Gaussian Process Q-Learning for Finite-Horizon Markov Decision Processes

Maximilian Bloor, Tom Savage, Calvin Tsay, Ehecatl Antonio Del Rio Chanona, Max Mowbray

Keywords: Finite Horizon Markov Decision Processes, Gaussian Process, Q-learning

Summary

Many real-world control and optimization problems require making decisions over a finite time horizon to maximize performance. This paper proposes a reinforcement learning framework that approximately solves the finite-horizon Markov Decision Process (MDP) by combining Gaussian Processes (GPs) with Q-learning. The method addresses two key challenges: the tractability of exact dynamic programming in continuous state-control spaces, and the need for sample-efficient state-action value function approximation in systems where data collection is expensive. Using GPs and backward induction, we construct state-action value function approximations that enable efficient policy learning with limited data. To handle the computational burden of GPs as data accumulate across iterations, we propose a subset selection mechanism that uses M-determinantal point processes to draw diverse, high-performing subsets. Theoretical analysis shows function class preserving property of the GPQL algorithm for convex MDPs with deterministic dynamics. The proposed method is evaluated on a linear quadratic regulator problem and online optimization of a non-isothermal semi-batch reactor. Improved learning efficiency is shown relative to the use of Deep Q-network.

Contribution(s)

1. The paper presents a framework for learning Gaussian process state-action value function approximations using Q-learning for deterministic finite horizon Markov Decision Processes.
   Context: Prior work has explored the use of Gaussian process models within the infinite horizon context (Engel et al., 2005; Grande et al., 2014; Chowdhary et al., 2014).
2. A subset selection strategy is proposed to ensure the online computational tractability of control inference using M-determinantal point processes to build a GP approximation of the state-action value function, that balances global coverage with local accurate modeling in highly performing regions of the state-control space (Kulesza et al., 2012; Moss et al., 2023)
   Context: Previous works take the approach of building variational approximations globally to the exact GP state-action value function approximation (Grande et al., 2014).

Gaussian Process Q-Learning for Finite-Horizon Markov Decision Processes

Maximilian Bloor[1], Tom Savage[1], Calvin Tsay[2], Ehecatl Antonio Del Rio Chanona[1], Max Mowbray[1]
{max.bloor22, m.mowbray}@imperial.ac.uk

[1]**Department of Chemical Engineering, Imperial College London, United Kingdom**
[2]**Department of Computing, Imperial College London, United Kingdom**

Abstract

Many real-world control and optimization problems require making decisions over a finite time horizon to maximize performance. This paper proposes a reinforcement learning framework that approximately solves the finite-horizon Markov Decision Process (MDP) by combining Gaussian Processes (GPs) with Q-learning. The method addresses two key challenges: the tractability of exact dynamic programming in continuous state-control spaces, and the need for sample-efficient state-action value function approximation in systems where data collection is expensive. Using GPs and backward induction, we construct state-action value function approximations that enable efficient policy learning with limited data. To handle the computational burden of GPs as data accumulate across iterations, we propose a subset selection mechanism that uses M-determinantal point processes to draw diverse, high-performing subsets. Theoretical analysis shows function class preserving property of the Gaussian Process Q-learning algorithm for convex MDPs with deterministic dynamics. The proposed method is evaluated on a linear quadratic regulator problem and online optimization of a non-isothermal semi-batch reactor. Improved learning efficiency is shown relative to the use of Deep Q-network.

1 Introduction

Sequential decision-making problems with finite time horizons appear across numerous domains. These problems are often naturally modeled as finite-horizon Markov Decision Processes (MDPs), where decisions must optimize the total cost incurred over a predetermined time window. Fed-batch process control and online optimization is a prominent example within the process systems engineering (PSE) community (Mesbah, 2016; Bradford et al., 2020), where optimal policy identification is inherently challenging due to difficulties in system model identification and uncertainty propagation over the decision horizon. While optimal solutions to these problems can theoretically be obtained through dynamic programming (DP), three significant challenges emerge in practice: (i) the intractability of exact DP in continuous state-action spaces, (ii) the lack of an exact descriptive process model, and (iii) the need for sample efficiency when data collection is expensive.

This paper introduces a reinforcement learning framework that addresses these challenges by combining the statistical modeling power of Gaussian Processes (GPs) with Q-learning. Previous works have predominantly explored the use of GPs to approximate state-space models (Kuss & Rasmussen, 2003; Deisenroth et al., 2015; Bradford et al., 2020; Mowbray et al., 2022), with relatively few taking a purely model-free approach. For example, Savage et al. (2021) exploited GP models for state-action value learning. However, the paper provides no mechanism for updating the underlying dataset, for example through a Monte Carlo or Q-learning estimator and therefore is not guaranteed

to identify the optimal policy. The GP temporal difference learning (GPTD) algorithm detailed in Engel et al. (2005) is tailored for prediction problems by explicitly modeling rewards and exploiting the structure of the Bellman equation, to identify the value function of a stationary policy that has generated the modeled data. The algorithm can be extended for on-policy control through the use of SARSA. Nevertheless, the authors do not exploit the uncertainty of the state-action value function posterior process to balance exploration and exploitation; and, as we argue subsequently, off-policy updates are preferable. Chowdhary et al. (2014) have explored the development of GP-based Q-learning algorithms; however convergence is dependent upon non-trivial selection of a regularization parameter. Grande et al. (2014) provide a GPQ learning algorithm with an update rule that assumes the observed rollout estimate and the approximation belong to a joint Gaussian distribution. This assumption likely only holds for a subset of MDPs.

Unlike traditional state-action value function approximation methods that rely on neural networks (Azizzadenesheli & Anandkumar, 2019) or linear models, GPs offer significant advantages: they provide epistemic and aleatoric uncertainty quantification that can be exploited for decision making through the use of bandit strategies, can incorporate prior domain knowledge through kernel selection, and inherently balance model complexity as data are collected.

The contributions of this work are as follows. We pose a framework for learning optimal control using Q-learning and GPs in finite-horizon MDPs with deterministic dynamics and justify the composition of the two. The computational challenges associated with the use of GPs are handled through implementation of a principled subset selection mechanism using M-determinantal point processes (M-DPPs) (Kulesza & Taskar, 2012; Moss et al., 2023). This mechanism enables efficient scaling of control inference with increasing dataset size, despite the cubic complexity of exact GP inference. We provide theoretical analysis showing the function class preserving property of the GPQL algorithm for convex MDPs with deterministic dynamics. Empirical validation on both a linear quadratic regulator (LQR) system and a non-isothermal semi-batch chemical reactor optimization problem demonstrate the algorithm's sample efficiency. This is a significant advantage in real-world chemical processes where experiments involve substantial costs and time investment.

In summary, we develop a GP-based Q-learning framework for finite-horizon MDPs, implement a practical approach using M-DPPs for strategic subset selection, and establish the convexity preserving property of the GPQL updates for finite-horizon convex MDPs. The rest of this paper is structured as follows: Section 2 presents preliminaries, Section 3 the GPQL algorithm, Section 4 the case studies, conclusions in Section 5 and analysis in Appendix A.1

2 Preliminaries

2.1 Finite-horizon Markov decision processes

Consider a finite-horizon Markov Decision Process (MDP) which is defined by the 5-tuple $\langle \mathcal{X}, \mathcal{U}, \mathcal{P}, \Phi, \mathcal{T} \rangle$. Specifically, the states are described by a compact state set, $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{n_x}$, with the controls restricted to a compact set, $\mathbf{u} \in \mathcal{U} \subset \mathbb{R}^{n_u}$. The underlying decision process adheres to discrete-time state transitions described by a conditional probability density function, $\mathcal{P} : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \to [0, \infty)$ over a time horizon, $\mathcal{T} = \{0, \ldots, T\}$. For convenience, in the following we denote the state-control pair, $\mathbf{z} = [\mathbf{x}, \mathbf{u}]^\intercal$, and corresponding set, $\mathcal{Z} = \mathcal{X} \times \mathcal{U}$. We restrict the presentation to consider deterministic state-transitions, such that one may define

$$\mathcal{P}(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{u}_t) = \prod_{i=1}^{n_x} \delta(x_{i,t+1} - f_i(\mathbf{x}_t, \mathbf{u}_t)) \,,$$

where $\delta(\cdot)$ indicates the dirac-delta function. $x_{i,t} \in \mathbb{R}$ the ith state component, and $\mathbf{f}(\mathbf{x}, \mathbf{u}) = [f_1(\mathbf{x}, \mathbf{u}), \ldots, f_{n_x}(\mathbf{x}, \mathbf{u})]^\intercal$ a discrete-time dynamic model of the system,

$$\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_t, \mathbf{u}_t) \,. \tag{1}$$

The objective of decision making is to select controls, according to a non-stationary state-feedback policy, $\pi(\cdot) = (\pi_0(\cdot), \ldots, \pi_{T-1}(\cdot))$, with $U_t \sim \pi_t(\mathbf{u}_t \mid \mathbf{x}_t)$, to minimize the expected sum of stage-costs allocated by a cost function, $\Phi(\mathbf{x}, \mathbf{u})$, incurred over a finite discrete-time horizon,

$$\begin{aligned} &\min_{\pi} \mathbb{E}_{\pi}\left[Q_t^{\pi}(X_0, U_0) \mid X_0 = \mathbf{x}_0\right] \\ &Q_t^{\pi}(\mathbf{x}, \mathbf{u}) := \Phi(\mathbf{x}, \mathbf{u}) + \mathbb{E}_{\pi}\left[\sum_{k=t+1}^{T-1} \Phi(X_k, U_k) \middle| U_k \sim \pi_k(\mathbf{u}_k \mid \mathbf{x}_k)\right] . \end{aligned} \tag{2}$$

Note that although the initial state, $\mathbf{x}_0$, is assumed to be known with certainty, in general the state itself is uncertain due to the definition of the policy as a conditional probability density function.

According to the principle of optimality, however, the optimal policy is a deterministic function of state and acts to greedily minimize the state-action value function at a given time within the horizon,

$$\pi^*(\mathbf{x}_t) \in \arg\min_{\pi} \; Q_t^{\pi}(\mathbf{x}_t, \pi_t(\mathbf{x}_t)) . \tag{3}$$

In principle, the optimal policy can be identified through dynamic programming (DP). However, DP becomes intractable in large or continuous state-control spaces and is reliant on a known model of the system, which is unavailable in general. This directs attention to the use of model-free approximate methods, such as Q-learning, which, in the tabular sense, aim to approximate the state-action values associated with state-control pairs in a given dataset,

$$\mathcal{D} = \{\mathcal{D}_t\}_{t=0:T-1} , \quad \mathcal{D}_t = \left\{\left(\mathbf{x}_t^{(\mathrm{i})}, \mathbf{u}_t^{(\mathrm{i})}, \mathbf{x}_{t+1}^{(\mathrm{i})}, \Phi(\mathbf{x}_t^{(\mathrm{i})}, \mathbf{u}_t^{(\mathrm{i})})\right)\right\}_{i=1:N} .$$

As the cardinality of dataset, N, increases, the approach provided by maintaining fixed point estimates of the state-action value function faces an explosion in memory cost. This challenge is typically handled through the construction of function approximators, e.g., deep Q-network Mnih et al. (2013).

2.2 Gaussian process

GPs are probabilistic models that describe the relationship between a finite number of evaluations of a multivariate, scalar-valued function via a joint multivariate Gaussian distribution. Consider the input locations, $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, corresponding noiseless function evaluations, $Y = \{f(\mathbf{z}_1), \ldots, f(\mathbf{z}_N)\}$, test location, $\mathbf{z}_*$, and function value, y_*. The zero-mean GP prior asserts,

$$\begin{bmatrix} Y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} K(Z, Z) & \mathbf{k}(Z, \mathbf{z}_*) \\ \mathbf{k}(Z, \mathbf{z}_*)^{\intercal} & k(\mathbf{z}_*, \mathbf{z}_*) \end{bmatrix}\right) \tag{4}$$

with $k(\mathbf{z}, \mathbf{z}') \in \mathbb{R}$ denoting the kernel function, $\mathbf{k}(Z, \mathbf{z}_*) \in \mathbb{R}^N$ the covariance of the function values between the input locations and test point, $k(\mathbf{z}_*, \mathbf{z}_*) \in \mathbb{R}$ the variance of the function value at the test point, and $K(Z, Z)$ denoting the gram matrix representative of the covariance of the function values evaluated at the input locations, Z. Bayesian inference gives the predictive posterior distribution:

$$\begin{aligned} \mu(\mathbf{z}_*) &= \mathbf{k}(Z, \mathbf{z}_*)^{\intercal} K(Z, Z)^{-1} Y \\ \mathtt{k}(\mathbf{z}_*, \mathbf{z}_*) &= k(\mathbf{z}_*, \mathbf{z}_*) - \mathbf{k}(Z, \mathbf{z}_*)^{\intercal} K(Z, Z)^{-1} \mathbf{k}(Z, \mathbf{z}_*) , \end{aligned} \tag{5}$$

with $y_* \sim \mathcal{N}\left(\mu(\mathbf{z}_*), \mathtt{k}(\mathbf{z}_*, \mathbf{z}_*)\right)$ providing a description of the posterior at $\mathbf{z}^*$. One may also be interested in a general description of the posterior process,

$$f(\cdot) \mid Y \sim \mathcal{GP}\left(\mu(\cdot), \mathtt{k}(\cdot, \cdot)\right) , \tag{6}$$

which may be used to approximate the state-action value function using the available data. A policy may then be defined to exploit the posterior process for decision-making, generalizing to continuous state-control spaces. Such a policy may be defined by greedily exploiting functions drawn from the posterior, which involves solving minimization problems. This motivates discussion regarding properties of the optimal solution map and value function.

2.3 Optimal value functions and solution maps of nonlinear programs

Consider the optimal (possibly point-to-set) solution map, $\mathbf{u}^* \in G^*(\mathbf{x})$ and optimal value function, $g^*(\mathbf{x})$, of the following parametric nonlinear program with constant constraint map (Fiacco & Ishizuka, 1990),

$$\begin{aligned} G^*(\mathbf{x}) &\coloneqq \arg\min_{\mathbf{u}\in\mathcal{U}} \; g(\mathbf{x},\mathbf{u}) \\ g^*(\mathbf{x}) &\coloneqq \min_{\mathbf{u}\in\mathcal{U}} \; g(\mathbf{x},\mathbf{u}) \,, \end{aligned} \tag{7}$$

with $g(\mathbf{x},\mathbf{u})$ indicating a multivariate, scalar valued objective function.

Assumption 1 *The nonlinear objective function is continuous in both* $\mathbf{x}$ *and* $\mathbf{u}$*, with* $(\mathbf{x},\mathbf{u}) \in \mathcal{Z}$*.*

Property 1 *Let Assumption 1 hold. The optimal value function,* $g^*(\mathbf{x})$*, is continuous (Fiacco & Ishizuka, 1990; Aubin & Frankowska, 1999), and the optimal solution map,* $G^*(\mathbf{x})$*, is upper semi-continuous on* $\mathbf{x} \in \mathcal{X}$ *(Sundaram, 1996).*

Strict convexity assumptions are typically required for the solution map to be single-valued and continuous (Fiacco & Ishizuka, 1990; Aubin & Frankowska, 1999).

3 Methodology

In the following section, we present a methodology that leverages GPs to approximate the state-action value function. This directs the presentation of the following problem statement.

3.1 Problem statement

Consider the finite-horizon MDP introduced in Section 2.1. To ensure that the state-action value function can be well approximated by a continuous function, we impose the following regularity conditions on the system dynamics and policies.

Assumption 2 (Regularity Conditions) *The underlying MDP satisfies the following conditions:* (i) *the state transition function* $\boldsymbol{f}_t : \mathcal{Z} \to \mathcal{X}$ *is continuous for all* $t \in \{0,\dots,T-1\}$*;* (ii) *any policy* $\pi(\boldsymbol{u} \mid \boldsymbol{x})$ *is a conditional probability density function continuous with respect to* $\boldsymbol{x}$ *with support on the compact set* $\mathcal{X}$*; and* (iii) *the cost function* $\Phi : \mathcal{Z} \to \mathbb{R}$ *is continuous over the compact set* $(\boldsymbol{x},\boldsymbol{u}) \in \mathcal{Z}$*.*

Proposition 1 (Continuity and Optimality Properties) *Under Assumption 2, the state-action value function* $Q_t^\pi : \mathcal{Z} \to \mathbb{R}_+$ *induced under any policy* π *is continuous for all* $t \in \{0,\dots,T-1\}$*. Furthermore, the optimal state-action value function* $Q_t^{\pi^*}(\boldsymbol{x},\boldsymbol{u})$ *is continuous on* $(\boldsymbol{x},\boldsymbol{u}) \in \mathcal{Z}$*, and the value function of the optimization problem* $\min_{\boldsymbol{u}\in\mathcal{U}} Q_t^\pi(\boldsymbol{x},\boldsymbol{u})$ *satisfies standard continuity properties from parametric optimization theory.*

3.2 Gaussian process Q-Learning (GPQL)

Having formalized the problem setting, we now direct our attention to the contribution of this work. Namely, we introduce Gaussian process Q-Learning (GPQL), which is well suited to solving the problems described, and adheres to the general framework for policy iteration.

3.2.1 Gaussian process state-action value function approximation

We propose to approximate the state-action value function, $Q_t^\pi(\cdot)$ using GP models and consider the construction of independent GP models for each time-step. Specifically, we assume a dataset, $\mathcal{D}_t \; \forall t$, which provides some discretization of the state-control space, $\mathbf{z} \in \mathcal{Z}$, from which state-action values may be estimated via fixed-point estimates, $\mathbf{Q}_t^\pi = \{Q_t^\pi(\mathbf{z}) \; \forall \mathbf{z} \in \mathcal{D}_t\}$, as in the tabular

setting. A joint prior distribution defines the predictive model of the state-action value, $Q^{\pi}_{t,*}$, at a new state-control pair, $\mathbf{z}_* \in \mathcal{Z}$,

$$\begin{bmatrix} \mathbf{Q}^{\pi}_t \\ Q^{\pi}_{t,*} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} K_t(Z,Z) & \mathbf{k}_t(Z,\mathbf{z}_*) \\ \mathbf{k}_t(Z,\mathbf{z}_*)^{\intercal} & k_t(\mathbf{z}_*,\mathbf{z}_*) \end{bmatrix} \right) . \tag{8}$$

Through conditioning the posterior process may be obtained, $Q^{\pi}_t(\cdot) \mid \mathbf{Q}^{\pi}_t \sim \mathcal{GP}\left(\mu^{\pi}_t(\cdot), \mathrm{k}^{\pi}_t(\cdot,\cdot)\right)$, as defined in Section 2.2. In principle, given knowledge of the underlying problem structure, one can select a kernel appropriately, such that the state-action value function lies within the class of functions well represented by the GP. Considerations for selection are outlined later in Section 3.2.3.

It is worth highlighting that the complexity of GP model inference scales cubically with the number of datapoints, N. In practice, we use the M-determinantal point process (M-DPP) to select a subset of the state-action value point estimates to build an approximation. A subset of $M \leq N$ points, Z_M, are sampled from $\mathcal{D}_t$, according to their probability under the M-DPP,

$$\begin{aligned} &\mathbb{P}(Z_M \subseteq Z) \propto |L_{Z_M}| \\ &\qquad |L_{Z_M}| = |K(Z_M, Z_M)| \cdot \prod_{\mathbf{z} \in Z_M} q(\mathbf{z})^2 , \end{aligned} \tag{9}$$

which is proportional to the trace of the Gram matrix weighted by the quality of the points as evaluated by $q(\mathbf{z})$, $\forall \mathbf{z} \in Z_M$. This is a strategy that was reported in Kulesza et al. (2012), and has since been utilized within the context of Bayesian optimization with sparse GPs (Moss et al., 2023). Intuitively, it selects points to balance performance under the quality function and coverage of the state-control space as evaluated by the kernel. Here, we utilize the strategy to simply build an exact GP estimator providing a local model in promising regions of the state-control space. The greedy method utilized in Chen et al. (2018); Moss et al. (2023) is implemented for convenience.

3.2.2 Thompson sampling for policy improvement

Within the course of approximate policy improvement, the GP model is exploited to generate a policy which balances exploration and exploitation. Specifically, a Thompson sampling (TS) policy, $\pi(\mathbf{u} \mid \mathbf{x})$, is deployed (Wilson et al., 2020). The policy is defined through minimization of state-action value functions, $Q^{\pi}_t(\cdot)$, sampled probabilistically from the GP posterior,

$$\pi_t(\mathbf{u} \mid \mathbf{x}) = \Pr\left(\mathbf{u} \in \arg\min_{\bar{\mathbf{u}} \in \mathcal{U}} Q^{\pi}_t(\mathbf{x}, \bar{\mathbf{u}}) \mid \mathbf{Q}^{\pi}_t \;\middle|\; Q^{\pi}_t(\cdot) \mid \mathbf{Q}^{\pi}_t \sim \mathcal{GP}\left(\mu^{\pi}_t(\cdot), \mathrm{k}^{\pi}_t(\cdot,\cdot)\right) \right) . \tag{10}$$

This enables one to exploit the GP-based state-action value function approximation to both explore and exploit the decision-space—effectively leveraging the quantification of epistemic uncertainty to generate data under a behavior policy, which may then be used to learn the target optimal policy.

Remark 1 *In principle, the TS behavior policy is upper semi-continuous in the state, and does not satisfy Assumption 2. However, this does not impact the state-action value function approximation problem posed as discussed in the subsequent section.*

3.2.3 Q-Learning with backward induction

Having defined means to both approximate the state-action value function using point-estimates and exploit it for decision-making, we now discuss updates of the point estimates, $\mathbf{Q}^{\pi}_t$. In any given rollout of the TS policy, one generates data that may be used to update the existing dataset,

$$\mathcal{D}_{\mathsf{n},t} = \left\{ \left(\mathbf{x}^{(\mathsf{n})}_t, \mathbf{u}^{(\mathsf{n})}_t, \Phi(\mathbf{x}^{(\mathsf{n})}_t, \mathbf{u}^{(\mathsf{n})}_t), \mathbf{x}^{(\mathsf{n})}_{t+1} \right) \right\} , \quad \mathcal{D}_t \leftarrow \mathcal{D}_t \cup \mathcal{D}_{\mathsf{n},t} \quad \forall t .$$

This new dataset may be used to update the GP posterior state-action value function approximations working from the final decision-step, given $\mathbf{Q}^{\pi}_{T-1} = \{\Phi(\mathbf{x}_{T-1}, \mathbf{u}_{T-1}),\ \forall (\mathbf{x}_{T-1}, \mathbf{u}_{T-1}) \in \mathcal{D}_{T-1}\}$,

and propagating information backwards over the horizon via the following update,

$$Q_t^\pi(\mathbf{x}_t, \mathbf{u}_t) \leftarrow Q_t^\pi(\mathbf{x}_t, \mathbf{u}_t) + \alpha \left(\Phi(\mathbf{x}_t, \mathbf{u}_t) + \min_{\mathbf{u}_{t+1} \in \mathcal{U}} \mu_{t+1}^\pi(\mathbf{x}_{t+1}, \mathbf{u}_{t+1}) - Q_t^\pi(\mathbf{x}_t, \mathbf{u}_t) \right) \quad \forall\, (\mathbf{x}_t, \mathbf{u}_t, \Phi(\mathbf{x}_t, \mathbf{u}_t), \mathbf{x}_{t+1}) \in \mathcal{D}_t, \quad \forall t < T-1\,, \tag{11}$$

with $\alpha \in (0, 1]$ denoting a Robbins Munro step-size. This is a deterministic update rule based on the mean posterior GP approximation of the future cost-to-go. In principle, in the deterministic finite horizon setting one may keep $\alpha = 1$; however, we found Robbins-Munro step sizes empirically preferable. The updated state-action value point-estimates are then used to construct updated GPs.

The following analysis is provided as a comment on the considerations for approximation.

Assumption 3 *The cost function, $\Phi : \mathcal{Z} \to \mathbb{R}$ is a strictly convex function draw from $\Phi \sim \mathcal{GP}(0, k(\cdot,\cdot))$ with convex kernel, $k \in \mathcal{K}_{PD}$, on the compact set, $\mathcal{Z} \times \mathcal{Z}$.*

Theorem 1 *Let Assumption 2 and 3 hold and ensure that the state-action value point estimates, $\boldsymbol{Q}_{T-1}^\pi$, are updated through (11). For the general case of nonlinear dynamics, the point estimates do not lie in the same class of functions as the cost function in Assumption 3.*

Theorem 2 *Assume an initial dataset generated through a policy in a finite horizon MDP satisfying Assumption 2. Provided the optimal future cost-to-go estimate from the posterior mean in (11) is the value function of a nonlinear program satisfying Property 1, the state-action value estimates are described by a continuous function for all timesteps.*

Theorems 1 and 2 motivate the use of the off-policy Q-learning update[1]. The implication is that, regardless of the behavior policy, the state-action value estimates yielded by (11) are well approximated by a continuous function, although one not necessarily within the same class as the cost function. The proof of Theorem 1 follows trivially from the assumption of nonlinear dynamics. The proof of Theorem 2 is general beyond GP approximation and follows the reasoning from Section 2.3. We note that, if the initial step size is set $\alpha = 1$, then the assumption for the initial state-action value estimates to have been generated by a policy satisfying the policy continuity assumption from Assumption 2 may be relaxed. The practical implications of Theorem 1 are relatively minimal given the establishment of representer theorems (Williams & Rasmussen, 2006) and approximation capacity of universal kernels (Micchelli et al., 2006). In the case of affine dynamics and convex cost the following remark can be stated.

Remark 2 *Assume time-varying or time-constant affine dynamics, and a cost function, $\Phi : \mathcal{Z} \to \mathbb{R}_+$, strictly convex on the compact space, $\mathcal{Z}$. Under the assumption of a strictly input convex kernel, positive definite Gram matrix, and restriction for all $\boldsymbol{Q}_{t+1}^\pi$ be strictly non-negative, then the optimal future cost-to-go is a convex function of the current state-control pair. Hence the iterates yielded by (11) are well approximated by a convex function.*

Remark 2 is exemplified by the case of the LQR, where the state-action value function is a quadratic function of the state and control, as is the cost function, for all time indices in the horizon. The restriction imposed on all $\mathbf{Q}_{t+1}^\pi$ to be non-negative can be enforced by simply bounding the iterates produced by (11). Algorithm 1 describes GPQL as proposed. In the Appendix A asymptotic analysis and probabilistic error bounds are provided for for the case of a convex MDP with deterministic dynamics, building upon Remark 2.

4 Experiments

We evaluate GPQL computationally on two control tasks: an LQR problem, and a non-isothermal semi-batch reactor, which highlights the practical applicability of our method to real-world systems.

[1]Note that these results do not necessarily hold for on-policy updates such as SARSA if the rollout policy does not have continuity properties, which is often the case.

Algorithm 1 Gaussian Process Q-Learning

Input: Dynamics, $\mathcal{P}(\cdot)$ with compact state set $\mathcal{X}$, compact control set $\mathcal{U}$, a cost function, $\Phi(\cdot)$, and discrete time horizon, $t \in \{0, \ldots, T-1\}$. A posterior Gaussian process state-action value function approximation $Q_t^\pi(\cdot) \mid \mathbf{Q}_t^\pi \sim \mathcal{GP}(\mu_t^\pi, \mathtt{k}_t^\pi(\cdot,\cdot)),\ \forall t$, estimated from initial data, $\mathbf{Q}_t^\pi$, and an initial policy π_t.
while evaluation budget available **do**
  Approximate Policy Evaluation: `Update` $Q_t^\pi(\cdot) \mid \mathbf{Q}_t^\pi \sim \mathcal{GP}(\mu_t^\pi, \mathtt{k}_t^\pi(\cdot,\cdot)),\ \forall t$
  Rollout the policy, $\pi(\cdot)$ by sampling $\mathcal{P}(\cdot)$
  Collect $(\mathbf{x}_t, \mathbf{u}_t, \Phi(\mathbf{x}_t, \mathbf{u}_t), \mathbf{x}_{t+1})$ and store in $\mathcal{D}_t,\ \forall t$.
  for $t \in \{T-1, \ldots, 0\}$ **do**
    Update the data $\mathbf{Q}_t^\pi$ according to (11).
    Update the posterior process $Q_t^\pi(\cdot) \mid \mathbf{Q}_t^\pi \sim \mathcal{GP}(\mu_t^\pi, \mathtt{k}_t^\pi(\cdot,\cdot))$
  end for
  Approximate Policy Improvement: `Update the policy,` π
  Define the Thompson Sampling policy, π, through the updated posterior processes via (10).
end while
Return: Optimized policy π^* and function approximation Q^*.

4.1 Linear Quadratic Regulator

The LQR problem provides an ideal benchmark for our approach as it satisfies the conditions in Remark 2 with its affine dynamics and strictly convex quadratic cost function. The quadratic cost structure ensures that the state-action value function belongs to the same convex function class throughout the learning process, making it well-suited for GP approximation with convex kernels. We formulate an LQR problem using a double integrator system (position and velocity control) with dynamics and cost matrices defined in Appendix C.1 with time horizon $T = 6$. For this system, we employed a positive definite convex kernel (specifically, a quadratic kernel) to align with the quadratic structure of the optimal Q-function and initialized with 5 episodes generated from a Sobol sequence on control trajectories. Figure 1 shows the final policy[2] that our GPQL algorithm

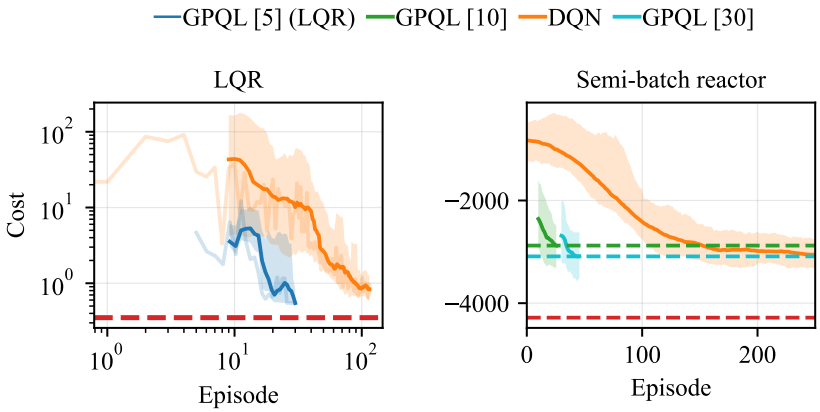

Figure 1: From left to right: Learning curve for the **LQR** case study for GPQL training. Learning curve for the **Semi-batch Reactor** case study. GPQL shown in blue for the LQR case with initial dataset size of 5, and in green and cyan for the Semi-batch Reactor case with initial dataset sizes of 10 and 30, respectively. DQN is orange and the oracle in dashed red. The shaded area shows a single standard deviation.

identifies consistently approaches the oracle's performance after collection of 25 further batches[3].

[2]The final policy greedily exploits the posterior mean function.
[3]Here a batch refers to a rollout of the policy within the system over the discrete time horizon.

We compare to a DQN (Mnih et al., 2013) implementation, which models the Q-function with a neural network. DQN requires approximately 105 batches to reach the same level of performance that GPQL achieves in a total of 30. Figure 2 illustrates the convergence by showing the state and

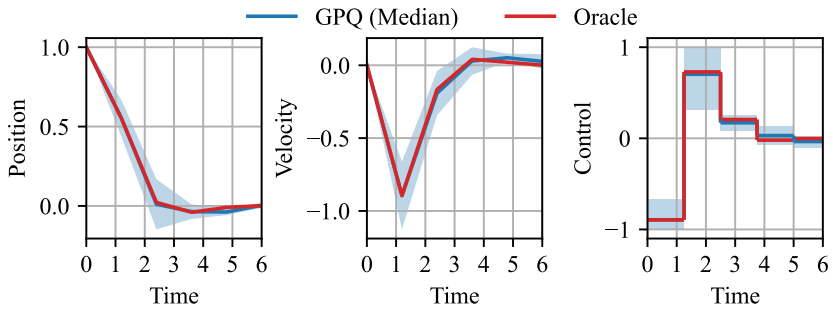

Figure 2: State and control trajectories for the LQR case study. Median states and controls from five repeats (blue) closely align with the oracle's optimal trajectories (red). Shaded areas represent one standard deviation.

control trajectories. The algorithm's learned policy (blue) closely aligns with the oracle's optimal trajectories (red) after training. This demonstrates that our method not only minimizes cost but also recovers the underlying optimal control structure of the LQR problem with relatively small datasets.

4.2 Semi-Batch Reactor

The second experiment involves a benchmark non-isothermal semi-batch reactor control problem (Bradford & Imsland, 2018; Bloor et al., 2024). The system expresses a series reaction mechanism: $\mathsf{A} \xrightarrow{\mathsf{k}_1} \mathsf{B} \xrightarrow{\mathsf{k}_2} 3\mathsf{C}$, where the first reaction is exothermic and the second is endothermic. The state vector $\mathbf{x} = [\mathsf{C_A}, \mathsf{C_B}, \mathsf{C_C}, \mathsf{T}, \mathsf{V}]^\intercal$ tracks species concentrations, temperature, and reactor volume, while control inputs $\mathbf{u} = [\mathsf{T_c}, \mathsf{F}]^\intercal$ represent cooling jacket temperature and feed flow rate. The objective is to maximize the final amount of product C—the cost function is detailed in the supplementary.

Due to the nonlinearity of this system, we employed a Matérn-5/2 kernel for our GP models. We evaluated GPQL with different initial dataset sizes (10 and 30 batches). We allow the TS policy to collect a further 25 batches and compare the final policy (greedily exploiting the posterior mean function) against an oracle nonlinear model predictive controller (NMPC) with perfect system model. Figure 1 shows the produced learning curves. With 30 initial batches, the starting performance is notably higher than with 10 batches, demonstrating the value of a larger initial dataset. An initial performance dip occurs in early training episodes due to the exploration behavior of TS.

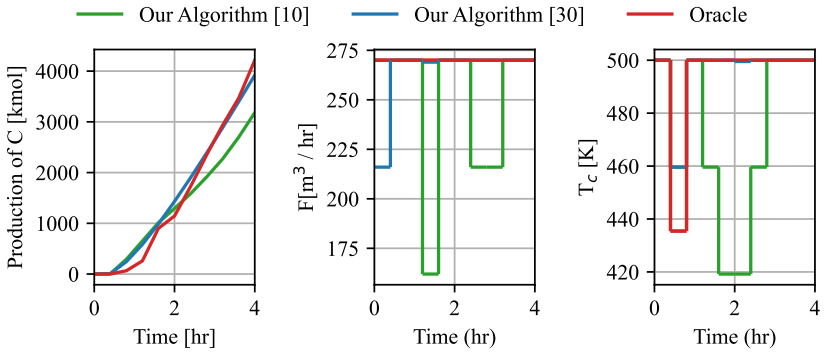

Figure 3: State and control trajectories for semi-batch reactor. The learned policy (blue) closely tracks the oracle (red) for key state variables and control inputs. Shaded areas represent one standard deviation.

Figure 3 compares the state and control trajectories to those of the oracle. The learned policy closely tracks the oracle except for early states and controls. As the initial dataset size increases to 30 batches, GPQL approaches the performance of the oracle, confirming that our method provides significant advantages in scenarios where experimental data is limited and costly.

5 Conclusions and Future Work

This paper introduces GPQL, an approach for solving finite-horizon MDPs by using GPs to approximate state-action value functions. Our method addresses the challenges of continuous state-control spaces while maintaining computational tractability through strategic subset selection. We establish theoretical guarantees on the continuity of state-action value functions and convergence properties, with particular analysis for convex MDPs. Our empirical evaluation on both LQR problems and a non-isothermal semi-batch reactor demonstrates that GPQL achieves near-optimal performance with limited data.

Future work includes extending our analysis to the case of uncertain dynamics. Additionally, investigating the application of the algorithm to MDPs with state chance constraints additionally imposed would enhance its practical utility for real-world control problems where state and control constraints are common.

A Analysis for Convex Markov Decision Process

While our GPQL framework applies generally, we provide stronger theoretical characterization for convex MDPs. This class includes LQR and many process control problems, where convexity is preserved through our GP approximation scheme.

A.1 Convex Markov Decision Process

We extend our analysis to the important special case where the underlying MDP exhibits convex structure. Consider the problem setting from Section 2 with additional assumptions that enable stronger theoretical guarantees.

Assumption 4 (Convex MDP Structure) *The cost function $\Phi : \mathcal{Z} \to \mathbb{R}_+$ is a positive definite convex function with Lipschitz constant L_Φ. The dynamics are affine such that $\boldsymbol{x}_{t+1} = \boldsymbol{f}_t(\boldsymbol{x}_t, \boldsymbol{u}_t)$ where $\boldsymbol{f}_t$ is an affine function for all $t \in \{0, \dots, T-1\}$, and the control set $\mathcal{U}$ is compact and convex.*

Under this assumption, the state-action value function $Q_t^\pi(\mathbf{x}, \mathbf{u})$ inherits convexity properties from the cost function through the additive structure and deterministic affine dynamics. The expectation operation preserves convexity, ensuring that Q_t^π remains convex in both state and control arguments.

Proposition 2 (Convex Structure Preservation) *Let Assumption 4 hold. The optimal solution mapping $\arg\min_{\boldsymbol{u} \in \mathcal{U}} Q_t^\pi(\boldsymbol{x}, \boldsymbol{u})$ is single-valued and continuous in $\boldsymbol{x}$, while the corresponding value function $\min_{\boldsymbol{u} \in \mathcal{U}} Q_t^\pi(\boldsymbol{x}, \boldsymbol{u})$ is convex in $\boldsymbol{x}$.*

This result follows from standard parametric optimization theory for convex programs, where strict convexity of the objective and compactness of the constraint set ensure continuity of the optimal solution mapping and convexity of the value function.

A.2 Gaussian Process Approximation Analysis

For the GP approximation component of our algorithm, we require the cost function to be drawn from a GP prior with appropriate kernel structure.

Assumption 5 (GP Prior Structure) *The cost function Φ is drawn from a GP with zero mean and strictly convex kernel $k(\cdot,\cdot) \in \mathcal{K}_{PD}$ on the compact set $\mathcal{Z} \times \mathcal{Z}$.*

Proposition 3 (GP Convexity Preservation) *Let the GP posterior process $\mathcal{GP}(\mu_t^\pi(\cdot), k_t^\pi(\cdot,\cdot))$ be constructed using a strictly input convex kernel with zero-mean prior, non-negative observed function values, and full-rank positive definite Gram matrix. Then the posterior mean $\mu_t^\pi(\boldsymbol{x},\boldsymbol{u})$ is strictly convex on $(\boldsymbol{x},\boldsymbol{u}) \in \mathcal{Z}$.*

This ensures that the optimization problems arising in the Thompson sampling policy remain well-posed with unique solutions.

Theorem 3 (Function Class Preservation) *Under Assumptions 4 and 5, together with Propositions 2 and 3, the iterative updates for state-action value point estimates generated by equation (11) maintain convex structure and belong to the same function class as the cost function throughout the learning process.*

This theoretical characterization establishes that GPQL preserves the underlying convex structure of the MDP, ensuring that the learned approximations retain the favorable optimization properties of the original problem. The result is particularly relevant for control applications where convexity guarantees are essential for reliable policy optimization.

The analysis demonstrates that backward induction through GP approximations maintains function class consistency when the underlying MDP exhibits convex structure, providing theoretical justification for the algorithm's effectiveness in this important special case.

B Additional Studies

B.1 Ablation of Subset Selection Strategy

To evaluate our proposed M-DPP subset selection mechanism, we compare four approaches for subset selection. The exact GP baseline uses the complete accumulated dataset without subset selection. Random subset selection samples datapoints from a uniform distribution. Uniform coverage selection chooses points with probability proportional to the kernel function alone.

The subset selection strategies use a maximum of 20 inducing points for both experiments, with additional comparisons at 30 points shown as dashed lines. The exact GP utilizes all available datapoints. This design evaluates whether principled subset selection achieves comparable performance to exact GP formulations while maintaining computational tractability. The ablation study demonstrates that M-DPP marginally outperforms uniform coverage in the LQR case, with both methods substantially exceeding the performance using the full dataset. For the semi-batch reactor, uniform coverage performs well at 20 inducing points while M-DPP performs better at 30 datapoints. Remarkably, both subset selection strategies consistently outperform the GP using the full dataset, likely because the the updates of to the Q-function becomes trapped in suboptimal policy regions as datasets grow, whereas M-DPP and uniform subset selection maintains better exploration-exploitation balance through diverse point selection.

B.2 Timing Analysis

The computational efficiency of different subset selection strategies was evaluated across both case studies to understand the trade-offs between accuracy and computational burden. Figures 5 and 6 present timing comparisons for four approaches: Exact GP, M-DPP subset selection, Random selection, and Uniform selection, conducted across five random seeds with a maximum subset size of 20 datapoints.

The offline phase analysis (left panels) reveals the fundamental scalability challenge of exact GPs. While subset-based methods maintain relatively constant fitting times regardless of dataset size,

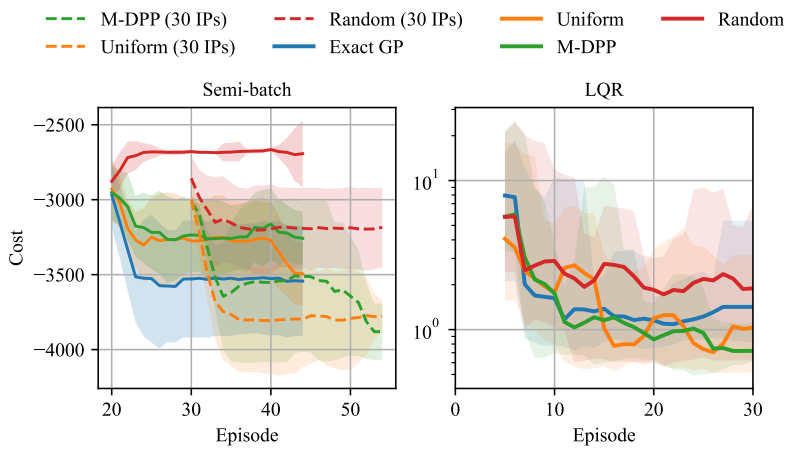

Figure 4: Ablation study of subset selection strategies. From left-to-right, the ablation study for semi-batch, and then the LQR case study. The Blue line is the full dataset (i.e. no subset selection), orange is a uniform coverage, green is the M-DPP, red is a random subset, and the dashed lines represent subsets with a maximum of 30 data points.

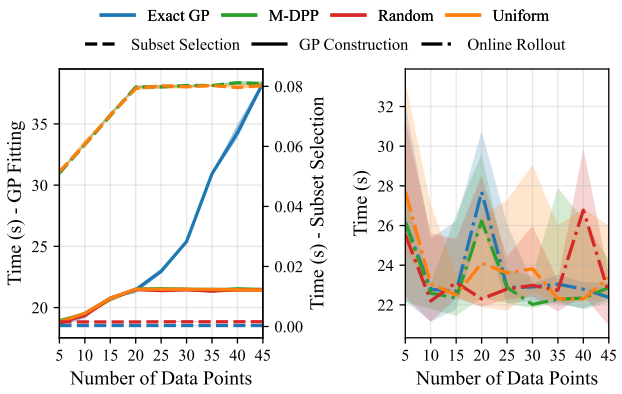

Figure 5: Computational timing comparison of subset selection strategies for the LQR case study across different dataset sizes. **Left panel:** Offline phase showing GP construction time (solid lines, left y-axis) and subset selection time (dashed lines, right y-axis). **Right panel:** Online rollout time (dash-dot lines). Four strategies are compared: Exact GP, M-DPP, Random selection, and Uniform selection. Lines show median timing across five seeds, with shaded regions indicating min-max ranges.

the exact GP approach exhibits the expected cubic growth in computational expense. The subset selection overhead (dashed lines, right y-axis) shows that M-DPP and uniform strategies both incur higher computational costs due to its kernel matrix computations compared to random and no selection strategies.

The online rollout phase (right panels, dash-dot lines) demonstrates that subset methods provide consistent performance benefits. Online computational requirements remain stable across all subset selection strategies, while the exact GP approach shows increasing computational burden as dataset sizes grow. The median timing across five seeds, with min-max ranges shown as shaded regions, demonstrates consistent computational patterns across different random initializations.

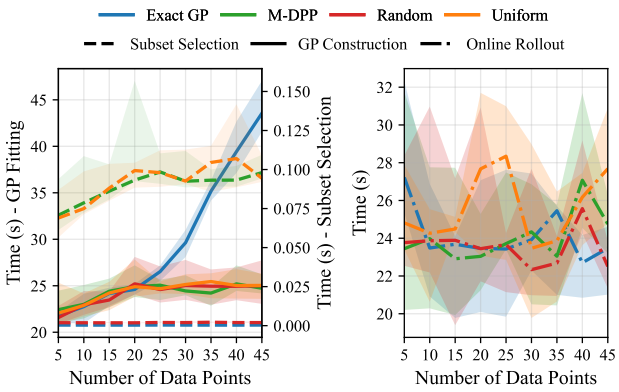

Figure 6: Computational timing comparison of subset selection strategies for the semi-batch reactor case study across different dataset sizes. **Left panel:** Offline phase showing GP construction time (solid lines, left y-axis) and subset selection time (dashed lines, right y-axis). **Right panel:** Online rollout time (dash-dot lines). Four strategies are compared: Exact GP, M-DPP, Random selection, and Uniform selection. Lines show median timing across five seeds, with shaded regions indicating min-max ranges.

Acknowledgments

The authors thank Benoît Chachuat and Mark van der Wilk for insightful and valuable discussions during the development of this work; and to Shiqiang Zhang and Jixiang Qing for their thoughts on presentation of the manuscript.

References

Jean-Pierre Aubin and Helene Frankowska. *Set-valued analysis*. Springer, 1999.

Kamyar Azizzadenesheli and Animashree Anandkumar. Efficient Exploration through Bayesian Deep Q-Networks, September 2019. URL http://arxiv.org/abs/1802.04412. arXiv:1802.04412 [cs].

Maximilian Bloor, José Torraca, Ilya Orson Sandoval, Akhil Ahmed, Martha White, Mehmet Mercangöz, Calvin Tsay, Ehecatl Antonio Del Rio Chanona, and Max Mowbray. Pc-gym: Benchmark environments for process control problems, 2024. URL https://arxiv.org/abs/2410.22093.

Eric Bradford and Lars Imsland. Economic stochastic model predictive control using the unscented kalman filter. *IFAC-PapersOnLine*, 51(18):417–422, 2018. 10th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2018.

Eric Bradford, Lars Imsland, Dongda Zhang, and Ehecatl Antonio Del Rio Chanona. Stochastic data-driven model predictive control using gaussian processes. *Computers & Chemical Engineering*, 139:106844, August 2020. ISSN 0098-1354. DOI: 10.1016/j.compchemeng.2020.106844. URL https://linkinghub.elsevier.com/retrieve/pii/S0098135419313080. Publisher: Elsevier BV.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in Neural Information Processing Systems*, 31, 2018.

Girish Chowdhary, Miao Liu, Robert Grande, Thomas Walsh, Jonathan How, and Lawrence Carin. Off-policy reinforcement learning with gaussian processes. *IEEE/CAA Journal of Automatica Sinica*, 1(3):227–238, 2014. DOI: 10.1109/JAS.2014.7004680.

Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian Processes for Data-Efficient Learning in Robotics and Control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):408–423, February 2015. ISSN 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.2013.218. URL http://ieeexplore.ieee.org/document/6654139/.

Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with Gaussian processes. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pp. 201–208, Bonn, Germany, 2005. ACM Press. DOI: 10.1145/1102351.1102377. URL http://portal.acm.org/citation.cfm?doid=1102351.1102377.

Anthony V Fiacco and Yo Ishizuka. Sensitivity and stability analysis for nonlinear programming. *Annals of Operations Research*, 27(1):215–235, 1990.

Robert Grande, Thomas Walsh, and Jonathan How. Sample efficient reinforcement learning with gaussian processes. In *International Conference on Machine Learning*, pp. 1332–1340. PMLR, 2014.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012. ISSN 1935-8237, 1935-8245. DOI: 10.1561/2200000044. URL http://arxiv.org/abs/1207.6083. arXiv:1207.6083 [stat].

Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Malte Kuss and Carl Rasmussen. Gaussian processes in reinforcement learning. *Advances in neural information processing systems*, 16, 2003.

Ali Mesbah. Stochastic Model Predictive Control: An Overview and Perspectives for Future Research. *IEEE Control Systems*, 36(6):30–44, December 2016. ISSN 1066-033X, 1941-000X. DOI: 10.1109/mcs.2016.2602087. URL https://ieeexplore.ieee.org/document/7740982/. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013. URL https://arxiv.org/abs/1312.5602.

Henry B Moss, Sebastian W Ober, and Victor Picheny. Inducing point allocation for sparse gaussian processes in high-throughput bayesian optimisation. In *International Conference on Artificial Intelligence and Statistics*, pp. 5213–5230. PMLR, 2023.

M. Mowbray, P. Petsagkourakis, E.A. Del Rio-Chanona, and D. Zhang. Safe chance constrained reinforcement learning for batch process control. *Computers & Chemical Engineering*, 157:107630, January 2022. ISSN 0098-1354. DOI: 10.1016/j.compchemeng.2021.107630. URL https://linkinghub.elsevier.com/retrieve/pii/S0098135421004087. Publisher: Elsevier BV.

Thomas Savage, Dongda Zhang, Max Mowbray, and Ehecatl Antonio Del Río Chanona. Model-free safe reinforcement learning for chemical processes using Gaussian processes. *IFAC-PapersOnLine*, 54(3):504–509, 2021. ISSN 24058963. DOI: 10.1016/j.ifacol.2021.08.292. URL https://linkinghub.elsevier.com/retrieve/pii/S240589632101065X.

Rangarajan K Sundaram. *A first course in optimization theory*. Cambridge university press, 1996.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

James Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Efficiently sampling functions from gaussian process posteriors. In *International Conference on Machine Learning*, pp. 10292–10302. PMLR, 2020.

Supplementary Materials

The following content was not necessarily subject to peer review.

C Experimental Setup

C.1 Linear Quadratic Regulator

$$\min_{\mathbf{u}_0,\ldots,\mathbf{u}_{T-1}} \quad \mathbf{x}_T^\intercal Q_T \mathbf{x}_T + \sum_{k=0}^{T-1} (\mathbf{x}_k^T Q \mathbf{x}_k + \mathbf{u}_k^T R \mathbf{u}_k) \tag{12}$$

$$\text{s.t} \quad \mathbf{x}_{k+1} = A\mathbf{x}_k + B\mathbf{u}_k \tag{13}$$

$$\mathbf{x}_0 = \mathbf{x}(0) \tag{14}$$

$$\mathbf{u}_L \leq \mathbf{u}_k \leq \mathbf{u}_U \tag{15}$$

$$\forall k \in \mathcal{T} \setminus \{T\} \tag{16}$$

where $\mathbf{x}_k \in \mathbb{R}^{n_x}$ represents the state vector, $\mathbf{u}_k \in \mathbb{R}^{n_u}$ is the control input vector, $A \in \mathbb{R}^{n_x \times n_x}$ and $B \in \mathbb{R}^{n_x \times n_u}$ define the discrete-time system dynamics, $Q \in \mathbb{R}^{n_x \times n_x}$ and $R \in \mathbb{R}^{n_u \times n_u}$ are positive semi-definite and positive definite cost matrices for state and control respectively, $Q_N \in \mathbb{R}^{n_x \times n_x}$ is the terminal state cost matrix, $\mathbf{u}_L$ and $\mathbf{u}_U$ are the lower and upper control bounds respectively, and N denotes the finite time horizon. In the LQR experiment, the following dynamics matrices are used $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Cost matrices are set as $Q = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 0.1 \end{bmatrix}$ and $R = 0.1$.

C.2 Semi-Batch Reactor

The dynamics of the semi-batch reactor are shown below:

$$\frac{d\mathsf{C_A}}{dt} = \frac{\mathsf{F}}{\mathsf{V}}(\mathsf{C_{A,in}} - \mathsf{C_A}) - \mathsf{k_1}\mathsf{C_A} \tag{17}$$

$$\frac{d\mathsf{C_B}}{dt} = -\frac{\mathsf{F}}{\mathsf{V}}\mathsf{C_B} + \mathsf{k_1}\mathsf{C_A} - \mathsf{k_2}\mathsf{C_B} \tag{18}$$

$$\frac{d\mathsf{C_C}}{dt} = -\frac{\mathsf{F}}{\mathsf{V}}\mathsf{C_C} + \mathsf{k_2}\mathsf{C_B} \tag{19}$$

$$\frac{d\mathsf{T}}{dt} = \frac{\mathsf{UA}(\mathsf{T_a} - \mathsf{T}) - \mathsf{F_{A0}}\mathsf{C_{P_A}}(\mathsf{T} - \mathsf{T_b})}{[\mathsf{C_A}\mathsf{C_{P_A}} + \mathsf{C_B}\mathsf{C_{P_B}} + \mathsf{C_C}\mathsf{C_{P_C}}]\mathsf{V} + \mathsf{N_{H_2SO_4}}\mathsf{C_{P_{H_2SO_4}}}} \tag{20}$$

$$+ \frac{[(\Delta\mathsf{H_{R1}})(-\mathsf{k_1}\mathsf{C_A}) + (\Delta\mathsf{H_{R2B}})(-\mathsf{k_2}\mathsf{C_B})]\mathsf{V}}{[\mathsf{C_A}\mathsf{C_{P_A}} + \mathsf{C_B}\mathsf{C_{P_B}} + \mathsf{C_C}\mathsf{C_{P_C}}]\mathsf{V} + \mathsf{N_{H_2SO_4}}\mathsf{C_{P_{H_2SO_4}}}} \tag{21}$$

$$\frac{d\mathsf{V}}{dt} = \mathsf{F} \tag{22}$$

The variables in the system are denoted as follows: $\mathsf{C_A}$, $\mathsf{C_B}$, and $\mathsf{C_C}$ represent the concentrations (mol/dm^3) of species A, B, and C; T is the reactor temperature (K); V is the liquid volume (m^3); F is the flow rate of pure A entering the reactor (m^3/h); $\mathsf{T_a}$ is the temperature of the heat exchanger (K); $\mathsf{C_{P_A}}$, $\mathsf{C_{P_B}}$, $\mathsf{C_{P_C}}$ are the specific heat capacities of components A, B, and C; $\Delta\mathsf{H_{R1A}}$ and $\Delta\mathsf{H_{R2B}}$ are the reaction enthalpies for the first and second reactions; $\mathsf{k_1}$ and $\mathsf{k_2}$ are temperature-dependent rate constants.

The cost function is defined,

$$\Phi(\mathbf{x}, \mathbf{u}) = \begin{cases} -\mathsf{C_C}(t+1)\mathsf{V}(t+1) & \text{if } t = T-1 \\ 0 & \text{else} \end{cases}. \tag{23}$$

D Deep Q-Network (DQN) Implementation Details

For comparison with our GPQL method, we implemented a standard DQN (Mnih et al., 2013) with the following configuration. The neural network architecture consists of fully connected layers with ReLU activations, taking concatenated state-action pairs as input and outputting scalar Q-values. To find the optimal action at each state, we optimize over the Q-function by sampling actions and selecting the one with minimum Q-value, similar to the approach used in our GPQL method.

Table 1: DQN Hyperparameters and Architecture

Parameter	**Value**
Network Architecture	[256, 256, 1]
Activation Function	ReLU
Learning Rate	3e-4
Batch Size	128
Replay Buffer Size	100,000
Target Network Update Frequency	Every 10 episodes
Action Optimization	10 random samples per state
Exploration (ϵ-greedy)	
Initial ϵ	1.0
Final ϵ	0.001
Decay Steps	500
Optimizer	Adam
Loss Function	MSE

The exploration strategy follows an exponential ϵ-greedy decay schedule: $\epsilon = \epsilon_{\text{final}} + (\epsilon_{\text{start}} - \epsilon_{\text{final}}) \exp(-\text{step}/\text{decay})$. Experience replay is used with uniform sampling from the replay buffer. The target network is updated via soft updates with $\tau = 0.005$.