Computers and Chemical Engineering 141 (2020) 106988

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Information-theoretic sensor planning for large-scale production surveillance via deep reinforcement learning

Ashutosh Tewari*, Kuang-Hung Liu, Dimitri Papageorgiou

ExxonMobil Research and Engineering Company 1545 US 22, Annandale, NJ: 08801, USA

ARTICLE INFO

Article history:
Received 11 November 2019
Revised 20 May 2020
Accepted 24 June 2020
Available online 4 July 2020

Keywords:
Active sensing
Deep reinforcement learning
Markov decision process
Production surveillance
Sensor resource management

ABSTRACT

Production surveillance is the task of monitoring oil and gas production from every well in a hydrocarbon field. Accurate surveillance is a basic necessity for several reasons that include improved resource management, better equipment health monitoring, reduced operational cost, and ultimately optimal hydrocarbon production. A key challenge in this task, especially for large fields with many wells, is the measurement of multiphase fluid flow using a limited number of noisy sensors of varying characteristics. Current surveillance practices are based on fixed utilization schedules of such flow sensors, which rarely change over time. Such a *passive* mode of sensing is completely agnostic to surveillance performance and thus often fails to achieve a desired accuracy. Here we propose an *active* surveillance approach, underpinned by the concept of *value of information*-based sensing. Borrowing some well-known concepts from Markov decision processes, reinforcement learning and artificial neural networks, we demonstrate that a practical active surveillance strategy can be devised, which can not only improve surveillance performance significantly, but also reduce usage of flow sensors.

1. Introduction

In many industrial applications there is a need to probe a dynamic system whose state (often hidden) is of interest for informed decision making. This is typically achieved by gathering necessary sensor data, and using a state estimation scheme such as the Kalman filter or its variants. To achieve this goal for application domains with limited sensing resources, one may also seek strategies that plan sensor utilization by optimally navigating some cost-benefit trade-off. Such strategies are often domain specific and cannot be trivially designed. Even in sensor rich applications, the absence of a systematic sensor planning strategy can lead to the collection of huge volumes of data, albeit with little information therein. Consider a large onshore hydrocarbon field such as the one shown in Fig. 1. The goal of production surveillance, in this setting, is to accurately monitor the individual production rates of thousands of wells. This is challenging for many reasons, such as disparate, multiphase flow-meters of varying fidelities and costs, well-comingling, evolving equipment and subsurface conditions, time varying sensor noise, and the sheer number of wells to be monitored continuously (Poulisse et al., 2006). The current surveillance practice in the *oil and gas* (O&G) industry is primarily based on predetermined sensor utilization schedules. As a result, it may be possible to monitor production at a coarser level (a well group), but gets extremely difficult to accurately monitor production from individual wells. In this work, we propose an active sensor planning strategy for production surveillance, where the sensors no longer adhere to a fixed utilization schedule. Instead, at every time step (e.g., a day) an optimal utilization plan is identified on the fly.

Our work makes the following three contributions. **First**, we tackle a practical problem of production surveillance, which is increasingly becoming important with the surge in onshore, unconventional oils fields with thousands of wells. This is a nascent area in O&G research, since historically the bulk of production came from onshore fields with only a few wells, which didn't necessitate sophisticated surveillance approaches. With the goal of large-scale deployment, we cast production surveillance as a sequential decision problem under uncertainty, thus opening the door to a rich and well-studied area in mathematics. **Second**, a sequential decision problem when formalized as a *Partially Observed Markov Decision Process* (POMDP) involves a crucial step of learning a policy function that allows fast and near optimal decision making in the field. To this end, we first reformulate a POMDP as a *Markov Decision Process* (MDP) in the belief space, and then appeal to deep neural networks to approximate the optimal policy function. The efficacy of our approach is demonstrated on a simulated testbed,

* Corresponding author.
E-mail addresses: ashutosh.tewari@exxonmobil.com (A. Tewari), kuang-hung.liu@exxonmobil.com (K.-H. Liu), dimitri.j.papageorgiou@exxonmobil.com (D. Papageorgiou).

https://doi.org/10.1016/j.compchemeng.2020.106988
0098-1354/

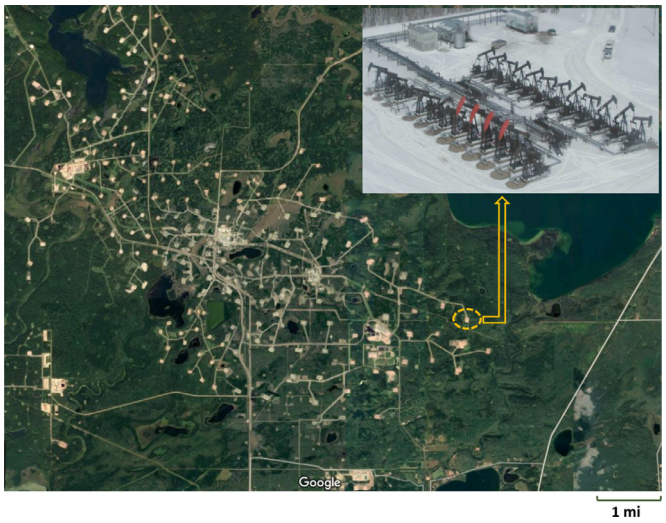


Fig. 1. Aerial view of a large North American oil field with thousands of wells. Each small dot in the image is a *well pad*, i.e., a cluster of wells designed to produce from a specific zone of a hydrocarbon reservoir in the subsurface. Hydrocarbon surveillance in this setting entails continuous monitoring of water, oil, and gas production from each individual well.

which shows improved surveillance performance both in terms of accuracy and cost. **Third**, grounded on the presented concepts, we seed the idea of *autonomous surveillance* as a new paradigm of surveillance in the O&G industry. With the global push for digitalization, recent advancements in the autonomous vehicles and low-cost multiphase flow sensors, autonomous surveillance is realizable in not so distant future.

The organization of this paper is as follows. A brief review on related works is presented in Section 1.1. Section 1.2 sets the notations to be used in the paper. Section 2 describes an experimental setup that we use as a testbed in this study, and highlights the challenges with surveillance in this setting. The formulation of active surveillance as a sequential decision problem is presented in Section 3, with necessary details on all of its crucial components. Results on the experimental testbed are included in Section 4 and claims are corroborated. Section 5 summarizes this work with some remarks on the future of production surveillance in O&G industry.

1.1. Related work

As described in Chapter 8 of Krishnamurthy (2016), the sensor planning framework studied in this paper belongs to a general class of problems known under several names including active sensing, measurement control, sensor scheduling, sensor control, sensor resource management, or simply sensor management. Throughout we will use sensor planning and sensor resource management interchangeably. For the sake of clarity, we adopt the definition given in Xiong and Svensson (2002) to formally define this general class of problems:

> Multi-sensor management is formally described as a system or process that seeks to manage or coordinate the usage of a suite of sensors or measurement devices in a dynamic, uncertain environment, to improve the performance of data fusion and ultimately that of perception.

Sensor resource management has a multi-decade tradition that includes successful commercial applications in radar, sonar, guidance, navigation, air traffic control, and space exploration dating back to the 1960s (Mallick et al., 2012). As a cornerstone in multi-target tracking, sensor management has been used successfully for years to allocate sensing resources in a sensor network to detect, surveil, track, and classify objects (or targets) in dynamic space, air, land, and sea environments.

Within the oil and gas community, sensor planning has not been widely adopted, an observation that motivates the importance and timeliness of this work. Satellite synthetic aperture radar has been used in "oil spill detection" to classify incidents of oil due to natural seeps, oil extraction, transportation, and consumption (Caruso et al., 2013). Similar applications have been used to detect and monitor gas flaring. Within the context of oil and gas production, only passive surveillance has been described in the academic literature. Poulisse et al. (2006) describe a FieldWare PRODUCTION UNIVERSE tool that estimates real time well production rates from simple field measurements. The tool's novelty lies in how it moves beyond traditional methods of surveillance by integrating bulk measurements from multiple wells.

Central to all sensor planning frameworks is the metric (or objective function) employed to weigh trade-offs between various sensor actions and the benefits from these actions. At their core, information-theoretic planning approaches employ information-theoretic metrics to choose sensing actions that favor information gain, for instance, actions that maximize prior to posterior entropy reduction. Note that. while the overall system may try to optimize a different metric, e.g., maximize expected oil production or target detection, a major advantage of an information-theoretic approach is that "it simplifies system design by separating it into two independent tasks: information collection and risk/reward optimization" Hero et al. (2008, p.34). Although this separation of tasks can be beneficial for most applications, it must be done judiciously for certain others. For example, Papageorgiou and Raykin (2007) demonstrate, for missile defense application, how methods greedily maximizing information gain can be inferior to methods using risk-based metrics as the former can divert scarce resources to improve classification of low-value objects instead of targeting classification of high-value ones.

Information-theoretic approaches to sensor planning have appeared in a number of research papers. Wang et al. (2005) employ information-theoretic metrics for sensor selection and sensor placement in sensor networks. The fusion of the selected sensor's observation with the prior target location distribution yields the greatest reduction of the entropy of the posterior target location distribution. Kreucher et al. (2005) proposed an approach to schedule agile sensors for multiple target tracking in a dynamic environment using Renyi entropy. Using a novel distributed approach, Yang et al. (2007) attempt to coordinate a decentralized network of mobile sensors tracking multiple moving targets. Each sensor agent maintains its own local set of track estimates and selects actions to maximize the expected information gain relative to its current uncertainty estimate, while communicating its information with a small number of neighbors. Ryan and Hedrick (2010) apply receding horizon control while maximizing information gain to schedule a single mobile sensor platform to track a moving target using a camera mounted on a fixed-wing unmanned aircraft. Jenkins and Castanón (2011) introduce an information-based sensor planning framework to classify a collection of objects based on their observed features. Using predictions of the information value from individual measurements, they task sensors to more efficiently collect additional observations. Likewise, La et al. (2014) employ a planning framework to map a scalar field, e.g., the concentration of an algae bloom or oil spill, using multiple distributed mobile sensors. The overall objective is to coordinate the sensor network to maintain a multivariate probability density function on the scalar field with reasonably low entropy.

A typical mathematical formalization of sensor planning problem is in terms of a *Partially Observed Markov Decision Process*

(POMDP). The term "partial" in POMDP indicates that the states of a system are hidden and can only be inferred indirectly via sensor measurements. Solving POMDPs is computationally hard, and a common solution approach is to seek optimal decisions in the *belief-space* (the space of all possible probability distributions over the hidden states) (Kaelbling et al., 1998). In other words, sensing decisions are based on the entire distribution over the states and not just the most likely ones. A hallmark of belief-space planning is that the dynamic system becomes observable in the belief-space and thus can be formulated as a *Markov Decision Process* (MDP). MDPs (a.k.a stochastic dynamic programming) have been extensively studied with numerous algorithmic solutions to them. In terms of sensor resource management, Washburn et al. (2002) applied an approximate dynamic programming algorithm based on the Gittins index rule for multi-armed bandit problems to a multi-target tracking application, to "take sensing actions that maximize the expected amount of information extracted from a scene of interest". The receding horizon control approach of Ryan and Hedrick (2010) is essentially an MDP subject to stochastic nonlinear models for both target motion and sensors. Particle filtering is used in an attempt to overcome the complex conditional entropy prediction step. Platt et al. (2010) cast a POMDP as a deterministic MDP on belief space along with a certain assumption on the observation process, and apply standard control techniques for sensor planning. In this sense, the proposed sensor planning approach closely mimics that of (Platt et al., 2010), except that instead of finding the optimal policy as a solution to an optimization problem, we approximate it using a deep neural network. The advantage being, once trained in an offline manner, the deployment of neural networks for real-time planning becomes trivial.

1.2. Notation

The lowercase letters are used for scalars, bold lowercase letters for vectors, uppercase letters for matrices/sets. The subscript is reserved for time indexing, and the superscripts are used to denote an element of a vector or a matrix. For example, $\boldsymbol{x}^i$ and X^{ij} denote the ith and the (i, j)th elements of a vector $\boldsymbol{x}$ and a matrix X, respectively. Likewise, $X^{i:}$ (or $X^{:\,i}$) represents the ith row (or column) of the matrix X. The cardinality of a set X is denoted as $|X|$. The operators $\mathbb{E}, P,$ and p denote expectation, probability distribution and probability density, respectively. Additionally, a list of symbols frequently appearing in the paper is included in Table 1, for quick reference.

2. Problem setup

Consider a setting where we are interested in inferring two-phase flow rates (oil and water) at time t from n wells using a subset of a total of m sensors. Compared to the number of wells to be monitored, the number of sensors is fewer, i.e. $m < n$. Additionally, the sensors have varying fidelities and costs with more accurate sensors typically being costlier. In hydrocarbon surveillance applications, these sensors are multiphase flow meters ranging from expensive *test-separators* that measure absolute flow rate of each phase, to *spin-cuts* which are manual measurements of water-to-fluid ratio taken at a well-head, to extremely crude *inferred* measurements of total liquid rate generated by artificial-lift systems. These sensors have widely different and non-uniform utilization schedules, and time varying measurement noise. Additionally, to reduce the instrumentation costs, the wells are typically grouped together (i.e., *comingled*) for expensive flow measurements, for example via test-separators. Many such well groups share a test-separator and take turns to utilize this shared resource based on a fixed schedule. There is often a provision to *shut-in* one or more wells while making a test-separator measurement, which incurs a cost (lost production) but improves the information content of the data.

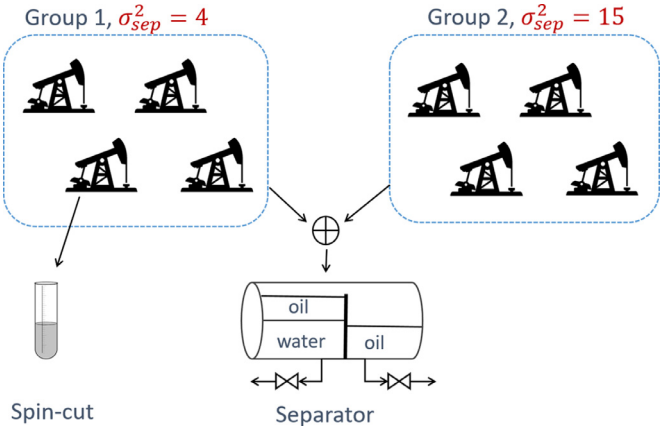


Fig. 2. Our small-scale surveillance setup to be used as a testbed to compare and contrast different surveillance methods. The setup includes two groups of four wells, and two types of sensor; the *test-separator* and the *spin-cut*. The former measures the absolute oil and water rates for a group of well with a utilization cost of \$10. The latter, however, measures the water-to-fluid ratio of each individual well and has a utilization cost of \$50. The cost numbers are rather arbitrary, and merely reflect the fact that the manual spin-cuts are more expensive than the automated test-separator measurements. Nevertheless, for actual applications these numbers need to be carefully set by the field manager.

Fig. 2 shows a surveillance setup, with some of the aforementioned characteristics, that will serve as a testbed to demonstrate our proof of concept. The setup consists of 8 wells (with 2 well groups), 1 test-separator and 1 operator (who takes spin-cut measurements). The deployment costs of test-separator and spin-cut are \$10 and \$50, respectively (see figure caption for details). Since, spin-cuts are precise measurements of water-to-fluid ratio, the measurement noise variance is set at 0.005 (note that this noise is on an entity $\in [0, 1]$). On the other hand, the test-separator is assumed to have a variable noise variance of 4 and 15 units on absolute flow rate measurements, depending on which well group is being measured (cf. Fig. 2). This is a common phenomenon in actual fields, where there are excessive gas producing wells, which negatively impact the accuracy of test-separator measurements as the gas infringes with the gravimetric separation of the multiphase flow. Finally, we assume that at every time step (a day in this study), one sensing action is taken. For the given setup, the space of all sensing actions constitutes of 39 mutually exclusive possibilities as enumerated by the flow-chart in Fig. 3.

Despite its relatively small size, the task of surveillance in this experimental testbed is very challenging for the following reasons; a) well-comingling generates data that does not inform individual well production rates, b) sensor noise is variable e.g. the test-separator measurement of group-2 is noisier than that of group-1, c) spin-cuts provide valuable information but are infrequent due to high utilization cost, and d) the lack of an *informed* method to coordinate shut-ins during test-separator measurements. The goal of our work is to devise a sensor planning strategy based on an information-theoretic optimality criterion that identifies the best information gathering sensing action at every time step (see Section 3.2 for details). Lastly, although our experimental testbed is small-scale, the methodology presented in this paper can be extended to a field with thousands of wells, multiple operators, and many test-separators or other types of multiphase flow sensors, including virtual sensors (Bikmukhametov and Jäschke, 2019), not considered here.

Table 1
List of frequently used symbols and their brief descriptions.

Symbol	Description
n	number of wells
m	number of flow sensor of various modalities
$\boldsymbol{x}_t \in \mathbb{R}_+^{2n}$	water and oil rates from n wells at time t
$\boldsymbol{y}_t$	sensor measurements at time t
$A \in \mathbb{R}^{2n\times 2n}$	linear operator that maps $\boldsymbol{x}_t \to \boldsymbol{x}_{t+1}$
h	measurement function that maps $\boldsymbol{x}_t \to \boldsymbol{y}_t$
H	linearized function of h at a certain point
Q_t	process noise covariance matrix at time t
R_t	diagonal measurement noise covariance matrix
$\rho_t : \mathbb{R}^{2n\times 2n} \to \mathbb{R}_+$	posterior density function of the state, $\boldsymbol{x}_t$, at time t
$\mathcal{B}$	belief-space where posterior density functions lies i.e. $\rho_t \in \mathcal{B}$
$\mathcal{A}$	discrete space of all possible sensing actions
$\pi : \mathcal{B} \to \mathcal{A}$	a decision function (policy) that maps belief-space to action space
π^*	an optimal policy with respect to POMDP's optimality criterion
$c : \mathbb{R}^{2n} \times \mathcal{A} \to \mathbb{R}_+$	surveillance cost of taking some sensing action at some state
$\tilde{\rho}_t$	Gaussian approximation of the ρ_t
$\boldsymbol{\mu}_t$, Σ_t	the mean and the covariance matrix of $\tilde{\rho}_t$
$\tilde{c}$	expected value of c with respect to $\tilde{\rho}$
Θ_t	joint representation of $\boldsymbol{\mu}_t$ and Σ_t i.e. $\Theta_t = \{\boldsymbol{\mu}_t, \Sigma_t\}$
$\beta \in \mathbb{R}_+$	scalar that controls the trade-off between sensing cost and the information metric
$\gamma \in [0, 1)$	the discount factor used to define a POMDP's infinite horizon cost
$\mathcal{Q}^\pi : \mathcal{B} \times \mathcal{A} \to \mathbb{R}_+$	infinite horizon cost associated with starting in some state, taking some action and following the policy π thereafter
$\tilde{\mathcal{Q}}$	neural network approximation of the $\mathcal{Q}$-function (referred to as a $\mathcal{Q}$-network)
$\boldsymbol{\xi}$	tunable parameters of a $\mathcal{Q}$-network

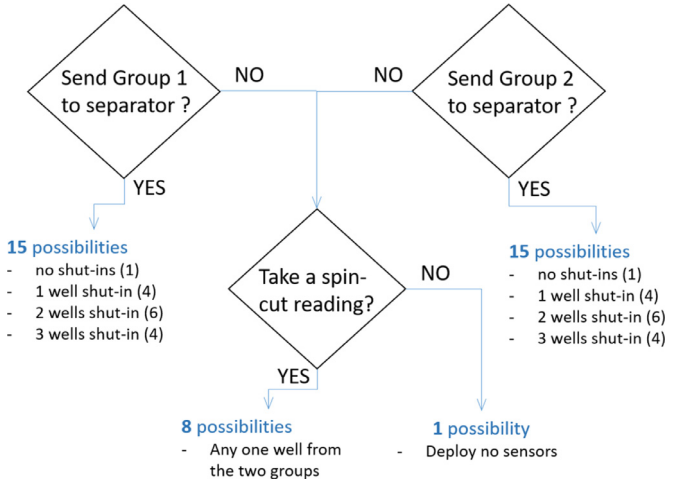


Fig. 3. A flowchart illustrating the origin of a 39-dimensional action space for the surveillance setup shown in Fig. 2. Each terminal node leads to a set of possible sensing actions, yielding a total of 39 mutually exclusive actions.

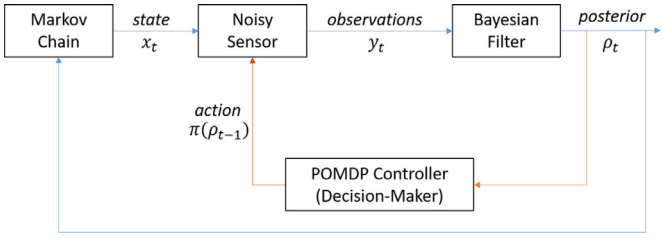


Fig. 4. A schematic of the POMDP formulation for the active-surveillance problem. The unobserved states $\boldsymbol{x}_t$ represent the per-well, per-phase (oil and water) flow rates, and the observations $\boldsymbol{y}_t$ represent the noisy field measurements. The outer loop corresponds to a Bayesian filter, which assimilates new measurements and updates the posterior distribution on $\boldsymbol{x}_t$ in a recursive fashion. The inner loop decides an optimal (with respect to a well-defined metric) sensing action to take for the next time step. The focus of this work is to develop a practical, mathematical approach for the inner control loop, while the outer loop follows the filtering approach proposed by Tewari et al. (2018).

3. Methodology

3.1. Formalizing active-surveillance as a POMDP

Let $\boldsymbol{x}_t \in \mathbb{R}_+^{2n}$ represent the production rates (of water and oil) from n wells, at time t, that need to be inferred from m noisy measurements arranged in a vector $\boldsymbol{y}_t$. We leverage the dynamic production model proposed by Tewari et al. (2018), wherein the rates admit a linear-Gaussian dynamics with time varying noise Q_t as shown in Eq. (1a). The measurements, however, can be a nonlinear function of the rates (such as spin-cuts), again with time varying noise as shown in Eq. (1b).

$$\boldsymbol{x}_t \sim \mathcal{N}(A\boldsymbol{x}_{t-1}, Q_t) \tag{1a}$$

$$\boldsymbol{y}_t \sim \mathcal{N}(h(\boldsymbol{x}_t), R_t) \tag{1b}$$

The symbol $\mathcal{N}$ above denotes a *multivariate-normal* distribution. Tewari et al. (2018) proposed a finite-dimensional Bayesian filter for this state space model. For efficient recursion, a *mixture of Gaussians* (MoG) was employed as the filtering density, and the posterior was sampled using the *Hamiltonion Monte Carlo* (HMC) algorithm (Hoffman and Gelman, 2014), at every time step. The need for sampling, in their framework, arose mainly because the posterior no longer enjoyed a closed-form representation due to nonlinear measurements and non-negativity constraints on the states (flow rates are positive). The HMC algorithm, a *Markov Chain Monte Carlo* (MCMC) method, utilizes gradient of the unnormalized posterior to design an efficient proposal mechanism that greatly speeds up posterior sampling. The samples so generated can then be fitted to a MoG distribution, to be used as the prior for the next time step.

In this work, we expand this *passive* surveillance framework to include an additional step that determines an *optimal* sensing action for the next time step. We cast this problem as a POMDP, the schematic of which is shown in Fig. 4. The outer loop depicts the aforementioned filtering framework as in Tewari et al. (2018). We focus on devising a decision function (a.k.a *policy* function)

$\pi : \mathcal{B} \to \mathcal{A}$ that identifies a sensing action (see the inner loop), given the current posterior $\rho_t \in \mathcal{B}$. The symbols $\mathcal{B}$ and $\mathcal{A}$ denote the belief space i.e. the space spanned by the posterior distributions, and the finite set of all possible sensing actions, respectively (39 in our test case, cf. Fig. 3). Therefore, a policy function maps a function space, $\mathcal{B}$, to a discrete action space. For a given policy π, a POMDP defines an *expected discounted cost* over an infinite horizon, starting from an initial state distribution ρ_0, as

$$\mathbb{E}\Big[c(\boldsymbol{x}_0, \pi(\rho_0)) + \gamma c(\boldsymbol{x}_1, \pi(\rho_1)) + \gamma^2 c(\boldsymbol{x}_2, \pi(\rho_2)) + \cdots\Big]; \quad \boldsymbol{x}_0 \sim \rho_0, \tag{2}$$

where $c(\boldsymbol{x}_t, \pi(\rho_t))$ is the instantaneous cost associated with the current state and the action taken by following the policy π. The discount factor $\gamma \in [0, 1)$ progressively down-weights future costs. The expectation ($\mathbb{E}[\cdot]$) is taken with respect to the joint distribution of states, $p(\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2 \cdots | \boldsymbol{y}^{\pi(\rho_0)}, \boldsymbol{y}^{\pi(\rho_1)}, \boldsymbol{y}^{\pi(\rho_2)} \cdots)$, under the specified dynamics and the sensor selection policy π. Since the total discounted cost (the bracketed term in Eq. (2)) is a linear combination of cost functionals at each time step, the expectation in (2) further simplifies to

$$\mathbb{E}[c(\boldsymbol{x}_0, \pi(\rho_0))] + \gamma \mathbb{E}[c(\boldsymbol{x}_1, \pi(\rho_1))] + \gamma^2 \mathbb{E}[c(\boldsymbol{x}_2, \pi(\rho_2))] + \ldots, \tag{3}$$

where each expectation is now taken with respect to the marginal distributions of the states $\boldsymbol{x}_t$ i.e. $p\big(\boldsymbol{x}_t | \boldsymbol{y}^{\pi(\rho_0)}, \boldsymbol{y}^{\pi(\rho_1)}, \ldots, \boldsymbol{y}^{\pi(\rho_t)}\big)$. Hence, any policy π is characterized by the resulting expected total discounted cost given by Eq. (3), and our goal is to find an optimal policy π^* which satisfies

$$\pi^* = \underset{\pi(\cdot)}{\operatorname{argmin}} \ \Big(\mathbb{E}[c(\boldsymbol{x}_0, \pi(\rho_0))] + \gamma \mathbb{E}[c(\boldsymbol{x}_1, \pi(\rho_1))] + \gamma^2 \mathbb{E}[c(\boldsymbol{x}_2, \pi(\rho_2))] + \cdots\Big), \tag{4}$$

for a given a surveillance cost functional $c(\boldsymbol{x}_t, \pi(\rho_t))$. An appropriately chosen cost functional, $c(\boldsymbol{x}_t, \pi(\rho_t))$, is crucial in a POMDP as it directly impacts the optimal policy that we wish to learn. In the next section, we explain the thought process behind defining this cost and elucidate why the resulting policy function is optimal from an information-theoretic viewpoint.

3.2. Surveillance cost

For surveillance applications, where the goal is to infer unknown states from noisy sensors, there are explicit and implicit cost components. The former is the tangible cost of deploying a sensor, which could be fixed or variable (some function of the state) and is usually straightforward to quantify; for example a spin-cut costs \$50 in our experimental testbed (see Fig. 2). The latter, however, is somewhat tricky to define and pertains to the uncertainty (or the information content) in our state estimates. One may rightly argue that a higher uncertainty, generally, should translate to a higher value of some quantifiable, physical cost. Nevertheless it is hard to establish, in practice, the "right" functional mapping from a measure of uncertainty to this true cost. Even more so, the "right" mapping would change from one application to the other. To elucidate this point, consider a specific problem of *choke*-optimization subject to a water-handling constraint. Such problems may arise when field production has to account for constraints imposed by facilities. For instance, if water-handling facility reaches a limit then the production has to be curtailed to meet that constraint. This is accomplished by scaling down the production of a well (known as *choking*) by a factor $\delta \in [0, 1]$. The optimal choke-factor for n wells ($\boldsymbol{\delta} = [\delta^1, \delta^2, \ldots, \delta^n]$) can be determined by casting an optimization problem to maximize expected oil production while maintaining a low probability of exceeding the water limit, i.e.

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta} \in [0,1]^n}{\operatorname{argmax}} \ \mathbb{E}[(\mathbf{1} - \boldsymbol{\delta})^T \boldsymbol{o}] \quad \text{s.t.} \quad P((\mathbf{1} - \boldsymbol{\delta})^T \boldsymbol{w} > l) \leq \epsilon = 0.05. \tag{5}$$

The objective function is the expected oil production for a given set of choke factors. For simplicity, it is assumed that the production rate varies linearly with the choke factor $\boldsymbol{\delta}$. The random variables $\boldsymbol{o}, \boldsymbol{w} \in \mathbb{R}^n_+$ are the oil and water production rates from n wells, and l is the water limit imposed by the facility. The problem in (5) is a stochastic optimization problem, since $\boldsymbol{o}$ and $\boldsymbol{w}$ are known only through their distributions. Fig. 5 shows an instance of these distributions for a case with $n = 3$ wells.

The constraint is a *chance*-constraint specifying the risk of exceeding the water limit in probabilistic terms. For a given distribution of oil and water rates, the optimization problem in (5) can be solved to obtain optimal choke factors. Of course, the uncertainty in the rate estimates have a profound effect on the solution. For instance, high variance estimates may cause the optimizer to yield high choke factors for high oil producers. Fig. 6 depicts the relationship between the uncertainty (total variance) and the cost (production loss) for two identical choke optimization problems, except with different water limit (l) values. Even though the true cost (production loss) increases with uncertainty, the dependence between the two entities is quite different for the two cases. This observation corroborates our previous claim that establishing the "right" mapping from the uncertainty to the true cost is very hard as it can vary significantly between different applications, or even within one application as shown in Fig. 6.

In light of the preceding discussion —and in addition to the fact that the goal here is to learn an optimal sensing policy that is not tied to any one subsequent decision problem— we adopt a simplified viewpoint and define the cost functional as

$$c(\boldsymbol{x}_t, \pi(\rho_t)) = \boldsymbol{f}^{\pi(\rho_t)} + M^{\pi(\rho_t)} \boldsymbol{x}_t + \beta (\boldsymbol{x}_t - \mathbb{E}[\boldsymbol{x}_t])^T (\boldsymbol{x}_t - \mathbb{E}[\boldsymbol{x}_t]). \tag{6}$$

The first term denotes the fixed cost associated with deploying a sensor. The vector $\boldsymbol{f} \in \mathbb{R}^{|\mathcal{A}|}_+$ encodes such fixed costs with $\boldsymbol{f}^i$ specifying the deployment cost of the ith sensor. Some of the sensing actions in our setup involve well shut-ins that incur temporary production loss. The second term captures this variable cost via a matrix $M \in \mathbb{R}^{\mathcal{A} \times 2n}$ whose ith row specifies a binary vector corresponding to the wells that are shut-in. Lastly, the third term assigns cost based on the deviation of the states from their expected value, thereby penalizing high variance (uncertain) states. The coefficient β provides a knob to adjust the cost associated with the uncertainty such that it is commensurate with the sensor deployment costs (the first two terms). A very large value of β would lead to policies which frequently deploy sensors and vice versa. We set this value to 0.1 for our analysis. Again, β is a design variables that needs to be carefully set by a field manager based on the field's sensing infrastructure. The inclusion of the third term in the surveillance cost makes the proposed sensor planning method information-theoretic in nature.

In order to use Eq. (6) in the optimization problem in (4), its expectation with respect to the posterior ρ_t needs to be computed, as done in Eq. (7) below

$$\mathbb{E}[c(\boldsymbol{x}_t, \pi(\rho_t))] = \boldsymbol{f}^{\pi(\rho_t)} + M^{\pi(\rho_t)} \mathbb{E}[\boldsymbol{x}_t] + \beta \ \mathbb{E}\Big[(\boldsymbol{x}_t - \mathbb{E}[\boldsymbol{x}_t])^T (\boldsymbol{x}_t - \mathbb{E}[\boldsymbol{x}_t])\Big]. \tag{7}$$

The first term remains unchanged as it denotes a deterministic, fixed measurement cost. The expectation of the other two terms is easy to compute, following certain assumptions, as shown in the next section.

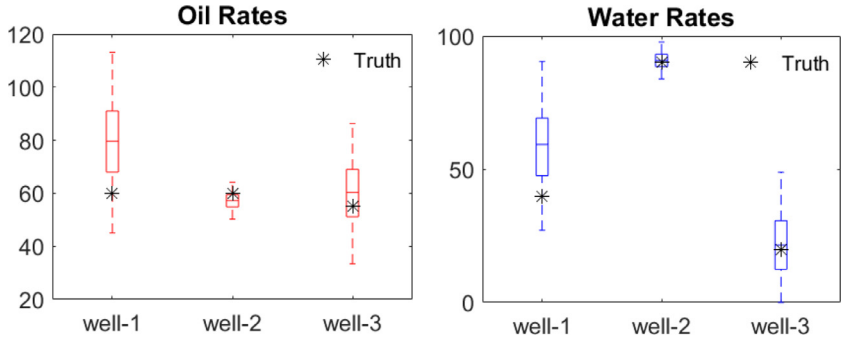


Fig. 5. Illustration of uncertainty in states (oil and water production rates) for a 3-well case. The distributions are plotted as box-plots with the box and outside bars showing the interquartile and the interdecile (p10-p90) ranges, respectively. The stochastic optimization problem in (5) when solved in this setting, with $l = 60$, yields the optimal choke factors as $\boldsymbol{\delta}^* = [0.35 \quad 1.0 \quad 0.0]$.

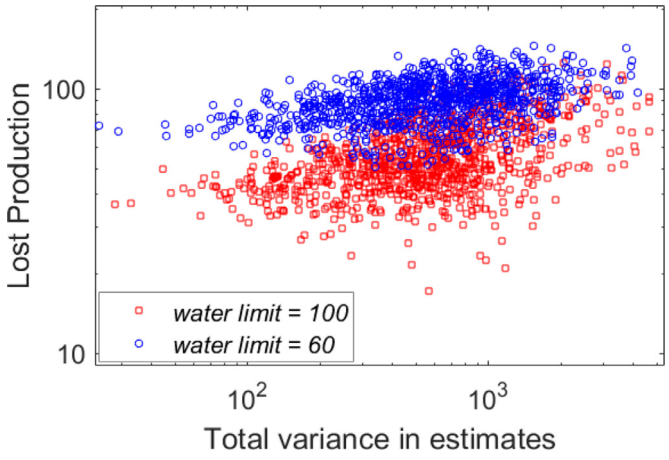


Fig. 6. The dependence of the production loss (due to optimal choking) on the total variance in the rate estimates. The scatter plot is obtained by first generating multiple scenarios similar to the one in Fig. 5, and then solving the stochastic optimization problem in (5) for each scenario, under two different water limit values (60 or 100). The production loss is calculated as $(\mathbf{1} - \boldsymbol{\delta}^*)^T \boldsymbol{o}_{true}$.

Remark: In the third term of the surveillance cost of Eq. (6), one could penalize a low-probability state (or negative log-posterior value of a state), instead of the deviation from the mean. This choice would replace $\mathbb{E}\big[(\boldsymbol{x}_t - \mathbb{E}[\boldsymbol{x}_t])^T(\boldsymbol{x}_t - \mathbb{E}[\boldsymbol{x}_t])\big]$ with $\mathbb{E}[-\log(\rho_t(\boldsymbol{x}_t))]$ in Eq. (7). The latter is known as *differential entropy*, a commonly used information-theoretic measure for continuous random variables. In Section 3.3 we remark on why differential entropy may not be desirable for our surveillance application.

3.3. POMDP to MDP

In an attempt to obtain an optimal policy that satisfies Eq. (4), we make some simplifying assumptions that enable us to recast the aforementioned POMDP as a MDP. The reformulation as an MDP allows us to seek an optimal policy as a solution to *Bellman's* dynamic programming recursive equation as discussed in Section 3.4. Before that, we make the following assumptions that facilitate POMDP to MDP reformulation:

- The posterior ρ_t at time t, which is a MoG, is approximated by a single Gaussian (say $\tilde{\rho}_t$) with mean ($\boldsymbol{\mu}_t$) and covariance (Σ_t) as the sufficient statistics. This approximation is straightforward with $\boldsymbol{\mu}_t$ and Σ_t being the convex combination (determined by the mixing proportions) of the individual components' mean vectors and covariance matrices.
- The nonlinear measurements are linearized and the non-negativity constraints on the states are removed. This approximation results in a linear-Gaussian system with a tractable filtering density (a Gaussian).
- The future sensor measurements values are assumed to coincide with their predictive mean value under the linear-Gaussian model resulting from the previous two approximations.

We now analyze the impact of these approximation. Due to the first approximation above, the expected value of the cost functional in Eq. (7) can be obtained as

$$\mathbb{E}[c(\boldsymbol{x}_t, \pi(\tilde{\rho}_t))] = \boldsymbol{f}^{\pi(\tilde{\rho}_t)} + M^{\pi(\tilde{\rho}_t):}\boldsymbol{\mu}_t + \mathrm{tr}(\Sigma_t), \tag{8}$$

where the expectation is now with respect to the approximated posterior, $\tilde{\rho}_t$. Note that the RHS of Eq. (8) can be completely specified in terms of the sufficient statistics $\{\boldsymbol{\mu}_t, \Sigma_t\}$ of ρ_t. Hence, the expected cost can be denoted as $\tilde{c}(\Theta_t, \pi(\Theta_t))$ i.e.

$$\tilde{c}(\Theta_t, \pi(\Theta_t)) = \boldsymbol{f}^{\pi(\Theta_t)} + M^{\pi(\Theta_t):}\boldsymbol{\mu}_t + \mathrm{tr}(\Sigma_t), \tag{9}$$

where $\Theta_t = \{\boldsymbol{\mu}_t, \Sigma_t\}$ is the joint representation of the sufficient statistics. The policy function, $\pi : \mathbb{R}^n_+ \times \mathbb{S}^n_+ \to \mathcal{A}$, now maps the joint space defined by positive real vectors and positive definite matrices to a discrete action space. On substituting Eq. (8) in the expression (3) and using the shorthand from Eq. (9), an MDP on Θ_t emerges, having a total discounted cost of

$$\tilde{c}(\Theta_0, \pi(\Theta_0)) + \gamma\tilde{c}(\Theta_1, \pi(\Theta_1)) + \gamma^2\tilde{c}(\Theta_2, \pi(\Theta_2)) + \cdots. \tag{10}$$

Finally, to complete the MDP formulation we need a transition function for the state Θ_t. To this end, we invoke the last two of the three aforementioned approximations, which yield a deterministic dynamics on the joint state Θ_t as

$$\boldsymbol{\mu}_t = A\boldsymbol{\mu}_{t-1}, \tag{11a}$$

$$\begin{aligned}\tilde{\Sigma}_t &= A\Sigma_t A^T,\\ \Sigma_t &= \tilde{\Sigma}_{t-1}\Big(I^{2n} - [H^{a_t:}]^T\big(H^{a_t:}\tilde{\Sigma}_{t-1}[H^{a_t:}]^T - R^{a_t a_t}\big)^{-1}H^{a_t:}\tilde{\Sigma}_{t-1}\Big)\end{aligned} \tag{11b}$$

where a_t is obtained as $a_t = \pi(\Theta_{t-1})$, I^{2n} is an identity matrix, and H a $|\mathcal{A}| \times 2n$ matrix obtained after linearization of the nonlinear observation model [$h(x)$ in (1b)] at the previous estimate's mean value $\boldsymbol{\mu}_{t-1}$. Eqs. (11a) and (11b) are, essentially, the Kalman filter update equations for our approximated linear-Gaussian system, where observations are made based on the following procedure: using the policy π, pick a sensing action for the next time step i.e. $a_t = \pi(\Theta_{t-1})$, then record the measurement as the predictive mean value i.e. $y_t = H^{a_t:}\big(A\boldsymbol{\mu}_{t-1}\big)$. A similar idea that defines an approximate dynamics on belief state was proposed by Platt et al. (2010), which they referred to as *planning assuming maximum likelihood observations*.

Having a completely specified MDP, with deterministic, nonlinear dynamics as shown in Eqs. (11a) & (11b), and a total discounted cost given by Eq. (10), our goal is to seek an optimal policy,

$$\pi^* = \underset{\pi(\cdot)}{\text{argmin}} \; [\tilde{c}(\Theta_0, \pi(\Theta_0)) + \gamma \tilde{c}(\Theta_1, \pi(\Theta_1)) + \gamma^2 \tilde{c}(\Theta_2, \pi(\Theta_2)) \cdots], \tag{12}$$

the details of which are presented in Sections 3.4 and 3.5. Before that we make following remarks to provide additional insights to a reader about the proposed approach.

1) The approximations that yield Eqs. (9), (11a) and (11b), only facilitate learning of an optimal sensor selection policy (inner loop in Fig. 4) by reformulating a POMDP as an MDP. The state estimation (outer loop in Fig. 4) is still carried out without any approximations using the filtering approach described in Tewari et al. (2018).

2) For systems that exhibit nonlinear dynamics, an additional approximation would be needed, in the form of linearization of the dynamical model at the current state estimate, to accomplish POMDP to MDP reformulation. This approximation is essential to yield closed-form transition functions for the MDP, without which it would be computationally very hard to learn an optimal policy function, π^*. Our application does not require this approximation as the production dynamics is linear to begin with.

3) Section 3.2 alluded to the possibility of using *differential entropy* $\left(\mathbb{E}[-\log(\rho_t(\boldsymbol{x}_t))]\right)$ in place of variance $\left(\mathbb{E}[(\boldsymbol{x}_t - \mathbb{E}[\boldsymbol{x}_t])^T(\boldsymbol{x}_t - \mathbb{E}[\boldsymbol{x}_t])]\right)$ in Eq. (7) as another information-theoretic metric. This substitution would yield $\log(\det(\Sigma_t))$ as the third term in Eq. (9) instead of $\text{tr}(\Sigma_t)$. As a result, an optimal policy would aim to take sensing actions geared towards reducing the determinant of the covariance matrices of the approximated posteriors $\hat{\rho}_t$. Such reduction can be effectively achieved by taking test-separator measurements (without shut-ins) because of the well comingling effect. However, the individual variance in the flow rates can still be quite high, which would not be desirable.

3.4. Bellman optimality criterion

The quality of any policy π can be quantified through a *state value* function, $\mathcal{V}^\pi : \mathcal{B} \to \mathbb{R}$, which measures the total discounted cost when an MDP starts from an arbitrary state Θ i.e.

$$\mathcal{V}^\pi(\Theta) = \sum_{t=0}^{\infty} \gamma^t \tilde{c}_t \mid \Theta_0 = \Theta, \tag{13}$$

where we write $\tilde{c}(\Theta_t, \pi(\Theta_t))$ as $\tilde{c}_t$ for brevity. In an MDP with stochastic dynamics, the value function is obtained as the expected value of the total discounted cost over all possible trajectories of the process. In our formulation, since the Markov process is deterministic (Eqs. (11a), (11b)), it dispenses with the computation of expectation. Likewise, a *state-action value* function of a policy π, $\mathcal{Q}^\pi : \mathcal{B} \times \mathcal{A} \to \mathbb{R}$, measures the total discounted cost (Eq. (14)) starting from an arbitrary state Θ, taking an arbitrary action a, and then following the policy i.e.

$$\mathcal{Q}^\pi(\Theta, a) = \sum_{t=0}^{\infty} \gamma^t \tilde{c}_t \mid \Theta_0 = \Theta, a_0 = a. \tag{14}$$

Methods that aim to find an optimal policy may either work with the *state value* function or the *state-action value* function. We refer to the latter as $\mathcal{Q}$-function, and choose to work with it for reasons noted in the ensuing discussion.

A common element that underpins all the methods that aim to find an optimal policy is the *Bellman* equation as shown below, which relates the value of $\mathcal{Q}$-function at time t, to the value of the same at the next time step, under the specified dynamics and the policy π i.e.

$$\mathcal{Q}^\pi(\Theta_t, a_t) = \tilde{c}(\Theta_t, a_t) + \gamma \mathcal{Q}^\pi(\Theta_{t+1}, \pi(\Theta_{t+1})). \tag{15}$$

Eq. (15) can be easily derived from Eq. (14) by separating the first term from the infinite series, and then representing the remaining infinite series with the new value of $\mathcal{Q}$-function, while the process evolves as per the specified policy and dynamics. An optimal policy π^* chooses the action with the minimum value of the corresponding Q-function at any time step i.e.

$$\pi^*(\Theta_t) = \underset{a \in \mathcal{A}}{\text{argmin}} \{\mathcal{Q}^{\pi^*}(\Theta_t, a)\}, \quad \text{or equivalently,} \tag{16a}$$

$$\mathcal{Q}^{\pi^*}(\Theta_t, \pi^*(\Theta_t)) = \min a \in \mathcal{A}\{\mathcal{Q}^{\pi^*}(\Theta_t, a)\}. \tag{16b}$$

Eq. (16b), when plugged in the Bellman Eq. (15) for an optimal policy π^*, yields the *Bellman optimality criterion* as

$$\mathcal{Q}^{\pi^*}(\Theta_t, a_t) = \tilde{c}(\Theta_t, a_t) + \gamma \;\; \min a \in \mathcal{A}\{\mathcal{Q}^{\pi^*}(\Theta_{t+1}, a)\}. \tag{17}$$

Note that the optimal policy π^* does not appear explicitly in Eq. (17). An important implication of this observation is that the $\mathcal{Q}$-function of π^* can be obtained even when the optimal policy is itself unknown. Refer to Lagoudakis (2010) for more details on this subject.

3.5. Optimal $\mathcal{Q}$-learning using neural networks

A large class of methods attempt to approximate the optimal $\mathcal{Q}$-function $\mathcal{Q}^{\pi^*}(\Theta, a)$ with a parametrized function $\hat{\mathcal{Q}}^{\pi^*}(\Theta, a;\ \boldsymbol{\xi})$. The approximation entails tuning of the parameters $\boldsymbol{\xi}$ in order to closely mimic the true function. Irrespective of the choice of the approximation architectures (Perceptron, polynomials, splines, radial basis functions, Gaussian process, etc.), the parameters $\boldsymbol{\xi}$ are iteratively updated by minimizing a sequence of loss functions that change at each iteration i i.e.

$$\mathcal{L}_i(\boldsymbol{\xi}_i) = ||\boldsymbol{y}_i - \hat{\mathcal{Q}}^{\pi^*}(\Theta_i, a_i;\ \boldsymbol{\xi}_i)||_2^2. \tag{18}$$

The target $\boldsymbol{y}_i = \tilde{c}(\Theta_i, a_i) + \gamma \;\; \min a \in \mathcal{A}\{\hat{\mathcal{Q}}^{\pi^*}(\Theta_{i+1}, a;\ \boldsymbol{\xi}_{i-1})\}$ is set based on the parameters' value, $\boldsymbol{\xi}_{i-1}$, from the previous iteration. Thereafter, the parameters can be updated using the gradient information as follows, where $\alpha \in (0, 1]$ is the learning rate

$$\boldsymbol{\xi}_{i+1} \leftarrow \boldsymbol{\xi}_i + \alpha\left(\boldsymbol{y}_i - \hat{\mathcal{Q}}^{\pi^*}(\Theta_i, a_i;\ \boldsymbol{\xi}_i)\right)\nabla_{\boldsymbol{\xi}_i}\hat{\mathcal{Q}}^{\pi^*}(\Theta_i, a_i;\ \boldsymbol{\xi}_i). \tag{19}$$

Note that the iteration i also has a connotation of time, which mean that the parameters are updated in an *online* fashion as the MDP evolves in time. An important distinction here from the traditional *supervised*-learning paradigm is that the targets ($\boldsymbol{y}_i$) are no longer fixed and depend on the unknown parameters $\boldsymbol{\xi}_i$. As a result of this circular dependence, designing learning algorithms with stable convergence behavior becomes trickier.

Remark: Algorithms that are based on update rules similar to the one in Eq. (19) fall in the category of *model-free* algorithms, because parameters are updated merely based on the samples $\langle \Theta_i, \tilde{c}_i, a_i, \Theta_{i+1} \rangle$ generated by interacting with the MDP. At no point during their execution, the learning algorithms are aware of the dynamical model that is driving the system forward in time.

Our approach to learn an optimal $\mathcal{Q}$-function is based on a neural network approximation architecture ($\mathcal{Q}$-network) along the line of work by Mnih et al. (2013). The authors in that work demonstrated that an optimal $\mathcal{Q}$-function (for the game of *Atari*) can be approximated using a highly nonlinear architecture, while ensuring a stable convergence behavior. The $\mathcal{Q}$-network has a feed forward architecture with *states* as the input and *actions* as the output. The middle layers impart the desired flexibility between the input and the output. One crucial benefit of this architecture (an output unit for each action) vs. other architectures (Riedmiller, 2005; Lange and Riedmiller, 2010) was that $\mathcal{Q}$-function value for each action is obtained with just one forward pass through the network.

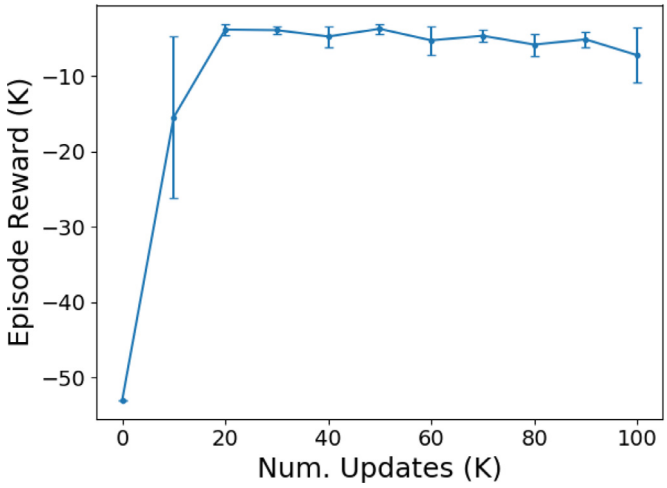


Fig. 7. Illustration of $\mathcal{Q}$-network training progression with the number of parameter updates. At each interval of 10k updates, $\mathcal{Q}$-networks are saved at 10 equidistant points. For instance, in the interval [20k-30k] the network is saved at 21k, 22k, 23k, ..., 30k updates. Each of the saved $\mathcal{Q}$-network is used to obtain a cumulative episode reward. The plot shows the mean cumulative rewards (of 10 such samples) and the corresponding errors bars ($\pm$ 1 standard deviation from the mean). In the later stages of training the cumulative reward converges to a value with small variance, suggesting a near-optimal $\mathcal{Q}$-network has been learned.

The other notable contribution by Mnih et al. (2013) is the concept of *experience-replay* that stabilizes and hastens the convergence of the $\mathcal{Q}$-network. The idea behind experience-replay is to archive the samples $\langle \Theta_t, \tilde{c}_t, a_t, \Theta_{t+1} \rangle$ from previous interactions with the MDP in a database $\mathcal{D}$. At every time-step, instead of updating the parameters using the most recent sample via Eq. (19), the parameters are updated using a few random samples from $\mathcal{D}$.

We adopt the same architecture ($\mathcal{Q}$-network) and learning algorithm (*experience-replay*) as in (Mnih et al., 2013). The input to the $\mathcal{Q}$-network is formed by vectorizing the state $\Theta = \{\boldsymbol{\mu}, C\}$, where C is the Cholesky factor of the covariance matrix Σ i.e. $\Sigma = CC^T$. Thus, for n wells ($2n$ flow rates) the input vector has $2n + n(2n + 1)$ elements. Since the states in our MDP are numerical, unlike the video frames of an Atari game, we replace the *convolutional* layers (a deep learning architecture suited for image inputs) with three fully-connected hidden layers with 100 nodes each. The output layer has the same number of nodes as the cardinality of the action space $\mathcal{A}$ ($|\mathcal{A}| = 39$ in our test case). The progression of the $\mathcal{Q}$-network training is shown in Fig. 7 where the x-axis shows the number of updates as in Eq. (19), and the y-axis represents the cumulative *episode* reward for the learned policy after every 10,000 updates (see figure caption for details). An episode is defined as a simulation run comprising 90 days of surveillance with a given $\mathcal{Q}$-network based policy. The reward is simply the negative surveillance cost. The wall-clock time for 100,000 updates is about 30 min for the $\mathcal{Q}$-network implemented in tensorflow (Abadi et al., 2015) on a 16-core CPU machine. As can be seen from Fig. 7, the training quickly converges to a robust policy. After ~ 20,000 updates the $\mathcal{Q}$-network stabilizes in terms of its ability to satisfy the Bellman optimality criterion in Eq. (17). An optimal $\mathcal{Q}$-network maximizes the cumulative reward (or minimizes the surveillance cost) of an episode.

4. Results

To assess the performance of different surveillance strategies, experiments were conducted in a simulated environment that provides us access to the necessary ground truth. Similar assessment in an actual oil field is not only be cost prohibitive but also disruptive in daily operations, especially for a proof of concept study. We use the surveillance testbed shown in Fig. 2 to run our experiments. The simulations are governed by the dynamic model described in Section 3 of the paper by Tewari et al. (2018). One experiment lasts for 90 days with the time step of one day. At every time step a sensing action is chosen (out of 39 possibilities), corresponding measurement is taken, and the prior distribution on flow rates is updated to posterior using the filtering framework proposed by Tewari et al. (2018). We compare 4 different surveillance methods listed in Table 2 (refer to the table caption for details).

Fig. 8 juxtaposes the surveillance output from the four methods, for one of the wells in groups 1. The simulation parameters (initial condition, dynamical model, sensor noise etc.) and the filtering approach were kept identical for all the methods. The only difference existed in sensor utilization plan, which is described in Table 2. Purely in terms of surveillance performance, the incumbent and the myopic sensing policies fared the worst, but for different reasons. The sensor data collected by the incumbent policy did not have information to resolves the flow rates per well (see the discussion in Section 2). The Myopic policy, on the other hand, turned out be too frugal in terms of sensor utilization (refer to Table 3). Quite often, a separator measurement does not have an immediate reward (in terms of information gain) unless some other measurements are taken in subsequent time steps. The myopic policy failed to see such long term impacts of taking a sensing action now, and hence, refrained from deploying sensors because of the immediate utilization costs. The smart-heuristic performed significantly better then the previous two policies, but was tangibly inferior to the Deep RL policy. Additionally, the Deep RL policy deployed the separators 25 fewer times than the smart-heuristic. This is important because a free sensing resource is of significant practical value, as it can either be used to monitor other well groups in the field, or can undergo a preventative maintenance procedure in its off-time. Another interesting observation relates to the Deep RL policy's usage of spin-cuts. Note that the spin-cuts are exclusively used for the group 2 wells. The policy learns that the separator measurements are ineffective to resolve the individual well flow rates for this group, because of high measurement noise (cf. Fig. 2). Therefore, the test-separator was summoned mostly for group 1 (25 times for group 1 vs. 5 times for group 2).

To obtain a comprehensive assessment of the four surveillance methods, we repeat the aforementioned experiment 100 times, with different initial conditions (flow rates at $t = 0$). For one experiment we have access to 180 estimates (90 each, for oil and water rates), and the corresponding ground truths. For each of these estimates, we compute the *bias* as the absolute deviation of the p50 value from the ground truth, and the *interdecile range* (IDR), the p90-p10 interval size. Thus, for 100 experiments, we end up with 18,000 pairs of bias and IDR values. Fig. 9 shows the box-plots of these values for the 4 surveillance methods. The performance of a surveillance method can be gauged by the resulting bias and IDR values, with smaller values being more desirable. The figure corroborates inferior performance of the incumbent and the myopic policies, for the reasons previously discussed. The difference is more notable in the box-plots of IDR values, indicating a large dispersion (uncertainty) in the rate estimates when these two surveillance methods are employed. Relatively, the performance of smart-heuristic and deep RL based methods are comparable, with the latter being marginally better.

A second set of experiments was conducted to distinguish smart-heuristic and deep RL sensor planning policies in terms of their ability to react to unexpected, real-world events in the field. It is a common practice to shutdown a well for an extended period of time (weeks) for maintenance. When such wells are brought back online, they exhibit immediate surge in production for reasons that include increased reservoir pressure and improved pump

Table 2
Description of different surveillance methods in terms of their sensor utilization plan. The *incumbent* method refers to the current practice of passive surveillance, in which sensors are deployed on a predetermined schedule. The *smart-heuristics* is a suggested improvement over the current practice, which aims to use a pattern of well shut-ins and spin-cuts in order to improve the information content in the measured data. The last two methods deploy sensors on-demand. However, the *Myopic* policy deploys a sensor with the lowest immediate cost. On the other hand, the proposed surveillance method (*Deep-RL policy*) deploys a sensor based on a longer term view of the surveillance cost.

Surveillance Method	Sensor Schedule	
	Test-separator	*Spin-cuts*
Incumbent	a well group is measured every 3rd day (no shut-ins)	a well is sampled once in 3 months randomly
Smart-heuristic	a well group is measured every 3rd day (a pattern is used for well shut-ins)	a well is sampled once in 3 months precisely
Myopic policy	on-demand	on-demand
Deep RL policy	on-demand	on-demand

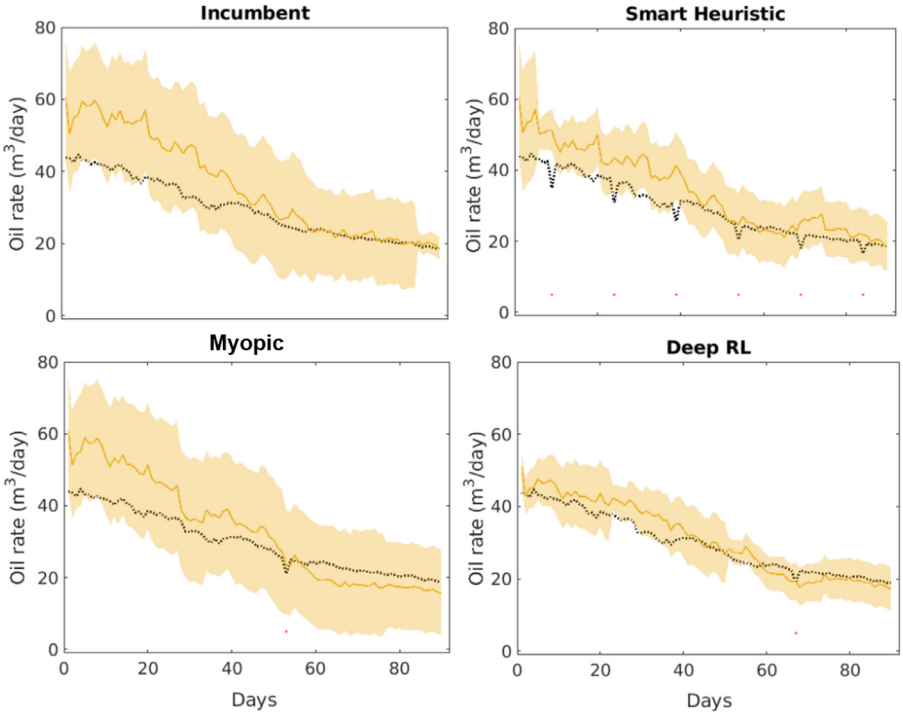


Fig. 8. Comparison of the four surveillance methods on the testbed shown in Fig. 2. Oil rates from a single well in group 1 are shown for this comparison. The dotted black line shows the ground truth, the solid line shows the p50 value of the estimated rate, and the band shows the interdecile range (p90-p10 interval). The momentary drops in true oil rates are due to planned well shut-ins during separator measurements. All the methods start with the same prior distributions on production rates on day 0, and use the same filtering approach to assimilate new sensor measurements.

Table 3
Number of times a sensor (separator or spin-cut) was deployed by a surveillance method during the course of 90 days. The numbers are broken down by the two well groups in the experimental testbed.

Surveillance Method	# Sep. meas. (group 1)	# Sep. meas. (group 2)	# SC meas. (group 1)	# SC meas. (group 2)
Incumbent	28	28	4	4
Smart-heuristics	27	27	4	4
Myopic policy	6	2	0	0
Deep RL policy	25	5	0	9

efficiency, among others. Fig. 10 shows the results on the same surveillance testbed (Fig. 2), but this time with well#2 from group 2 shut-in for a duration of two weeks. The ground truth rates (dotted back line) show a production jump at the end of the shut-in period (day 40). The deep RL policy clearly outperforms the smart-heuristic policy in terms of accurately tracking the production rates post shut-in. During the shut-in, the uncertainty in well#2's production grows with time, in the absence of measurements, in accordance with the dynamics specified by Eq. (11a). At the end of the shut-in, the Deep RL policy correctly perceives the increased cost (due to increased uncertainty) and summons a spin-cut for well#2. On the other hand, the smart-heuristic policy, being based on a fixed schedule, does not have the ability to react in such situations.

Fig. 11 shows the box-plots of the bias and IDR values by repeating the extended shut-in experiments several times, as done previously in Fig. 9. We omit the box-plots for the *incumbent* and *myopic* policies as they have been deemed inferior in the previous experiment.

Remark: The proposed sensor planning method is easily scalable to existing large hydrocarbon fields such as the one shown in Fig. 1. These fields are typically partitioned into zones (called well pads) with 15–20 wells, and equipped with their own dedicated sensing resources. Thus, sensor planning in such fields decomposes into as many independent problems as the number of

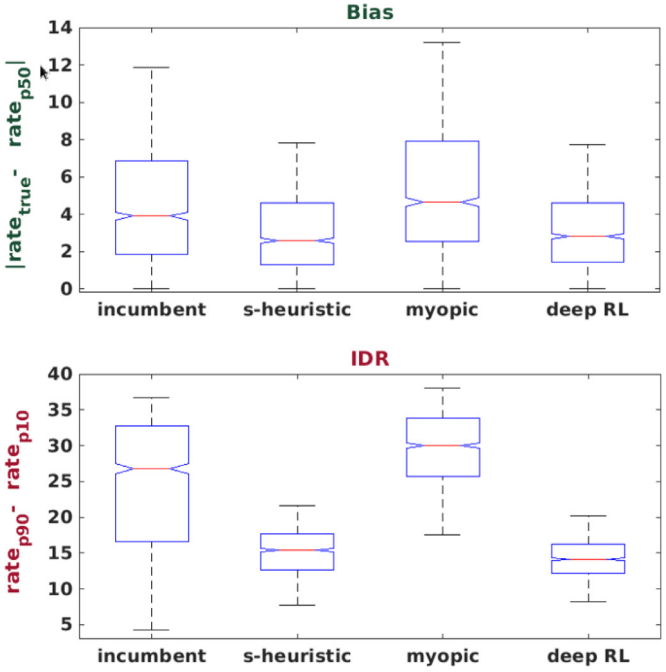


Fig. 9. Comparison of the performance of the 4 surveillance methods via the box-plots of the corresponding *bias* and *interdecile range* (IDR) values. The former is computed as the absolute deviation of the p50 value of the flow estimate from the ground truth. The latter, a standard statistical dispersion measure, is the size of p90-p10 interval. These values are obtained at each time step of an experiment, over 100 independent experiments with different initial conditions. A good surveillance performance is characterized by low bias and IDR values.

well pads. As a result, separate optimal policies, π^*, can be independently learned (following the proposed method) and deployed for every pad with ease.

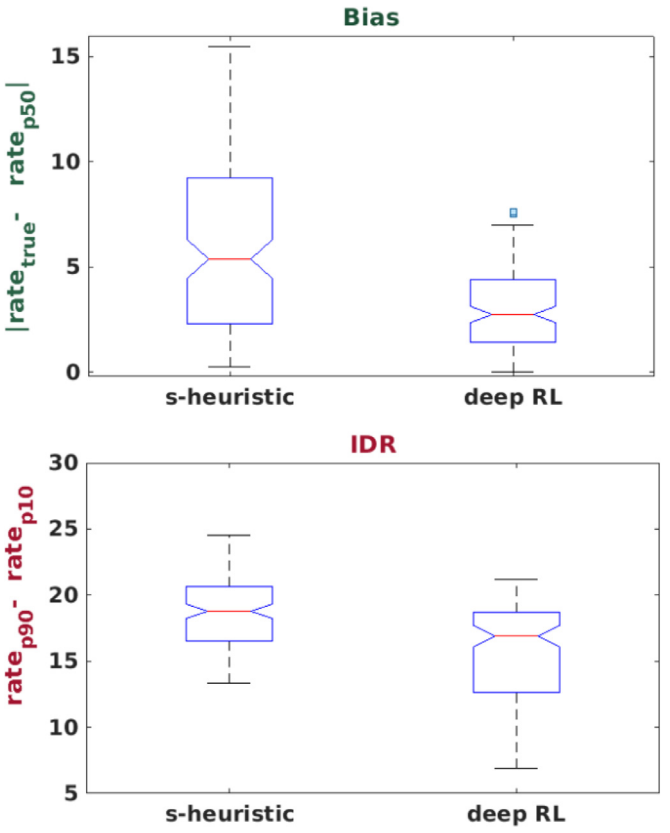
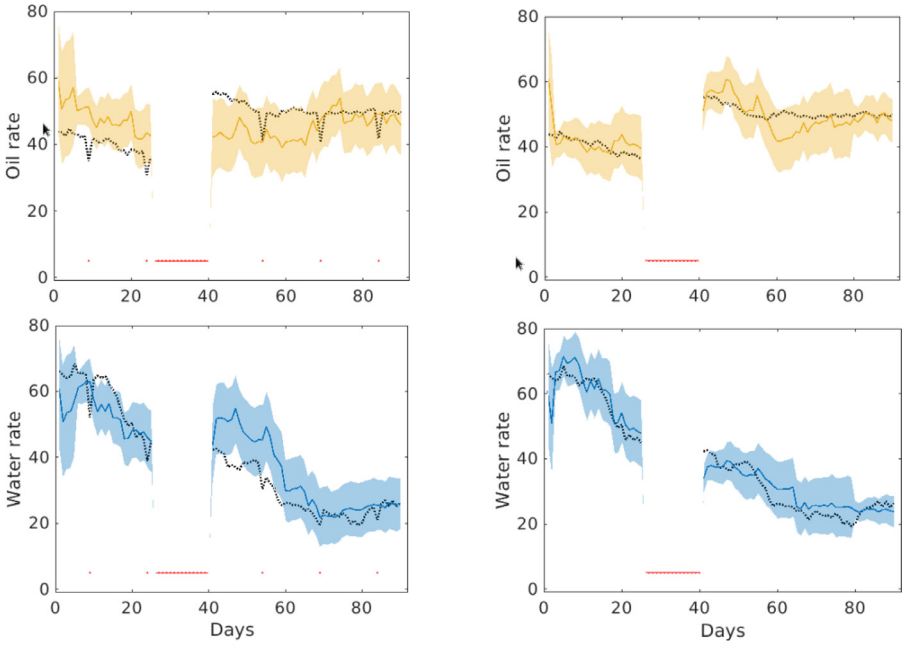


Fig. 11. Comparison of the performance of *smart-heuristic* and *deep RL* based surveillance methods via the box-plots of the corresponding *bias* and *IDR* values. These values are obtained at each time step of an experiment, over 100 independent experiments with different initial and shut-in conditions. The latter is specified by the well, the location and the total duration (1 to 3 week) of chosen shut-in event.



(a) Smart-Heuristic (b) Deep RL

Fig. 10. Surveillance results with *smart-heuristic* (Fig. 10a) and *deep RL* (Fig. 10b) sensor scheduling policies for an extended shut-in scenario. One of the wells from group 2 (of the experimental testbed in Fig. 2) is shut down for a period of 2 weeks. When the well becomes operational again, the deep RL policy immediately schedules a spin-cut measurement that leads to better tracking of the post shut-in production. Smart-heuristic does not possess this reactive ability unless the response is hard coded (a difficult and laborious task).

Fig. 12. An illustration of the concept of *autonomous* surveillance in a large onshore field. The map shows a zoomed out region of the field in Fig. 1. The solid circular markers represent ground robots working collaboratively to surveil the field, and the dotted arrows show their optimized path. A planning method, such as the one proposed in this paper, is needed to route the robots so that a desired surveillance performance can be achieved.

5. Conclusion

In this work we outline an information-theoretic approach for production surveillance in oil fields, a task which is often hampered by inadequate sensing resources and continuously evolving subsurface conditions. The approach is grounded on the principle of information-based planning that actively deploys sensors when and where the need is the most, while being cognizant of sensing cost. A proof of concept, in a simulated environment, clearly demonstrates advantages over the current practice of passive surveillance. Additionally, the proposed sensor planning approach is conceptually portable to other application areas of relevance to the oil & gas industry such as seismic surveys, methane leak detection, loss-prevention systems, and seabed seep detection.

The future of production surveillance is likely to be driven by low oil prices that continue to push for technologies that are not only cost-effective but also more agile and accurate. Conventional surveillance methods that rely heavily on test-separators as the primary multiphase flow metering device are inherently sluggish for real-time monitoring because of the lengthy process of gravimetric phase separation. More importantly, test-separators can prove to be prohibitive for large onshore fields because of the high associated capital and operational costs. The advancements in non-intrusive multiphase flow metering technology, in the last decade (Hansen et al., 2019), along with increasing applications of robotics in O&G industry (Frost&Sullivan, 2017; Munoz, 2018; McBride, 2017) can pave the way for *autonomous* surveillance technology with minimal continuous human intervention. It is easy to envision a large onshore field, such as the one shown in Fig. 12, with its surveillance needs met by a fleet of ground robots that are equipped with multiphase flow meters, and adept at wellhead operations. Sensor planning in this setting amounts to continuous path planning of the ground robots to keep the surveillance cost small. The mathematical framework presented in the paper is extendable to this setting, and is akin to the widely studied area of robotic motion planning in other domains such as aerospace and defense.

Declaration of Competing Interest

The authors declare that they do not have any financial or non-financial conflict of interests

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.compchemeng.2020.106988.

CRediT authorship contribution statement

Ashutosh Tewari: Conceptualization, Investigation, Methodology, Software, Writing - original draft. **Kuang-Hung Liu:** Software, Writing - review & editing. **Dimitri Papageorgiou:** Investigation, Writing - original draft, Writing - review & editing.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., David, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Morre, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Bikmukhametov, T., Jäschke, J., 2019. First principles and machine learning virtual flow metering: a literature review. J. Pet. Sci. Eng. 106487.

Caruso, M.J., Migliaccio, M., Hargrove, J.T., Garcia-Pineda, O., Graber, H.C., 2013. Oil spills and slicks imaged by synthetic aperture radar. Oceanography 26 (2), 112–123.

Frost&Sullivan, 2017. Application of drones and robots in oil and gas industry – a cost effective and safe method of inspection and surveillance. WebLink.

Hansen, L.S., Pedersen, S., Durdevic, P., 2019. Multi-phase flow metering in offshore oil and gas transportation pipelines: trends and perspectives. Sensors (9).

Hero, A.O., Kreucher, C.M., Blatt, D., 2008. Information theoretic approaches to sensor management. In: Foundations and Applications of Sensor Management. Springer, pp. 33–57.

Hoffman, M.D., Gelman, A., 2014. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo.. J. Mach. Learn. Res. 15 (1), 1593–1623.

Jenkins, K.L., Castanón, D.A., 2011. Information-based adaptive sensor management for sensor networks. In: Proceedings of the 2011 American Control Conference. IEEE, pp. 4934–4940.

Kaelbling, L.P., Littman, M.L., Cassandra, A.R., 1998. Planning and acting in partially observable stochastic domains. Artif. Intell. 101 (1–2), 99–134.

Kreucher, C., Kastella, K., Hero Iii, A.O., 2005. Sensor management using an active sensing approach. Signal Process. 85 (3), 607–624.

Krishnamurthy, V., 2016. Partially Observed Markov Decision Processes. Cambridge University Press.

La, H.M., Sheng, W., Chen, J., 2014. Cooperative and active sensing in mobile sensor networks for scalar field mapping. IEEE Trans. Syst. Man Cybern. 45 (1), 1–12.

Lagoudakis, M.G., 2010. Value Function Approximation. Springer US, Boston, MA, pp. 1011–1021.

Lange, S., Riedmiller, M., 2010. Deep auto-encoder neural networks in reinforcement learning. In: The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–8.

Mallick, M., Krishnamurthy, V., Vo, B.-N., 2012. Integrated Tracking, Classification, and Sensor Management. Wiley Online Library.

McBride, C. D., 2017. Robots and the future of oil and gas workforce. WebLink.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing Atari with deep reinforcement learning. arXiv:1312.5602WebLink.

Munoz, J.-M., 2018. The argos project: autonomous robots for gas and oil sites. https://www.ep.total.com/en/innovations/research-development/ground-robots-key-future-architectures/.

Papageorgiou, D., Raykin, M., 2007. A risk-based approach to sensor resource management. In: Advances in Cooperative Control and Optimization. Springer, pp. 129–144.

Platt, R., Tedrake, R., Kaelbling, L., Lozano-Perez, T., 2010. Belief space planning assuming maximum likelihood observations. In: Robotics Science and Systems Conference (RSS).

Poulisse, H., Van Overschee, P., Briers, J., Moncur, C.E., Goh, K.-C., 2006. Continuous well production flow monitoring and surveillance. In: Intelligent Energy Conference and Exhibition. Society of Petroleum Engineers.

Riedmiller, M., 2005. Neural fitted q iteration – first experiences with a data efficient neural reinforcement learning method. In: Machine Learning: ECML 2005. Springer Berlin Heidelberg, pp. 317–328.

Ryan, A., Hedrick, J.K., 2010. Particle filter based information-theoretic active sensing. Rob. Auton. Syst. 58 (5), 574–584.

Tewari, A., de Waele, S., Subrahmanya, N., 2018. Enhanced production surveillance using probabilistic dynamic models. Int. J. Progn. Health Manage. 9 (19), 1737–1760.
Wang, H., Yao, K., Estrin, D., 2005. Information-theoretic approaches for sensor selection and placement in sensor networks for target localization and tracking. J. Commun. Netw. 7 (4), 438–449.
Washburn, R.B., Schneider, M., Fox, J., 2002. Stochastic dynamic programming based approaches to sensor resource management. In: Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002.(IEEE Cat. No. 02EX5997), vol. 1. IEEE, pp. 608–615.
Xiong, N., Svensson, P., 2002. Multi-sensor management for information fusion: issues and approaches. Inf. Fusion 3 (2), 163–186.
Yang, P., Freeman, R.A., Lynch, K.M., 2007. Distributed cooperative active sensing using consensus filters. In: Proceedings 2007 IEEE International Conference on Robotics and Automation. IEEE, pp. 405–410.