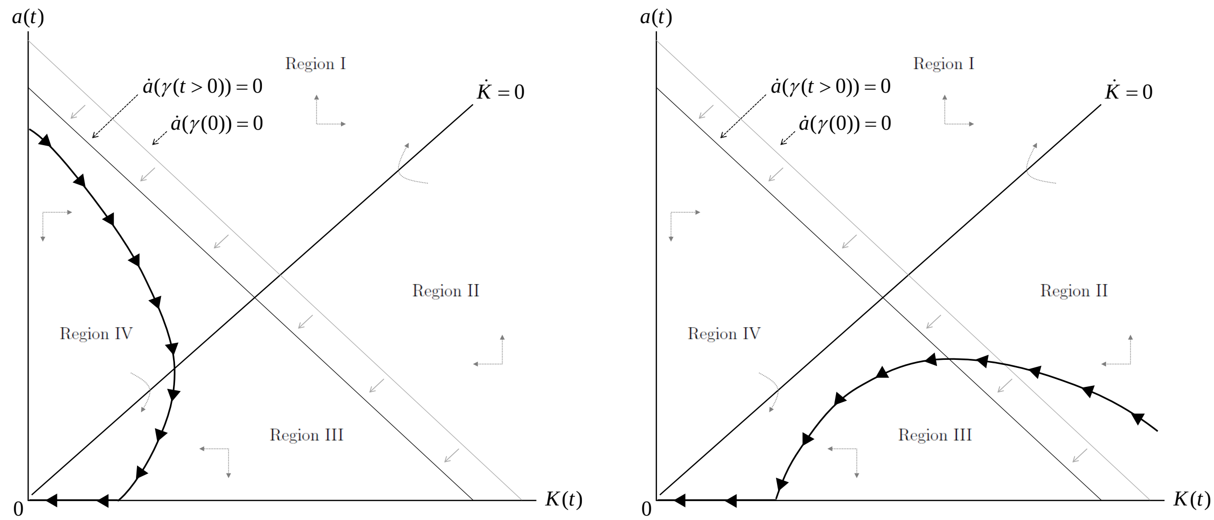
As above, we provisionally assume that drilling occurs ($a(t) > 0$) and that the flow constraint binds ($F(t) = K(t)$) to facilitate analysis using a two-dimensional phase diagram. The chief difficulty with the fully general case arises from the fact that, with scarcity, there are now two state variables to consider: the capacity $K(t)$ and the number of remaining wells $R(t)$. This fact complicates the graphical analysis because the two-dimensional phase diagram is no longer time stationary. The $\dot{K}(t) = 0$ locus is stationary and is still given by equation (23) above. With $\gamma > 0$, however, the $\dot{a}(t) = 0$ locus is given by:

$$\dot{a} = 0: \; a(t) = d^{-1}\left(\frac{XU'(F(t))}{r+\lambda} - \frac{\lambda\gamma_0}{r+\lambda}e^{rt}\right), \tag{24}$$

where we again must have $F(t) = K(t)$ everywhere on this locus. The final term in (24) causes the $\dot{a}(t) = 0$ locus to shift downward over time (if $d(a)$ is linear, the rate of downward shift will increase exponentially at rate r). The non-stationarity of the $\dot{a}(t) = 0$ locus implies that there is no steady-state equilibrium in this model.

Figure 7 sketches two potential equilibrium drilling and extraction paths in the general model: the figure's left panel has an initial condition in which $K_0 = 0$, while the right panel has an initial condition with a large capacity. Both panels depict the movement over time in the $\dot{a}(t) = 0$ locus downward and to the left. As was the case for figure 6, production

Figure 7: Phase diagrams with a scarce supply of wells



Note: This figure shows phase diagrams for the general model in which wells are scarce and therefore $\gamma(t) > 0$. The left panel sketches a potential equilibrium drilling and extraction path starting from $K_0 = 0$, while the right panel sketches a path starting from a large inherited capacity. See text for details.

must be constrained in regions III and IV, on the $\dot{a}(t) = 0$ locus, and on the segment of the $\dot{K}(t) = 0$ locus below the $\dot{a}(t) = 0$ locus. Production may be unconstrained in region II and along the $a(t) = 0$ axis, and region I may not be entered.

We formalize the restrictions optimality imposes on the drilling and extraction paths in two theorems. The first makes use of only monotonicity and continuity assumptions on inverse oil demand and drilling costs and an assumption that drilling costs are sufficiently small that drilling will actually occur. This theorem is therefore quite general, as it even allows for cases in which $P(0)$ is infinite. Making the realistic assumption that $P(0)$ is finite yields the additional result that drilling must cease in finite time (since θ is bounded above by $P(0)$).

Theorem 1. *Assume that the marginal cost of drilling $d(a)$ is continuously differentiable, with $d'(a) > 0$ and $d(0) > 0$. Assume R_0 is finite and strictly greater than zero and that $K_0 \geq 0$. Assume that the inverse demand for oil is given by $U'(F) = P(F)$, with $P(F)$*

continuously differentiable, $P'(F) < 0$, and $P(F) \to 0$ as $F \to \infty$. Assume that the inverse demand curve and drilling supply curve satisfy the following: $XP(0)/(\lambda + r) > d(0)$. Then the following rules hold: (1) all available wells will be drilled in the limit as $t \to \infty$, and if $K_0 = 0$ drilling will begin instantly; (2) at any time that drilling ceases, production must be constrained for a measurable period afterward, and drilling can never begin again; (3) the drilling rate and capacity cannot both be weakly increasing (region I in the phase diagrams can never be entered), nor can the drilling and extraction rates both be weakly increasing; (4) starting from $K_0 = 0$, drilling must initially be decreasing over time; and (5) once capacity is strictly decreasing over time, it cannot subsequently weakly increase, and once extraction is strictly decreasing over time, it cannot subsequently weakly increase.

The formal proof of theorem 1 is given in appendix B. The theorem's intuition, however, is straightforward. Rule (1) follows from the fact that a non-zero level of drilling must be profitable when capacity is sufficiently small. Rule (2) follows from the idea that it would be sub-optimal to drill a costly well and then not fully utilize that well's capacity immediately after drilling. Rule (3) follows from the fact that, as illustrated by the phase diagram, region I cannot be escaped. Thus, if ever the equilibrium path should enter this region, the drilling rate would be shocked to zero the moment reserves are exhausted, violating the necessary condition given by equation (9).[36] Rule (4) immediately follows from rule (3), and rule (5) comes from rule (3) and the fact that it is not possible to cross from region III to region IV in the phase diagrams.

Theorem 1 implies that, if $K_0 = 0$, drilling and extraction must initially be in region IV and then transition to region III. In both of these regions, production must be constrained. However, it is not clear whether or not production may enter region II from region III (since the $\dot{a}(t) = 0$ locus is shifting inward), and if so whether production is unconstrained while in region II. It is also not clear whether, once drilling ceases (as it must if $P(0)$ is finite),

[36] In appendix B, we show that rule (3) must hold even if reserves are infinite: entering region I must ultimately lead to either a violation of the transversality condition or a discontinuitiy in $F(t)$, neither of which is possible in the optimum.

production is constrained thereafter.

A modest assumption on the shape of the demand curve allows us to make substantial progress in resolving these remaining questions. We assume that the inverse elasticity of demand, $\eta(F) \equiv -F\frac{P'(F)}{P(F)}$ is weakly increasing in F. This property is satisfied by nearly all single-product demand curves used in applied work. In particular, it is easy to verify that this property is satisfied by any inverse demand curve of the form $P(F) = \alpha - \beta F^{\delta}$, with either $\alpha > d(0)(r+\lambda)/X, beta > 0, \delta > 0$ or $\alpha \leq 0, \beta < 0, \delta < 0$. The first set of parameters encompasses a wide array of concave and convex demands with a finite $P(0)$, while the second set of parameters allows for $P(0)$ to be infinite (and if $\alpha = 0$, demand is constant elasticity). Given demand satisfying this property, the final result of theorem 2 is that if $K_0 = 0$, then drilling is always weakly decreasing and production is constrained along the entire optimal path.

Theorem 2. *In addition to the assumptions of theorem 1, assume that the inverse elasticity of demand* $\eta(F) \equiv -F\frac{P'(F)}{P(F)}$ *is weakly increasing in* F*. Then the following rules hold: (6) once the rate of drilling strictly declines it can never subsequently increase, even weakly; (7) drilling will end in finite time if and only if* $\lambda\eta(F) < r$ *for* $F > 0$ *sufficiently small; (8) if drilling ceases in finite time, then all subsequent production will always be constrained; and (9) if we additionally have* $K_0 = 0$*, then production is constrained along the entire optimal path.*

The formal proof for theorem 2 is given in appendix C. Intuitively, the functional form assumption on demand precludes increases in the log-log convexity of the inverse demand curve when capacity is declining. Without such an increase in convexity, there is no incentive for drilling activity to increase once capacity is declining (rule (6)). Moreover, once drilling ceases and production is constrained (per rule (2) from theorem 1), there is no incentive to subsequently produce below the constraint (rules (7) and (8)). Rule (9), the main result of the theorem, is implied directly by rules (6)–(8).

The left panel of figure 7 depicts an equilibrium path starting from $F(0) = K_0 = 0$

that follows the trajectory established by theorems 1 and 2, under the assumption that $P(0)$ is finite. The rate of drilling is initially high but decreases over time as the extraction rate builds (and the oil price falls). Eventually, once the drilling rate becomes sufficiently small, the extraction rate will decrease and the oil price will rise. Thus, there is a peak in production. Ultimately, all wells well be drilled in finite time, and the extraction rate will then decline to zero (along the $K(t)$ axis). Throughout the drilling and extraction program, oil production is capacity constrained.

The right panel illustrates another example with an inherited capacity sufficiently large that the optimal program begins in region II. In this case, production may initially be below the constraint, even if the inverse demand curve satisfies the conditions of theorem 2. In particular, if demand is sufficiently inelastic at K_0 that producing at the constraint at K_0 causes the price to rise at a rate sufficiently greater than r for a sufficiently long time, it may be optimal to produce below the constraint while the rate of drilling increases (so that $\dot{d}(a(t))/d(a(t)) = r$). Ultimately, however, production must return to the constraint (permanently),[37] $\dot{a}(t)$ must decrease and eventually turn negative, drilling must ultimately cease, and extraction will gradually decay to zero.

The impact of unexpected demand shocks on the equilibrium path will be qualitatively similar to that of the no-scarcity case discussed above, though the phase diagram does not permit a graphical analysis. For instance, a sufficiently large positive demand shock that occurs when production is constrained will result in a backwardated expected price path. To see this, consider necessary condition (9). Suppose there is a vertical shift of magnitude Z in demand $U'(F)$ but no change in the drilling path. In this case, the entire path for $\theta(t)$ will increase by $Z/(r + \lambda)$, per condition (16). This shift implies that under the original drilling path, $\theta(t)X - d(a(t))$ must now be rising more slowly than the rate of interest, which is sub-optimal. The only way to restore optimality is for the rate of drilling to jump upward

[37] For a measurable period after the instant at which production becomes constrained, the oil price must actually rise at a rate strictly greater than r, since we must have $P(t) - \lambda\theta(t) > 0$ when production is constrained, $\dot{\theta}(t)/\theta(t) = r$ at the instant production becomes constrained, and $\dot{\theta}(t)$ is continuous (per lemma 6 in appendix B).

upon the impact of the shock. If the shock is large enough, the jump in the drilling rate will be sufficient to cause the rate of extraction to increase, yielding backwardation. Similarly, a negative demand shock will cause the rate of drilling to jump down and can result in significant contango, with prices expected to rise at a rate potentially greater than r.

The effects of a negative demand shock relate to the result from theorem 2 that the optimal program includes a period of unconstrained production only if $K_0 > 0$. In particular, unconstrained production requires that K_0 be sufficiently large that the optimal program begins in region II of the phase diagram or on the $a(t) = 0$ axis (demand must also be sufficiently inelastic). Historically, we must of course have $K_0 = 0$. However, the $K_0 > 0$ condition can be viewed as a situation that would occur if the optimal program were interrupted by an unanticipated, negative demand shock. A large enough shock could cause the new optimal program to begin in region II or on the $a(t) = 0$ axis, where production may be unconstrained. The fact that the data from Texas are consistent with production always being constrained suggest that such a sufficiently large negative demand shock did not occur during the sample period.

4.4 Special case of equilibrium dynamics with scarcity

Here, we present an analytically tractable case of the general model by assuming that the marginal drilling cost is constant up to a constraint. This case is particularly useful to consider because it allows us to express drilling costs on a per-barrel basis, facilitating a direct comparison to standard Hotelling results.[38]

Under this assumption, the marginal cost of drilling cost takes the following form:

$$d(a(t)) = \bar{d}, \text{ for } a(t) \leq \bar{a}$$

$$d(a(t)) = \infty, \text{ for } a(t) > \bar{a},$$

[38] While we have not yet proven that, with $d'(a) = 0$ for low values of a, production must always be constrained in the specification we consider (linear oil demand), our analysis below strongly suggests that this is the case. The proof will be included in a future version of the paper.

where $\bar{a}$ is the constraint on drilling. Since the drilling rate is bounded, flow cannot jump instantaneously (or "pulse"); instead, time must elapse for flow to increase. We also assume that the inherited flow of oil is relatively small ($F_0 < \bar{a}X/\lambda$) and that $U''' \geq 0$. With this formulation, necessary condition (9) is replaced by:

$$a(t) \geq 0, \theta(t)X - [\bar{d} + \mu(t)] - \gamma_0 e^{rt} \leq 0, c.s. \tag{25}$$

The shadow price on the capacity constraint, denoted $\mu(t)$, is zero whenever the drilling rate is below the constraint ($a(t) < \bar{a}$); whenever $\mu(t)$ is strictly positive, drilling must be at its maximum feasible rate ($a(t) = \bar{a}$). Thus, the full marginal cost of drilling is given by $\bar{d}+\mu(t)$, which we can interpret as the price to rent a drilling rig in a competitive market, inclusive of the rent on scarce capacity.[39]

Overall, the dynamics in this model are consistent with section 4.3 above in that the rate of drilling weakly declines over time and the extraction path is characterized by a peak in production. Specifically, the optimal program consists of three intervals, and the boundaries between them are determined endogenously. During the first interval, drilling is set at the maximum feasible rate ($a(t) = \bar{a}$, $\mu(t) > 0$), with the flow of oil rising at a decreasing rate ($\dot{F}(t) > 0, \ddot{F}(t) < 0$). During the second interval, drilling proceeds at a slower rate ($a(t) \in (0, \bar{a})$, $\mu(t) = 0$). with the flow of oil falling at an increasing rate ($\dot{F}(t) < 0, \ddot{F}(t) < 0$) and drilling decreasing over time ($\dot{a}(t) < 0$). Thus, oil flow is kinked at the first boundary, implying a discrete drop in the rate of drilling at the boundary. Finally, during the third interval, drilling drops to zero ($a(t) = 0$) and oil flow decays toward zero exponentially ($\dot{F}(t) < 0, \ddot{F}(t) > 0$). Thus, oil flow is also kinked at the second boundary. The kinks in oil flow—and the discrete jumps in drilling rate—occur in this special case because equation (9) holds for a range of drilling rates when the drilling capacity constraint

[39] Feng, Zhao and Kling (2002) study the optimal time path of carbon sequestration using a setup that is mathematically similar to ours, with the stock of carbon in the atmosphere playing the role of our oil flow from existing wells and the fixed amount of land suitable for carbon sequestration activities playing the role of our fixed stock of wells to be drilled. Like us, they derive analytical results for the case of constant marginal costs (of sequestration) up to a period capacity constraint.

is not binding.

In appendix E, we establish these properties formally and describe an algorithm for determining the boundaries between time intervals and the shadow value on the resource constraint (γ_0). If the stock of wells (R_0) is higher, then the corresponding shadow value falls, resulting in a longer first interval, uniformly higher flow during the second interval, and later termination of the second interval.

Figure 8 depicts a simulation of this model for the case of quadratic utility (satisfying $U'(0) < \infty, U''' \geq 0$) and assuming $F_0 = 0$. The reader can verify that this figure illustrates the properties described above for the time paths of oil prices (panel a), production (panel b), and drilling (panel c). For comparison, the figure also depicts Hotelling's time path for prices and production (in panels a and b). Lastly, the figure depicts our paper's key necessary condition for optimality, which is that the discounted revenue stream minus marginal cost of drilling a well must rise at the rate of interest while drilling occurs (panel d). Since revenues in this case are constrained by the upper bound on oil prices, drilling must eventually cease.

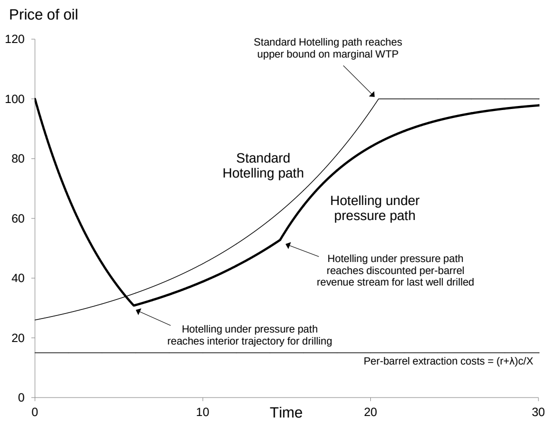
Throughout this section, we have assumed a bounded oil price ($U'(0) < \infty$) and a constrained drilling rate ($\bar{a} < \infty$). If the oil price were unbounded, then the equilibrium would be similar but without the final interval: drilling would never cease entirely. If instead there were no constraint on drilling, then the first phase would disappear. At the first instant, wells would be drilled at an infinite rate (a "pulse") so that equation (18) would hold immediately at $t = 0$ and would continue to hold until drilling ceased.

5 The role of costly above-ground storage

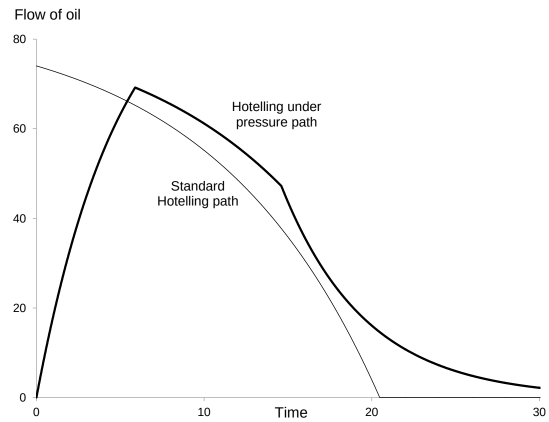
We have shown that our dynamics imply that it is possible, on the equilibrium path, for price to temporarily rise at a rate faster than the rate of interest while production is constrained. However, this result has not taken into account the possibility that oil may be stored above-ground. This section therefore asks how the availability of costly above-ground storage would

Figure 8: Hotelling and Hotelling-under-pressure paths

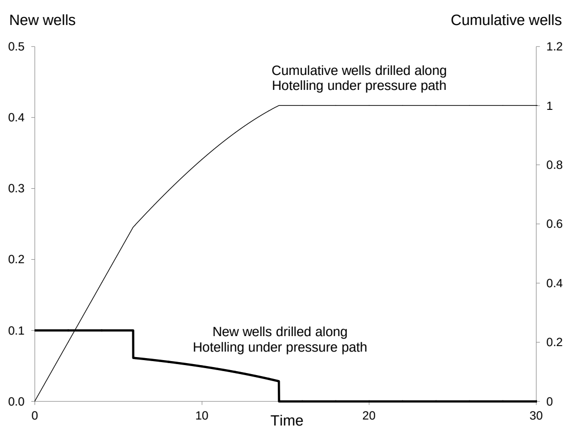
(a) Oil prices



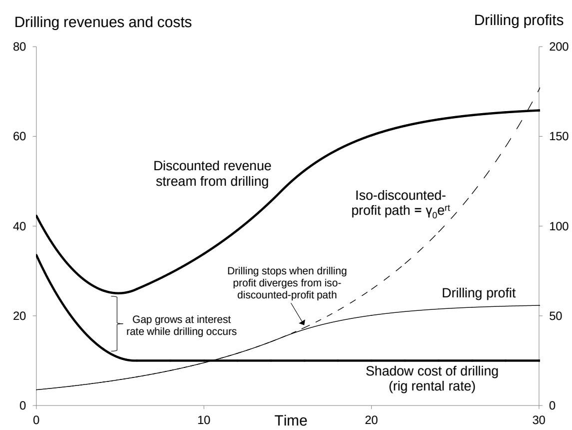
(b) Oil production



(c) Drilling activity



(d) Drilling incentives



Note: This figure illustrates the optimal time paths of drilling, oil production, and prices for the case of a linear demand curve and constant marginal drilling costs up to a capacity constraint. Figure assumes $\lambda = 0.2$, $X = 200$, $R = 1$, and $S = RX/\lambda = 1000$ for the oil reserve, an intercept of $a = 100$ and slope of $b = -1$ for the linear demand function, a marginal drilling cost of $c = 10,000$ with capacity constraint of $\bar{a} = 0.1$, initial oil flow of $F_0 = 0$, and discount rate of $r = 0.1$. See text for details.

affect our conclusions.[40]

For tractability, we assume that storage costs take the form of a continuous payment equal to m percent of the value of the stored oil. In this case, a stockpiler would be indifferent

[40]It is clear that, if storage were costless and limitless, then the oil price could never rise more quickly than r. Real-world logistical costs and storage constraints argue against this hypothetical, as does the fact that we observe instances in futures market data when the price is expected to rise more quickly than r.

regarding buying or selling oil if the oil price is expected to rise at the constant percentage rate of $m + r$.

We will show that this model still accommodates the possibility that the oil price can rise more quickly than r on the equilibrium path while production is constrained. In fact, it must be the case that over any measurable interval over which storage is occurring (and therefore the oil price is rising at the rate $m + r$), production must be constrained. Why? In order for production to be unconstrained, θ must be rising at the rate r. But then equation 7 implies that the oil price must also be rising at r, an immediate contradiction.

Hence, if aboveground storage is possible, the equilibrium may involve intervals in which price rises in percentage terms at rate $r + m$ and production must be at capacity throughout each interval. Such a steep rise in expected prices actually occurred in winter 1998–1999 (see figure 1) and was in fact accompanied by above-ground storage (see figure 11 in appendix A).

6 Conclusion

Standard Hotelling models take for granted that extractors have precise control over the rate at which production flows to market. Our analysis of crude oil drilling and production in Texas shows this assumption to be inconsistent with the technology and cost structure of the crude oil extraction industry. Oil is not extracted barrel-by-barrel. Instead, extractors drill wells, and the maximum flow from these wells is geologically constrained by the pressure underground, which is roughly proportional to the volume of recoverable oil remaining.

We develop a new model of exhaustible resource extraction that accommodates these important features of the crude oil extraction industry. Our model replicates several salient facts observed in the real world: (1) production from pre-existing wells steadily declines over time and does not respond to oil price shocks; (2) drilling of new wells and drilling rig rental rates strongly co-vary with oil prices; (3) local oil-producing regions and fields exhibit production peaks; and (4) expected future oil prices can be backwardated after positive demand shocks and can rise faster than the interest rate (temporarily) following negative demand shocks. Canonical Hotelling models cannot explain any of these facts. While the literature that allows for expansions in capacity with convex adjustment costs can explain production peaks and price backwardation, this literature fails to rationalize why production from existing wells can decline over time but at the same time can be insensitive to price shocks. Such models also do not permit an expectation that prices will rise faster than the rate of interest.

Our model could be extended in several logical ways. First, we assume a single continuum of homogenous wells, all located within the same, static market for drilling rigs. Instead, we could divide the continuum of wells across two or more regions—say Texas and Alaska, or onshore and offshore—each with its own supply of drilling rigs. Since different locations typically have their own geological features, it would also be natural to consider variation in drilling costs, production decline rates, and resource stocks across regions—or across individual wells within the same region. Some of the heterogeneity in decline rates may in fact be endogenous to the rate and form of drilling and resource development. Second, while the stock of drilling rigs and crews in a given region is fixed in the short run, this stock can change over time, as new rigs are built and crews are trained, as old rigs are scrapped and workers retire, or as existing rigs and crews are moved from one region to another. These drilling industry dynamics could be incorporated to enrich the model. Third, uncertainty about future oil demand could be incorporated into the model, in which case each undrilled well could be characterized as a real option. We leave these extensions to future work.

References

Adelman, M. A., "Mineral Depletion, with Special Reference to Petroleum," *The Review of Economics and Statistics*, 1990, *72* (1), 1–10.

Alquist, Ron and Lutz Kilian, "What Do We Learn From the Price of Crude Oil Futures?," *Journal of Applied Econometrics*, June/July 2010, *25* (4), 539–573.

Berck, Peter and Michael Roberts, "Natural Resource Prices: Will They Ever Turn Up?," *Journal of Environmental Economics and Management*, 1996, *31*, 65–78.

Black, Geoffrey and Jeffrey T. LaFrance, "Is Hotelling's Rule Relevant to Domestic Oil Production?," *Journal of Environmental Economics and Management*, 1998, *36*, 149–169.

Cairns, Robert, "The Green Paradox of the Economics of Exhaustible Resources," *Energy Policy*, forthcoming.

Cairns, Robert D. and Graham A. Davis, "Adelman's Rule and the Petroleum Firm," *The Energy Journal*, 2001, *22* (3), 31–54.

Dahl, Carol and Mine Yucel, "Testing Alternative Hypotheses of Oil Producer Behavior," *The Energy Journal*, 1991, *0* (4), 117–138.

Darcy, Henry, *Les Fontaines Publiques de la Ville de Dijon* 1856.

Davis, Graham A. and Robert D. Cairns, "Simple Analytics of Valuing Producing Petroleum Reserves," *The Energy Journal*, 1998, *19* (4), 133–142.

Devarajan, Shantayanan and Anthony C. Fisher, "Hotelling's "Economics of Exhaustible Resources": Fifty Years Later," *Journal of Economic Literature*, 1981, *19* (1), 65–73.

Feng, Hongli, Jinhua Zhao, and Catherine L Kling, "The time path and implementation of carbon sequestration," *American Journal of Agricultural Economics*, February 2002, *84* (1), 134–149.

Gaudet, Gerard, "Natural resource economics under the rule of Hotelling," *Canadian Journal of Economics/Revue Canadienne d'Economique*, 2007, *40* (4), 1033–1059.

— **and Stephen Salant**, "The Hotelling Model with Multiple Demands," in Robert Halvorsen and David Layton, eds., *Handbook on the Economics of Natural Resources*, Edward Elgar, 2014.

Griffin, James M, "OPEC Behavior: A Test of Alternative Hypotheses," *American Economic Review*, December 1985, *75* (5), 954–63.

Halvorsen, Robert and Tim R. Smith, "On Measuring Natural Resource Scarcity," *Journal of Political Economy*, 1984, *92* (5), 954–964.

Hamilton, James D., "Understanding Crude Oil Prices," *The Energy Journal*, 2009, *30* (2), 179–206.

— , "Oil Prices, Exhaustible Resources, and Economic Growth," in Roger Fouquet, ed., *Handbook of Energy and Climate Change*, Cheltenham, United Kingdom: Edward Elgar Publishing, 2013.

Hogan, William W., "A Dynamic Putty–Semi-putty Model of Aggregate Energy Demand," *Energy Economics*, 1989, *11* (1), 53–69.

Holland, Stephen P., "Modeling Peak Oil," *The Energy Journal*, 2008, *29* (2), 61–79.

Hotelling, Harold, "The Economics of Exhaustible Resources," *Journal of Political Economy*, 1931, *39* (2), 137–175.

Hyne, Norman J., *Nontechnical Guide to Petroleum Geology, Exploration, Drilling, and Production*, 2nd edition ed., Tulsa, Oklanhoma: PennWell Corporation, 2001.

Jones, Clifton T., "OPEC Behaviour Under Falling Prices: Implications for Cartel Stability," *The Energy Journal*, 1990, *0* (3), 117–130.

Kellogg, Ryan, "The Effect of Uncertainty on Investment: Evidence from Texas Oil Drilling," *American Economic Review*, forthcoming.

Kilian, Lutz, "Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market," *American Economic Review*, 2009, *99* (3), 1053–1069.

_ **and Bruce Hicks**, "Did Unexpectedly Strong Economic Growth Cause the Oil Price Shock of 2003–2008?," *Journal of Forecasting*, August 2013, *32*, 385–394.

Kline, Patrick, "Understanding Sectoral Labor Market Dynamics: An Equilibrium Analysis of the Oil and Gas Field Services Industry," November 2008. Unpublished Manuscript.

Krautkraemer, Jeffrey A., "Nonrenewable Resource Scarcity," *Journal of Economic Literature*, 1998, *36* (4), 2065–2107.

Lee, Dwight, "Price Controls, Binding Constraints, and Intertemporal Economic Decision Making," 1978.

Léonard, Daniel and Ngo van Long, *Optimal Control Theory and Static Optimization in Economics* number 9780521331586. In 'Cambridge Books.', Cambridge University Press, November 1992.

Livernois, John R. and Russell S. Uhler, "Extraction Costs and the Economics of Nonrenewable Resources," *Journal of Political Economy*, 1987, *95* (1), 195–203.

Mason, Charles F. and Klaas van't Veld, "Hotelling Meets Darcy: A New Model of Oil Extraction," June 2013. Paper presented at the 20th Annual Conference of the European Association of Environmental and Resource Economists (Toulouse, France).

Miller, Merton H. and Charles W. Upton, "A Test of the Hotelling Valuation Principle," *Journal of Political Economy*, 1985, *93* (1), 1–25.

Nystad, Arild N., "Rate sensitivity and the optimal choice of production capacity of petroleum reservoirs," *Energy Economics*, 1987, *9* (1), 37–45.

Pindyck, Robert S., "The Optimal Exploration and Production of Nonrenewable Resources," *Journal of Political Economy*, October 1978, *86* (5), 841–61.

Ramcharran, Harri, "Oil Production Responses to Price Changes: An Empirical Application of the Competitive Model to OPEC and Non-OPEC Countries," *Energy economics*, 2002, *24* (2), 97–106.

Rao, Nirupama S., "Taxation and the Extraction of Exhaustible Resources: Evidence from California Oil Production," April 2010. MIT-CEEPR Working Paper 10-006.

Salant, Stephen W., "The Economics of Natural Resource Extraction: A Primer for Development Economists," *World Bank Research Observer*, February 1995, *10* (1), 93–111.

Slade, Margaret E., "Trends in Natural-resource Commodity Prices: An Analysis in the Time Domain," *Journal of Environmental Economics and Management*, 1982, *9*, 122–137.

_ **and Henry Thille**, "Whither Hotelling: Tests of the Theory of Exhaustible Resources," *Annual Review of Resource Economics*, 09 2009, *1* (1), 239–259.

Smith, James L., "Modeling the Impact of Taxes on Petroleum Exploration and Development," 2012. IMF Workig Paper No. 12/278.

Smith, V. Kerry, "Natural Resource Scarcity: A Statistical Analysis," *Review of Economics and Statistics*, 1979, *61*, 423–427.

Society of Petroleum Evaluation Engineers, "Survey of Economic Parameters Used in Property Evaluations," 1995.

Solow, Robert, "The Economics of Resources or the Resources of Economics," *American Economic Review Papers and Proceedings*, 1974, *64*, 1–14.

Switzer, Sheldon and Stephen Salant, "Optimal Capacity Expansion by Foresighted Extractors and Producers (Expansion Optimale de Dapacité par des Exploitants Prévoyants d'une Ressource Renouvelable)," in G. Gaudet and P. Lasserre, eds., *Problemes de*

Ressources Naturelles et Theorie Economique, University of Laval Press, Laval, Quebec, 1986.

Thompson, Andrew C., "The Hotelling Principle, Backwardation of Futures Prices and the Values of Developed Petroleum Reserves—the Production Constraint Hypothesis," *Resource and Energy Economics*, 2001, *23*, 133–156.

A Additional empirical results

This empirical appendix has two parts. First, we show that the primary features of figure 2—the deterministic production decline and the lack of response to price shocks—hold in subsamples of production from relatively high-volume leases and from wells drilled in-sample. Second, we present results that rule out alternative explanations for the lack of price response.

A.1 Production decline in high-volume leases and wells drilled in-sample

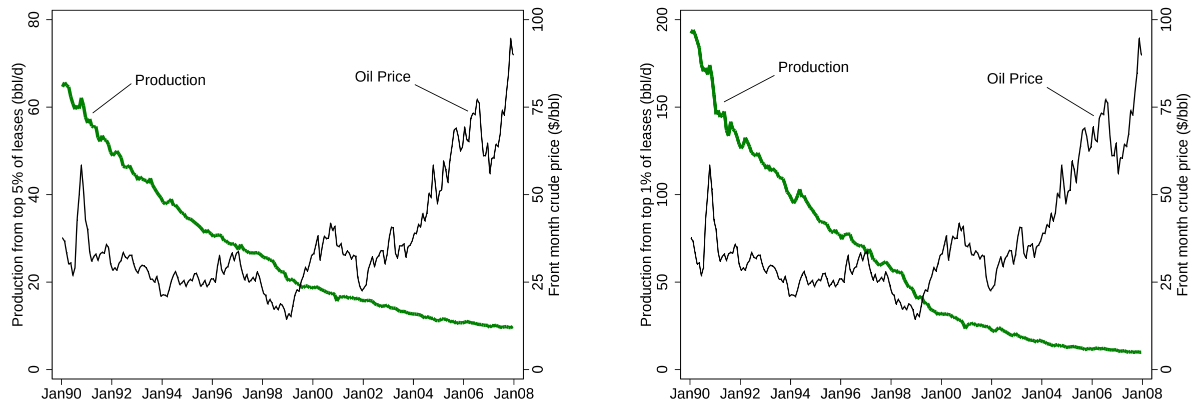
Figure 2 presents average monthly production from all active oil leases in Texas for which there was no rig activity from 1990–2007. Average lease-level production in Texas is quite low, raising the question of whether our empirical results extend to higher-volume fields that might be found elsewhere in the world.

Figure 9 presents production data from subsamples of relatively high-volume leases in Texas. For each lease in the dataset, we obtain its total production by summing its production rate over the entire 1990–2007 sample. We then assign each lease to its appropriate percentile based on total production. The left panel of figure 9 includes leases within the top 5% of total production, and the right panel includes leases in the top 1%. The average top 1% lease produces nearly 200 bbl/day at the start of the sample, far larger than the overall average initial production of about 8 bbl/day shown in figure 2. Nonetheless, in both panels of figure 9, production declines deterministically and exhibits essentially no response to price signals. The production declines are steeper in figure 9 than in figure 2, suggesting that either high-volume leases have relatively high decline rates or that leases typically have decline curves that are hyperbolic rather than exponential.

We next examine production from wells drilled during the sample. To undertake this analysis, we first match drilling records, which come from TRRC drilling permits, to the TRRC lease-level production data. This match must be done based on lease names. Because

Figure 9: Production from existing wells in high-volume Texas leases

(a) Top 5% of leases **(b)** Top 1% of leases



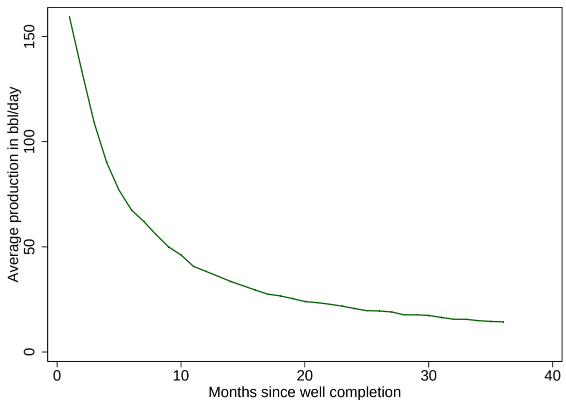
Note: This figure presents crude oil front month prices and daily average oil production from leases on which there was no rig activity (so that all production comes from pre-existing wells). All prices are real \$2007. The left panel (a) includes leases that are in the top 5% of total production in the 1990–2007 sample, while the right panel (b) includes leases that are in the top 1%.

naming conventions vary across the two datasets, we are able to match only 67.2% of drilled oil wells to a lease.

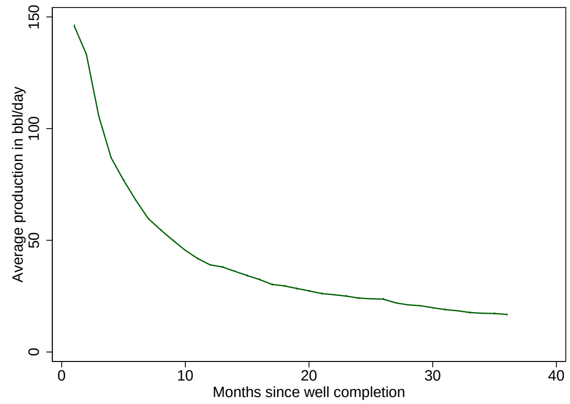
Because production can only be observed at the lease-level rather than at the well-level, we isolate the sample to drilled wells that are the only producing well on their lease for three years after the well was completed. The remaining sample consists of 4,105 wells drilled between 1990 and 2004 (inclusive). We break these 15 years into five 3-year periods. Each panel of figure 10 then plots, for wells drilled within a particular period, the average production for the first three years of the wells' lives. For each period of drilling, production from drilled wells declines deterministically over time and again does not indicate any price responsiveness. Production from wells drilled in later periods is less, on average, than production from wells drilled in early periods, suggesting that firms sensibly drill highly productive wells before drilling less productive wells. Incorporating well-level heterogeneity into the theoretical model is therefore likely to be a fruitful path for future research.

Figure 10: Production from wells drilled during the sample period

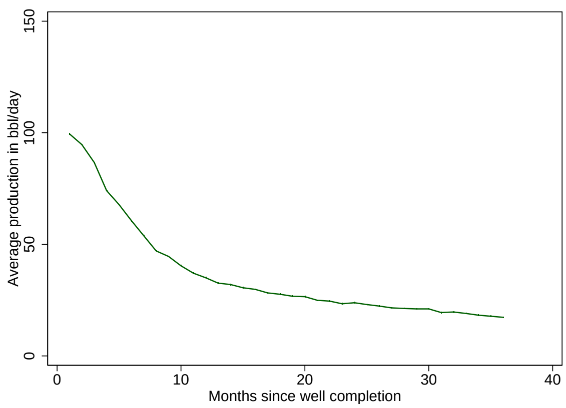
(a) Wells drilled from 1990–1992



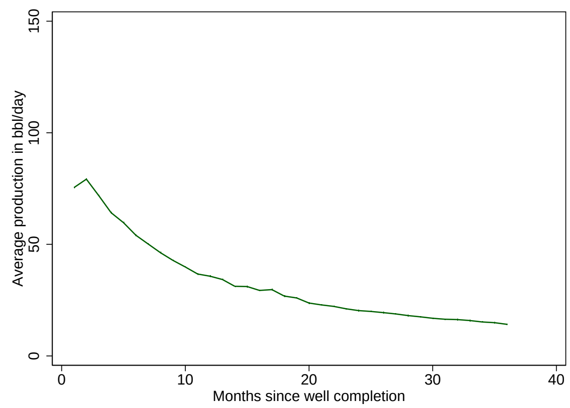
(b) Wells drilled from 1993–1995



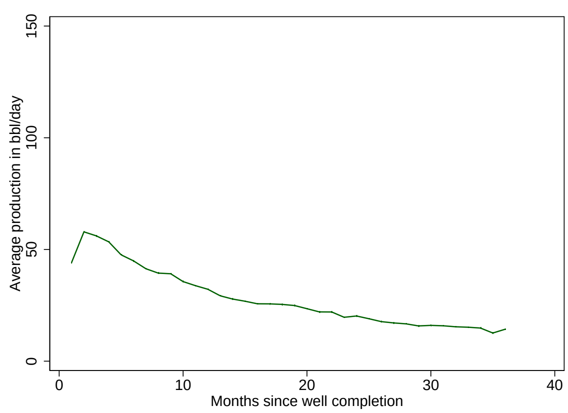
(c) Wells drilled from 1996–1998



(d) Wells drilled from 1999–2001



(e) Wells drilled from 2002–2004



Note: Each panel plots average monthly production for wells drilled during the indicated time interval. See text for details.

A.2 Ruling out alternative explanations for production's lack of response to price incentives

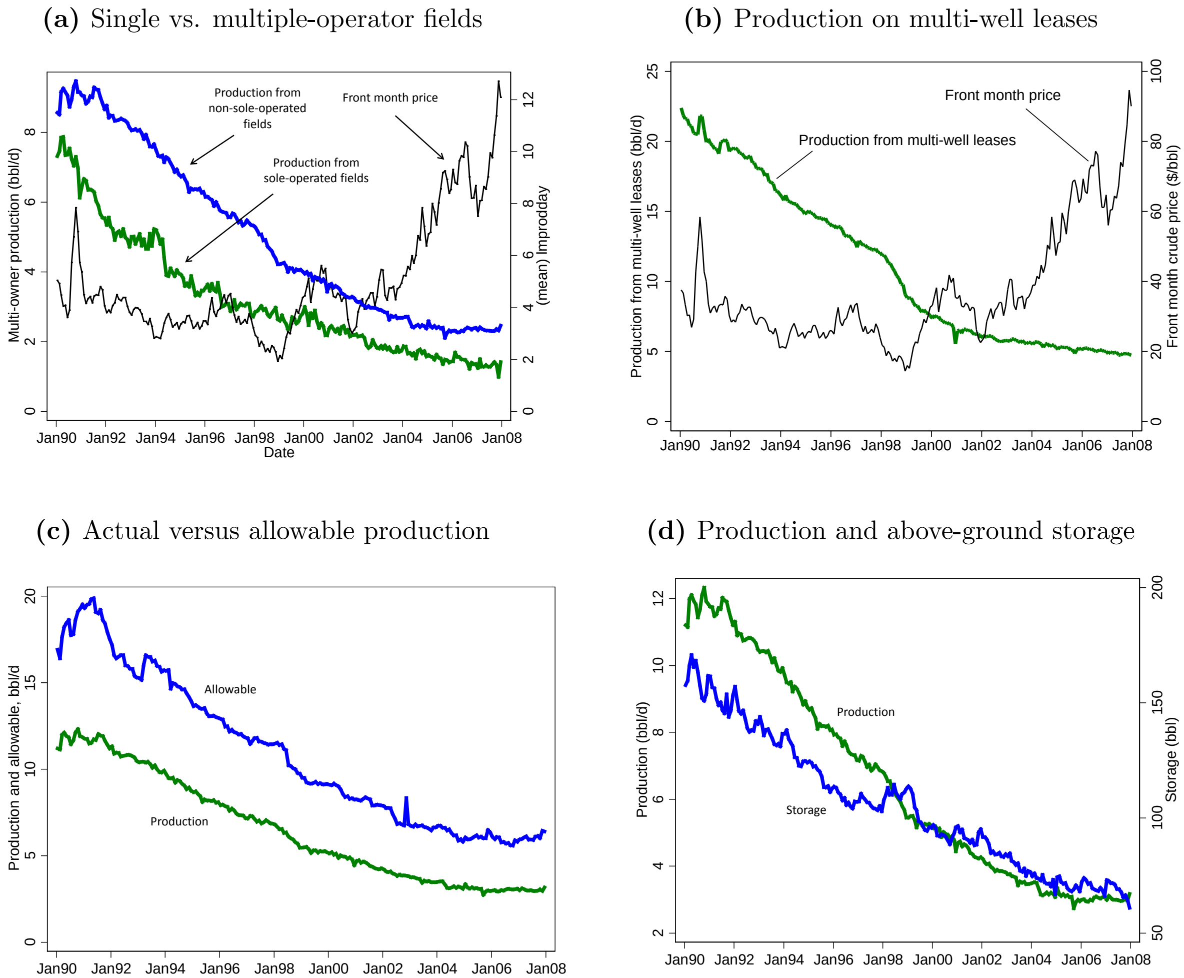
There are several potential alternative explanations for the lack of price response among existing oil wells that we need to rule out. First, races to oil created by common-pool externalities in oil fields with multiple lease-holders would diminish the incentive that individual lease-holders have to defer production when prices are expected to rise, since much of the deferred production would be extracted by others instead. Figure 11(a) shows, however, that the long, downward trend in oil production manifests both for oil fields with multiple operators, as well as for oil fields with just a single operator. If anything, fields with multiple operators showed a bigger decline in production around 1998, perhaps because a larger share of leases on these fields have marginal production.

Second, a condition of many leases is that the lease holder produce oil; a firm that drops lease-level production to zero may therefore risk losing the lease. Figure 11(b) shows, however, that production from multi-well leases, on which producers may shut in at least some of their wells without risk of losing the lease, does not respond to price signals.

Third, oil production in Texas is subject to maximum allowable production quotas—or "allowables"—as determined by the Texas Railroad Commission. This system dates to the East Texas Oil Boom of the 1930s when races to oil led to overproduction and collapsing world oil prices. Whether originally intended to end the race to oil, or simply to cartelize the Texas oil industry and boost prices, this system persists to this day, and very lease in our data has a monthly allowable, including on fields with just a single operator. One obvious concern is that these maximum production quotas are binding, leading to the lack of price response. Figure 11(c) shows, however, that the average production for leases in our main sample is well below the average allowable production. Thus, the allowables are not binding and therefore cannot explain the lack of price response that we observe in our data.

Fourth and finally, one possible concern is that the decision makers whose behavior we observe in our data could earn profits by delaying production during periods of extreme

Figure 11: Graphical evidence ruling out alternative explanations



(a) Single vs. multiple-operator fields

(b) Production on multi-well leases

(c) Actual versus allowable production

(d) Production and above-ground storage

Note: Panel (a) shows average daily production on fields with multiple operators as well as on fields with just a single operator, along with the spot price of crude oil. Panel (b) shows average production for leases with multiple wells in our main sample. Panel (c) shows average actual production as well as average allowable production for leases in our main sample. Panel (d) shows average oil production and average above-ground storage of crude oil on leases in our main sample. See text for details.

contango but do not, perhaps because they are myopic or simply do not understand the potential to lock-in guaranteed profits by deferring production and taking a short position in the futures market. Figure 11(d) shows, however, that above-ground storage of crude oil on these leases increased notably during the 1998 period of extreme contango. Lease-holders responded to these price incentive by accumulating inventories above-ground, deferring sales—not extraction—to take advantage of the expected increase in prices.

B Proofs for rules (1)–(5) pertaining to the general case

This appendix provides the proof for theorem 1, which we reproduce here:

Theorem. *Assume that the marginal cost of drilling $d(a)$ is continuously differentiable, with $d'(a) > 0$ and $d(0) > 0$. Assume R_0 is finite and strictly greater than zero and that $K_0 \geq 0$. Assume that the inverse demand for oil is given by $U'(F) = P(F)$, with $P(F)$ continuously differentiable, $P'(F) < 0$, and $P(F) \to 0$ as $F \to \infty$. Assume that the inverse demand curve and drilling supply curve satisfy the following: $XP(0)/(\lambda + r) > d(0)$. Then the following rules hold: (1) all available wells will be drilled in the limit as $t \to \infty$, and if $K_0 = 0$ drilling will begin instantly; (2) at any time that drilling ceases, production must be constrained for a measurable period afterward, and drilling can never begin again; (3) the drilling rate and capacity cannot both be weakly increasing (region I in the phase diagrams can never be entered), nor can the drilling and extraction rates both be weakly increasing; (4) starting from $K_0 = 0$, drilling must initially be decreasing over time; and (5) once capacity is strictly decreasing over time, it cannot subsequently weakly increase, and once extraction is strictly decreasing over time, it cannot subsequently weakly increase.*

The proof proceeds via the following series of lemmas.

Lemma 1. *If $K(t) > 0$, then $F(t) = 0$ is never optimal.*

Proof. Suppose by contradiction that $F(t) = 0$. If we assume $P(0) = \infty$, then equation (7) is violated, leading to an immediate contradiction. So assume $P(0) < \infty$ instead. Since $F(t) = 0 < K(t)$, we must have $\phi = 0$ (equation 8). Equations (7) and (13) then imply that $\dot{P}(t) = rP(t)$. But this also leads to a contradiction, for $P(F)$ is maximized at $F = 0$, which implies that it is impossible both for $F(t) = 0$ and for $\dot{P}(t)/P(t) = r$ (since the price must weakly decline after time t). □

Lemma 2. *Capacity $K(t)$ must go to zero in the limit as $t \to \infty$.*

Proof. Suppose that, by contradiction, there is some $\hat{K} > 0$ such that $K(t) \geq \hat{K}$ for all t. By lemma 1, we must have $F(t) > 0$ for all t. Thus, equation (7) must hold with equality, so that it and equation (13) together imply that $\theta(t) > 0$ for all t. Moreover, the lower bound $\hat{K}$ on $K(t)$ implies that production must ultimately be unconstrained forever (since the number of wells to drill is finite, $F(t)$ must become arbitrarily close to zero, otherwise $K(t)$ would fall below $\hat{K}$). Therefore, $\theta(t)$ must rise at r forever. However, this contradicts the transversality condition (14) when $\lim_{t\to\infty} K(t) > 0$. Thus, it must be that $\lim_{t\to\infty} K(t) = 0$ (i.e., it is sub-optimal to leave valuable capacity unused). □

Lemma 3. *Drilling will always occur at the optimum, and optimality requires complete exhaustion of reserves in the limit as $t \to \infty$. Moreover, if $K_0 = 0$, drilling will begin immediately at $t = 0$.*

Proof. Suppose by contradiction that there is no drilling ($a(t) = 0$ for all $t \geq 0$).

Because $P(F)$ is continuous, because $K(t)$ goes to zero in the limit, and because we have assumed that $P(0)X/(\lambda + r) > d(0)$, it must be true that for t sufficiently large, $\frac{P(K(t))X}{\lambda+r} > d(0)$. Suppose we are at such a time t, and consider the profitability of drilling at a rate $\epsilon > 0$ for an interval $\Delta \approx 0$ and then producing the new wells at their constraint for all remaining time.

The cost of drilling these wells is given by:

$$\int_0^\Delta e^{-rs} D(\epsilon)\, ds < \Delta D(\epsilon) \leq \epsilon \Delta d(\epsilon), \tag{26}$$

where the last inequality comes from the fact that $d(a)$ is upward sloping.

Assuming that production begins the instant drilling stops and then continues at the constraint forever, the discounted revenue from drilling these wells (which add $X\epsilon\Delta$ to capacity) is given by:

$$X\epsilon\Delta e^{-r\Delta}\int_0^\infty P(K(t+\Delta+s))e^{-(\lambda+r)s}\,ds \tag{27}$$

$$\geq X\epsilon\Delta P(K(t)+X\epsilon\Delta)e^{-r\Delta}\int_0^\infty e^{-(\lambda+r)s}\,ds = \frac{X\epsilon\Delta P(K(t)+X\epsilon\Delta)e^{-r\Delta}}{\lambda+r}, \tag{28}$$

where the inequality comes from the facts that the maximum possible flow rate is $K(t)+X\epsilon\Delta$ and $P(F)$ is monotonically decreasing, so that the initial price will be lower than the price at all subsequent times. Then, by the continuity of $P(F)$ and $d(a)$ and from the fact that $\frac{XP(K(t))}{\lambda+r} > d(0)$, there must exist an ϵ and Δ sufficiently small that drilling is profitable at t.

To see that reserves must be zero in the limit, suppose by contradiction that $\lim_{t\to\infty} R(t) > 0$. In the limit, $a(t)$ must approach zero (so long as reserves are finite), which implies (from the arguments above) that $K(t)$ must become arbitrarily close to zero. Again, per the above arguments, drilling must therefore be profitable in the limit. Thus, it cannot be profitable to leave behind unused reserves, so $\lim_{t\to\infty} R(t) = 0$.

Finally, if $K_0 = 0$, the above condition tells us that drilling is profitable at $t = 0$. There is no incentive to delay, since the system is fixed with $K(t) = 0$ and $F(t) = 0$ until drilling occurs. Thus, drilling will begin immediately.

This lemma completes the proof of rule (1). □

Lemma 4. *$\gamma(t) > 0$.*

Proof. Suppose otherwise. From above, we know that both $a(t)$ and $K(t)$ approach zero in the limit. Thus, it must be the case that, eventually, $\frac{XP(K(t))}{\lambda+r} > d(a(t))$ will hold forever. Recall that $\theta(t)$ denotes the shadow value of capacity (which is in turn equal to λ times the stock of remaining oil in the drilled wells). This value must be, at minimum, the discounted revenue stream that would be generated by producing the capacity at the constraint forever (per the solution to equation (16)), since producing below the constraint is only optimal when it increases discounted revenues and therefore the value of capacity. As $K(t)$ goes to zero, the discounted revenue from a marginal unit of capacity must be at least $P(K(t))/(\lambda + r)$.

Thus, for sufficiently large t, we must have $\theta(t)X > d(a(t))$, which violates equation (9) if $\gamma(t) = 0$. $\square$

Lemma 5. *$K(t) > 0\ \forall t > 0$, and therefore $F(t) > 0\ \forall t > 0$.*

Proof. Suppose $K(t) > 0$ for some t. Equation (12) specifies that, if $a(s) = 0\ \forall s > t$, then $K(s)$ declines at a rate that is no greater than that for exponential decay (since $F(s) \leq K(s)$. Such a decline curve will never reach zero in finite time. Moreover, any drilling activity will only further increase $K(s)$. Thus, once $K(t) > 0$ for some t, then $K(s) > 0$ for all $s > t$.

All that remains is to show that K becomes positive immediately. If $K_0 > 0$, the proof is complete. For the $K_0 = 0$ case, note from above that this case results in drilling taking place at $t = 0$. Suppose $a(0) = \infty$ (a pulse of drilling). In this case, we must have $K(0) > 0$ (since $F(0)$ is bounded above), and the proof is complete. On the other hand, suppose the initial rate of drilling is finite and occurs over (at least) an initial time period of duration Δ. In this case, it suffices to show that $\exists t_1 > 0$ s.t. $\dot{K}(t) > 0\ \forall t \in (0, t_1)$, where $t_1 < \Delta$. Intuitively, this must be the case because the first-order effect of $a(t) > 0$ outweighs the second order effect from $F(t) \geq 0$. Formally, we show this by noting that $\dot{K}(t)$ is given by:

$$\dot{K}(t) = a(t)X - \lambda F(t) \geq a(t)X - \lambda K(t) \tag{29}$$

$$= a(t)X - \lambda \int_0^t (a(s)X - \lambda F(s))\, ds. \tag{30}$$

For t sufficiently small, $\dot{K}(t)$ must be positive. Thus, $\exists t_1 > 0$ s.t. $\forall t \in (0, t_1), K(t) > 0$, and the proof is complete. $\square$

Lemma 6. *$\theta(t)$, $\gamma(t)$, $a(t)$, $K(t)$, $F(t)$, $P(t)$, $\phi(t)$, $\dot{\gamma}(t)$, $\dot{K}(t)$, and $\dot{\theta}(t)$, are continuous for all $t > 0$, as is $\dot{a}(t)$ when $a(t) > 0$.*

Proof. Equation (11) immediately implies that both $\gamma(t)$ and $\dot{\gamma}(t)$ are continuous.

$\phi(t)$ is bounded below by zero and cannot be infinite due to equation (7) and the fact that $F(t) > 0$. $\theta(t)$ is similarly bounded for the same reason. Equation (13) then implies that $\dot{\theta}(t)$ exists and therefore that $\theta(t)$ is continuous.

Due to equation (9), the continuity of $\theta(t)$, combined with the assumption that $d'(a) > 0$, implies that $a(t)$ must also be continuous $\forall t > 0$.

The continuity of $a(t)$, combined with the boundedness of $F(t)$, then imply via equation (12) that $K(t)$ is continuous.

Now suppose $F(t)$ is not continuous. If $F(t)$ is below the constraint both before and after the discontinuity ("jump"), this will violate equation (7) (which must hold with equality) because $\theta(t)$ is continuous and because $\phi(t)$ must equal zero both before and after the jump. If $F(t)$ jumps down from the constraint to a point below the constraint, this too is a contradiction: for equation (7) to hold with equality, the jump down in $F(t)$ must be matched with a jump up in $\phi(t)$ (because $P(F)$ is strictly monotonically decreasing). But $\phi(t)$ must be zero after the jump. Finally, there is a contradiction if $F(t)$ jumps up from below the constraint to the constraint. To hold equation (7) with equality, the jump up in $F(t)$ must be matched with a jump down in $\phi(t)$, but $\phi(t)$ must be zero before the jump.

The continuity of $F(t)$ immediately implies the continuity of $P(t)$, since $P(F)$ is continuous. Equation (7) then implies that $\phi(t)$ must also be continuous.

The results above, combined with equations (12) and (13) imply that $\dot{K}(t)$ and $\dot{\theta}(t)$ are continuous.

Finally, the results above combined with equation (9) and the assumption that $d(a)$ is continuously differentiable imply that $\dot{a}(t)$ is continuous when $a(t) > 0$. □

Lemma 7. *At any time that drilling ceases, production must be constrained for a measurable period afterward. Moreover, once drilling ceases, it may never begin again.*

Proof. Let $\hat{t}$ denote the time that drilling stops. To see that production must be constrained at this time, suppose by contradiction that it is not. In that case, we must have $\dot{\theta}(\hat{t}) = r\theta(\hat{t})$ (by equation (13)), further implying by equation (9) and the continuity of $\dot{theta}(t)$ that

$\dot{d}(a(\hat{t})) = rd(a(\hat{t}))$. $d(a)$ is strictly monotonically increasing and $d(0) > 0$, implying that the drilling rate must be rising at $\hat{t}$. But this is a contradiction, since $a(\hat{t}) = 0$ and the continuity of $a(t)$ implies that $a(t)$ must be decreasing at $\hat{t}$.

So production is constrained at $\hat{t}$, which is equivalent to $P(\hat{t}) > \lambda\theta(\hat{t})$. Because $P(t)$ and $\theta(t)$ are continuous, it must therefore be the case that production is constrained for a measurable time period after $\hat{t}$.

Intuitively, it is sub-optimal to stop drilling and immediately produce below the constraint because drilling is costly, and this combination of actions effectively wastes the capacity generated by the final amount of drilling.

To see that drilling cannot re-start once it has stopped, note that $\theta(t)$ is rising at a rate slower than r immediately after drilling stops and that it may never rise more quickly than r. Thus, $\theta(t)X - \gamma(t)$ will never be strictly greater than zero after drilling has stopped, and by equation (9) drilling may therefore not occur again.

This lemma completes the proof of rule (2). □

Lemma 8. *The drilling rate and capacity cannot both be strictly increasing (region I in the phase diagrams can never be entered). This lemma holds even with infinite reserves.*

Proof. Suppose not. That is, suppose $\dot{a}(t) > 0$ and $\dot{K}(t) > 0$ simultaneously for some time t. We will show that, no matter how drilling and production evolve from this point, a contradiction will result.

First, suppose production is constrained at t. In this case, equation (18) holds, and we have:

$$\dot{a}(t) = \frac{X}{d'(a(t))}\left[\frac{(r+\lambda)d(a(t))}{X} - P(F(t)) + \frac{\lambda\gamma_0}{X}e^{rt}\right]. \tag{31}$$

If $\dot{a}(t) > 0$ and production is constrained, we will continue to have $\dot{a} > 0$ for subsequent times so long as production is constrained. If production is constrained forever, then the rate of drilling will be shocked to zero at the time when reserves are exhausted, which contradicts the continuity of $a(t)$ (lemma 6).[41] Thus we cannot have $\dot{a}(t) > 0$ and $\dot{K}(t) > 0$

[41] Note that even with infinite reserves, this path will still yield a contradiction. Why? As K and F increase

simultaneously with production at the constraint forever.

So suppose instead that production is unconstrained at t (logic similar to the below applies if we allow production to fall below the constraint at some later time). In this case, we must have $\dot{P}/P = r$ and $\dot{\theta}/\theta = r$, implying via equation (9) that $\dot{d}(a)/d(a) = r$, in turn implying that $\dot{a}$¿0. Moreover, so long as production is unconstrained, we will continue to have $\dot{a}$¿0, and if production is unconstrained forever, we will again have a discontinuity in a once reserves are exhausted.[42] On the other hand, returning production to the constraint must result in a discontinuity in F, which also contradicts lemma 6. Why? Unconstrained production implies that $\dot{F} < 0$. But note that since we had $\dot{K}(t)$¿0, we must continue in this case to have $\dot{K} > 0$ after time t. Thus, the only way to return production to the constraint is to induce a discontinuity in F.

Thus, if at any time $\dot{a}(t) > 0$ and $\dot{K}(t) > 0$, there must eventually be a discontinuity in either a or F, either of which is a contradiction. □

Lemma 9. *The drilling rate and capacity cannot both be weakly increasing. This lemma holds even with infinite reserves.*

Proof. Suppose not. Lemma 8 above handled the case in which both $\dot{a}(t) > 0$ and $\dot{K}(t) > 0$, so here we only need to consider cases involving equalities. If $\dot{K}(t) = 0$ and $\dot{a}(t) > 0$, then equation (12) implies that $K(t)$ must strictly increase immediately after t, contradicting lemma 8. If $\dot{a}(t) = 0$, it must be the case that production is constrained at t, and equation 31 implies that a must begin rising immediately after t, contradicting lemma 8. □

Lemma 10. *The drilling and extraction rates cannot both be strictly increasing. This lemma holds even with infinite reserves.*

Proof. Suppose not. That is, suppose $\dot{a}(t) > 0$ and $\dot{F}(t) > 0$ simultaneously for some time t. $\dot{F}(t) > 0$ implies that production is constrained at t. This situation cannot continue

without bound, the price will eventually go to zero, implying that θ must also go to zero by equation (7). But then by equation (9) we cannot have strictly positive drilling, a contradiction.

[42] Again note that even with infinite reserves, this path will still yield a contradiction. Why? K is forever increasing on this path, and θ is rising forever at the rate of interest. These contradict the transversality condition (14).

forever for the same reason given in lemma 8. If production falls below the constraint, there will again be a contradiction (a discontinuity in either a or F) for the same reason given in lemma 8. $\square$

Lemma 11. *The drilling and extraction rates cannot both be weakly increasing. This lemma holds even with infinite reserves.*

Proof. Here again, we only need to consider cases involving equalities. If $\dot{F}(t) = 0$, it must be that $\dot{K}(t) = 0$, and combined with $\dot{a}(t) \geq 0$, we have the same situation as 9. Similarly, if $\dot{a}(t) = 0$ and $\dot{F}(t) \geq 0$, it must be that $\dot{K}(t) \geq 0$, and we have the same situation as 9.

This result completes the proof of rule (3). $\square$

Lemma 12. *Starting from $K_0 = 0$, drilling must initially be strictly decreasing over time.*

Proof. We showed in lemmas 3 and 5 that drilling must begin immediately if $K_0 = 0$ and that capacity must therefore initially be strictly increasing. We just showed that both drilling and capacity cannot simultaneously be weakly increasing; thus, drilling must initially be strictly decreasing. This proves rule (4). $\square$

Lemma 13. *Once capacity is strictly decreasing over time, it cannot subsequently increase.*

Proof. Suppose by contradiction that $K(t)$ transitions from being strictly decreasing to increasing. Because $\dot{K}$ is continuous, any such transition must involve a moment in which $\dot{K} = 0$. Let time $\hat{t}$ denote this moment.

First, we must have $\dot{F}(\hat{t}) \leq 0$, since either production is constrained, in which case $\dot{F}(\hat{t}) = 0$, or production is unconstrained, in which case $\dot{F}(\hat{t}) < 0$.

Second, we cannot have $\dot{a}(\hat{t}) \geq 0$, since this would contradict lemma 9.

Third, we cannot have $\dot{F}(\hat{t}) = 0$, since this must lead to a contradiction. Since $\dot{a}(\hat{t}) < 0$, then $\dot{K}$ must be strictly negative after time $\hat{t}$ and strictly positive before time $\hat{t}$, contradicting the premise that $\dot{K}$ transitions from strictly negative to positive at $\hat{t}$.

The only remaining possibility is that production is unconstrained at $\hat{t}$, with $\dot{a}(\hat{t}) < 0$. This, however, is an immediate contradiction, since we must have $\dot{a} > 0$ if production is unconstrained. □

Lemma 14. *Once production is strictly decreasing over time, it cannot subsequently increase.*

Proof. Suppose not, and let $\hat{t}$ denote the time at which production transitions to increasing. First, production must be constrained at $\hat{t}$, and $\dot{K}(\hat{t})$ must be weakly greater than zero. Second, production must also be constrained just prior to $\hat{t}$, since otherwise either $\dot{K}(t)$ or $F(t)$ must be discontinuous at $\hat{t}$, a contradiction. But then it must be the case that $\dot{K}(t)$ is strictly negative just before $\hat{t}$ and positive afterward, contradicting lemma 13. This contradiction completes the proof of rule (5). □

C Proofs for rules (6)–(9) pertaining to the general case (including sufficient conditions for a binding flow constraint)

This appendix provides the proof for theorem 2, which we reproduce here:

Theorem. *In addition to the assumptions of theorem 1, assume that the inverse elasticity of demand $\eta(F) \equiv -F\frac{P'(F)}{P(F)}$ is weakly increasing in F. Then the following rules hold: (6) once the rate of drilling strictly declines it can never subsequently increase, even weakly; (7) drilling will end in finite time if and only if $\lambda\eta(F) < r$ for $F > 0$ sufficiently small; (8) if drilling ceases in finite time, then all subsequent production will always be constrained; and (9) if we additionally have $K_0 = 0$, then production is constrained along the entire optimal path.*

The proof proceeds in the following steps:

Lemma 15. *Once the rate of drilling strictly declines it can never subsequently weakly increase.*

Proof. We will again proceed by contradiction. We already showed in rule (3) that the rate of drilling and the rate of extraction (along with capacity) cannot be both simultaneously increasing. Thus, in order to have $\dot{a}(t)$ transition from being strictly negative to weakly positive, it must be the case that capacity and production are strictly declining both during the transition and during any measurable interval in which $\dot{a}(t) \geq 0$. Since we also know that $a(t)$ goes to zero in the limit as $t \to \infty$, we must additionally have a subsequent transition in which $\dot{a}(t)$ changes from weakly positive to strictly negative.

Let t_1 denote the time of the first transition, when $\dot{a}(t) < 0$ before t_1 and $\dot{a}(t) \geq 0$ after t_1. Let $t_2 \geq t_1$ denote the time of the second transition, when $\dot{a}(t) \geq 0$ before t_2 and $\dot{a}(t) < 0$ after t_2. We will first consider the case in which $t_2 > t_1$. At both t_1 and t_2, $\dot{a} = 0$, so production is constrained at both times, and via equation 31, the following two equations must hold:

$$\frac{(r+\lambda)d(a(t_1))}{X} + \frac{\lambda\gamma_0}{X}e^{rt_1} = P(F(t_1)) \tag{32}$$

$$\frac{(r+\lambda)d(a(t_2))}{X} + \frac{\lambda\gamma_0}{X}e^{rt_2} = P(F(t_2)) \tag{33}$$

Where $a(t_2) \geq a(t_1)$ and $P(F(t_2)) > P(F(t_1))$.

Moreover, since $\dot{a}(t) < 0$ before t_1 and $\dot{a}(t) \geq 0$ after t_1, it must be that when we totally differentiate equation (31) at t_1 and divide through by $P(t_1)$, we have:

$$\frac{r\lambda\gamma_0 e^{rt_1}}{XP(F(t_1))} \geq \frac{\dot{P}(t_1)}{P(t_1)}. \tag{34}$$

On the other hand, at t_2 we must have:

$$\frac{r\lambda\gamma_0 e^{rt_2}}{XP(F(t_2))} \leq \frac{\dot{P}(t_2)}{P(t_2)}. \tag{35}$$

Given our functional form assumption for $P(F)$, it must be the case that $\dot{P}(t_2)/P(t_2) \le \dot{P}(t_1)/P(t_1)$. Why? Note that, when production is constrained and no drilling is taking place, the rate at which $P(t)$ increases is given by $\dot{P}(t)/P(t) = \lambda\eta(K(t))$. Since $K(t_2) < K(t_1)$, it must be that $\dot{P}(t_2)/P(t_2) \le \dot{P}(t_1)/P(t_1)$.

Thus, in order for equations 34 and 35 to hold, we must have the price rises, on average, weakly faster than r between t_1 and t_2. However, this means that equations 32 and 33 cannot hold, since $d(a(t))$ cannot rise faster than r and must be rising strictly more slowly than r in the neighborhoods of t_1 and t_2, which implies that it must be rising strictly more slowly than r, on average, between t_1 and t_2. Thus, we have a contradiction.

Finally, consider the case in which $\dot{a}(t) = 0$ only for the instant t_1. This situation is the limiting case from the lemma above, in which $t_2 \to t_1$. As $t_2 \to t_1$, we have $\dot{P}(t_2)/P(t_2) \to \dot{P}(t_1)/P(t_1)$. Thus, in the limit, equations 34 and 35 yield that $\dot{P}(t_1)/P(t_1)$ must equal r. This result again yields a contradiction as equations 32 and 33 cannot hold in the limit. This result completes the proof of rule (6). $\square$

Lemma 16. *Suppose that, for some $F^* > 0$, if $F < F^*$ then $\lambda\eta(F) < r$ (note that such an F^* must exist for any inverse demand function that satisfies the conditions of the theorem and has $P(0)$ finite; however, the converse is not true (the existence of F^* does not imply that $P(0)$ finite)). Then drilling must stop in finite time.*

Proof. Suppose, by contradiction, than drilling does not stop in finite time. We know that both $a(t)$ and $K(t)$ must approach zero in the limit, so there must be a point at which $K(t) < F^*$ and both $a(t)$ and $K(t)$ are decreasing forever (by rules (6) and (5), respectively). Because $a(t)$ is decreasing, production must be constrained forever.

Note that, when production is constrained and no drilling is taking place, the rate at which $P(t)$ increases is given by $\dot{P}(t)/P(t) = \lambda\eta(t)$. Thus, when $K < F^*$, it must be the case that $P(t)$ is rising strictly more slowly than r (with or without drilling activity). Necessary conditions (7) and (13) then jointly imply that $\theta(t)$ must also be rising strictly more slowly than r. Then, because $\gamma(t)$ rises at r, we must have that for t sufficiently large,

$\gamma(t) > \theta(t)X$. This is a contradiction, however, since we must have $\gamma(t) < \theta(t)X$ for drilling to occur, per equation (9). □

Lemma 17. *Suppose that for all $F > 0$, $\lambda\eta(F) \geq r$ (note that this condition implies that $P(0)$ is infinite, though again the converse is not true). Then drilling will not stop in finite time.*

Proof. Suppose, by contradiction, that drilling stops in finite time. Denote the time drilling stops by $\hat{t}$. First, consider the case in which the inequality is strict; that is, for all $F > 0$, $\lambda\eta(F) > r$. At $\hat{t}$, it cannot be optimal to produce at the constraint. Why? Doing so causes $P(t)$ to rise at the rate $\lambda\eta(t) > r$. Moreover, at any future time, price must rise at a rate weakly greater than r (with a strict inequality any time production is constrained). Thus, it is profitable to deviate at $\hat{t}$ by producing less at $\hat{t}$ and producing more later. However, producing below the constraint immediately after drilling contradicts rule (2), so we have a contradiction.

Now consider the case in which $\exists F^* > 0$ such that, for all $F < F^*$, $\lambda\eta(F) = r$. If drilling ceases at some capacity level $K(\hat{t}) > F^*$, then it will be optimal to immediately produce below the constraint, and we again have the contradiction above. What if drilling ceases at a capacity level $K(\hat{t}) \leq F^*$? In this case, producing at the constraint causes $P(t)$ to rise at r. Rule (2) tells us that production must be at the constraint at $\hat{t}$ and for a measurable period afterward. Note that production must in fact always be at the constraint in this case: the continuity of $F(t)$ implies that if production were to fall below the constraint, $P(t)$ would have to at least briefly rise faster than r, which cannot happen when production is unconstrained.

Thus, from $\hat{t}$ onward, it must be that production is at the constraint and that $P(t)$ rises at r forever. In order for equations (7) and (13) to both hold forever, it must also be the case that $\theta(t)$ rises at r forever, and in particular that it rises at r at $\hat{t}$. But this leads to a contradiction: if both $\theta(t)$ and $\gamma(t)$ are rising at r at $\hat{t}$, then by equation (9) $d(a(t))$ must also be rising at r at $\hat{t}$. However, the fact that $a(\hat{t}) = 0$ and the continuity of $a(t))$ imply that $a(t)$ must be decreasing at $\hat{t}$. Thus, if for all $F > 0$, $\lambda\eta(F) \geq r$, then drilling will not stop in finite time. This result and the previous lemma together prove rule (7). $\square$

Lemma 18. *Suppose that, for some $F^* > 0$, if $F < F^*$ then $\lambda\eta(F) < r$, and therefore that drilling stops in finite time. Suppose that at some time t^*, drilling has stopped and $K(t^*) \leq F^*$ (such a time must exist since $K(t)$ goes to zero in the limit). Then production must be constrained at t^*.*

Proof. Suppose, by contradiction, that production is unconstrained at t^*. $P(t)$ must then rise at rate r, so we have $P'(F)\dot{F}/P(F) = r$. Rearranging and substituting in the inverse demand elasticity, we have $\dot{F} = -rF/\eta(F)$. Meanwhile, $\dot{K} = -\lambda F$. Because $r/\eta(F) > \lambda$, we have that F declines faster than K, so that production must be unconstrained forever, declining at rate $r/\eta(F)$. Thus, total cumulative production equals $F(t^*) \int_{t^*}^{\infty} e^{-rt/\eta(F(t))}\, dt$. This value is strictly less than the remaining stock of oil at t^*, $K(t^*)/\lambda$. Thus, capacity does not go to zero in the limit, violating lemma 2. $\square$

Lemma 19. *If drilling ceases in finite time, then all subsequent production will always be constrained.*

Proof. Let $F^* > 0$ denote the maximum value of F such that $\lambda\eta(F) < r\ \forall F < F^*$. Let $\hat{t}$ again denote the time at which drilling stops, and $K(\hat{t})$ denote capacity at this time. If $K(\hat{t}) \leq F^*$, then by lemma 18 above production must be constrained at $\hat{t}$ and all subsequent t.

What if $K(\hat{t}) > F^*$? We know from rule (5) that production must be constrained for a measurable period after $\hat{t}$ and that, from lemma 18, production must be constrained once capacity falls to F^* (denote the time at which this occurs by t^*). Suppose by contradiction that, at some time $\tilde{t}$ strictly greater than $\hat{t}$ and strictly less than t^*, production becomes unconstrained.

First, note that because $\tilde{t} < t^*$, it must be that $\lambda\eta(K(\tilde{t})) \geq r$ and therefore that, when production is constrained at times t just before $\tilde{t}$, $P(t)$ must be rising at a rate weakly greater than r.

Second, at $\tilde{t}$, we must have $P(\tilde{t}) - \lambda\theta(\tilde{t}) = 0$ by equation (7). In addition, we have that $\dot{\theta}(\tilde{t}) = r\theta(\tilde{t})$ and (from lemma 6) that $\dot{\theta}(t)$ is continuous. Further, for times t just before $\tilde{t}$, production is constrained, so $P(t) - \lambda\theta(t) > 0$.

The above facts generate a contradiction. To have $P(t) - \lambda\theta(t) > 0$ for t just prior to $\tilde{t}$, $P(\tilde{t}) - \lambda\theta(\tilde{t}) = 0$, $\dot{\theta}(\tilde{t}) = r\theta(\tilde{t})$, and $\dot{\theta}(t)$ continuous, it must be the case that $\dot{P}(t)/P(t) < r$. However, at such times, we have also shown that $P(t)$ must be rising at a rate weakly greater than r. Thus, it cannot be the case that production is unconstrained following the cessation of drilling.

Intuitively, the inverse demand assumption in the theorem implies that the incentive to produce below the constraint increases with K. In addition, optimality requires that drilling cease at a time at which there is no incentive to produce below the constraint. Capacity can only decline from this time onward, so production must be constrained forever.

This result proves rule (8). □

Lemma 20. *If we additionally have $K_0 = 0$, then production is constrained along the entire optimal path.*

Proof. We have already proven that if drilling ceases in finite time, then production must always be constrained afterward. It remains to show that production is constrained while drilling is occurring. With $K_0 = 0$, we know from rule (1) that drilling will begin instantly. Rule (4) tells us that drilling must initially be decreasing over time, and rule (6) tells us that the rate of drilling can never increase. Thus, while drilling is occurring, the rate of drilling is always decreasing. This implies that we cannot have $d(a(t))$ increasing at the rate of interest, so production must be constrained during drilling. This result proves rule (9) and completes the theorem. □

D Standard Hotelling result

In this section, we briefly restate the standard Hotelling result before illustrating the conditions under which the path of Hotelling's planner and that of our planner coincide.

Assume that oil flow $F(t)$ at time t generates instantaneous utility flow of $U(F(t))$, with $U(0) = 0$, $U'(\cdot) > 0$, and $U''(\cdot) < 0$. Assume that a total of S units of oil can be extracted at rate $F(t) \geq 0$, which is under the complete control of the oil extractor, and that the cost of this extraction is $C(F(t)) = cF(t)$, where $c \in [0, U'(0))$ is the constant marginal cost of extraction. Assume a social planner discounts utility and costs continuously at exogenous rate r and, if wealth maximizing agents are involved, they also discount profit flows at rate r. To maximize the discounted utility, the planner chooses $F(t)$ so that marginal utility less marginal cost of extraction grows exponentially at the rate of interest whenever oil flow is strictly positive:

$$F(t) \geq 0,\ U'(F(t)) - c - \gamma_0 e^{rt} \leq 0,\ \text{with complementary slackness (c.s.).} \tag{36}$$

where $F(t)$ is the amount of oil extracted and consumed at time t and γ_0 is an undetermined multiplier. Thus, quantity flows according to: $F(t) = U'^{-1}\left(\gamma_0 e^{rt} + c\right)$, where U'^{-1} is the inverse of the first derivative of $U(\cdot)$. In addition, the resource stock must get used up either in finite time or asymptotically: $\int_0^\infty F(t)dt = S$, which uniquely determines γ_0 and therefore the time path of extraction and marginal utilities. If we assume that marginal utility is unbounded at zero ($U'(0) = \infty$), then marginal utility must rise forever and the resource stock will only be exhausted in the limit. If we instead assume that marginal utility is bounded at zero ($c < U'(0) < \infty$), then the resource stock will be exhausted in finite time at the precise instant that the rising marginal utility path reaches its upper bound. This is not only the planner's optimal extraction path but it is the aggregate extraction path that emerges in the competitive equilibrium of a decentralized market.

There are two reasons why a planner with the extraction technology described above in the main text would be unlikely to generate Hotelling's extraction path. First, oil flow in our model is constrained such that, even if the planner drilled every well immediately and produced at the maximum possible rate, oil would flow forever. Thus, price cannot rise at the rate of interest whenever oil is flowing, as Hotelling's path requires, unless marginal utility is also able to rise forever. Second, in our model, the incentive to drill in a given period depends, in part, on the cost of drilling in *other* periods, as captured by the $d'(a(t))\dot{a}(t)$ term in equation (18). One implication is that oil flow in our model can *increase* over intervals during which the marginal cost of drilling is falling, whereas the cost of extraction typically increases with production in standard models.

To overcome the first of these threats, we must assume that marginal utility is unbounded: $U'(0) = \infty$. To overcome the second, we must assume that the marginal cost of drilling is constant: $d(a) = \bar{d}$ for all $a \geq 0$. In this case, the imputed per-barrel cost of drilling is well-defined and is given by $c = \bar{d}(r + \lambda)/X$, which we assume is equivalent to the per-barrel extraction cost faced by Hotelling's planner. Given these two assumptions, condition (18) implies that the marginal utility of oil flow minus the per-barrel marginal cost of extraction rises at the rate of interest:

$$U'(F(t)) - c = \frac{\lambda \gamma_0}{X} e^{rt}. \tag{37}$$

Note, however, that we have implicitly assumed that the planner starts drilling wells at the outset of the planning period and never stops, producing at the constraint throughout, so that condition (18) always applies. For this to be the case, however, we need two other conditions to hold. First, the flow from drilled wells must decay sufficiently fast, so that the planner is able to achieve the Hotelling path via judicious control of the drilling rate while producing at her constraint. Second, the initial capacity constraint on oil flow cannot be too high, for otherwise the planner would delay in drilling the first well—and may even produce below her constraint initially. If either of these conditions fails, then the planner is geologically constrained to an inferior path.

To illustrate the first of these two conditions, we assume for simplicity that drilling costs are zero ($c = 0$) and that utility from oil flow takes the constant elasticity form: $U(F) = \alpha F^\beta$ for $\alpha > 0, \beta \in (0, 1)$. We also assume provisionally that $\lambda(1-\beta) > r$ and that $S = (F_0 + R_0 X)/\lambda$ is the total resource stock. In this case, the optimal program is given by:

$$\theta(t) = \frac{\gamma_0 e^{rt}}{X}, \text{ where } \gamma_0 = \frac{\alpha\beta X}{\lambda}\left(\frac{rS}{1-\beta}\right)^{\beta-1} \tag{38}$$

$$F(t) = \frac{rS}{1-\beta} e^{-\frac{r}{1-\beta}t} \tag{39}$$

$$a(t) = \frac{\left(\lambda - \frac{r}{1-\beta}\right)}{X} F(t) \tag{40}$$

$$\gamma(t) = \gamma_0 e^{rt}. \tag{41}$$

Note that the planning period begins with pulse of drilling such that oil flow totaling $rS/(1-\beta) - F_0$ is immediately added to the inherited flow. Equation (40) implies that $a(t) > 0$ for all $t \geq 0$, since $F(t) > 0$ and since we have provisionally assumed that $\lambda(1-\beta) - r > 0$.[43] It is straightforward to verify that this program satisfies each of the necessary conditions (7)–(13) and is therefore optimal for our planner.[44] Since $a(t) > 0$, it achieves the same discounted utility as Hotelling's planner.

However, our planner cannot always accomplish this feat. Suppose that the rate of decay from drilled wells is too low, with $\lambda(1-\beta) < r$. To achieve the result in (18) that $U'(F(t)) = \frac{\lambda\gamma_0}{X}e^{rt}$, the planner must set $\dot{F}(t) = \frac{rU'(F(t))}{U''(F(t))} = -\frac{rF(t)}{1-\beta}$. But then equation (12) implies that $a(t)X = \lambda F(t) - \frac{rF(t)}{1-\beta} + \lambda F(t)$. Substituting and simplifying we conclude that $a(t)X = \frac{F(t)(\lambda(1-\beta)-r)}{1-\beta}$, which violates nonnegativity of $a(t)$. Intuitively, the planner is geologically constrained to a price path that rises more slowly than the rate of interest. Even if she drills all of the wells immediately and produces at the maximum possible rate, oil cannot be extracted quickly enough to satisfy Hotelling's rule.

[43] Since we have assumed that drilling and extraction are both costless in this example, any alternative drilling path such that the production path in (39) is feasible is also optimal. For any positive drilling cost, however, the planner would produce at the constraint and defer drilling until necessary.

[44] Equations (38) and (41) imply that the marginal condition (9) holds with equality. Equations (38) and (39) imply that condition (7) holds. Equation (41) ensures that (11) holds. Finally, equations (39) and (40) ensure that (12) holds.

Suppose instead that the inherited oil flow is too high, with $F_0 > rS/(1-\beta)$. In this case, in which it must be that $\lambda(1-\beta) > r$, the planner will choose not to drill any wells initially, and production will decline at rate λ until drilling commences (when $F_t = rS/(1-\beta)$). During this time, price will rise at a rate greater than r, and our planner will therefore not achieve the same utility as Hotelling's planner.[45]

To summarize, when the extraction technology involves drilling wells rather than producing barrels, we should not expect the Hotelling path to be optimal, unless four conditions hold: (1) the marginal cost of drilling a well is constant, (2) marginal utility is unbounded at zero, (3) the inherited rate of oil flow is not too high, and (4) the decay in flow from drilled wells is sufficiently fast. While the latter two conditions seem reasonable (given that new wells are constantly being drilled in the real world) the former two conditions are not tenable. Our analysis of the Texas data shows clearly that marginal costs rise with the rate of drilling, while the viability of alternative fuels at *current* oil prices argues against an unbounded oil price.

E Details for endogenous price model with scarcity

This appendix formally derives the properties of the optimal drilling and extraction path in the problem posed in section 4.4, under the assumption that production is always constrained (a future version of this paper will seek to verify that this assumption holds in this special case with $d'(a) = 0$ for a sufficiently small).

[45] Since the planner can produce below the production constraint, the standard Hotelling path can still be achieved. However, this possibility requires that drilling is costless. Otherwise, it can be shown that price would initially rise at the rate of interest (while production is below the constraint and drilling is zero), would then rise faster than the rate of interest (after the constraint starts to bind while drilling remains at zero), and finally would rise more slowly than at the rate of interest (after drilling turns positive with the constraint continuing to bind).

Marginal drilling costs are given by:

$$d(a(t)) = \bar{d}, \text{ for } a(t) \leq \bar{a}$$
$$d(a(t)) = \infty, \text{ for } a(t) > \bar{a},$$

where $\bar{a}$ is the capacity constraint on the drilling rate. Since the drilling rate is bounded, the rate of oil flow cannot jump instantaneously (or "pulse"). Time must elapse for flow to increase. We assume that the inherited flow of oil is relatively small ($F_0 < \bar{a}X/\lambda$) and that $U''' \geq 0$.

With this formulation, necessary condition (9) is replaced by:

$$a(t) \geq 0, \theta(t)X - [\bar{d} + \mu(t)] - \gamma_0 e^{rt} \leq 0, c.s. \tag{42}$$

The shadow price on the capacity constraint, denoted $\mu(t)$, is zero whenever the drilling rate is below the constraint ($a(t) < \bar{a}$); whenever $\mu(t)$ is strictly positive, drilling must be at its maximum feasible rate ($a(t) = \bar{a}$). Thus, the full marginal cost of drilling is given by $\bar{d}+\mu(t)$. We can interpret $\bar{d}+\mu(t)$ as the price to rent a drilling rig in a competitive market inclusive of the rent on scarce capacity.

The optimal program consists of three intervals. During the first interval, for $t \in [0, \hat{t}]$, drilling is set at the maximum feasible rate ($a(t) = \bar{a}$) with $\mu(t) > 0$ in condition (42). During the second interval, for $t \in (\hat{t}, \hat{T}]$, drilling proceeds at a slower rate ($a(t) \in (0, \bar{a})$) with $\mu(t) = 0$ in condition (42). Finally, during the third interval, for $t \in (\hat{T}, \infty)$, the drilling rate jumps down to zero ($a(t) = 0$) and oil flow decays exogenously. The boundaries between these three time intervals, given by $\hat{t}$ and $\hat{T}$, are determined endogenously.

How do drilling and oil flow evolve during these three intervals? During the first interval, drilling is set at the maximum rate and the flow of oil increases monotonically from its initial level of F_0 at a decreasing rate. To verify that $\dot{F}(t) > 0$ and $\ddot{F}(t) < 0$, note that equation

(12) simplifies to the following when $a(t) = \bar{a}$

$$F(t) = \frac{\bar{a}X}{\lambda} + k_0 e^{-\lambda t} \text{ for } t = [0, \hat{t}), \tag{43}$$

where $k_0 = F_0 - \bar{a}X/\lambda < 0$, given our assumption on inherited oil flow. Since $e^{-\lambda t}$ is a strictly decreasing, strictly convex function and $k_0 < 0$, $F(t)$ is a strictly increasing, strictly concave function for $t \in [0, \hat{t})$, as was to be proved.

During the second interval, the flow of oil decreases over time and this decrease grows in magnitude monotonically as time elapses ($\dot{F}(t) < 0, \ddot{F}(t) < 0.$). In addition, drilling must strictly decrease over time ($\dot{a}(t) < 0$). To establish these conclusions, differentiate condition (18) with respect to time to obtain:

$$U''(F(t))\dot{F}(t) = r\frac{\lambda\gamma_0}{X}e^{rt}, \tag{44}$$

recalling here and throughout that marginal drilling costs are constant. Use (44) to eliminate $\dot{F}(t)$ from the equation of motion in (12). Use (18) to eliminate γ_0, and finally solve for $a(t)X$:

$$a(t)X = \frac{r[U'(F(t)) - \frac{r+\lambda}{X}\bar{d}]}{U''(F(t))} + \lambda F(t). \tag{45}$$

Note that the first term on the right-hand side of equation (45) is negative. Subtracting $\lambda F(t)$ from both sides and recalling equation (12), it is clear that $\dot{F}(t) < 0$. Moreover, our assumption that $U'''(\cdot) \geq 0$ implies that the negative first term on the right-hand side of equation (45) grows in magnitude over time. This implies that $\ddot{F}(t) < 0$ and that $\dot{a} < 0$ for $t \in (\hat{t}, \hat{T}]$, as was to be proved.

During the third and final interval, drilling jumps down to zero. Denote the oil flow at the moment when drilling stops as $\hat{F} = F(\hat{T})$. Oil flow decays exponentially from $\hat{F}$ at the exogenous rate λ. Hence, $\dot{F}(t) < 0$ and $\ddot{F}(t) > 0$ for $t \in (\hat{T}, \infty)$.

How do drilling and oil flow behave at the boundaries? To begin, denote the number of wells drilled up through time t as $A(t)$. Since the drilling rate is bounded below by zero and above by $\bar{a} > 0$, there can be no upward or downward jumps in $A(t)$. The function is continuous although, as we will see, not differentiable everywhere. In the first interval, $A(t)$ rises linearly: $A(t) = \bar{a}t$ for $t \in [0, \hat{t})$. In the second interval,

$$A(t) = \bar{a}\hat{t} + \int_{\hat{t}}^{t} a(s)ds, \tag{46}$$

where $a(t)$ is given by (45) above. In the third interval, $A(t) = A(\hat{T}) = \bar{a}\hat{t} + \int_{\hat{t}}^{\hat{T}} a(s)ds$. The rate of drilling can jump up or down in this problem because, when $D(a(t)) = \bar{d}a(t)$, the Hamiltonian is linear in the control variable $a(t)$. However, the jumps are finite since $a(t)$ is bounded and cannot produce discontinuities in $A(t)$. They do, however, produce kinks—not only in $A(t)$ but also in $F(t)$. Kinks in both functions must occur at the boundary between the first and second interval ($\hat{t}$) and again at the boundary between the second and third interval ($\hat{T}$).

Now consider what occurs at the first of the two boundaries. Since $\dot{F}(t)$ is strictly positive throughout the first interval and strictly negative throughout the second interval, $F(t)$ must be kinked at the boundary between the two intervals ($\hat{t}$). The left-derivative there is strictly positive while the right-derivative is strictly negative. For this to occur, the rate of drilling must jump down at the end ($\hat{t}$) of the first interval: $\bar{a} > a(\hat{t}^{+})$. Since the left-derivative of $A(t)$ at $\hat{t}$ is $\bar{a}$ and the right-derivative is $a(\hat{t}^{+}), A(t)$ is also kinked at $\hat{t}$.

Consider next what occurs at the second of the two boundaries. Drilling activity jumps down to zero at $\hat{T}$. Since $\dot{A}(t) = a(t)$, there must be a kink at $\hat{T}$ in $A(t)$—in particular, the left derivative of $A(\hat{T})$ is strictly positive and the right derivative is zero. Since $\dot{F}(t) = a(t)X - \lambda F(t)$ and $F(t)$ is continuous, there is a kink at $\hat{T}$ in $F(t)$—its right derivative is negative and its left derivative is larger (strictly positive or at least not as negative).

Figure 8 depicts a simulation of our model under the assumption that $F_0 = 0$ and utility is quadratic (satisfying $U'(0) < \infty, U''' \geq 0$). The reader can verify that this figure illustrates the properties established above.

Finally, how do we determine the optimal program in any given instance? In particular, how do we determine the boundaries between intervals ($\hat{t}$ and $\hat{T}$) and the shadow value on the resource constraint (γ_0)? The first interval ends at $\hat{t}$. Since $\mu(t) = 0$ at the transition, condition (18) holds at the transition, and so the transition time as a function of γ_0 is given implicitly by the following expression:

$$U'\left(\frac{\bar{a}X}{\lambda} + k_0 e^{-\lambda\hat{t}}\right) - \frac{\lambda\gamma_0}{X}e^{r\hat{t}} = \frac{(r+\lambda)\bar{d}}{X}. \tag{47}$$

Meanwhile, it can be shown that oil flow at the moment the last well is drilled, denoted by $\hat{F}$, is independent of γ_0 and is defined implicitly by:

$$\frac{U'(\hat{F}) - \frac{r\bar{d}}{X}}{\lambda} = \int_0^\infty U'\left(\hat{F}e^{-\lambda t}\right)e^{-(r+\lambda)t}dt. \tag{48}$$

Briefly, from the moment the last well is drilled, oil flows exogenously. Thus, the marginal value of oil flow from that moment onward is a deterministic function of oil flow (the right-hand side). At the same time, the necessary conditions combine to yield the marginal value of oil flow prior to that moment as a function of oil flow (the left-hand side). Both conditions must hold simultaneously at the boundary, pinning down oil flow at the moment the last well is drilled.[46]

[46]Let $\hat{T}$ denote the endogenously chosen time that the last well is drilled. Imposing exogenously that the planner stop drilling at her optimal $\hat{T}$ will obviously not affect her maximized utility. Moreover, we can treat the utility accruing after $\hat{T}$ as a "scrap value" that depends on oil flow $\hat{F}$ at $\hat{T}$. Altering the program to obtain some different oil flow at $\hat{T}$ cannot raise the value of the scrap value at $\hat{T}$ plus the utility accruing before $\hat{T}$. Thus, at $\hat{T}$, the standard endpoint condition for the case of a free state variable and fixed terminal time with a scrap value applies:

$$\theta(\hat{T}) = e^{r\hat{T}}\frac{\partial}{\partial\hat{F}}\int_{s=\hat{T}}^\infty e^{-rs}U(\hat{F}e^{-\lambda(s-\hat{T})})ds.$$

The right-hand side is the benefit at $\hat{T}$ of endowing the final interval with marginally greater oil flow; the left-hand side is the marginal cost in terms of foregone net utility earned prior to $\hat{T}$. Differentiating, we obtain:

$$\theta(\hat{T}) = \int_{s=\hat{T}}^\infty U'\left(F(\hat{T})e^{-\lambda(s-\hat{T})}\right)e^{-(r+\lambda)(s-\hat{T})}ds.$$

These observations lead to a straightforward iterative method to determine γ_0, $\hat{t}$, and $\hat{T}$. The procedure is as follows. First, guess some $\gamma_0 \geq 0$. Second, compute the implied flow over the first two intervals. Third, assume that drilling stops when oil flow reaches $\hat{F}$, which is independent of the initial guess of γ_0. Fourth, compute the flow throughout the final interval. Fifth, check whether the cumulative flow over the three intervals matches the total resource stock ($S = (F_0 + R_0X)/\lambda$). Sixth, stop if the two sums match; otherwise, revise γ_0, recognizing that a higher γ_0 will result in a smaller cumulative oil flow. There can be at most one γ_0 that results from this solution algorithm, and every necessary condition is satisfied by this solution.

Cumulative oil flow is calculated as follows. Oil flow during the first interval is given by equation (43). The first interval ends at $\hat{t}$, which is defined implicitly by equation (47). Since the left-hand side of this equation is strictly decreasing in $\hat{t}$ there can be at most one solution to this equation. For future reference, note that if γ_0 were higher, the first interval would end sooner (smaller $\hat{t}$). Oil flow during the second interval is given by equation (18). Since the left-hand side of this equation is decreasing in $F(t)$, the equation uniquely defines $F(t)$ at every moment during the second interval. For future reference, note that if γ_0 were higher, the flow at every instant during the second interval would be uniformly smaller. Assume that the second interval ends when the flow drops to the $\hat{F}$ threshold defined above. Finally, oil flow during the final interval decays exponentially at rate λ from an initial level of $\hat{F}$, which is independent of γ_0.

Either the cumulative flow of oil over the three intervals matches $S = (F_0 + R_0X)/\lambda$ or it does not. If the cumulative flow of oil is larger than S, then we have not identified the

But from equation (18) and equation (42) we can write $\theta(\hat{T})$ as follows:

$$\theta(\hat{T}) = \frac{\bar{d}}{X} + \frac{\gamma_0 e^{r\hat{T}}}{X} = \frac{U'(F(\hat{T}) - \frac{r\bar{d}}{X}}{\lambda}.$$

Thus, eliminating $\theta(\hat{T})$ using these last two equations, we conclude:

$$\frac{U'(\hat{F}) - \frac{r\bar{d}}{X}}{\lambda} = \int_0^\infty U'\left(\hat{F}e^{-\lambda t}\right) e^{-(r+\lambda)t} dt.$$

optimum and should raise γ_0. The first interval (during which drilling is set to the constraint) will be shorter, while the second interval will have uniformly lower oil flow and will reach $\hat{F}$ sooner. The final interval will have unchanged cumulative oil flow. Hence, cumulative oil flow will be strictly smaller. By similar arguments, if cumulative flow is smaller than S, then we should lower γ_0. One implication is that if R_0 increased exogenously, then the equilibrium γ_0 must fall. This would result in a longer first interval, uniformly higher flow during the second interval, a later termination of the second interval, and an unchanged flow during the final interval.

Throughout this section, we have assumed that marginal utility is bounded ($U'(0) < \infty$) and that the rate of drilling is constrained ($\bar{a} < \infty$). If marginal utility were unbounded at the origin, then the equilibrium would be similar to what we have just described but without the final interval: drilling would never cease completely. This case differs from that in the appendix above because the drilling rate is constrained. If instead there were no constraint on the rate of drilling, then the first phase would disappear. At the first instant, the planner would drill wells at an *infinite* rate (a "pulse") so that equation (18) would hold immediately at $t = 0$ and would continue to hold until drilling ceased.