Submitted to the Annals of Statistics
arXiv: arXiv:1708.08396

CONTROLLED SEQUENTIAL MONTE CARLO

By Jeremy Heng*, Adrian N. Bishop†, George Deligiannidis‡ and Arnaud Doucet‡

ESSEC Business School, CSIRO and University of Technology Sydney†, and Oxford University‡*

Sequential Monte Carlo methods, also known as particle methods, are a popular set of techniques for approximating high-dimensional probability distributions and their normalizing constants. These methods have found numerous applications in statistics and related fields; e.g. for inference in non-linear non-Gaussian state space models, and in complex static models. Like many Monte Carlo sampling schemes, they rely on proposal distributions which crucially impact their performance. We introduce here a class of controlled sequential Monte Carlo algorithms, where the proposal distributions are determined by approximating the solution to an associated optimal control problem using an iterative scheme. This method builds upon a number of existing algorithms in econometrics, physics, and statistics for inference in state space models, and generalizes these methods so as to accommodate complex static models. We provide a theoretical analysis concerning the fluctuation and stability of this methodology that also provides insight into the properties of related algorithms. We demonstrate significant gains over state-of-the-art methods at a fixed computational complexity on a variety of applications.

1. Introduction. Sequential Monte Carlo (SMC) methods have found a wide range of applications in many areas of statistics as they can be used, among others things, to perform inference for dynamic non-linear non-Gaussian state space models [32, 41, 33] but also for complex static models [39, 9, 13]; see [8, 17, 31] for recent reviews of this active area. Although these methods are supported by theoretical guarantees [12], the number of samples required to achieve a desired level of precision of the corresponding Monte Carlo estimators can be prohibitively large in practice, especially so for high-dimensional problems.

The present work is one means to address the computational difficulties with SMC in offline inference settings. In particular, we leverage ideas from optimal control theory and we seek novel SMC methods that achieve a desired level of precision at a fraction of the computational cost of state-of-the-

MSC 2010 subject classifications: Primary 62M05; secondary 62F12, 62M10

Keywords and phrases: State space models, annealed importance sampling, normalizing constants, optimal control, approximate dynamic programming, reinforcement learning

art algorithms. We introduce a class of algorithms that will be referred to as controlled SMC, under which the sequence of SMC proposal distributions are related naturally with an associated optimal control problem. The cost functional is the Kullback–Leibler divergence from the sought after proposals to the target distributions and we may account for an arbitrary current proposal estimate. With this formulation, the optimal proposal distributions are specified by the optimal control policy of a related dynamic programming recursion. In general, this dynamic programming recursion is intractable. However, by making this connection, we can then exploit an array of methods and procedures for so-called approximate dynamic programming (ADP). Broadly speaking, a single iteration of our proposed methodology involves: 1) based on a current sequence of proposal distributions, running a SMC method to obtain a collection of samples that approximate the sequence of SMC target distributions; 2) using these samples as support points, we approximate intractable backward recursions using regression to compute a new policy that specifies a new sequence of approximately optimal proposal distributions. Continuing in this manner allows us to further refine the proposal distributions and improve our approximation of the target distributions, via a novel iteration of SMC and ADP.

Prior influential work in [42] proposed a motivating method in the context of importance sampling, for computing the marginal likelihood in state space models. In this contribution, the sequential structure which defines the marginal likelihood is exploited and proposal distributions are defined by a sequence of parameterized Markov transition kernels. A criterion based on the variance of importance weights is introduced to optimize these parameters and an iterative procedure with fixed random numbers is proposed. The work of [47] extends [42] by employing these optimized proposal distributions within a SMC methodology. In particular, [47] identified the appropriate importance weights one should use for resampling, which is crucial to ensure that the variance of the marginal likelihood estimator remains controlled. Moreover, [47] also recommends relaxing the use of common random variables across iterations. Recent work in [23], again with a focus on discrete time state space models, may be viewed as an extension of [47] where resampling is performed at every iteration, instead of just the last. The resulting algorithm is numerically much more stable than [42, 47]. Although the iterative procedures in [42, 47, 23] are similar in spirit to our proposed methodology, the main and important difference is that all these works employ an optimality criterion, to learn proposal distributions, that is not adjusted across iterations to account for any improvements made in prior iterations. Finally, we highlight related ideas in [29, 45] where the focus is partially observed diffusion models and the algorithms proposed therein are based on other strategies to learn a parameterized additive control directly. Such ideas have also been exploited in physics to perform rare event simulation for diffusions [40].

Our work extends these contributions in the following ways. Firstly, these preceding works [42, 47, 23] consider only state space models. In contrast, the methodology proposed here allows us to perform inference for static models; a direct extension of these prior methods [42, 47, 23] to static models is infeasible, as it leads to algorithms which are not implementable.

Secondly, in contrast to the methodology in [42, 47, 23], the Kullback–Leibler optimality criterion at each iteration in our approach, is dependent on the approximately optimal proposal distributions computed at the preceding iteration; i.e. we seek to minimize the residual discrepancy between any previously estimated proposals and the target distributions. This difference allows us to elucidate the effect each iteration in our method has on refining proposal distributions and improves algorithmic performance as illustrated in Section 6.2. See also [49, 50] for related iterative procedures in continuous-time optimal control approximation.

The controlled SMC methodology is one of the main contributions of this work. Another contribution is to provide a detailed theoretical analysis of various aspects of our methodology. In Proposition 5.1, we provide a backward recursion that characterizes the error of policies estimated using our ADP procedure. This error is given naturally in terms of function approximation errors with finite samples and the stability properties of the dynamic programming recursion defining the optimal policy, which is addressed in Proposition 5.2. These results show that we can obtain good approximations of the optimal policy and hence the optimal proposal distributions, if the function classes employed are 'rich' enough and the number of samples used to learn policies is sufficiently large. In Theorem 5.1, we then establish a central limit theorem for our ADP algorithm as the number of samples used in the policy learning goes to infinity. This reveals that the algorithm concentrates around an idealized ADP algorithm and provides a precise characterization of how Monte Carlo errors correlate over time. These preceding results concern a single iteration of our proposed method and may be applied to the existing algorithms discussed above, e.g. [42, 47, 23]. Using the notion of iterated random functions, we introduce a novel framework in Theorem 5.2 to understand the asymptotic behaviour of our algorithm as the number of iterations converges to infinity. This elucidates the need for iterating the ADP procedure and provides insight into the number of iterations required in practice. The discussion surrounding Theorem 5.2 also emphasizes a key difference between the newly proposed method and existing work in [42, 47, 23]. After the first version of this work appeared, a similar approach was developed for generic stochastic control problems in [24]. Our results hold under strong assumptions but appear to capture our experimental results remarkably well.

The rest of this paper is organized as follows. We introduce SMC methods in the framework of Feynman-Kac models [12] in Section 2 and twisted variants in Section 3, as this affords us the generality to cover both state space models and static models. We then identify the optimal policy that induces an optimal SMC method in Section 4.1. We describe general methods to approximate the optimal policy in Section 4.2 and develop an iterative scheme to refine policies in Section 4.3. The proposed methodology is illustrated on a neuroscience application in Section 4.4. We present the results of our analysis in Section 5 and conclude with applications in Sections 6-7. All proofs are given in the Supplementary Material which also includes three additional applications. MATLAB code to reproduce all numerical results is available online[1].

2. Motivating models and sequential Monte Carlo.

2.1. *Notation.* We first introduce notation used throughout the article. Given integers $n \leq m$ and a sequence $(x_t)_{t\in\mathbb{N}}$, we define the set $[n:m] = \{n,\dots,m\}$ and write the subsequence $x_{n:m} = (x_n,\dots,x_m)$. When $n < m$, we use the convention $\prod_{t=m}^{n} x_t = 1$. Let $(\mathsf{E},\mathcal{E})$ be an arbitrary measurable space. We denote the set of all finite signed measures by $\mathcal{S}(\mathsf{E})$, the set of all probability measures by $\mathcal{P}(\mathsf{E}) \subset \mathcal{S}(\mathsf{E})$, and the set of all Markov transition kernels on $(\mathsf{E},\mathcal{E})$ by $\mathcal{M}(\mathsf{E})$. Given $\mu,\nu \in \mathcal{P}(\mathsf{E})$, we write $\mu \ll \nu$ if μ is absolutely continuous w.r.t. ν and denote the corresponding Radon-Nikodym derivative as $\mathrm{d}\mu/\mathrm{d}\nu$. For any $x \in \mathsf{E}$, δ_x denotes the Dirac measure at x. The set of all real-valued, $\mathcal{E}$-measurable, lower bounded, bounded or continuous functions on E are denoted by $\mathcal{L}(\mathsf{E})$, $\mathcal{B}(\mathsf{E})$ and $\mathcal{C}(\mathsf{E})$ respectively. Given $\gamma \in \mathcal{S}(\mathsf{E})$ and $M \in \mathcal{M}(\mathsf{E})$, we define $(\gamma \otimes M)(\mathrm{d}x,\mathrm{d}y) = \gamma(\mathrm{d}x)M(x,\mathrm{d}y)$ and $(M \otimes \gamma)(\mathrm{d}x,\mathrm{d}y) = M(y,\mathrm{d}x)\gamma(\mathrm{d}y)$ as finite signed measures on the product space $\mathsf{E} \times \mathsf{E}$, equipped with the product σ-algebra $\mathcal{E} \times \mathcal{E}$. Given $\gamma \in \mathcal{S}(\mathsf{E})$, $M \in \mathcal{M}(\mathsf{E})$, $\varphi \in \mathcal{B}(\mathsf{E})$, $\xi \in \mathcal{B}(\mathsf{E}\times\mathsf{E})$, we define the integral $\gamma(\varphi) = \int_{\mathsf{E}} \varphi(x)\gamma(\mathrm{d}x)$, the signed measure $\gamma M(\cdot) = \int_{\mathsf{E}} \gamma(\mathrm{d}x)M(x,\cdot) \in \mathcal{S}(\mathsf{E})$ and functions $M(\varphi)(\cdot) = \int_{\mathsf{E}} \varphi(y)M(\cdot,\mathrm{d}y) \in \mathcal{B}(\mathsf{E})$, $M(\xi)(\cdot) = \int_{\mathsf{E}} \xi(\cdot,y)M(\cdot,\mathrm{d}y) \in \mathcal{B}(\mathsf{E})$.

[1] Link: https://github.com/jeremyhengjm/controlledSMC

2.2. *Feynman-Kac models.* We begin by introducing Feynman-Kac models [12] and defer a detailed discussion of their applications to Sections 2.3-2.4. Consider a non-homogeneous Markov chain of length $T+1 \in \mathbb{N}$ on a measurable space $(\mathsf{X}, \mathcal{X})$, associated with an initial distribution $\mu \in \mathcal{P}(\mathsf{X})$, and a collection of Markov transition kernels $M_t \in \mathcal{M}(\mathsf{X})$ for $t \in [1:T]$. We denote the law of the Markov chain on path space X^{T+1}, equipped with the product σ-algebra $\mathcal{X}^{T+1}$, with

$$\mathbb{Q}(\mathrm{d}x_{0:T}) = \mu(\mathrm{d}x_0) \prod_{t=1}^{T} M_t(x_{t-1}, \mathrm{d}x_t) \tag{1}$$

and denote expectations w.r.t. $\mathbb{Q}$ by $\mathbb{E}_{\mathbb{Q}}$, whereas we write $\mathbb{E}_{\mathbb{Q}}^{t,x}$ for conditional expectations on the event $X_t = x \in \mathsf{X}$. Given a sequence of strictly positive functions $G_0 \in \mathcal{B}(\mathsf{X})$ and $G_t \in \mathcal{B}(\mathsf{X} \times \mathsf{X})$ for $t \in [1:T]$, we define the Feynman-Kac path measure

$$\mathbb{P}(\mathrm{d}x_{0:T}) = Z^{-1} G_0(x_0) \prod_{t=1}^{T} G_t(x_{t-1}, x_t)\, \mathbb{Q}(\mathrm{d}x_{0:T}) \tag{2}$$

where $Z := \mathbb{E}_{\mathbb{Q}}\left[G_0(X_0) \prod_{t=1}^{T} G_t(X_{t-1}, X_t)\right]$ denotes the normalizing constant. Equation (2) can be understood as the probability measure obtained by repartitioning the probability mass of $\mathbb{Q}$ with the potential functions $(G_t)_{t\in[0:T]}$.

To examine the time evolution of (2), we define the following sequence of positive signed measures $\gamma_t \in \mathcal{S}(\mathsf{X})$ for $t \in [0:T]$ by

$$\gamma_t(\varphi) = \mathbb{E}_{\mathbb{Q}}\left[\varphi(X_t) G_0(X_0) \prod_{s=1}^{t} G_s(X_{s-1}, X_s)\right] \tag{3}$$

and their normalized counterparts $\eta_t \in \mathcal{P}(\mathsf{X})$ by

$$\eta_t(\varphi) = \gamma_t(\varphi)/Z_t \tag{4}$$

for $\varphi \in \mathcal{B}(\mathsf{X})$, $t \in [0:T]$, where $Z_t := \gamma_t(\mathsf{X})$. Equations (3) and (4) are known as the unnormalized and normalized (updated) Feynman-Kac models respectively [12, Definition 2.3.2]. These models are determined by the triple $\left\{\mu, (M_t)_{t\in[1:T]}, (G_t)_{t\in[0:T]}\right\}$, which depends on the specific application of interest. The measure η_T is the terminal time marginal distribution of $\mathbb{P}$ and $Z = Z_T = \mu(G_0) \prod_{t=1}^{T} \eta_{t-1}(M_t(G_t))$.

2.3. *State space models.* Consider an X-valued hidden Markov chain $(X_t)_{t\in[0:T]}$, whose law on $(\mathsf{X}^{T+1}, \mathcal{X}^{T+1})$ is given by

$$\mathbb{H}(\mathrm{d}x_{0:T}) = \nu(\mathrm{d}x_0) \prod_{t=1}^{T} f_t(x_{t-1}, \mathrm{d}x_t)$$

where $\nu \in \mathcal{P}(\mathsf{X})$ and $f_t \in \mathcal{M}(\mathsf{X})$ for $t \in [1:T]$. The Y-valued observations $(Y_t)_{t\in[0:T]}$ are assumed to be conditionally independent given $(X_t)_{t\in[0:T]}$ and the conditional distribution of Y_t has a strictly positive density $g_t(X_t,\cdot)$ with $g_t \in \mathcal{B}(\mathsf{X}\times\mathsf{Y})$ for $t \in [0:T]$. Here $\left\{\nu, (f_t)_{t\in[1:T]}, (g_t)_{t\in[0:T]}\right\}$ can potentially depend on unknown static parameters $\theta \in \Theta$, but this is notationally omitted for simplicity. Given access to a realization $y_{0:T} \in \mathsf{Y}^{T+1}$ of the observation process, statistical inference for these models relies on the marginal likelihood of $y_{0:T}$ given θ,

$$Z(y_{0:T}) = \mathbb{E}_{\mathbb{H}}\left[\prod_{t=0}^{T} g_t(X_t, y_t)\right],$$

and/or the smoothing distribution, i.e. the conditional distribution of $X_{0:T}$ given $Y_{0:T} = y_{0:T}$ and θ

$$\mathbb{P}(\mathrm{d}x_{0:T}|y_{0:T}) = Z(y_{0:T})^{-1}\prod_{t=0}^{T} g_t(x_t, y_t)\,\mathbb{H}(\mathrm{d}x_{0:T}). \tag{5}$$

If we set $\mathbb{Q} \in \mathcal{P}(\mathsf{X}^{T+1})$ defined in (1) equal to $\mathbb{H}$, we recover the Feynman-Kac path measure representation (2) by defining $G_t(x_{t-1}, x_t) = g_t(x_t, y_t)$ for all $t \in [0:T]$. However, this representation is not unique. Indeed any $\mathbb{Q}$ satisfying $\mathbb{H} \ll \mathbb{Q}$ provides a Feynman-Kac path measure representation of (2) by defining the potentials

$$G_0(x_0) = \frac{\mathrm{d}(\nu\cdot g_0)}{\mathrm{d}\mu}(x_0), \quad G_t(x_{t-1}, x_t) = \frac{\mathrm{d}(f_t\cdot g_t)(x_{t-1},\cdot)}{\mathrm{d}M_t(x_{t-1},\cdot)}(x_t), \quad t \in [1:T].$$

As outlined in [17], most SMC algorithms available at present correspond to the same basic mechanism applied to different Feynman-Kac representations of a given target probability measure. The bootstrap particle filter (BPF) presented in [22] corresponds to $\mathbb{Q} = \mathbb{H}$, i.e. $M_t(x_{t-1}, \mathrm{d}x_t) = f_t(x_{t-1}, \mathrm{d}x_t)$ for $t \in [1:T]$, while the popular 'fully adapted' auxiliary particle filter (APF) of [41] uses $M_t(x_{t-1}, \mathrm{d}x_t) = \mathbb{P}(\mathrm{d}x_t|x_{t-1}, y_t) \propto f_t(x_{t-1}, \mathrm{d}x_t)g_t(x_t, y_t)$.

As a motivating example, we consider a model for $T+1 = 3000$ measurements collected from a neuroscience experiment [48]. The observation $y_t \in \mathsf{Y} = [0:M]$ at each time instance $t \in [0:T]$, shown in the left panel of Figure 1, represents the number of activated neurons over $M = 50$ repeated experiments and is modelled as a binomial distribution with probability of success $p_t \in [0,1]$. We will write its probability mass function as $y_t \mapsto \mathrm{Bin}(y_t; M, p_t)$. To model the time varying behaviour of activation probabilities, it is assumed that $p_t = \kappa(X_t)$ where $\kappa(u) := (1+\exp(-u))^{-1}$ for $u \in \mathbb{R}$ is the logistic link function and $(X_t)_{t\in[0:T]}$ is a real-valued first-order autoregressive process. This corresponds to a time homogeneous state space model on $\mathsf{X} = \mathbb{R}$, equipped with its Borel σ-algebra $\mathcal{X} = \mathfrak{B}(\mathbb{R})$, with $\nu = \mathcal{N}(0,1)$,

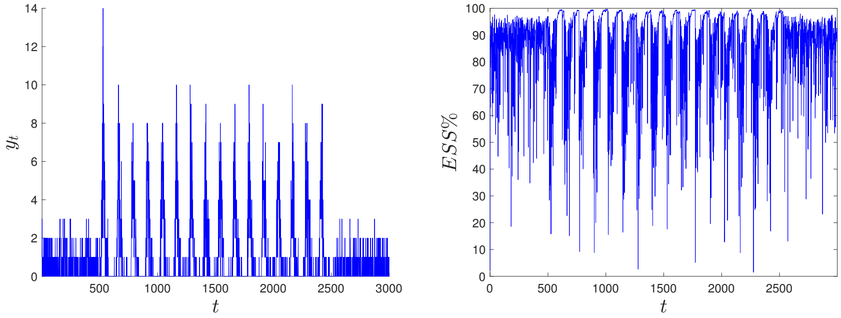

FIG 1. *Number of activated neurons over $M = 50$ repeated experiments with time (*left*) and effective sample size of bootstrap particle filter with $N = 1024$ particles (*right*) for the neuroscience model with parameters $\alpha = 0.99$ and $\sigma^2 = 0.11$.*

$f(x_{t-1}, \mathrm{d}x_t) = \mathcal{N}(x_t; \alpha x_{t-1}, \sigma^2)\mathrm{d}x_t$, and $g(x_t, y_t) = \mathrm{Bin}(y_t; M, \kappa(x_t))$ for $t \in [1:T]$, where we denote the Gaussian distribution on $\mathbb{R}^d$ with mean vector $\xi \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ by $\mathcal{N}(\xi, \Sigma)$ and its Lebesgue density by $x \mapsto \mathcal{N}(x; \xi, \Sigma)$. The parameters of this model to be inferred from data are $\theta = (\alpha, \sigma^2) \in [0, 1] \times \mathbb{R}_+$.

2.4. *Static models.* Suppose we are interested in sampling from a target distribution $\eta(\mathrm{d}x) = Z^{-1}\gamma(\mathrm{d}x) \in \mathcal{P}(\mathsf{X})$ and/or estimating its normalizing constant $Z = \gamma(\mathsf{X})$. To facilitate inference, we introduce a sequence of probability measures $(\eta_t)_{t\in[0:T]}$ in $\mathcal{P}(\mathsf{X})$ that bridges a simple distribution $\eta_0 = \mu$ to the target distribution $\eta_T = \eta$ with $\eta \ll \mu$. Our implementation in Section 7 adopts the geometric path [18, 39, 13]

$$(6) \qquad \gamma_t(\mathrm{d}x) := \mu(\mathrm{d}x)\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\mu}(x)\right)^{\lambda_t}, \qquad \eta_t(\mathrm{d}x) := \gamma_t(\mathrm{d}x)/Z_t, \quad t \in [0:T],$$

where $Z_t := \gamma_t(\mathsf{X})$ and $(\lambda_t)_{t\in[0:T]} \in [0, 1]^{T+1}$ is an increasing sequence satisfying $\lambda_0 = 0$ and $\lambda_T = 1$; see [13, Section 2.3.1] for choices in other inference settings. In order to define $\mathbb{Q}$, we introduce a sequence of 'forward' Markov transition kernels $M_t \in \mathcal{M}(\mathsf{X})$ for $t = [1:T]$ where $\eta_{t-1}M_t$ approximates η_t. One expects the distribution $\hat{\eta} = \eta_0 M_1 \cdots M_T$ of samples drawn from a non-homogeneous Markov chain with initial distribution η_0 and transition kernels $(M_t)_{t\in[1:T]}$ to be close to $\eta_T = \eta$. However, importance sampling cannot be employed to correct for the discrepancy between $\hat{\eta}$ and η, as $\hat{\eta}$ is typically analytically intractable.

SMC samplers described in [13] circumvent this difficulty by performing importance sampling on path space $(\mathsf{X}^{T+1}, \mathcal{X}^{T+1})$ using an artificial extended target distribution of the form

$$\mathbb{P}(\mathrm{d}x_{0:T}) = \eta(\mathrm{d}x_T) \prod_{t=1}^{T} L_{t-1}(x_t, \mathrm{d}x_{t-1}),$$

where $L_t \in \mathcal{M}(\mathsf{X})$ for $t \in [0 : T-1]$ is a sequence of auxiliary 'backward' Markov transition kernels. Assuming that we have $L_{t-1} \otimes \gamma_t \ll \gamma_{t-1} \otimes M_t$ with strictly positive and bounded Radon-Nikodym derivative for all $t \in [1 : T]$, the Feynman-Kac path measure representation (2) can be recovered by defining

$$(7) \qquad G_0(x_0) = 1, \quad G_t(x_{t-1}, x_t) = \frac{\mathrm{d}(L_{t-1} \otimes \gamma_t)}{\mathrm{d}(\gamma_{t-1} \otimes M_t)}(x_{t-1}, x_t), \quad t \in [1 : T].$$

Under these potentials, the normalized Feynman-Kac models (4) act as the sequence of bridging distributions $(\eta_t)_{t\in[0:T]}$ in this setting. In annealed importance sampling (AIS) [39] and the sequential sampler proposed in [9], one selects $M_t \in \mathcal{M}(\mathsf{X})$ as a Markov chain Monte Carlo (MCMC) kernel that is η_t-invariant and $L_{t-1} \in \mathcal{M}(\mathsf{X})$ as its time reversal, i.e. $L_{t-1} \otimes \eta_t = \eta_t \otimes M_t$, so the potentials in (7) simplify to

$$(8) \qquad G_0(x_0) = 1, \quad G_t(x_{t-1}) = \frac{\gamma_t(x_{t-1})}{\gamma_{t-1}(x_{t-1})}, \quad t \in [1 : T].$$

3. Twisted models and sequential Monte Carlo.

3.1. *Twisted Feynman-Kac models.* SMC methods can perform poorly when the discrepancy between $\mathbb{P}$ and $\mathbb{Q}$ is large. The right panel of Figure 1 illustrates that this is the case when we employ BPF on the neuroscience application in Section 2.3: the effective sample size (ESS), a common criterion used to assess the quality of a particle approximation [33, p. 34–35], falls below 20% when the data change abruptly. This is because the kernel $M_t(x_{t-1}, \mathrm{d}x_t) = f_t(x_{t-1}, \mathrm{d}x_t)$ used to sample particles at time t does not take the observations into account. Better performance could be obtained using observation-dependent kernels. Indeed, in the context of state space models, the smoothing distribution (5) can be written as $\mathbb{P}(\mathrm{d}x_{0:T}|y_{0:T}) = \mathbb{P}(\mathrm{d}x_0|y_{0:T}) \prod_{t=1}^{T} \mathbb{P}(\mathrm{d}x_t|x_{t-1}, y_{t:T})$ with

$$(9)$$
$$\mathbb{P}(\mathrm{d}x_0|y_{0:T}) = \frac{\nu(\mathrm{d}x_0)\psi_0^*(x_0)}{\nu(\psi_0^*)}, \quad \mathbb{P}(\mathrm{d}x_t|x_{t-1}, y_{t:T}) = \frac{f_t(x_{t-1}, \mathrm{d}x_t)\psi_t^*(x_t)}{f_t(\psi_t^*)(x_{t-1})}, \quad t \in [1 : T],$$

where the kernel $f_t(x_{t-1}, \cdot)$ is *twisted* using the so-called backward information filter [5, 6], given by $\psi_t^*(x_t) = \mathbb{P}(y_{t:T}|x_t)$, for $t \in [0 : T]$.

The backward information filter can also be defined using the backward recursion

$$\begin{aligned}\psi_T^*(x_T) &= g_T(x_T, y_T), \\ \psi_t^*(x_t) &= g_t(x_t, y_t) f_{t+1}(\psi_{t+1}^*)(x_t), \quad t \in [0 : T-1].\end{aligned} \tag{10}$$

We can exploit this to obtain an approximation $\hat{\psi}_t(x_t), t \in [0 : T]$ using regression [42, 47, 23]. We can then sample particles at time t using a proposal $M_t^{\hat{\psi}}(x_{t-1}, \mathrm{d}x_t) \propto f_t(x_{t-1}, \mathrm{d}x_t)\hat{\psi}_t(x_t)$ that approximates $\mathbb{P}(\mathrm{d}x_t | x_{t-1}, y_{t:T})$.

Abstracting the above discussion from state space models to general Feynman–Kac models, where the potential G_t might depend on both x_{t-1} and x_t, motivates the following definitions.

Definition 3.1. *(Admissible policies) A sequence of functions $\psi = (\psi_t)_{t\in[0:T]}$ is an admissible policy if these functions are strictly positive and satisfy $\psi_0 \in \mathcal{B}(\mathsf{X})$, $\psi_t \in \mathcal{B}(\mathsf{X} \times \mathsf{X})$ for all $t \in [1 : T]$. The set of all admissible policies will be denoted as Ψ.*

Definition 3.2. *(Twisted path measures) Given a policy $\psi \in \Psi$ and a path measure $\mathbb{F} \in \mathcal{P}(\mathsf{X}^{T+1})$ of the form $\mathbb{F}(\mathrm{d}x_{0:T}) = \nu(\mathrm{d}x_0) \prod_{t=1}^T K_t(x_{t-1}, \mathrm{d}x_t)$ for some $\nu \in \mathcal{P}(\mathsf{X})$ and $K_t \in \mathcal{M}(\mathsf{X})$ for $t \in [1 : T]$, the ψ-twisted path measure of $\mathbb{F}$ is defined as $\mathbb{F}^\psi(\mathrm{d}x_{0:T}) = \nu^\psi(\mathrm{d}x_0) \prod_{t=1}^T K_t^\psi(x_{t-1}, \mathrm{d}x_t)$ where*

$$\nu^\psi(\mathrm{d}x_0) := \frac{\nu(\mathrm{d}x_0)\psi_0(x_0)}{\nu(\psi_0)}, \quad K_t^\psi(x_{t-1}, \mathrm{d}x_t) := \frac{K_t(x_{t-1}, \mathrm{d}x_t)\psi_t(x_{t-1}, x_t)}{K_t(\psi_t)(x_{t-1})}, \quad t \in [1 : T]. \tag{11}$$

For any policy $\psi \in \Psi$, since $\mathbb{P} \ll \mathbb{Q} \ll \mathbb{Q}^\psi$ by positivity of ψ, we can rewrite the measure $\mathbb{P}$ defined in (2) as

$$\mathbb{P}(\mathrm{d}x_{0:T}) = Z^{-1} G_0^\psi(x_0) \prod_{t=1}^T G_t^\psi(x_{t-1}, x_t)\, \mathbb{Q}^\psi(\mathrm{d}x_{0:T}) \tag{12}$$

where the twisted potentials associated with the twisted path measure $\mathbb{Q}^\psi$ are given by

$$\begin{aligned} G_0^\psi(x_0) &:= \frac{\mu(\psi_0) G_0(x_0) M_1(\psi_1)(x_0)}{\psi_0(x_0)}, \\ G_t^\psi(x_{t-1}, x_t) &:= \frac{G_t(x_{t-1}, x_t) M_{t+1}(\psi_{t+1})(x_t)}{\psi_t(x_{t-1}, x_t)}, \quad t \in [1 : T-1], \\ G_T^\psi(x_{T-1}, x_T) &:= \frac{G_T(x_{T-1}, x_T)}{\psi_T(x_{T-1}, x_T)}. \end{aligned} \tag{13}$$

Note from (12) that $Z = \mathbb{E}_{\mathbb{Q}^\psi}\left[G_0^\psi(X_0) \prod_{t=1}^T G_t^\psi(X_{t-1}, X_t)\right]$ by construction, whereas the triple $\left\{\mu^\psi, (M_t^\psi)_{t\in[1:T]}, (G_t^\psi)_{t\in[0:T]}\right\}$ induces the ψ-twisted

Feynman-Kac models given by

$$\gamma_t^{\psi}(\varphi) = \mathbb{E}_{\mathbb{Q}^{\psi}}\left[\varphi(X_t)G_0^{\psi}(X_0)\prod_{s=1}^{t} G_s^{\psi}(X_{s-1}, X_s)\right], \quad \eta_t^{\psi}(\varphi) = \gamma_t^{\psi}(\varphi)/Z_t^{\psi}, \tag{14}$$

for $\varphi \in \mathcal{B}(\mathsf{X})$, $t \in [0:T]$, where $Z_t^{\psi} := \gamma_t^{\psi}(\mathsf{X})$. For $t \in [0:T-1]$, the marginal distributions of the twisted model are given by

$$\eta_t^{\psi}(\mathrm{d}x_t) = \eta_t(\mathrm{d}x_t)M_{t+1}(\psi_{t+1})(x_t)Z_t/Z_t^{\psi} \tag{15}$$

and do not generally coincide with the ones of the original model (4). However, we stress that they coincide at time T as

$$Z = Z_T^{\psi} = \mu^{\psi}(G_0^{\psi})\prod_{t=1}^{T}\eta_{t-1}^{\psi}(M_t^{\psi}(G_t^{\psi})). \tag{16}$$

To illustrate the effect of twisting models in the static setting of Section 2.4, rewriting the twisted potentials (13) using (15) as

$$G_0^{\psi}(x_0) = \frac{\mathrm{d}\eta_0^{\psi}}{\mathrm{d}\mu^{\psi}}(x_0), \quad G_t^{\psi}(x_{t-1}, x_t) = \frac{\mathrm{d}(L_{t-1}\otimes\gamma_t^{\psi})}{\mathrm{d}(\gamma_{t-1}^{\psi}\otimes M_t^{\psi})}(x_{t-1}, x_t), \quad t \in [1:T],$$

shows that this corresponds to employing the same backward kernels $(L_t)_{t\in[0:T-1]}$, but altered bridging distributions $(\eta_t^{\psi})_{t\in[0:T]}$, initial distribution μ^{ψ} and forward kernels $(M_t^{\psi})_{t\in[1:T]}$.

3.2. *Twisted sequential Monte Carlo.* Consider a policy $\psi \in \Psi$ such that sampling from the initial distribution $\mu^{\psi} \in \mathcal{P}(\mathsf{X})$ and the transition kernels $(M_t^{\psi})_{t\in[1:T]}$ in $\mathcal{M}(\mathsf{X})$ is feasible and evaluation of the twisted potentials (13) is tractable. We can now construct the ψ-twisted SMC method as simply the standard sampling-resampling SMC algorithm applied to ψ-twisted Feynman-Kac models [17]. The resulting algorithm provides approximations of the probability measures $(\eta_t^{\psi})_{t\in[0:T]}$, normalizing constant Z and path measure $\mathbb{P}$, by simulating an interacting particle system of size $N \in \mathbb{N}$. An algorithmic description is detailed in Algorithm 1, where $\mathcal{R}(w_1, \ldots, w_N)$ refers to a resampling operation based on a vector of unnormalized weights $(w_n)_{n\in[1:N]} \in \mathbb{R}_+^N$. For example, this is the categorical distribution on $[1:N]$ with probabilities $(w_n/\sum_{m=1}^{N} w_m)_{n\in[1:N]}$, when multinomial resampling is employed; other lower variance and adaptive resampling schemes can also be considered [19]. All simulations presented in this article employ the systematic resampling scheme.

Given the output of the algorithm, i.e. an array of X-valued position variables $(X_t^n)_{(t,n)\in[0:T]\times[1:N]}$ and an array of $[1:N]$-valued ancestor variables

```
Algorithm 1 $\psi$-twisted sequential Monte Carlo
Input: number of particles $N \in \mathbb{N}$ and policy $\psi \in \Psi$.
  1. At time $t = 0$ and particle $n \in [1:N]$:
     (a) sample $X_0^n \sim \mu^{\psi}$;
     (b) sample ancestor index $A_0^n \sim \mathcal{R}\big(G_0^{\psi}(X_0^1), \ldots, G_0^{\psi}(X_0^N)\big)$.
  2. For time $t \in [1:T]$ and particle $n \in [1:N]$:
     (a) sample $X_t^n \sim M_t^{\psi}(X_{t-1}^{A_{t-1}^n}, \cdot)$;
     (b) sample ancestor index $A_t^n \sim \mathcal{R}\Big(G_t^{\psi}(X_{t-1}^{A_{t-1}^1}, X_t^1), \ldots, G_t^{\psi}(X_{t-1}^{A_{t-1}^N}, X_t^N)\Big)$.
Output: trajectories $(X_t^n)_{(t,n)\in[0:T]\times[1:N]}$ and ancestries $(A_t^n)_{(t,n)\in[0:T]\times[1:N]}$.
```

$(A_t^n)_{(t,n)\in[0:T]\times[1:N]}$, we have a particle approximation of η_t^{ψ} given by the weighted random measure

$$\eta_t^{\psi,N} = \sum_{n=1}^{N} W_t^{\psi,n} \delta_{X_t^n}, \quad W_t^{\psi,n} := \frac{G_t^{\psi}(X_{t-1}^{A_{t-1}^n}, X_t^n)}{\sum_{m=1}^{N} G_t^{\psi}(X_{t-1}^{A_{t-1}^m}, X_t^m)},$$

for $t \in [1:T]$ (similar expression for $t = 0$) and an unbiased estimator of Z resembling the form of (16)

$$Z^{\psi,N} = \left\{ \frac{1}{N} \sum_{n=1}^{N} G_0^{\psi}(X_0^n) \right\} \prod_{t=1}^{T} \left\{ \frac{1}{N} \sum_{n=1}^{N} G_t^{\psi}(X_{t-1}^{A_{t-1}^n}, X_t^n) \right\}. \tag{17}$$

With stored trajectories [26], we can also form a particle approximation of $\mathbb{P}$ with $\mathbb{P}^{\psi,N} = N^{-1} \sum_{n=1}^{N} \delta_{X_{0:T}^n}$, where $X_{0:T}^n$ denotes the path obtained by tracing the ancestral lineage of particle X_T^n, i.e. $X_{0:T}^n := (X_t^{B_t^n})_{t\in[0:T]}$ with $B_T^n := A_T^n$ and $B_t^n := A_t^{B_{t+1}^n}$ for $t \in [0 : T-1]$. Many convergence results are available for these approximations as the size N of the particle system increases [12]. However, depending on the choice of $\psi \in \Psi$, the quality of these approximations may be inadequate for practical values of N; for example, the large variance of (17) often hinders its use within particle MCMC schemes [1] and the approximation $\mathbb{P}^{\psi,N}$ could degenerate quickly with T. The choice of an optimal policy is addressed in the following section.

4. Controlled sequential Monte Carlo.

4.1. *Optimal policies.* Suppose we have an arbitrary current policy $\psi \in \Psi$, initially given by a sequence of constant functions. We would like to twist the path measure $\mathbb{Q}^{\psi} \in \mathcal{P}(\mathsf{X}^{T+1})$ further with a policy $\phi \in \Psi$, so that the resulting twisted path measure $(\mathbb{Q}^{\psi})^{\phi} \in \mathcal{P}(\mathsf{X}^{T+1})$ is in some sense ‘closer’ to the target Feynman-Kac measure $\mathbb{P}$. Note from Definition 3.2 that $(\mathbb{Q}^{\psi})^{\phi} = \mathbb{Q}^{\psi\cdot\phi}$, where $\psi \cdot \phi = (\psi_t \cdot \phi_t)_{t\in[0:T]}$ denotes element-wise multiplication, is simply the $(\psi \cdot \phi)$-twisted path measure of $\mathbb{Q}$. From (13), the corresponding twisted potentials are given by

$$\begin{aligned} G_0^{\psi\cdot\phi}(x_0) &= \frac{\mu^{\psi}(\phi_0)G_0^{\psi}(x_0)M_1^{\psi}(\phi_1)(x_0)}{\phi_0(x_0)}, \\ G_t^{\psi\cdot\phi}(x_{t-1},x_t) &= \frac{G_t^{\psi}(x_{t-1},x_t)M_{t+1}^{\psi}(\phi_{t+1})(x_t)}{\phi_t(x_{t-1},x_t)}, \quad t\in[1:T-1], \\ G_T^{\psi\cdot\phi}(x_{T-1},x_T) &= \frac{G_T^{\psi}(x_{T-1},x_T)}{\phi_T(x_{T-1},x_T)}. \end{aligned} \tag{18}$$

The choice of ϕ that optimally refines an arbitrary policy ψ is given by the following optimality result.

Proposition 4.1. *For any $\psi \in \Psi$, under the policy $\phi^* = (\phi_t^*)_{t\in[0:T]}$ defined recursively as*

$$\begin{aligned} \phi_T^*(x_{T-1},x_T) &= G_T^{\psi}(x_{T-1},x_T), \\ \phi_t^*(x_{t-1},x_t) &= G_t^{\psi}(x_{t-1},x_t)M_{t+1}^{\psi}(\phi_{t+1}^*)(x_t), \quad t\in[1:T-1], \\ \phi_0^*(x_0) &= G_0^{\psi}(x_0)M_1^{\psi}(\phi_1^*)(x_0), \end{aligned} \tag{19}$$

the refined policy $\psi^* := \psi \cdot \phi^*$ satisfies the following properties:

1. *the twisted path measure $\mathbb{Q}^{\psi^*}$ coincides with the Feynman-Kac path measure $\mathbb{P}$;*
2. *the normalized Feynman-Kac model $\eta_t^{\psi^*}$ is the time t-marginal distribution of $\mathbb{P}$ and its normalizing constant $Z_t^{\psi^*} = Z$ for all $t \in [0:T]$;*
3. *the normalizing constant estimator $Z^{\psi^*,N} = Z$ almost surely for any $N \in \mathbb{N}$.*

Moreover, if $G_0^{\psi} \in \mathcal{B}(\mathsf{X})$ and $G_t^{\psi} \in \mathcal{B}(\mathsf{X}\times\mathsf{X})$ for $t \in [1:T]$ then $\phi^* \in \Psi$.

This proposition implies that SMC sampling with the optimal ψ^*-twisted version of Algorithm 1 ensures that the normalizing constant estimator is constant over the entire time horizon, and is equal to the desired normalizing constant. This follows because the SMC weights themselves are almost surely constant; one can see this by substituting the optimal choice (19) into (18). The variance of the SMC weights and the constancy of the normalizing constant estimator can both be used (as described later) as measures of performance evaluation or adaptive tuning.

In a state space context, (19) corresponds to the recursion satisfied by the backward information filter introduced in (10) when $\psi \in \Psi$ are constant functions, i.e. $\mu^{\psi} = \mu = \nu$ and $M_t^{\psi} = M_t = f_t, t \in [1 : T]$; see, e.g., [5, 6].

As it can be shown that ϕ^* is the optimal policy of an associated Kullback–Leibler optimal control problem (Supplementary Material, Section 5), we shall refer to it as the optimal policy w.r.t. $\mathbb{Q}^{\psi}$, although the optimality properties in Proposition 4.1 only identify a policy up to normalization factors. An application of this result gives us the optimal policy $\psi^* = \psi \cdot \phi^*$ w.r.t. $\mathbb{Q}$, which is admissible if the original potentials $(G_t)_{t \in [0:T]}$ are bounded[2].

4.2. *Approximate dynamic programming.* Equation (19) may be viewed as a dynamic programming backward recursion. The optimal policy ϕ^* w.r.t. $\mathbb{Q}^{\psi}$ will give rise to an optimally controlled SMC algorithm via a $\psi^* = (\psi \cdot \phi^*)$-twisted version of Algorithm 1. In all but simple cases, the recursion (19) defining ϕ^* is intractable. We now exploit the connection to optimal control by adapting numerical methods (i.e. approximate dynamic programming) for finite horizon control problems [2, p. 329–331] to our setup. The resulting methodology approximates ϕ^* by combining function approximation and iterating the backward recursion (19).

In the following, we will approximate $V_t^* := -\log \phi_t^*, t \in [0 : T]$ as this corresponds to learning the optimal value functions of the associated control problem. Compared to learning optimal policies directly, as considered in [23], the latter choice is often more desirable as computing in logarithmic scale offers more numerical stability and the minimization is additionally analytically tractable in important scenarios. Moreover, this allows us to relate regression errors to performance properties of the resulting twisted SMC method in the next section.

Let $(X_t^n)_{(t,n) \in [0:T] \times [1:N]}$ and $(A_t^n)_{(t,n) \in [0:T] \times [1:N]}$ denote the trajectories and ancestries, obtained by running a ψ-twisted SMC. At time T, to approximate $V_T^* := -\log \phi_T^* = -\log G_T^{\psi}$, we consider the least squares problem

$$\hat{V}_T = \arg \min_{\varphi \in \mathsf{F}_T} \sum_{n=1}^{N} \left(\varphi(X_{T-1}^{A_{T-1}^n}, X_T^n) + \log G_T^{\psi}(X_{T-1}^{A_{T-1}^n}, X_T^n) \right)^2, \tag{20}$$

where F_T is a pre-specified function class. An approximation of ϕ_T^* can then be obtained by taking $\hat{\phi}_T := \exp(-\hat{V}_T)$. To iterate the backward recursion $\phi_{T-1}^* = G_{T-1}^{\psi} M_T^{\psi}(\phi_T^*)$, we set $\xi_{T-1} := G_{T-1}^{\psi} M_T^{\psi}(\hat{\phi}_T)$ by plugging in the

[2] For ease of presentation, the notion of admissibility adopted in Definition 3.1 is more stringent than necessary as non-admissible optimal policies can still lead to valid optimal SMC methods.

approximation $\hat{\phi}_T \approx \phi_T^*$ and consider the least squares problem

(21)
$$\hat{V}_{T-1} = \arg\min_{\varphi \in \mathsf{F}_{T-1}} \sum_{n=1}^{N} \left(\varphi(X_{T-2}^{A_{T-2}^n}, X_{T-1}^n) + \log \xi_{T-1}(X_{T-2}^{A_{T-2}^n}, X_{T-1}^n) \right)^2,$$

where F_{T-1} is another function class to be specified. As before, we form the approximation $\hat{\phi}_{T-1} := \exp(-\hat{V}_{T-1})$. Continuing in this manner until time 0 gives us an approximation $\hat{\phi} = (\hat{\phi}_t)_{t\in[0:T]}$ of ϕ^*. We shall refer to this procedure as the approximate dynamic programming algorithm and provide a detailed description in Algorithm 2.

Restricting the function classes $(\mathsf{F}_t)_{t\in[0:T]}$ to contain only lower bounded functions ensures that the estimated policy $\hat{\phi}$ lies in Ψ, hence the refined policy $\psi \cdot \hat{\phi}$ also lies in Ψ. We defer a detailed discussion on the choice of function classes and shall assume for now this is such that under the refined policy $\psi \cdot \hat{\phi} \in \Psi$, sampling from initial distribution $\mu^{\psi\cdot\hat{\phi}} \in \mathcal{P}(\mathsf{X})$, transition kernels $(M_t^{\psi\cdot\hat{\phi}})_{t\in[1:T]}$ in $\mathcal{M}(\mathsf{X})$ is feasible and evaluation of twisted potentials $(G_t^{\psi\cdot\hat{\phi}})_{t\in[0:T]}$ is tractable.

As the size of the particle system N increases, it is natural to expect $\hat{\phi}$ to converge (in a suitable sense) to a policy defined by an idealized algorithm that performs the least squares approximations in (20)-(21) using L^2-projections. This will be established in Section 5.2 for a common choice of function class. It follows that the quality of $\hat{\phi}$, as an approximation of the optimal policy ϕ^*, will depend on the number of particles N and the 'richness' of chosen function classes $(\mathsf{F}_t)_{t\in[0:T]}$. A more precise characterization of the ADP error in terms of approximate projection errors will be given in Section 5.1.

4.3. *Policy refinement.* If the recursion (19) could be performed exactly, no policy refinement would be necessary as we would initialize ψ as a policy of constant functions and obtain the optimal policy $\psi^* = \phi^*$ w.r.t. $\mathbb{Q}$. This will not be possible in practical scenarios. Given a current policy $\psi \in \Psi$, we employ ADP and obtain an approximation $\hat{\phi}$ of the optimal policy ϕ^* w.r.t. $\mathbb{Q}^\psi$. The residuals from the corresponding least squares approximations (20)-(21) are given by

$$\varepsilon_T^\psi := \log \hat{\phi}_T - \log G_T^\psi, \quad \varepsilon_t^\psi := \log \hat{\phi}_t - \log G_t^\psi - \log M_{t+1}^\psi(\hat{\phi}_{t+1}), \quad t \in [0:T-1].$$

From (18), these residuals are related to twisted potentials of the refined policy $\psi \cdot \hat{\phi}$ via

$$\log G_0^{\psi\cdot\hat{\phi}} = \log \mu^\psi(\hat{\phi}_0) - \varepsilon_0^\psi, \quad \log G_t^{\psi\cdot\hat{\phi}} = -\varepsilon_t^\psi, \quad t \in [1:T]. \tag{22}$$

Algorithm 2 Approximate dynamic programming

```
Input: policy $\psi \in \Psi$ and output of $\psi$-twisted SMC method (Algorithm 1).
    1. Initialization: set $M_{T+1}^{\psi}(\hat{\phi}_{T+1})(X_T^n) = 1$ for $n \in [1:N]$.
    2. For time $t \in [1:T]$:
        (a) set $\xi_t(X_{t-1}^{A_{t-1}^n}, X_t^n) = G_t^{\psi}(X_{t-1}^{A_{t-1}^n}, X_t^n) M_{t+1}^{\psi}(\hat{\phi}_{t+1})(X_t^n)$ for $n \in [1:N]$;
        (b) fit $\hat{V}_t = \arg\min_{\varphi \in \mathsf{F}_t} \sum_{n=1}^N \left(\varphi(X_{t-1}^{A_{t-1}^n}, X_t^n) + \log \xi_t(X_{t-1}^{A_{t-1}^n}, X_t^n)\right)^2$;
        (c) set $\hat{\phi}_t = \exp(-\hat{V}_t)$.
    3. At time $t = 0$:
        (a) set $\xi_0(X_0^n) = G_0^{\psi}(X_0^n) M_1^{\psi}(\hat{\phi}_1)(X_0^n)$ for $n \in [1:N]$;
        (b) fit $\hat{V}_0 = \arg\min_{\varphi \in \mathsf{F}_0} \sum_{n=1}^N (\varphi(X_0^n) + \log \xi_0(X_0^n))^2$;
        (c) set $\hat{\phi}_0 = \exp(-\hat{V}_0)$.
Output: policy $\hat{\phi} = (\hat{\phi}_t)_{t \in [0:T]} \in \Psi$.
```

Using this relation, we can monitor the efficiency of ADP via the variance of SMC weights in the $(\psi \cdot \hat{\phi})$-twisted version of Algorithm 1. It follows from (22) that the Kullback–Leibler divergence from $(\mathbb{Q}^{\psi})^{\hat{\phi}}$ to $\mathbb{P}$ is at most

$$|\log \mu^{\psi}(\hat{\phi}_0) - \log Z| + \|\varepsilon_0^{\psi}\|_{L^1(\mathbb{P}_0)} + \sum_{t=1}^{T} \|\varepsilon_t^{\psi}\|_{L^1(\mathbb{P}_{t-1,t})} \tag{23}$$

where $\| \cdot \|_{L^1}$ denotes the L^1-norm w.r.t. the one time $(\mathbb{P}_t)_{t \in [0:T]}$ and two time $(\mathbb{P}_{t,s})_{(t,s) \in [0:T-1] \times [t+1:T]}$ marginal distributions of $\mathbb{P}$. This shows how performance of $(\psi \cdot \hat{\phi})$-twisted SMC depends on the quality of the ADP approximation of the optimal policy w.r.t. $\mathbb{Q}^{\psi}$.

If we further twist the path measure $\mathbb{Q}^{\psi \cdot \hat{\phi}}$ by a policy $\hat{\zeta} \in \Psi$, the subsequent ADP procedure defining $\hat{\zeta}$ would consider the least squares problem

$$-\log \hat{\zeta}_T := \arg \min_{\varphi \in \mathsf{F}_T} \sum_{n=1}^{N} \left(\varphi(X_{T-1}^{A_{T-1}^n}, X_T^n) - \varepsilon_T^{\psi}(X_{T-1}^{A_{T-1}^n}, X_T^n)\right)^2, \tag{24}$$

at time T, and for $t \in [1:T-1]$

(25)

$$-\log \hat{\zeta}_t := \arg \min_{\varphi \in \mathsf{F}_t} \sum_{n=1}^{N} \left(\varphi(X_{t-1}^{A_{t-1}^n}, X_t^n) - (\varepsilon_t^{\psi} - \log M_{t+1}^{\psi \cdot \hat{\phi}}(\hat{\zeta}_{t+1}))(X_{t-1}^{A_{t-1}^n}, X_t^n)\right)^2,$$

where $(X_t^n)_{(t,n) \in [0:T] \times [1:N]}$ and $(A_t^n)_{(t,n) \in [0:T] \times [1:N]}$ denote the output of $(\psi \cdot \hat{\phi})$-twisted SMC. Equations (24)-(25) reveal that it might be beneficial to have an iterative scheme to refine policies as this allows repeated least squares fitting of residuals, in the spirit of L^2-boosting methods [7]. Moreover, it follows from (22)-(23) that errors would not accumulate over iterations. The resulting iterative algorithm, summarized in Algorithm 3, will be referred to as the controlled SMC method (cSMC). The overall computational complexity is of order $I \times T \times (NC_{\text{sample}}C_{\text{evaluate}} + C_{\text{approx}})$, where $C_{\text{sample}}(d)$ is the cost of sampling from each initial distribution or transition kernel in (11), $C_{\text{evaluate}}(d)$ is the cost of evaluating each twisted potential in (13), and $C_{\text{approx}}(N, d)$ is the cost of each least squares approximation[3]. The first iteration of the algorithm would coincide with that of [23] for state space models, if regressions were computed on the natural scale; subsequent iterations differ in policy refinement strategy. To maintain a coherent terminology, we will refer to the standard SMC method and ψ^*-twisted SMC method as the uncontrolled and optimally controlled SMC methods respectively. From the output of the algorithm, we can estimate $\mathbb{P}$ with $\mathbb{P}^{\psi^{(I)},N}$ and its normalizing constant Z with $Z^{\psi^{(I)},N}$ as explained in Section 3.2.

It is possible to consider performance monitoring and adaptive tuning for the SMC sampling and the iterative policy refinement. Recalling the relationship between residuals and twisted potentials (22), we note that monitoring the variance of the SMC weights, using for example the ESS, allows us to evaluate the effectiveness of the ADP algorithm and to identify time instances when the approximation is inadequate. We can also deduce if the estimated policy is far from optimal by comparing the behaviour of the normalizing constant estimates across time with those when the optimal policy is applied, as detailed in Proposition 4.1. When implementing Algorithm 3, the number of iterations $I \in \mathbb{N}$ can be pre-determined using preliminary runs or chosen adaptively until successive policy refinement yields no improvement in performance. For example, one can iterate policy refinement until the ESS across time achieves a desired minimum threshold and/or there is no improvement in ESS across iterations; see Section 6.2 for a numerical illustration. In Section 5.3, under appropriate regularity assumptions, we show that this iterative scheme generates a geometrically ergodic Markov chain on Ψ and characterize its unique invariant distribution. For all numerical examples considered in this article, we observe that convergence happens very rapidly, so only a small number of iterations is necessary.

[3] The dependence of these costs on their arguments will depend on the specific problem of interest and the choice of function classes. As an example, C_{approx} will depend linearly on N in the case of linear least squares regression.

```
Algorithm 3 Controlled sequential Monte Carlo
Input: number of particles $N \in \mathbb{N}$ and iterations $I \in \mathbb{N}$.
  1. Initialization: set $\psi^{(0)}$ as constant one functions.
  2. For iterations $i \in [0 : I - 1]$:
      (a) run $\psi^{(i)}$-twisted SMC method (Algorithm 1);
      (b) perform ADP (Algorithm 2) with SMC output to obtain policy $\hat{\phi}^{(i+1)}$;
      (c) construct refined policy $\psi^{(i+1)} = \psi^{(i)} \cdot \hat{\phi}^{(i+1)}$.
  3. At iteration $i = I$:
      (a) run $\psi^{(I)}$-twisted SMC method (Algorithm 1).
Output: trajectories $(X_t^n)_{(t,n)\in[0:T]\times[1:N]}$ and ancestries $(A_t^n)_{(t,n)\in[0:T]\times[1:N]}$ from
$\psi^{(I)}$-twisted SMC method.
```

4.4. *Illustration on neuroscience model.* We now apply our proposed methodology on the neuroscience model introduced in Section 2.3. We take BPF as the uncontrolled SMC method, i.e. we set $\mu = \nu$ and $M_t = f$ for $t \in [1 : T]$. Under the following choice of function classes

$$\mathsf{F}_t = \left\{ \varphi(x_t) = a_t x_t^2 + b_t x_t + c_t : (a_t, b_t, c_t) \in \mathbb{R}^3 \right\}, \quad t \in [0 : T], \tag{26}$$

the policy $\psi^{(i)} = (\psi_t^{(i)})_{t\in[0:T]}$ at iteration $i \in [1 : I]$ of Algorithm 3 has the form

$$\psi_t^{(i)}(x_t) = \exp\left(-a_t^{(i)} x_t^2 - b_t^{(i)} x_t - c_t^{(i)} \right), \quad t \in [0 : T],$$

where $a_t^{(i)} := \sum_{j=1}^{i} a_t^j, b_t^{(i)} := \sum_{j=1}^{i} b_t^j, c_t^{(i)} := \sum_{j=1}^{i} c_t^j$ for $t \in [0 : T]$ and $(a_t^{j+1}, b_t^{j+1}, c_t^{j+1})_{t\in[0:T]}$ denotes the coefficients estimated using linear least squares at iteration $j \in [0 : I - 1]$. Exact expressions of the twisted initial distribution, transition kernels and potentials, required to implement cSMC are given in Section 9.3 of Supplementary Material.

Figure 2 illustrates that the parameterization (26) provides a good approximation of the optimal policy. We note (left panel) the significant improvement of ESS across iterations, and see how this may be used as a measure of performance evaluation. In the right panel, we can also deduce how far the estimated policy is from optimality by observing the behaviour of normalizing constant estimates as discussed previously. Indeed, while the uncontrolled SMC approximates $Z_t = Z_t^{\psi^{(0)}} = p(y_{0:t})$, the controlled SMC scheme approximates $Z_t^{\psi^*} = p(y_{0:T})$ for all $t \in [0 : T]$.

Moreover, we see from the left panel of Figure 3 that the improvement in performance is reflected in the estimated policy's ability to capture abrupt changes in the data. This plot also demonstrates the effect of policy refine-

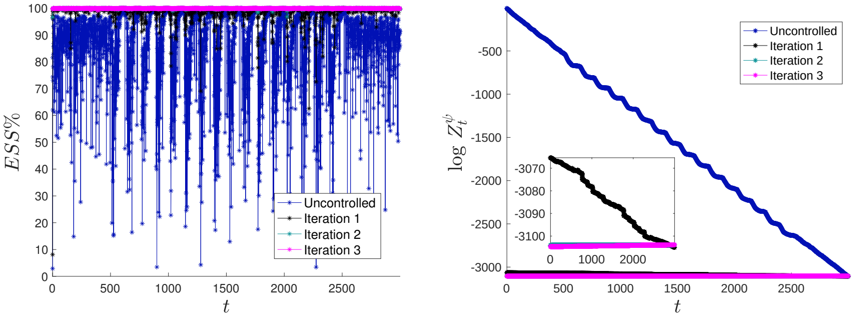

FIG 2. *Comparison of uncontrolled and controlled SMC methods in terms of effective sample size (*left*) and normalizing constant estimation (*right*) on the neuroscience model introduced in Section 2.3. The parameters are $\alpha = 0.99, \sigma^2 = 0.11$ and the algorithmic settings of cSMC are $I = 3, N = 128$.*

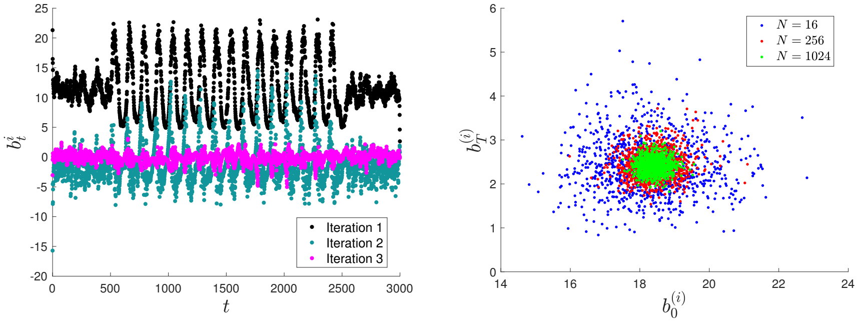

FIG 3. *Applying controlled SMC method on the neuroscience model introduced in Section 2.3: coefficients estimated at each iteration with $N = 128$ particles (*left*) and invariant distribution of coefficients with various number of particles (*right*).*

ment: by refitting residuals from previous iterations (24)-(25), the magnitude of estimated coefficients decreases with iterations as the residuals can be adequately approximated by simpler functions. Lastly, in the right panel of Figure 3, we illustrate the invariant distribution of coefficients estimated by cSMC using a long run of $I = 1000$ iterations, with the first 10 iterations discarded as burn-in. These plots show that the distribution concentrates as the size of the particle system N increases, which is consistent with our findings presented in Section 5.3.

5. Analysis. This section considers several theoretical aspects of the proposed methodology, and may be skipped without affecting the methodological developments thus far and the experimental results that follow.

5.1. *Policy learning.* The goal of this section is to characterize the error of ADP (Algorithm 2) for learning the optimal policy (19) in terms of regression errors. We first define, for any $\mu \in \mathcal{P}(\mathsf{E})$, the set $\mathfrak{L}^2(\mu)$ of $\mathcal{E}$-measurable functions $\varphi : \mathsf{E} \to \mathbb{R}^d$ such that $\|\varphi\|_{L^2(\mu)} := (\int_{\mathsf{E}} |\varphi(x)|^2 \mu(\mathrm{d}x))^{1/2} < \infty$, and $L^2(\mu)$ as the set of equivalence classes of functions in $\mathfrak{L}^2(\mu)$ that agree μ-almost everywhere. To simplify notation, we introduce some operators.

Definition 5.1. *(Bellman operators) Given $\psi \in \Psi$ such that $G_0^\psi \in \mathcal{B}(\mathsf{X})$ and $G_t^\psi \in \mathcal{B}(\mathsf{X}\times\mathsf{X})$ for $t \in [1:T]$, we define the operators $Q_t^\psi : L^2(\nu_{t+1}^\psi) \to L^2(\nu_t^\psi)$ for $t \in [0:T-1]$ as*

$$Q_0^\psi(\varphi)(x) = G_0^\psi(x) M_1^\psi(\varphi)(x), \quad \varphi \in L^2(\nu_1^\psi),$$
$$Q_t^\psi(\varphi)(x,y) = G_t^\psi(x,y) M_{t+1}^\psi(\varphi)(y), \quad \varphi \in L^2(\nu_{t+1}^\psi),$$

where $\nu_0^\psi := \mu^\psi \in \mathcal{P}(\mathsf{X})$ and $\nu_t^\psi := \eta_{t-1}^\psi \otimes M_t^\psi \in \mathcal{P}(\mathsf{X}\times\mathsf{X})$ for $t \in [1:T]$. For notational convenience define $Q_T^\psi(\varphi)(x,y) = G_T^\psi(x,y)$ for any φ (take ν_{T+1}^ψ as an arbitrary element in $\mathcal{P}(\mathsf{X}\times\mathsf{X})$).

Although these operators are typically used to define unnormalized predictive Feynman-Kac models [12, Proposition 2.5.1], we shall adopt terminology from control literature and refer to them as Bellman operators. It can be shown that these Bellman operators are well-defined and are in fact bounded linear operators – see Proposition 5.2. In this notation, we can rewrite (19) more succinctly as

$$\phi_T^* = G_T^\psi, \quad \phi_t^* = Q_t^\psi \phi_{t+1}^*, \quad t \in [0:T-1]. \tag{27}$$

To understand how regression errors propagate in time, for $-1 \le s \le t \le T$, we define the Feynman-Kac semigroup $Q_{s,t}^\psi : L^2(\nu_{t+1}^\psi) \to L^2(\nu_{s+1}^\psi)$ associated to a policy $\psi \in \Psi$ as

$$Q_{s,t}^\psi(\varphi) = \begin{cases} \varphi, & s = t, \\ Q_{s+1}^\psi \circ \cdots \circ Q_t^\psi(\varphi), & s < t, \end{cases} \tag{28}$$

for $\varphi \in L^2(\nu_{t+1}^\psi)$. To describe regression steps taken to approximate the intractable recursion (27), we introduce the following operations.

Definition 5.2. *(Logarithmic projection) On a measurable space $(\mathsf{E}, \mathcal{E})$, let $\nu \in \mathcal{P}(\mathsf{E})$, $\xi : \mathsf{E} \to \mathbb{R}_+$ be a $\mathcal{E}$-measurable function such that $-\log\xi \in L^2(\nu) \cap \mathcal{L}(\mathsf{E})$, and $\mathsf{F} \subset \mathcal{L}(\mathsf{E})$ be a closed linear subspace of $L^2(\nu)$. We define the (F, ν)-projection operator $P^\nu : \mathcal{B}(\mathsf{E}) \to \mathcal{B}(\mathsf{E})$ as*

$$P^\nu \xi = \exp\Big(- \arg\min_{\varphi \in \mathsf{F}} \|\varphi + \log\xi\|_{L^2(\nu)}^2 \Big). \tag{29}$$

The projection theorem gives existence of a unique $P^{\nu}\xi$. We have chosen to define $-\log P^{\nu}\xi$ as the orthogonal projection of $-\log\xi$ onto F, as this corresponds to learning the optimal value functions of the associated control problem. Since projections are typically intractable, a practical implementation will involve a Monte Carlo approximation of (29).

Definition 5.3. *(Approximate projection) Following notation in Definition 5.2, given a consistent approximation ν^N of ν, i.e. $\nu^N(\varphi) \to \nu(\varphi)$ almost surely for any $\varphi \in L^1(\nu)$, we define the approximate (F,ν)-projection operator $P^{\nu,N} : \mathcal{B}(\mathsf{E}) \to \mathcal{B}(\mathsf{E})$ as the (F,ν^N)-projection operator. We additionally assume that the function class F is such that $P^{\nu,N}\xi$ is a random function for all $\xi \in \mathcal{B}(\mathsf{E})$.*

If $\psi \in \Psi$ is the current policy, we use the output of ψ-twisted SMC (Algorithm 1) to learn the optimal policy ϕ^*, through the empirical measures

$$(30)\qquad \nu_0^{\psi,N} = \frac{1}{N}\sum_{n=1}^{N}\delta_{X_0^n}, \quad \nu_t^{\psi,N} = \frac{1}{N}\sum_{n=1}^{N}\delta_{\left(X_{t-1}^{A_{t-1}^n}, X_t^n\right)}, \quad t \in [1:T],$$

which are consistent approximations of $(\nu_t^{\psi})_{t\in[0:T]}$ [12], defined in Definition 5.1. Given pre-specified closed and linear function classes $\mathsf{F}_0 \subset L^2(\nu_0^{\psi}) \cap \mathcal{L}(\mathsf{X})$, $\mathsf{F}_t \subset L^2(\nu_t^{\psi}) \cap \mathcal{L}(\mathsf{X}^2)$, $t \in [1:T]$, we denote the approximate $(\mathsf{F}_t,\nu_t^{\psi})$-projection operator by $P_t^{\psi,N}$ for $t \in [0:T]$. We can now write our ADP algorithm detailed in Algorithm 2 succinctly as

$$(31)\qquad \hat{\phi}_T = P_T^{\psi,N} G_T^{\psi}, \quad \hat{\phi}_t = P_t^{\psi,N} Q_t^{\psi}\hat{\phi}_{t+1}, \quad t \in [0:T-1].$$

The following result characterizes how well (31) can approximate (27).

Proposition 5.1. *Suppose that we have a policy $\psi \in \Psi$, number of particles N and closed, linear function classes $\mathsf{F}_0 \subset L^2(\nu_0^{\psi}) \cap \mathcal{L}(\mathsf{X})$, $\mathsf{F}_t \subset L^2(\nu_t^{\psi}) \cap \mathcal{L}(\mathsf{X}^2)$, $t \in [1:T]$ such that:*

*[**A1**] the Feynman-Kac semigroup defined in (28) satisfies*

$$(32)\qquad \|Q_{s,t}^{\psi}(\varphi)\|_{L^2(\nu_{s+1}^{\psi})} \le C_{s,t}^{\psi}\|\varphi\|_{L^2(\nu_{t+1}^{\psi})}, \quad -1 \le s < t \le T-1,$$

for some $C_{s,t}^{\psi} \in [0,\infty)$ and all $\varphi \in L^2(\nu_{t+1}^{\psi})$;

*[**A2**] the approximate $(\mathsf{F}_t,\nu_t^{\psi})$-projection operator satisfies*

$$\sup_{\xi\in\mathsf{S}_t^{\psi}} \mathbb{E}^{\psi,N}\|P_t^{\psi,N}\xi - \xi\|_{L^2(\nu_t^{\psi})} \le e_t^{\psi,N} < \infty$$

where $\mathsf{S}_t^{\psi} := \{Q_t^{\psi}\exp(-\varphi) : \varphi \in \mathsf{F}_{\mathsf{t+1}}\}$ for $t \in [0:T-1]$ and $\mathsf{S}_T^{\psi} := \{G_T^{\psi}\}$.

Then the policy $\hat{\phi} \in \Psi$ generated by ADP algorithm (31) satisfies

$$\mathbb{E}^{\psi,N}\|\hat{\phi}_t - \phi_t^*\|_{L^2(\nu_t^\psi)} \leq \sum_{u=t}^{T} C_{t-1,u-1}^{\psi} e_u^{\psi,N}, \quad t \in [0:T], \tag{33}$$

where $C_{t-1,t-1}^{\psi} = 1$ and $\mathbb{E}^{\psi,N}$ denotes expectation w.r.t. the law of the ψ-twisted SMC method (Algorithm 1).

Equation (33) reveals how function approximation errors propagate backwards in time. If the choice of function class is 'rich' enough and the number of particles is sufficiently large, then these errors can be kept small and ADP provides a good approximation of the optimal policy. If the number of particles is taken to infinity, the projection errors are driven solely by the choice of function class (as the latter dictates $e_t^{\psi,\infty}$). Moreover, observe that these errors are also modulated by stability constants of the Feynman-Kac semigroup in (32). We now establish the inequality (32). For $\varphi \in \mathcal{B}(\mathsf{E})$, we write its supremum norm as $\|\varphi\|_\infty = \sup_{x\in\mathsf{E}} |\varphi(x)|$.

Proposition 5.2. *Suppose $\psi \in \Psi$ is such that $G_0^\psi \in \mathcal{B}(\mathsf{X})$, $G_t^\psi \in \mathcal{B}(\mathsf{X} \times \mathsf{X})$ for $t \in [1:T]$ and let $\delta := \max_{t\in[0:T]} \|G_t^\psi\|_\infty$ (and $Z_{-1}^\psi := 1$). Then (32) holds with*

$$\begin{aligned} C_{s,t}^{\psi} &= \left(Z_t^\psi / Z_s^\psi \prod_{u=s+1}^{t} \|G_u^\psi\|_\infty\right)^{1/2} \\ &\leq \left(Z_t^\psi / Z_s^\psi\right)^{1/2} \delta^{(t-s)/2}, \quad -1 \leq s < t \leq T-1. \end{aligned} \tag{34}$$

For the case $G_t^\psi(x,y) = G_t^\psi(y)$ for all $x, y \in \mathsf{X}$ and $t \in [1:T]$, if we assume additionally for each $t \in [1:T]$ that:
[A3] *there exist $\sigma_t^\psi \in \mathcal{P}(\mathsf{X})$ and $\kappa_t^\psi \in (0,\infty)$ such that for all $x \in \mathsf{X}$ we have*

$$M_t^\psi(x, \mathrm{d}y) \leq \kappa_t^\psi \sigma_t^\psi(\mathrm{d}y). \tag{35}$$

Then inequality (32) holds with

$$C_{s,t}^{\psi} = \left[\kappa_{s+2}^\psi \|G_{s+1}^\psi\|_\infty \, \sigma_{s+2}^\psi\big(Q_{s+1,t}^\psi(1)\big) \frac{Z_t^\psi}{Z_s^\psi}\right]^{1/2}, \quad -1 \leq s < t \leq T-1. \tag{36}$$

The assumption of bounded potentials is typical in similar analyses of ADP errors [21, Section 8.3.3] and stability of SMC methods [12]. The second part of Proposition 5.2 shows that it is possible to exploit regularity properties of the transition kernels to obtain better constants $C_{s,t}^\psi$. Conditions such as (35) are common in the filtering literature, see for example [14, Eq. (9)] and [12, ch. 4].

5.2. *Limit theorems.* We now study the asymptotic behaviour of the ADP algorithm (31), with a current policy $\psi \in \Psi$, as the size of the particle system N grows to infinity. For a common choice of function class, we will establish convergence to a policy $\tilde{\phi} = (\tilde{\phi}_t)_{t\in[0:T]}$, defined by the idealized algorithm

$$\tilde{\phi}_T = P_T^{\psi} G_T^{\psi}, \quad \tilde{\phi}_t = P_t^{\psi} Q_t^{\psi} \tilde{\phi}_{t+1}, \quad t \in [0 : T-1], \tag{37}$$

where P_t^{ψ} denotes the $(\mathsf{F}_t, \nu_t^{\psi})$-projection operator for $t \in [0 : T]$. In particular, we consider logarithmic projections that are defined by linear least squares approximations; this corresponds to function classes of the form

$$\mathsf{F}_t := \left\{ \Phi_t^T \beta : \beta \in \mathbb{R}^M \right\}, \quad t \in [0 : T], \tag{38}$$

where $\Phi_0 \subset L^2(\nu_0^{\psi}) \cap \mathcal{L}(\mathsf{X})$, $\Phi_t \subset L^2(\nu_t^{\psi}) \cap \mathcal{L}(\mathsf{X}^2)$, $t \in [1 : T]$ are vectors of $M \in \mathbb{N}$ pre-specified basis functions. We will treat M as fixed in our analysis and refer to [21, Theorem 8.2.4] for results on how M should increase with N to balance the tradeoff between enriching (38) and the need for more samples to achieve the same estimation precision. We denote by $\tilde{\phi} := (\tilde{\phi}_t)_{t\in[0:T]}$ the policy generated by the idealized algorithm (37) where $\tilde{\phi}_t := \exp(-\Phi_t^T \beta_t^{\psi})$, β_t^{ψ} being the corresponding least squares estimate. This result builds upon the central limit theorem for particle methods established in [10, 12, 30].

THEOREM 5.1. *Consider the ADP algorithm (31) with current policy $\psi \in \Psi$, under linear least squares approximations (38). Under appropriate regularity conditions, for all $x \in \mathsf{X}^{2T+1}$, the estimated policy $\hat{\phi}(x)$ converges in probability to the policy $\tilde{\phi}(x)$ as $N \to \infty$. Moreover, for all $x \in \mathsf{X}^{2T+1}$,*

$$\sqrt{N} \left(\hat{\phi}(x) - \tilde{\phi}(x) \right) \xrightarrow{\mathrm{d}} \mathcal{N} \left(0_{(T+1)}, \Omega^{\psi}(x) \right) \tag{39}$$

for some $\Omega^{\psi} : \mathsf{X}^{2T+1} \to \mathbb{R}^{(T+1)\times(T+1)}$, where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution and $0_p = (0, \ldots, 0)^T \in \mathbb{R}^p$ is the zero vector.

A precise mathematical statement of this result and its proof are given in Section 3 of Supplementary Material. Note that the proof relies on a technical central limit theorem on path space that can be deduced in the case of multinomial resampling from [12, Theorem 9.7.1]. The exact form of Ω^{ψ} reveals how errors correlate over time and suggests that we may expect the variance of the estimated policy to be larger at earlier times, due to the inherent backward nature of the ADP approximation.

5.3. *Iterated approximate dynamic programming.* We provide here a theoretical framework to understand the qualitative behaviour of policy $\psi^{(I)}$, estimated by Algorithm 3, as the number of iterations I grows to infinity. This offers a novel perspective of iterative algorithms for finite horizon optimal control problems and may be of general interest.

To do so, we require the set of all admissible policies to be a complete separable metric space. This follows if we impose that X is a compact metric space and work with $\Psi := \mathcal{C}(\mathsf{X}) \prod_{t=1}^{T} \mathcal{C}(\mathsf{X} \times \mathsf{X})$, equipped with the metric $\rho(\varphi, \xi) := \sum_{t=0}^{T} \|\varphi_t - \xi_t\|_\infty$ for $\varphi = (\varphi_t)_{t\in[0:T]}, \xi = (\xi_t)_{t\in[0:T]} \in \Psi$; non-compact state spaces can also be accommodated with a judicious choice of metric (see e.g. [4, p. 380]).

We begin by writing the iterative algorithm with $N \in \mathbb{N}$ particles as an iterated random function $F^N : \mathsf{U} \times \Psi \to \Psi$, defined by $F_U^N(\psi) = \psi \cdot \hat{\phi}$, where $\hat{\phi}$ is the output of ADP algorithm (31) and $U \in \mathsf{U}$ encodes all uniform random variables needed to simulate a ψ-twisted SMC method (Algorithm 1). As the uniform variables $(U^{(I)})_{I\in\mathbb{N}}$ used at every iteration are independent and identically distributed, iterating F^N defines a Markov chain $(\psi^{(I)})_{I\in\mathbb{N}}$ on Ψ. We will write $\mathbb{E}$ to denote expectation w.r.t. the law of $(U^{(I)})_{I\in\mathbb{N}}$ and $\pi^{(I)} \in \mathcal{P}(\Psi)$ to denote the law of $\psi^{(I)}$. Similarly, we denote the iterative scheme with exact projections by $F : \Psi \to \Psi$, defined as $F(\psi) = \psi \cdot \tilde{\phi}$, where $\tilde{\phi}$ is the output of the idealized ADP algorithm (37). We denote by $\varphi^* \in \Psi$ a fixed point (if it exists) of F, i.e. $F(\varphi^*) = \varphi^*$. The following is based on results developed in [15].

Theorem 5.2. *Assume that the iterated random function F^N satisfies:*

[A4] $\mathbb{E}\left[\rho(F_U^N(\varphi_0), \varphi_0)\right] < \infty$ *for some* $\varphi_0 \in \Psi$,

[A5] *there exists a measurable function* $L^N : \mathsf{U} \to \mathbb{R}_+$ *with* $\mathbb{E}\left[L_U^N\right] < \alpha$ *for some* $\alpha \in [0, 1)$ *such that* $\rho(F_U^N(\varphi), F_U^N(\xi)) \le L_U^N \rho(\varphi, \xi)$ *for all* $\varphi, \xi \in \Psi$.

Then the Ψ-valued Markov chain $(\psi^{(I)})_{I\in\mathbb{N}}$ generated by Algorithm 3 admits a unique invariant distribution $\pi \in \mathcal{P}(\Psi)$ and

$$\varrho(\pi^{(I)}, \pi) \le C(\psi^{(0)}) r^I, \quad I \in \mathbb{N}, \tag{40}$$

for some $C : \Psi \to \mathbb{R}_+$ and $r \in (0, 1)$, where ϱ denotes the Prohorov metric on $\mathcal{P}(\Psi)$ induced by the metric ρ. If we suppose in addition that:

[A6] *for each $\psi \in \Psi$, $\rho(F_U^N(\psi), F(\psi)) \le N^{-1/2} E_U^{\psi,N}$ where $(E_U^{\psi,N})_{N\in\mathbb{N}}$ is a uniformly integrable sequence of non-negative random variables with finite mean that converges in distribution to a limiting distribution with support on $\mathbb{R}_+$, then we also have that*

$$\mathbb{E}_\pi\left[\rho(\psi, \varphi^*)\right] \le N^{-1/2} \mathbb{E}\left[E_U^{\varphi^*,N}\right] (1 - \alpha)^{-1} \tag{41}$$

where φ^* is a fixed point of F and $\mathbb{E}_\pi$ denotes expectation w.r.t. $\psi \sim \pi$.

Assumption A5 requires the ADP procedure to be sufficiently regular: i.e. for two policies $\varphi, \xi \in \Psi$ that are close, given the same uniform random variables U to simulate a φ-twisted and ξ-twisted SMC method, the policies $\hat{\varphi}$ (w.r.t. $\mathbb{Q}^{\varphi}$) and $\hat{\xi}$ (w.r.t. $\mathbb{Q}^{\xi}$) estimated by (31) should also be close enough to keep the Lipschitz constant L_U^N small. Assumption A6 is necessary to quantify the Monte Carlo error involved when employing approximate projections and can be deduced for example using the central limit theorem in (39). See Section 4 of the Supplementary Material for a discussion on when and why contraction occurs, and a simple example where Assumptions A4-A6 are verified.

The first part of Theorem 5.2, which establishes existence of a unique invariant distribution and geometric convergence to the latter, follows from standard theory on iterated random functions; see, e.g., [15]. The second conclusion of Theorem 5.2, which provides a characterization of the limiting distribution, is to the best of our knowledge novel. The fixed point φ^* can be interpreted as a policy for which subsequent refinement using exact (i.e. with $N \to \infty$) projections onto the same function classes yields no change.

6. Application to state space models.

6.1. *Neuroscience model.* We return to the neuroscience model introduced in Section 2.3 and explore cSMC's utility as a smoother, with algorithmic settings described in Section 4.4, in comparison to the forward filtering backward smoothing (FFBS) procedure of [16, 31]. We consider an approximation of the maximum likelihood estimate (MLE) $(\alpha, \sigma^2) = (0.99, 0.11)$ as parameter value and the smoothing functional $x_{0:T} \mapsto M(\kappa(x_0), ..., \kappa(x_T))$ whose expectation represents the expected number of activated neurons at each time. Although BPF's particle approximation of the smoothing distribution degenerates quickly in time, cSMC with $I = 3$ iterations offers a marked improvement: for example, the number of distinct ancestors at the initial time is on average 63 times that of BPF. We use $N = 1024$ particles for cSMC and select the number of particles in FFBS to match compute time. The results, displayed in the left panel of Figure 4, show some gains over FFBS and especially so at later times.

We then investigate the relative variance of the log-marginal likelihood estimates obtained using cSMC and BPF in a neighbourhood of the approximate MLE. As the marginal likelihood surface is rather flat in α, we fix $\alpha = 0.99$ and vary $\sigma^2 \in \{0.01, 0.02, \ldots, 0.2\}$. We use $I = 3$ iterations, $N = 128$ particles for cSMC and $N = 5529$ particles for BPF to match computational cost. The results, reported in the right panel of Figure 4, demonstrate that while the relative variance of estimates produced by BPF increases exponentially as σ^2 decreases, that of cSMC is stable across the values of σ^2 considered.

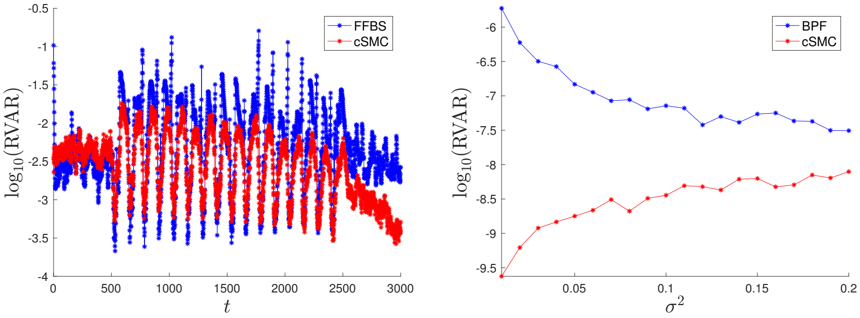

FIG 4. *Assessing performance on the neuroscience model introduced in Section 2.3 based on* 100 *independent repetitions of each algorithm: sample relative variance of smoothing expectation (*left*) and log-marginal likelihood estimates (*right*).*

Lastly, we perform Bayesian inference on the unknown parameters $\theta = (\alpha, \sigma^2)$ and compare the efficiency of cSMC and BPF within a particle marginal Metropolis–Hastings (PMMH) algorithm [1]. We specify a uniform prior on $[0,1]$ for α and an independent inverse-Gamma prior distribution $\mathcal{IG}(1, 0.1)$ for σ^2. Initializing at $\theta = (0.99, 0.11)$, we run two PMMH chains $(\theta_k^{\text{cSMC}})_{k\in[0:K]}$, $(\theta_k^{\text{BPF}})_{k\in[0:K]}$ of length $K = 100,000$. Both chains are updated using an independent Gaussian random walk proposal with standard deviation $(0.002, 0.01)$, but rely on cSMC or BPF to produce unbiased estimates of the marginal likelihood when computing acceptance probabilities. To ensure a fair comparison, we use $I = 3$ iterations and $N = 128$ particles for cSMC which matches the compute time taken by BPF with $N = 5529$ particles, so that both PMMH chains require very similar computational cost. The autocorrelation functions of each PMMH chain, shown in Figure 5, reveal that the $(\theta_k^{\text{BPF}})_{k\in[0:K]}$ chain has poorer mixing properties. These differences can be summarized by the effective sample size, computed as the length of the chain K divided by the estimated integrated autocorrelation time for each parameter of interest, which was found to be $(4356, 2442)$ for $(\theta_k^{\text{BPF}})_{k\in[0:K]}$ and $(20973, 13235)$ for $(\theta_k^{\text{cSMC}})_{k\in[0:K]}$.

6.2. *The Lorenz-96 model.* Following [38], we consider the Lorenz-96 model [34] in a low noise regime, i.e. the Itô process $\xi(s) = (\xi_i(s))_{i\in[1:d]}, s \geq 0$ defined as the weak solution of the stochastic differential equation:

$$d\xi_i = (-\xi_{i-1}\xi_{i-2} + \xi_{i-1}\xi_{i+1} - \xi_i + \alpha)\,dt + \sigma_f dB_i, \quad i \in [1:d], \tag{42}$$

where indices should be understood modulo d, $\alpha \in \mathbb{R}$ is a forcing parameter, $\sigma_f^2 \in \mathbb{R}_+$ is a noise parameter and $B(s) = (B_i(s))_{i\in[1:d]}, s \geq 0$ is a d-dimensional standard Brownian motion. The initial condition is taken as

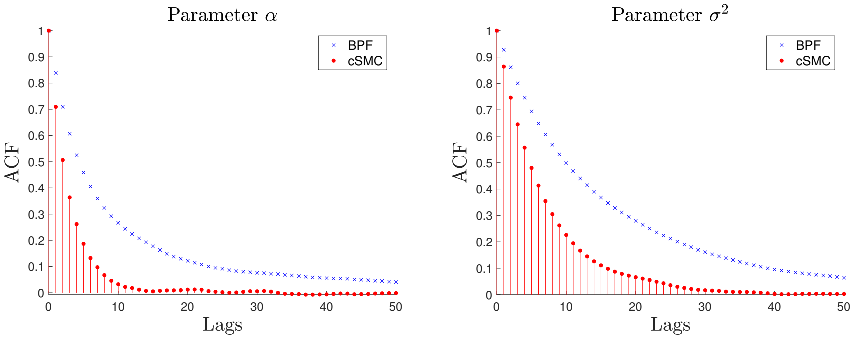

FIG 5. *Autocorrelation functions of PMMH chains, with marginal likelihood estimates produced by cSMC or BPF, for parameters of the neuroscience model introduced in Section 2.3.*

$\xi(0) \sim \mathcal{N}(0_d, \sigma_f^2 I_d)$. We assume that the process is observed at a regular time grid of size $h > 0$ according to $Y_t \sim \mathcal{N}(H\xi(s_t), R), s_t = th, t \in [0 : T]$, and consider the partially observed case where $H_{ii} = 1$ for $i = 1, \ldots, p$ and 0 otherwise with $p = d - 2$.

As discussed in [38], an efficient discretization scheme in this low noise regime [36, ch. 3] is given by adding Brownian increments to the output of a high-order numerical integration scheme on the drift of (42). Incorporating time discretization gives a time homogenous state space model on $(\mathsf{X}, \mathcal{X}) = (\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$ with $\nu = \mathcal{N}(0_d, \sigma_f^2 I_d)$, $f(x_{t-1}, \mathrm{d}x_t) = \mathcal{N}(x_t; q(x_{t-1}), \sigma_f^2 h I_d)\mathrm{d}x_t$ and $g(x_t, y_t) = \mathcal{N}(y_t; Hx_t, R)$ for $t \in [1 : T]$, where $y_{0:T} \in \mathsf{Y}^{T+1} = (\mathbb{R}^p)^{T+1}$ is a realization of the observation process and $q : \mathsf{X} \to \mathsf{X}$ denotes the mapping induced by a fourth order Runge–Kutta (RK4) method on $[0, h]$. We will take noise parameters as $\sigma_f^2 = 10^{-2}, R = \sigma_g^2 I_p$, observe the process for 10 time units, i.e. set $h = 0.1$, $T = 100$ and implement RK4 with a step size of 10^{-2}. For this application, we can employ the fully adapted APF as uncontrolled SMC method [41], i.e. set $\mu = \nu^\psi$ and $M_t = f^\psi$ for $t \in [1 : T]$ with policy $\psi_t = g, t \in [0 : T]$.

Our ADP approximation will utilize the function classes

$$\mathsf{F}_t = \left\{ \varphi(x_t) = x_t^T A_t x_t + x_t^T b_t + c_t : (A_t, b_t, c_t) \in \mathbb{S}_d \times \mathbb{R}^d \times \mathbb{R} \right\}, t \in [0 : T], \tag{43}$$

where $\mathbb{S}_d = \{A \in \mathbb{R}^{d \times d} : A = A^T\}$. Under this parameterization, the policy $\psi^{(i)} = (\psi_t^{(i)})_{t \in [0:T]}$ at iteration $i \in [1 : I]$ of Algorithm 3 is given by

$$-\log \psi_t^{(i)}(x_t) = x_t^T A_t^{(i)} x_t + x_t^T b_t^{(i)} + c_t^{(i)}, \quad t \in [0 : T], \tag{44}$$

where $A_t^{(i)} := \sum_{j=1}^i A_t^j, b_t^{(i)} := \sum_{j=1}^i b_t^j, c_t^{(i)} := \sum_{j=1}^i c_t^j$ for $t \in [0 : T]$

and $(A_t^{j+1}, b_t^{j+1}, c_t^{j+1})_{t\in[0:T]}$ denotes coefficients estimated using linear least squares at iteration $j \in [0 : I-1]$. Having APF as uncontrolled SMC is also equivalent to taking BPF as uncontrolled with an initial policy $\psi^{(0)} = (\psi_t^{(0)})_{t\in[0:T]}$ of the form (44) with $A_t^{(0)} := \frac{1}{2}\sigma_g^{-2}H^T H$, $b_t^{(0)} := -\sigma_g^{-2}Hy_t$ and $c_t^{(0)} := \frac{1}{2}\sigma_g^{-2}y_t^T y_t + \frac{1}{2}p\log(2\pi) + \frac{1}{2}d\log(\sigma_g^2)$ for $t \in [1 : T]$. For $A \in \mathbb{S}_d$, the notation $A \succ 0$ refers to A being positive definite. If the constraints $(\sigma_f^{-2}I_d + 2A_0^{(i)})^{-1} \succ 0$, $(\sigma_f^{-2}h^{-1}I_d + 2A_t^{(i)})^{-1} \succ 0, t \in [1 : T]$ are satisfied or imposed[4], then sampling from the twisted initial distribution and transition kernels is feasible and evaluation of the corresponding potentials is also tractable; see Section 9.2 of Supplementary Material for exact expressions. The diagnostics discussed in Section 4.4 indicate that (44) provides an adequate approximation of the optimal policy by adapting to the chaotic behaviour of the Lorenz system.

We begin by comparing the relative variance of the log-marginal likelihood estimates obtained by cSMC and APF, as α takes values in a regular grid between 2.5 to 8.5. We consider $d = 8$ and simulate observations under the model with $\alpha = 4.8801, \sigma_g^2 = 10^{-4}$. We employ $N = 512$ particles and the following adaptive strategy within cSMC: perform policy refinement until the minimum ESS over time is at least 90%, terminating at a maximum of 4 iterations. To ensure a fair comparison, the number of particles used in APF is chosen to match computation time. The results, plotted in the left panel of Figure 6, show that cSMC offers significant variance reduction across all values of α considered. Moreover, we see from the right panel of Figure 6 that the adaptive criterion allows us to adaptively increase the number of iterations as we move away from the data generating parameter. We then compare cSMC against the iterated APF [23, Algorithm 4] when function approximations are performed in the logarithmic scale (43). Using $N = 512$ particles and $I = 3$ iterations with the fully adapted APF as initialization for both algorithms, the sample variance of cSMC log-marginal likelihood estimates at $\alpha = 4.8801$, based on $1,000$ independent repetitions, was smaller than iterated APF at each iteration $i \in [1 : 3]$, with a relative ratio of $\{0.99, 0.94, 0.92\}$, respectively.

Next we consider configurations $(d, \sigma_g^2) \in \{8, 16, 32, 64\} \times \{10^{-4}, 10^{-3}, 10^{-2}\}$ with $\alpha = 4.8801$ and generate observations under the model. We use $I = 1$ iteration for cSMC in all configurations and increase the number of particles N with d for both algorithms. As before, N is chosen so that both methods require the same compute time to ensure a fair comparison. The relative

[4] In our numerical implementation, we find that these constraints are already satisfied when the step size h is sufficiently small. Otherwise, they can be imposed by projecting onto the set of real symmetric positive definite matrices using the algorithm in [25].

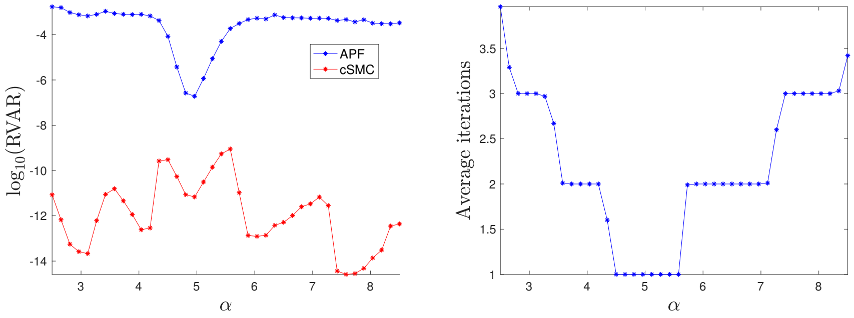

FIG 6. *Lorenz-96 model of Section 6.2 with data generating parameter $\alpha = 4.8801$: sample relative variance of log-marginal likelihood estimates based on* 100 *independent repetitions of each algorithm (*left*), average number of iterations taken by cSMC with adaptation (*right*).*

variance of both methods are reported in Table 1. These results indicate several order of magnitude gains over APF in all configurations considered.

7. Application to static models.

We now detail how the proposed methodology can be applied to static models described in Section 2.4. The framework introduced in [13] generalizes the AIS method of [39] and the sequential sampler of [9] by allowing arbitrary forward and backward kernels instead of being restricted to MCMC kernels. This degree of freedom is useful here as sampling from twisted MCMC kernels and computing integrals w.r.t. these kernels is typically impossible.

7.1. *Setup.*

We consider the Bayesian framework where the target distribution of interest is a posterior distribution $\eta(\mathrm{d}x) = Z^{-1}\mu(\mathrm{d}x)\ell(x,y)$ defined on $(\mathsf{X},\mathcal{X}) = (\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$, given by a Bayes update with a prior distribution $\mu \in \mathcal{P}(\mathsf{X})$ and a likelihood function $\ell : \mathsf{X}\times\mathsf{Y} \to \mathbb{R}_+$. In applications, the marginal likelihood $Z(y) := \int_{\mathsf{X}} \mu(\mathrm{d}x)\ell(x,y)$ of observations $y \in \mathsf{Y}$ is often also a quantity of interest. Assuming η has a strictly positive and continuously differentiable density $x \mapsto \eta(x)$ w.r.t. Lebesgue measure on $\mathbb{R}^d$, we select the forward kernel M_t related to the transition kernel of an unadjusted Langevin algorithm (ULA) [44, 43] targeting η_t defined in (6). For e.g., we will define $M_t(x_{t-1}, \mathrm{d}x_t) = \mathcal{N}(x_t; x_{t-1} + \frac{1}{2}h\Gamma\nabla\log\eta_t(x_{t-1}), h\Gamma)\mathrm{d}x_t$ where $h > 0$ denotes the step size, and Γ is a positive definite pre-conditioning matrix (which in the simplest case may be the identity $\Gamma = I_d$).

Under appropriate regularity conditions, for sufficiently small h, M_t admits an invariant distribution that is close to η_t [35]. Moreover, as the corresponding Langevin diffusion is η_t-reversible, this suggests that M_t will

				Observation noise		
				$\sigma_g^2 = 10^{-4}$	$\sigma_g^2 = 10^{-3}$	$\sigma_g^2 = 10^{-2}$
			N	$\log_{10}(\text{RVAR})$	$\log_{10}(\text{RVAR})$	$\log_{10}(\text{RVAR})$
Algorithm	APF	$d = 8$	1382	-6.7263	-5.6823	-4.4061
		$d = 16$	2027	-7.4056	-5.9009	-4.4719
		$d = 32$	4034	-7.5943	-5.4901	-4.1039
		$d = 64$	$11,468$	-7.5173	-5.3765	-3.1057
	cSMC	$d = 8$	512	-11.1252	-10.4173	-8.66563
		$d = 16$	512	-11.8899	-11.1011	-9.29596
		$d = 32$	1024	-12.5804	-11.8622	-9.6577
		$d = 64$	4096	-13.5959	-12.7691	-9.74631

TABLE 1

Algorithmic settings and performance of APF and cSMC for each dimension d and observation noise σ_g^2 considered. Notationally, N refers to the number of particles and RVAR *is the sample relative variance of log-marginal likelihood estimates over* 100 *independent repetitions of each method.*

also be approximately η_t-reversible for small h. This prompts the choice of backward kernel $L_{t-1}(x_t, \mathrm{d}x_{t-1}) = M_t(x_t, \mathrm{d}x_{t-1})$, in which case, we expect the potentials (7) to be close to (8) when the step size is small. We have limited the scope of this article to overdamped Langevin dynamics; future work could consider the use of generalized Langevin dynamics and other non-reversible dynamics.

7.2. *Log-Gaussian Cox point process.* We end with a challenging high dimensional application of Bayesian inference for log-Gaussian Cox point processes on a dataset[5] concerning the locations of 126 Scots pine saplings in a natural forest in Finland [37, 11, 20]. The actual square plot of 10×10 square metres is standardized to the unit square and locations are plotted in the left panel of Figure 7. We then discretize $[0,1]^2$ into a 30×30 regular grid. Given a latent intensity process $\Lambda = (\Lambda_m)_{m \in [1:30]^2}$, the number of points in each grid cell $Y = (Y_m)_{m \in [1:30]^2} \in \mathbb{N}^{30^2}$ are modelled as conditionally independent and Poisson distributed with means $a\Lambda_m$, where $a = 1/30^2$ is the area of each grid cell. The prior distribution for Λ is specified by $\Lambda_m = \exp(X_m)$, $m \in [1:30]^2$, where $X = (X_m)_{m \in [1:30]^2}$ is a Gaussian process with constant mean $\mu_0 \in \mathbb{R}$ and exponential covariance function $\Sigma_0(m,n) = \sigma^2 \exp(-|m-n|/(30\beta))$ for $m, n \in [1:30]^2$. We will adopt the parameter values $\sigma^2 = 1.91$, $\beta = 1/33$ and $\mu_0 = \log(126) - \sigma^2/2$ estimated by [37]. This application corresponds to dimension $d = 900$, a prior distribution $\mu = \mathcal{N}(\mu_0 1_d, \Sigma_0)$ with $1_d = (1, \ldots, 1)^T \in \mathbb{R}^d$ and likelihood function $\ell(x,y) = \prod_{m \in [1:30]^2} \exp\left(x_m y_m - a\exp(x_m)\right)$, where $y = (y_m)_{m \in [1:30]^2} \in \mathsf{Y} = \mathbb{N}^d$ is the

[5] The dataset can be found in the `R` package `spatstat` as `finpines`.

given dataset.

For this application, cSMC relies on pre-conditioned ULA moves with the choice of $\Gamma^{-1} = \Sigma_0^{-1} + a\exp(\mu_0 + \sigma^2/2)I_d$ considered in [20]. As the above choice of pre-conditioning captures the curvature of the posterior distribution, we adopt the following function classes

$$\begin{aligned}(45)\quad \mathsf{F}_0 &= \left\{\varphi(x_0) = x_0^T A_0 x_0 + x_0^T b_0 + c_0 : (A_0, b_0, c_0) \in \mathbb{S}_d \times \mathbb{R}^d \times \mathbb{R}\right\}, \\ \mathsf{F}_t &= \Big\{\varphi(x_{t-1}, x_t) = x_t^T A_t x_t + x_t^T b_t + c_t - (\lambda_t - \lambda_{t-1})\log \ell(x_{t-1}, y) \\ &\qquad : (A_t, b_t, c_t) \in \mathbb{S}_d \times \mathbb{R}^d \times \mathbb{R}\Big\}, \quad t \in [1:T],\end{aligned}$$

where $(A_t)_{t\in[0:T]}$ are restricted to diagonal matrices to reduce the computational overhead involved in estimating large number of coefficients for a problem of this scale. The rationale for approximating the x_{t-1} dependency in $\psi_t^*(x_{t-1}, x_t), t \in [1:T]$ is based on the argument that the potentials (7) would be close to that of AIS (8) for sufficiently small step size h. We refer to Section 8.1 of Supplementary Material for exact expressions required to implement cSMC. As before, the diagnostics discussed in Section 4.4 reveal that such a parameterization offers an adequate approximation of the optimal policy.

We select as competing algorithms: 1) standard AIS with pre-conditioned Metropolis-adjusted Langevin algorithm (MALA) moves; and, 2) an adaptive (pre-conditioned) AIS. For both cSMC and standard AIS, we adopt the geometric path (6) with $\lambda_t = t/T$ and fix the number of time steps as $T = 20$. We use $N = 4096$ particles, $I = 3$ iterations for cSMC and 5 times more particles for standard AIS to ensure that our comparison is performed at a fixed computational complexity. Using pilot runs, we chose a step size of 0.4 for MALA to achieve suitable acceptance probabilities, and a smaller step size of 0.05 for ULA as this improves the approximation in (45). For the adaptive AIS algorithm, we also adopt (6) with λ_t adapted so that the ESS% is maintained above 80% [27, 46, 51] and with an adaptive step size chosen to ensure an acceptance probability within the range of 30% to 50% at each time step [27, 3]. Since the runtime of adaptive AIS is random, we choose the number of particles to ensure the averaged computational cost matches that of cSMC and standard AIS; this is typically on the order of 2 times as many particles as cSMC.

The results obtained show that standard AIS performs poorly in this scenario, providing high variance estimates of the log-marginal likelihood compared to each iteration of cSMC, as displayed in the right panel of Figure 7. Adaptive AIS performs better than standard AIS but it is still outperformed by cSMC. The sample variance of log-marginal likelihood estimates is 573

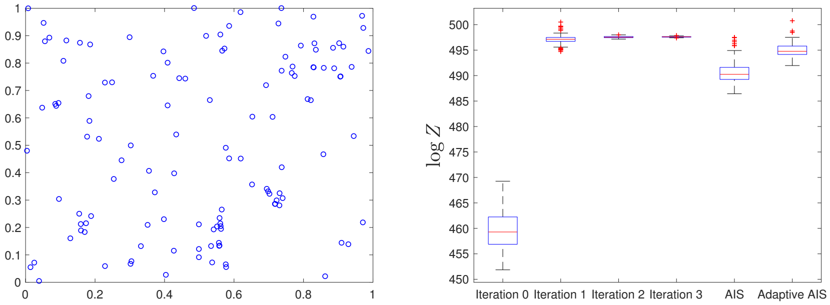

FIG 7. *Locations of* 126 *Scots pine saplings in a natural forest in Finland (*left*) and log-marginal likelihood estimates obtained with* 100 *independent repetitions of cSMC, standard AIS, and adaptive AIS (*right*).*

times smaller for the last iteration of cSMC compared to standard AIS, and it is 200 times smaller compared to adaptive AIS. The mean squared error[6] of adaptive AIS algorithm is 920 times larger than that of cSMC.

SUPPLEMENTARY MATERIAL

Supplementary Material for 'Controlled Sequential Monte Carlo' (URL). The supplement contains proofs of all results, a detailed description of the connection to Kullback–Leibler control, three more applications employing other flexible function classes, and some model specific expressions.

References.

[1] C. Andrieu, A. Doucet and R. Holenstein. Particle Markov chain Monte Carlo (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):357–385, 2010.

[2] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic Programming*. Athena Scientific, 1996.

[3] A. Beskos, A. Jasra, N. Kantas and A. Thiery. On the convergence of adaptive sequential Monte Carlo methods. *Annals of Applied Probability*, 26(2):1111–1146, 2016.

[4] K. Bichteler. *Stochastic Integration with Jumps*. Cambridge University Press, 2002.

[5] Y. Bresler. Two-filter formula for discrete-time non-linear Bayesian smoothing. *International Journal of Control*, 43(2):629–641, 1986.

[6] M. Briers, A. Doucet and S. Maskell. Smoothing algorithms for state-space models. *Annals of the Institute of Statistical Mathematics*, 62(1):61–89, 2010.

[7] P. Bühlmann and B. Yu. Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

[8] R. Chen, L. Ming and J. S. Liu. Lookahead strategies for sequential Monte Carlo. *Statistical Science*, 28(1):69–94, 2013.

[6] Computed by taking reference to an estimate obtained using many repetitions of a SMC sampler with a large number of particles.

[9] N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

[10] N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Annals of Statistics*, 32(6):2385–2411, 2004.

[11] O. F. Christensen, G. O. Roberts and J. S. Rosenthal. Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):253–268, 2005.

[12] P. Del Moral. *Feynman-Kac Formulae.* Springer, 2004.

[13] P. Del Moral, A. Doucet and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[14] P. Del Moral and A. Guionnet. Central limit theorem for nonlinear filtering and interacting particle systems. *Annals of Applied Probability*, 9(2):275–297, 1999.

[15] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review.* 41(1):45–76, 1999.

[16] A. Doucet, S. J. Godsill and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[17] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In *Handbook of Nonlinear Filtering* (editors D. Crisan and B. L. Rozovsky), Oxford University Press, 656–704, 2011.

[18] A. Gelman and X. L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

[19] M. Gerber, N. Chopin and N. Whiteley. Negative association, ordering and convergence of resampling methods. *Annals of Statistics*, to appear, 2019.

[20] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[21] E. Gobet. *Monte-Carlo Methods and Stochastic Processes: From Linear to Non-Linear.* CRC Press, 2016.

[22] N. J. Gordon, D. Salmond and A. F. M. Smith. A novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEE Proceedings on Radar and Signal Processing*, 140:107–113, 1993.

[23] P. Guarniero, A. M. Johansen and A. Lee. The iterated auxiliary particle filter. *Journal of the American Statistical Association*, 112(520):1636–1647, 2017.

[24] A. Gupta, R. Jain and P. Glynn. A fixed point theorem for iterative random contraction operators over Banach spaces. arXiv:1804.01195, 2018.

[25] N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

[26] P. E. Jacob, L. M. Murray and S. Rubenthaler. Path storage in the particle filter. *Statistics and Computing*, 25(2):487–496, 2015.

[27] A. Jasra, D. A. Stephens, A. Doucet and T. Tsagaris. Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, 2011.

[28] H. J. Kappen, V. Gómez and M. Opper. Optimal control as a graphical model inference problem. *Machine Learning*, 87(2):159–182, 2012.

[29] H. J. Kappen and H. C. Ruiz. Adaptive importance sampling for control and inference. *Journal of Statistical Physics*, 162(5):1244–1266, 2016.

[30] H. R. Künsch. Recursive Monte Carlo filters: algorithms and theoretical analysis. *Annals of Statistics*, 33(5):1983–2021, 2005.

[31] H. R. Künsch. Particle filters. *Bernoulli*, 19(4):1391–1403, 2013.

[32] J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal*

of the American Statistical Association, 93(443):1032–1044, 1998.

[33] J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, 2001.

[34] E. N. Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on Predictability*, Vol. 1, 1996.

[35] J. C. Mattingly, A. M. Stuart and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185–232, 2002.

[36] G. N. Milstein and M. V. Tretyakov. *Stochastic Numerics for Mathematical Physics.* Springer, 2004.

[37] J. Møller, A. R. Syversveen and R. P. Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

[38] L. M. Murray, S. Singh, P. E. Jacob and A. Lee. Anytime Monte Carlo. arXiv:1612.03319, 2016.

[39] R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[40] T. Nemoto, F. Bouchet, R. L. Jack and V. Lecomte. Population-dynamics method with a multicanonical feedback control. *Physical Review E*, 93(6):062123, 2016.

[41] M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association,* 94(446):590–599, 1999.

[42] J. F. Richard and W. Zhang. Efficient high-dimensional importance sampling. *Journal of Econometrics*, 141(2):1385–1411, 2007.

[43] G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, 2002.

[44] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[45] H. C. Ruiz and H. J. Kappen. Particle smoothing for hidden diffusion processes: adaptive path integral smoother. *IEEE Transactions on Signal Processing*, 65(12):3191–3203, 2017.

[46] C. Schäfer and N. Chopin. Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing*, 23(2):163–184, 2013.

[47] M. Scharth and R. Kohn. Particle efficient importance sampling. *Journal of Econometrics*, 190(1):133–147, 2016.

[48] S. Temereanca, E. N. Brown and D. J. Simons. Rapid changes in thalamic firing synchrony during repetitive whisker stimulation. *Journal of Neuroscience*, 28(44):11153–11164, 2008.

[49] E. A. Theodorou and E. Todorov. Relative entropy and free energy dualities: Connections to path integral and KL control. *In Proceedings 51st IEEE Conference on Decision and Control (CDC),* 1466-1473, 2012.

[50] S. Thijssen and H. J. Kappen. Path integral control and state-dependent feedback. *Physical Review E*, 91(3):032104, 2015.

[51] Y. Zhou, A. M. Johansen and J. A. D. Aston. Towards automatic model comparison: An adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.

5 Nepal Park
Singapore 139408, Singapore
E-mail: heng@essec.edu

15 Broadway
Sydney, NSW 2007, Australia
E-mail: adrian.bishop@uts.edu.au

24-29 St Giles'
Oxford OX1 3LB, UK
E-mail: deligian@stats.ox.ac.uk; doucet@stats.ox.ac.uk

Submitted to the Annals of Statistics
arXiv: arXiv:1708.08396

SUPPLEMENTARY MATERIAL FOR 'CONTROLLED SEQUENTIAL MONTE CARLO'

By Jeremy Heng[*], Adrian N. Bishop[†], George Deligiannidis[‡] and Arnaud Doucet[‡]

ESSEC Business School[*]*, CSIRO and the University of Technology Sydney*[†] *and Oxford University*[‡]

1. Proofs of Section 4.1.

Proof of Proposition 4.1. By Fubini's theorem, ϕ^* is well-defined as the integrals in (19) exist since $Z = \mathbb{E}_{\mathbb{Q}^\psi}\left[G_0^\psi(X_0)\prod_{t=1}^T G_t^\psi(X_{t-1},X_t)\right]$ is finite, and is admissible if the potentials $(G_t^\psi)_{t\in[0:T]}$ are bounded. From (12), the first t^{th}-marginal distribution and time t^{th}-marginal distribution of $\mathbb{P}$ are given by

(S1)
$$\mathbb{P}(\mathrm{d}x_{0:t}) = Z^{-1}\mu^\psi(\mathrm{d}x_0)G_0^\psi(x_0)\left\{\prod_{s=1}^{t-1} M_s^\psi(x_{s-1},\mathrm{d}x_s)G_s^\psi(x_{s-1},x_s)\right\}M_t^\psi(x_{t-1},\mathrm{d}x_t)\phi_t^*(x_{t-1},x_t)$$

and

$$\mathbb{P}(\mathrm{d}x_t) = Z^{-1}Z_t^\psi\eta_t^\psi(\mathrm{d}x_t)M_{t+1}^\psi(\phi_{t+1}^*)(x_t) \tag{S2}$$

respectively, for $t\in[0:T]$. The representation (Property 1)

$$\mathbb{P}(\mathrm{d}x_{0:T}) = \left(\mu^\psi\right)^{\phi^*}(\mathrm{d}x_0)\prod_{t=1}^T (M_t^\psi)^{\phi^*}(x_{t-1},\mathrm{d}x_t) = \mathbb{Q}^{\psi^*}(\mathrm{d}x_{0:T})$$

follows from (S1)-(S2) by noting that $\mu^\psi(\phi_0^*) = Z$ and

$$\mathbb{P}(\mathrm{d}x_t|x_{0:t-1}) = \frac{M_t^\psi(x_{t-1},\mathrm{d}x_t)\phi_t^*(x_{t-1},x_t)}{M_t^\psi(\phi_t^*)(x_{t-1})}$$

for $t\in[1:T]$. Under the refined policy $\psi^* := \psi\cdot\phi^*$, it follows from (18) that

$$G_0^{\psi^*}(x_0) = Z, \quad G_t^{\psi^*}(x_{t-1},x_t) = 1, \quad t\in[1:T],$$

hence Property 3 follows from the form of the estimator (17) and $Z_t^{\psi^*} = Z$ for all $t\in[0:T]$. Using the latter, (15), and (S2) establishes Property 2. □

To build some intuition, we provide a characterization of the optimal policy in a specific setting which guides the choice of the function classes considered in Section 7.2.

PROPOSITION 1. *For any policy $\psi \in \Psi$ such that the corresponding twisted potentials $(G_t^{\psi})_{t\in[0:T]}$ and transition densities of $(M_t^{\psi})_{t\in[1:T]}$ are log-concave on their domain of definition, the optimal policy $\phi^* = (\phi_t^*)_{t\in[0:T]}$ w.r.t. $\mathbb{Q}^{\psi}$ is a sequence of log-concave functions.*

PROOF OF PROPOSITION 1. For $t = T$, log-concavity of $\phi_T^* = G_T^{\psi}$ follows by assumption. For $t \in [0 : T-1]$, we proceed with an inductive argument on the backward recursion (19). Assuming that ϕ_{t+1}^* is log-concave, note that $x_t \mapsto M_{t+1}^{\psi}(\phi_{t+1}^*)(x_t)$ is log-concave since the product $(x_t, x_{t+1}) \mapsto \phi_{t+1}^*(x_t, x_{t+1}) M_{t+1}^{\psi}(x_t, x_{t+1})$ is and log-concavity is preserved by marginalization. Hence ϕ_t^* is log-concave as the product of log-concave functions is also log-concave. □

2. Proofs of Section 5.1.

PROOF OF PROPOSITION 5.1. We begin by noting the semigroup property

$$Q_{s,u}^{\psi}(\varphi) = Q_{s,t}^{\psi} \circ Q_{t,u}^{\psi}(\varphi), \quad 0 \le s < t < u \le T,$$

where we recall that we have defined $Q_T^{\psi}(\varphi) = G_T^{\psi}$ for any φ for notational convenience.

Define the approximate Bellman operators as $\hat{Q}_t^{\psi}\varphi = P_t^{\psi,N} Q_t^{\psi}\varphi$ for $\varphi \in L^2(\nu_{t+1}^{\psi}), t \in [0 : T]$. The measures ν_t^{ψ} for $t \in [0 : T]$ have been introduced in Definition 5.1. By defining $\hat{\phi}_{T+1} = 1$ for notational convenience and using (27), we obtain the following telescoping decomposition

$$\hat{\phi}_t - \phi_t^* = \sum_{u=t}^{T} Q_{t-1,u-1}^{\psi} \circ \hat{Q}_u^{\psi}(\hat{\phi}_{u+1}) - Q_{t-1,u-1}^{\psi} \circ Q_u^{\psi}(\hat{\phi}_{u+1}).$$

Hence by the triangle inequality, we have

$$\|\hat{\phi}_t - \phi_t^*\|_{L^2(\nu_t^{\psi})} \le \sum_{u=t}^{T} \|Q_{t-1,u-1}^{\psi} \circ \hat{Q}_u^{\psi}(\hat{\phi}_{u+1}) - Q_{t-1,u-1}^{\psi} \circ Q_u^{\psi}(\hat{\phi}_{u+1})\|_{L^2(\nu_t^{\psi})}$$

for any $t \in [0 : T]$. Under Assumption A1, (28) are linear bounded operators, hence

$$\|\hat{\phi}_t - \phi_t^*\|_{L^2(\nu_t^{\psi})} \le \sum_{u=t}^{T} C_{t-1,u-1}^{\psi} \|P_u^{\psi,N} Q_u^{\psi}(\hat{\phi}_{u+1}) - Q_u^{\psi}(\hat{\phi}_{u+1})\|_{L^2(\nu_t^{\psi})}.$$

Taking expectations and applying Assumption A2 yields (33). □

Proof of Proposition 5.2. It follows from (14) that for any $r \in [1:T]$ and $\varphi \in L^1(\eta_r^\psi)$ we have

(S3)
$$\eta_0^\psi(\varphi) = \frac{\mu^\psi(G_0^\psi \varphi)}{\mu^\psi(G_0^\psi)}, \quad \eta_r^\psi(\varphi) = \frac{\eta_{r-1}^\psi(M_r^\psi(G_r^\psi \varphi))}{\eta_{r-1}^\psi(M_r^\psi(G_r^\psi))}, \quad \eta_{r-1}^\psi(M_r^\psi(G_r^\psi)) = \frac{Z_r^\psi}{Z_{r-1}^\psi}.$$

Now for $r \in [1:T-1]$ and $\varphi \in L^2(\nu_{r+1}^\psi)$, using Jensen's inequality and the above identity

$$\begin{aligned}
\|Q_r^\psi(\varphi)\|_{L^2(\nu_r^\psi)}^2 &= \int_{\mathsf{X}^2} G_r^\psi(x,y)^2 M_{r+1}^\psi(\varphi)^2(y)\eta_{r-1}^\psi(\mathrm{d}x)M_r^\psi(x,\mathrm{d}y)\\
&\le \|G_r^\psi\|_\infty \int_{\mathsf{X}^2} G_r^\psi(x,y) M_{r+1}^\psi(\varphi^2)(y)\eta_{r-1}^\psi(\mathrm{d}x)M_r^\psi(x,\mathrm{d}y)\\
&= \|G_r^\psi\|_\infty \eta_{r-1}^\psi(M_r^\psi(G_r^\psi)) \int_{\mathsf{X}} M_{r+1}^\psi(\varphi^2)(y)\eta_r^\psi(\mathrm{d}y)\\
&= \frac{Z_r^\psi}{Z_{r-1}^\psi}\|G_r^\psi\|_\infty \|\varphi\|_{L^2(\nu_{r+1}^\psi)}^2.
\end{aligned}$$

The result for $r = 0$ follows the same arguments. Letting $\varphi \in L^2(\nu_{t+1}^\psi)$, whence $Q_{s+1,t}^\psi(\varphi) \in L^2(\nu_{s+2}^\psi)$, the above bound with $r = s+1$ implies that

$$\|Q_{s,t}^\psi(\varphi)\|_{L^2(\nu_{s+1}^\psi)}^2 = \|Q_{s+1}^\psi Q_{s+1,t}^\psi(\varphi)\|_{L^2(\nu_{s+1}^\psi)}^2 \le \frac{Z_{s+1}^\psi}{Z_s^\psi}\|G_{s+1}^\psi\|_\infty \|Q_{s+1,t}^\psi(\varphi)\|_{L^2(\nu_{s+2}^\psi)}^2.$$

Iterating we establish (34).

When $G_r^\psi(x,y) = G_r^\psi(y)$ for all $x, y \in \mathsf{X}$ and $r \in [1:T]$,

$$\begin{aligned}
\|Q_{s,t}^\psi(\varphi)\|_{L^2(\eta_s^\psi M_{s+1}^\psi)}^2 &= \int \left[Q_{s,t}^\psi(\varphi)(x)\right]^2 \eta_s^\psi M_{s+1}^\psi(\mathrm{d}x)\\
&= \int \left[\frac{Q_{s,t}^\psi(\varphi)(x)}{Q_{s,t}^\psi(1)(x)}\right]^2 (Q_{s,t}^\psi(1))^2(x)\eta_s^\psi M_{s+1}^\psi(\mathrm{d}x)\\
&\le \int \frac{Q_{s,t}^\psi(\varphi^2)(x)}{Q_{s,t}^\psi(1)(x)} (Q_{s,t}^\psi(1))^2(x)\eta_s^\psi M_{s+1}^\psi(\mathrm{d}x),
\end{aligned}$$

by Jensen's inequality applied to the Markov operator $\varphi \mapsto Q_{s,t}^\psi(\varphi)/Q_{s,t}^\psi(1)$. From Assumption A3 in (35), and the boundedness of $(G_t^\psi)_{t\in[0:T]}$ it follows that

$$Q^{\psi}_{s,t}(1)(x) = G^{\psi}_{s+1}(x)\int M^{\psi}_{s+2}(x,\mathrm{d}y)Q^{\psi}_{s+1,t}(1)(y) \leq \kappa^{\psi}_{s+2}\|G^{\psi}_{s+1}\|_{\infty}\ \sigma^{\psi}_{s+2}(Q^{\psi}_{s+1,t}(1)) < \infty.$$

Therefore we can write

$$\begin{aligned}
&\|Q^{\psi}_{s,t}(\varphi)\|^2_{L^2(\eta^{\psi}_s M^{\psi}_{s+1})} \leq \int Q^{\psi}_{s,t}(\varphi^2)(x)Q^{\psi}_{s,t}(1)(x)\eta^{\psi}_s M^{\psi}_{s+1}(\mathrm{d}x)\\
&\leq \kappa^{\psi}_{s+2}\,\|G^{\psi}_{s+1}\|_{\infty}\,\sigma^{\psi}_{s+2}\big(Q^{\psi}_{s+1,t}(1)\big)\int Q^{\psi}_{s,t}(\varphi^2)(x)\eta^{\psi}_s M^{\psi}_{s+1}(\mathrm{d}x)\\
&\leq \kappa^{\psi}_{s+2}\,\|G^{\psi}_{s+1}\|_{\infty}\,\sigma^{\psi}_{s+2}\big(Q^{\psi}_{s+1,t}(1)\big)\,\eta^{\psi}_s M^{\psi}_{s+1}(Q^{\psi}_{s,t}(1))\int \frac{Q^{\psi}_{s,t}(\varphi^2)(x)}{\eta^{\psi}_s M^{\psi}_{s+1}(Q^{\psi}_{s,t}(1))}\eta^{\psi}_s M^{\psi}_{s+1}(\mathrm{d}x)\\
&= \left[\kappa^{\psi}_{s+2}\,\|G^{\psi}_{s+1}\|_{\infty}\,\sigma^{\psi}_{s+2}\big(Q^{\psi}_{s+1,t}(1)\big)\frac{Z^{\psi}_t}{Z^{\psi}_s}\right]\|\varphi\|^2_{L^2(\eta^{\psi}_t M^{\psi}_{t+1})},
\end{aligned}$$

since one can check that for any function f

$$\frac{\eta^{\psi}_s M^{\psi}_{s+1}Q^{\psi}_{s,t}(f)}{\eta^{\psi}_s M^{\psi}_{s+1}Q^{\psi}_{s,t}(1)} = \eta^{\psi}_t M^{\psi}_{t+1}(f), \qquad \eta^{\psi}_s M^{\psi}_{s+1}Q^{\psi}_{s,t}(1) = \frac{Z^{\psi}_t}{Z^{\psi}_s}. \qquad \square$$

3. Proofs of Section 5.2. Given $\gamma \in \mathcal{S}(\mathsf{E})$ and matrix-valued $\varphi : \mathsf{E} \to \mathbb{R}^{p\times d}$ with $\varphi_{i,j} \in \mathcal{B}(\mathsf{E})$ for all $i \in [1:p], j \in [1:d]$, we extend the definition of $\gamma(\varphi)$ element-wise, i.e. $\gamma(\varphi)_{i,j} = \gamma(\varphi_{i,j})$. Assuming that the Gram matrices

$$A^{\psi,N}_t := \nu^{\psi,N}_t(\Phi_t\Phi_t^T), \quad t \in [0:T], \tag{S4}$$

are invertible, under (38) the estimated policy has the form $\hat{\phi}_t = \exp(-\Phi_t^T\beta^{\psi,N}_t), t \in [0:T]$, where the least squares estimators $\beta^{\psi,N}_t = (A^{\psi,N}_t)^{-1}b^{\psi,N}_t, t \in [0:T]$ are defined by the backward recursion

$$\begin{aligned}
b^{\psi,N}_T &= -\nu^{\psi,N}_T(\log G^{\psi}_T \cdot \Phi_T),\\
b^{\psi,N}_t &= -\nu^{\psi,N}_t(\{\log G^{\psi}_t + \log M^{\psi}_{t+1}(\exp(-\Phi^T_{t+1}(A^{\psi,N}_{t+1})^{-1}b^{\psi,N}_{t+1}))\}\ \Phi_t),
\end{aligned} \tag{S5}$$

for $t \in [0:T-1]$. To prove the claims in Theorem 5.1, we first establish convergence of $\beta^{\psi,N}_t$ to $\beta^{\psi}_t := (A^{\psi}_t)^{-1}b^{\psi}_t$, given by the Gram matrix $A^{\psi}_t := \nu^{\psi}_t(\Phi_t\Phi_t^T)$ and vector b^{ψ}_t defined by the backward recursion

$$\begin{aligned}
b^{\psi}_T &= -\nu^{\psi}_T(\log G^{\psi}_T \cdot \Phi_T),\\
b^{\psi}_t &= -\nu^{\psi}_t(\{\log G^{\psi}_t + \log M^{\psi}_{t+1}(\exp(-\Phi^T_{t+1}(A^{\psi}_{t+1})^{-1}b^{\psi}_{t+1}))\}\ \Phi_t),
\end{aligned} \tag{S6}$$

for $t \in [0:T-1]$.

PROPOSITION 2. *Consider ADP algorithm (31), with current policy $\psi \in \Psi$, under linear least squares approximations (38) with basis functions $(\Phi_t)_{t\in[0:T]}$ chosen so that:*

[A7] *the Gram matrices $(A_t^\psi)_{t\in[0:T]}$ are invertible;*

[A8] *the function $x \mapsto M_t^\psi(\exp(-\Phi_t^T\beta))(x)$ is $\mathcal{X}$-measurable for all $\beta \in \mathbb{R}^M, t \in [1:T]$ and the integrals in (S6) are finite;*

[A9] *for each $t \in [0:T-1]$, there exist a $\mathcal{X}$-measurable function $C_t : \mathsf{X} \to \mathbb{R}_+$ and a continuous function $\delta_t : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\nu_t^\psi(C_t|\Phi_t|) < \infty$ and $\lim_{x\to 0}\delta_t(x) = 0$ respectively such that*

$$\left|\log M_{t+1}^\psi(\exp(-\Phi_{t+1}^T\beta))(x) - \log M_{t+1}^\psi(\exp(-\Phi_{t+1}^T\beta'))(x)\right| \leq C_t(x)\delta_t(|\beta-\beta'|)$$

holds for all $x \in \mathsf{X}$ and $\beta, \beta' \in \mathbb{R}^M$. As $N \to \infty$, the least squares estimators $\beta^{\psi,N} := (\beta_t^{\psi,N})_{t\in[0:T]}$ converge in probability to $\beta^\psi := (\beta_t^\psi)_{t\in[0:T]}$;

[A10] *(i) for each $t \in [0:T-1]$, the function $\beta \mapsto \log M_{t+1}^\psi(\exp(-\Phi_{t+1}^T\beta))(x)$ is continuously differentiable for all $x \in \mathsf{X}$;*

(ii) its gradient $x \mapsto g_{t+1}^\psi(\beta, x) := \nabla_\beta \log M_{t+1}^\psi(\exp(-\Phi_{t+1}^T\beta))(x)$ is $\mathcal{X}$-measurable for all $\beta \in \mathbb{R}^M$, satisfies $\nu_t^\psi(|\Phi_t g_{t+1}^\psi(\beta_{t+1}^\psi, \cdot)^T|) < \infty$ and for each $t \in [0:T-1]$, there exists a positive, $\mathcal{X}$-measurable function $C_t' : \mathsf{X} \to \mathbb{R}_+$ satisfying $\nu_t^\psi(C_t'|\Phi_t|) < \infty$ such that

$$\left|g_{t+1}^\psi(\beta, x) - g_{t+1}^\psi(\beta', x)\right| \leq C_t'(x)|\beta - \beta'|$$

holds for all $x \in \mathsf{X}$ and $\beta, \beta' \in \mathbb{R}^M$;

[A11] *the vector-valued function $\xi^\psi = (\xi_t^\psi)_{t\in[0:T]} : \mathsf{X}^{2T+1} \to \mathbb{R}^{(T+1)M}$ defined componentwise as*

(S7)

$$\xi_t^\psi = -(A_t^\psi)^{-1}\{\log G_t^\psi + \log M_{t+1}^\psi(\exp(-\Phi_{t+1}^T\beta_{t+1}^\psi))\}\Phi_t - (A_t^\psi)^{-1}\Phi_t\Phi_t^T\beta_t^\psi, \quad t \in [0:T-1],$$
$$\xi_T^\psi = -(A_T^\psi)^{-1}(\log G_T^\psi \cdot \Phi_T + \Phi_T\Phi_T^T\beta_T^\psi),$$

satisfies $\xi^\psi \in L^2(\nu^\psi)$ with $\nu^\psi := \otimes_{t=0}^T \nu_t^\psi \in \mathcal{P}(\mathsf{X}^{2T+1})$ and the following central limit theorem

$$\sqrt{N}\left(\nu^{\psi,N}(\xi^\psi) - \nu^\psi(\xi^\psi)\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0_{(T+1)M}, \Gamma^\psi\right) \tag{S8}$$

with $\nu^{\psi,N} := \otimes_{t=0}^T \nu_t^{\psi,N}$.

Then we have

$$\sqrt{N}\left(\beta^{\psi,N} - \beta^\psi\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0_{(T+1)M}, \Sigma^\psi\right) \tag{S9}$$

where $\Sigma^\psi = U^\psi \Gamma^\psi (U^\psi)^T$ is given by a block upper triangular matrix $U^\psi \in \mathbb{R}^{(T+1)M\times(T+1)M}$ defined by blocks of size $M \times M$

$$U^\psi_{s,t} = \begin{cases} \prod_{u=s-1}^{t-2} E^\psi_u, & s < t, \\ I_M, & s = t, \\ 0_{M\times M}, & s > t, \end{cases} \tag{S10}$$

for $s,t \in [1 : T+1]$, with $E^\psi_t := -(A^\psi_t)^{-1}\nu^\psi_t(\Phi_t g^\psi_{t+1}(\beta^\psi_{t+1}, \cdot)^T), t \in [0 : T-1]$ and $0_{M\times M}$ as the $M \times M$ matrix of zeros.

Proof of Proposition 2. Note that for each $t \in [0 : T]$, by the strong law of large numbers (LLN) for the particle approximation $\nu^{\psi,N}_t$ (see [4]) $A^{\psi,N}_t \to A^\psi_t$ almost surely as $N \to \infty$, therefore using continuity of matrix inversion and the continuous mapping theorem, we have $(A^{\psi,N}_t)^{-1} \to (A^\psi_t)^{-1}$ almost surely. Using continuity of the spectral matrix norm and another application of the continuous mapping theorem, we see that the minimum eigenvalue of $A^{\psi,N}_t$ converges to that of A^ψ_t, which is strictly positive under Assumption A7. Hence for sufficiently large values of N, we have invertibility of $A^{\psi,N}_t$ with probability one.

Starting with time $t = T$, by LLN $b^{\psi,N}_T \to b^\psi_T$ in probability, so by Slutsky's lemma it follows that $\beta^{\psi,N}_T \to \beta^\psi_T$ in probability. Consider the difference

$$\beta^{\psi,N}_T - \beta^\psi_T = (A^{\psi,N}_T)^{-1}(b^{\psi,N}_T - A^{\psi,N}_T\beta^\psi_T) = ((A^\psi_T)^{-1} + o_p(1))\,(b^{\psi,N}_T - A^{\psi,N}_T\beta^\psi_T).$$

Since $(A^\psi_T)^{-1}(b^{\psi,N}_T - A^{\psi,N}_T\beta^\psi_T) = \nu^{\psi,N}_T(\xi^\psi_T)$ and $\nu^\psi_T(\xi^\psi_T) = 0_M$ with ξ^ψ_T defined in (S7), it follows from (S8) that $b^{\psi,N}_T - A^{\psi,N}_T\beta^\psi_T = O_p(N^{-1/2})$. Therefore

$$\beta^{\psi,N}_T - \beta^\psi_T = \nu^{\psi,N}_T(\xi^\psi_T) + o_p(N^{-1/2}) \tag{S11}$$

and applying the central limit theorem (CLT) in Assumption A11 gives

$$\sqrt{N}\left(\beta^{\psi,N}_T - \beta^\psi_T\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0_M, \Gamma^\psi_{T+1,T+1}\right)$$

where $\Gamma^\psi_{T+1,T+1} \in \mathbb{R}^{M\times M}$ refers to the lowest right block of Γ^ψ.

We now argue inductively: for time $t \in [0 : T-1]$, we decompose $b^{\psi,N}_t = c^{\psi,N}_t + d^{\psi,N}_t$ where

$$c^{\psi,N}_t := -\nu^{\psi,N}_t(\{\log G^\psi_t + \log M^\psi_{t+1}(\exp(-\Phi^T_{t+1}\beta^\psi_{t+1}))\}\Phi_t),$$
$$d^{\psi,N}_t := \nu^{\psi,N}_t(\{\log M^\psi_{t+1}(\exp(-\Phi^T_{t+1}\beta^\psi_{t+1})) - \log M^\psi_{t+1}(\exp(-\Phi^T_{t+1}\beta^{\psi,N}_{t+1}))\}\Phi_t).$$

Assumption A8 implies $c_t^{\psi,N} \to b_t^{\psi}$ in probability. If $\beta_{t+1}^{\psi,N} \to \beta_{t+1}^{\psi}$ in probability, by Assumption A9 we have

$$|d_t^{\psi,N}| \leq \nu_t^{\psi,N}(C_t|\Phi_t|)\delta_t(|\beta_{t+1}^{\psi,N} - \beta_{t+1}^{\psi}|) = o_p(1),$$

hence $\beta_t^{\psi,N} \to \beta_t^{\psi}$ in probability. We now examine the difference

$$\beta_t^{\psi,N} - \beta_t^{\psi} = ((A_t^{\psi})^{-1} + o_p(1))\,(c_t^{\psi,N} + d_t^{\psi,N} - A_t^{\psi,N}\beta_t^{\psi}). \tag{S12}$$

Since $(A_t^{\psi})^{-1}(c_t^{\psi,N} - A_t^{\psi,N}\beta_t^{\psi}) = \nu_t^{\psi,N}(\xi_t^{\psi})$ and $\nu_t^{\psi}(\xi_t^{\psi}) = 0_M$ with ξ_t^{ψ} defined in (S7), it follows from (S8) that $c_t^{\psi,N} - A_t^{\psi,N}\beta_t^{\psi} = O_p(N^{-1/2})$. To study the term $d_t^{\psi,N}$, we use Assumption A10(i) and apply Taylor's theorem to obtain

$$d_t^{\psi,N} = -\nu_t^{\psi,N}((\beta_{t+1}^{\psi,N} - \beta_{t+1}^{\psi})^T g_{t+1}^{\psi}(\beta_{t+1}^{\psi}, \cdot)\Phi_t) + r_t^{\psi,N}$$

with remainder

$$r_t^{\psi,N} = -\nu_t^{\psi,N}\left((\beta_{t+1}^{\psi,N} - \beta_{t+1}^{\psi})^T \left[g_{t+1}^{\psi}(\tilde{\beta}_{t+1}^{N}, \cdot) - g_{t+1}^{\psi}(\beta_{t+1}^{\psi}, \cdot)\right]\Phi_t\right)$$

for some $\tilde{\beta}_{t+1}^{N}$ lying on the line segment between $\beta_{t+1}^{\psi,N}$ and β_{t+1}^{ψ}. Applying Assumption A10(ii) we have that

$$\begin{aligned}\left|r_t^{\psi,N}\right| &\leq |\tilde{\beta}_{t+1}^{\psi,N} - \beta_{t+1}^{\psi}||\beta_{t+1}^{\psi,N} - \beta_{t+1}^{\psi}|\nu_t^{\psi,N}\left(C_t'(\cdot)|\Phi_t|\right)\\ &\leq |\beta_{t+1}^{\psi,N} - \beta_{t+1}^{\psi}|^2\nu_t^{\psi,N}\left(C_t'(\cdot)|\Phi_t|\right)\\ &= |\beta_{t+1}^{\psi,N} - \beta_{t+1}^{\psi}|^2\left[\nu_t^{\psi}\left(C_t'(\cdot)|\Phi_t|\right) + o_p(1)\right]\end{aligned}$$

where the second inequality follows from the definition of $\tilde{\beta}_{t+1}^{N}$ and the final equality by the LLN. By the inductive hypothesis we have that

$$\sqrt{N}\left(\beta_{t+1}^{\psi,N} - \beta_{t+1}^{\psi}\right) \xrightarrow{\mathtt{d}} \mathcal{N}\left(0_M, \Sigma_{t+1,t+1}^{\psi}\right)$$

for some $\Sigma_{t+1,t+1}^{\psi} \in \mathbb{R}^{M\times M}$, and since by assumption $\nu_t^{\psi}\left(C_t'(\cdot)|\Phi_t|\right) < \infty$ we conclude that $r_t^{\psi,N} = O_p(N^{-1})$. From Assumption A10(ii) and the LLN we conclude that $d_t^{\psi,N} = O_p(N^{-1/2})$ and we can thus write

$$(A_t^{\psi})^{-1}d_t^{\psi,N} = E_t^{\psi}(\beta_{t+1}^{\psi,N} - \beta_{t+1}^{\psi}) + o_p(N^{-1/2})$$

where $E_t^{\psi} := -(A_t^{\psi})^{-1}\nu_t^{\psi}(\Phi_t g_{t+1}^{\psi}(\beta_{t+1}^{\psi}, \cdot)^T)$. Combining these observations with (S12) gives

$$\beta_t^{\psi,N} - \beta_t^{\psi} - E_t^{\psi}(\beta_{t+1}^{\psi,N} - \beta_{t+1}^{\psi}) = \nu_t^{\psi,N}(\xi_t^{\psi}) + o_p(N^{-1/2}). \tag{S13}$$

Stacking (S13) for $t \in [0 : T-1]$ and (S11) as a $(T+1)M$-dimensional vector yields

$$\zeta^{\psi,N} := \begin{pmatrix} (\beta_0^{\psi,N} - \beta_0^{\psi}) - E_0^{\psi}(\beta_1^{\psi,N} - \beta_1^{\psi}) \\ (\beta_1^{\psi,N} - \beta_1^{\psi}) - E_1^{\psi}(\beta_2^{\psi,N} - \beta_2^{\psi}) \\ \vdots \\ (\beta_{T-1}^{\psi,N} - \beta_{T-1}^{\psi}) - E_{T-1}^{\psi}(\beta_T^{\psi,N} - \beta_T^{\psi}) \\ \beta_T^{\psi,N} - \beta_T^{\psi} \end{pmatrix} = \nu^{\psi,N}(\xi^{\psi}) + o_p(N^{-1/2}).$$

Noting that the block matrix U^{ψ} defined in (S10) is such that $U^{\psi}\zeta^{\psi,N} = \beta^{\psi,N} - \beta^{\psi}$ for any $N \in \mathbb{N}$, (S9) follows from the CLT in Assumption A11 and an application of the continuous mapping theorem. □

We first make some remarks about the assumptions required in Proposition 2. Assumptions A7 and A8 ensure that the least squares estimators converge to a well-defined limit. Assumptions A9 and A10 are made to deal with the intractability of the function $(\beta, x) \mapsto \log M_{t+1}^{\psi}(\exp(-\Phi_{t+1}^T\beta))(x)$, which can be verified when its form is known. Lastly, Assumption A11, which asserts existence of a path central limit theorem for the function (S7), can be deduced in the case of multinomial resampling from [4, Theorem 9.7.1]. In the following, we will write $A_{s,t} \in \mathbb{R}^{M\times M}$ to denote the $s, t \in [1 : T+1]$ submatrix of a block matrix $A \in \mathbb{R}^{(T+1)M\times(T+1)M}$.

Theorem 1. *Consider ADP algorithm (31), with current policy $\psi \in \Psi$, under linear least squares approximations (38) with basis functions $(\Phi_t)_{t\in[0:T]}$ chosen so that Assumptions A7-A11 in Proposition 2 are satisfied. Then as $N \to \infty$, for all $x \in \mathsf{X}^{2T+1}$, the estimated policy $\hat{\phi}(x)$ converges in probability to the policy $\tilde{\phi}(x)$ generated by the idealized algorithm (37). Moreover, for all $x \in \mathsf{X}^{2T+1}$, we have*

$$\sqrt{N}\left(\hat{\phi}(x) - \tilde{\phi}(x)\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0_{(T+1)}, \Omega^{\psi}(x)\right), \tag{S14}$$

where $\Omega^{\psi} : \mathsf{X}^{2T+1} \to \mathbb{R}^{(T+1)\times(T+1)}$ is given by

$$\Omega_{s,t}^{\psi} = \tilde{\phi}_s\tilde{\phi}_t\Phi_s^T \sum_{k=s}^{T+1}\sum_{\ell=t}^{T+1} U_{s,k}^{\psi}\Gamma_{k,\ell}^{\psi}(U_{\ell,t}^{\psi})^T\Phi_t \tag{S15}$$

for $s, t \in [1 : T+1]$.

Proof of Theorem 1. Appealing to the continuous mapping theorem allows us to conclude from Proposition 2 that $\hat{\phi}_t$ converges (pointwise) in probability to $\tilde{\phi}_t := \exp(-\Phi_t^T \beta_t^{\psi}), t \in [0:T]$. Applying the delta method on (S9) establishes that the (pointwise) fluctuations satisfy (S14), where $\Omega_{s,t}^{\psi} = \tilde{\phi}_s \tilde{\phi}_t \Phi_s^T \Sigma_{s,t}^{\psi} \Phi_t$ for $s,t \in [1:T{+}1]$. The form of the asymptotic variance (S15) follows from the block upper triangular structure of (S10). □

4. Proofs of Section 5.3.

Proof of Theorem 5.2. Under Assumptions A4 and A5, existence of a unique invariant distribution $\pi \in \mathcal{P}(\Psi)$ and geometric convergence (40) follow from [5, Theorem 1.1]. Let φ^* denote a fixed point of F and define the backward process $\varphi^{(I)} = F_{U^{(1)}}^N \circ \cdots \circ F_{U^{(I)}}^N(\varphi^*)$ for $I \in \mathbb{N}$. Noting from [5, Proposition 1.1] that the limit $\varphi^{(\infty)} := \lim_{I\to\infty} \varphi^{(I)}$ does not depend on φ^* and is distributed according to π, we shall construct the random policy $\psi \sim \pi$ by taking $\psi = \varphi^{(\infty)}$.

By the triangle inequality,

$$\rho(\psi, \varphi^*) \leq \rho(\varphi^{(\infty)}, \varphi^{(I)}) + \rho(\varphi^{(I)}, \varphi^*) \tag{S16}$$

for any $I \in \mathbb{N}$. To examine the first term in (S16), we consider the decomposition in the proof of [5, Proposition 5.1]:

$$\rho(\varphi^{(I+J)}, \varphi^{(I)}) \leq \sum_{i=0}^{J-1} \prod_{j=1}^{I+i} L_{U^{(j)}}^N \rho(F_{U^{(I+i+1)}}^N(\varphi^*), \varphi^*)$$

for $I, J \in \mathbb{N}$. By the monotone convergence theorem, taking the limit $J \to \infty$ gives

$$\mathbb{E}\left[\rho(\varphi^{(\infty)}, \varphi^{(I)})\right] \leq \sum_{i=0}^{\infty} \prod_{j=1}^{I+i} \mathbb{E}\left[L_{U^{(j)}}^N\right] \mathbb{E}\left[\rho(F_{U^{(I+i+1)}}^N(\varphi^*), \varphi^*)\right].$$

Under Assumptions A4 and A5, it follows that $\zeta := \mathbb{E}\left[\rho(F_U^N(\varphi^*), \varphi^*)\right] < \infty$ since by the triangle inequality

$$\begin{aligned}\rho(F_U^N(\varphi^*), \varphi^*) &\leq \rho(F_U^N(\varphi^*), F_U^N(\varphi_0)) + \rho(F_U^N(\varphi_0), \varphi_0) + \rho(\varphi_0, \varphi^*) \\ &\leq (1 + L_U^N)\rho(\varphi^*, \varphi_0) + \rho(F_U^N(\varphi_0), \varphi_0).\end{aligned}$$

Applying Assumption A5, the triangle inequality and the fact that $\varphi^{(I)} \to \varphi^{(\infty)}$ as $I \to \infty$ establishes that

$$\mathbb{E}\left[\rho(\varphi^{(\infty)}, \varphi^{(I)})\right] \leq \sum_{j=0}^{\infty} \mathbb{E}\left[\rho(\varphi^{(I+j)}, \varphi^{(I+j+1)})\right] \leq \zeta\alpha^I(1-\alpha)^{-1}$$

and hence

$$\lim_{I\to\infty} \mathbb{E}\left[\rho(\varphi^{(\infty)}, \varphi^{(I)})\right] = 0. \tag{S17}$$

For the second term in (S16), using the fact that φ^* is a fixed point of F, the triangle inequality and Assumptions A5 and A6

$$\begin{aligned}
\rho(\varphi^{(I)}, \varphi^*) &= \rho(F^N_{U^{(1)}} \circ \cdots \circ F^N_{U^{(I)}}(\varphi^*), F(\varphi^*)) \\
&\leq \sum_{i=1}^{I} \rho(F^N_{U^{(1)}} \circ \cdots \circ F^N_{U^{(i)}}(\varphi^*), F^N_{U^{(1)}} \circ \cdots \circ F^N_{U^{(i-1)}} \circ F(\varphi^*)) \\
&\leq \sum_{i=1}^{I} \prod_{j=1}^{i-1} L^N_{U^{(j)}} \rho(F^N_{U^{(i)}}(\varphi^*), F(\varphi^*)) \\
&\leq N^{-1/2} \sum_{i=1}^{I} \prod_{j=1}^{i-1} L^N_{U^{(j)}} E^{\varphi^*,N}_{U^{(i)}}
\end{aligned}$$

with the convention that $(F^N_{U^{(1)}} \circ F^N_{U^{(0)}})(\varphi) = \varphi$. Taking expectations and the limit $I \to \infty$ gives

$$\lim_{I\to\infty} \mathbb{E}\left[\rho(\varphi^{(I)}, \varphi^*)\right] \leq N^{-1/2}\mathbb{E}\left[E^{\varphi^*,N}_{U}\right](1-\alpha)^{-1}. \tag{S18}$$

Combining (S16), (S17) and (S18) allows us to conclude (41). □

The following discussion offers some insights into when and why contraction (Assumption A5) happens. Let ρ denote a metric under which the set of all admissible policies Ψ is a complete separable metric space. Let ψ^* denote the optimal policy w.r.t. $\mathbb{Q}$ that we want to approximate. Given two policies $\varphi, \xi \in \Psi$, by triangle inequality, the ADP algorithm $F^N : \mathsf{U} \times \Psi \to \Psi$ satisfies

$$\begin{aligned}
\rho(F^N_U(\varphi), F^N_U(\xi)) &\leq \rho(F^N_U(\varphi), F(\varphi)) + \rho(F^N_U(\xi), F(\xi)) + \rho(F(\varphi), F(\xi)) \\
&\leq \rho(F^N_U(\varphi), F(\varphi)) + \rho(F^N_U(\xi), F(\xi)) + \rho(F(\varphi), \psi^*) + \rho(F(\xi), \psi^*) \qquad \text{(S19)}
\end{aligned}$$

where $F : \Psi \to \Psi$ denotes the idealized ADP algorithm with exact projections. We consider the first and second terms of (S19) that concern the Monte Carlo error of the ADP algorithm. Under Assumption A6, we have

$$\rho(F^N_U(\varphi), F(\varphi)) + \rho(F^N_U(\xi), F(\xi)) \leq N^{-1/2}\left(E^{\varphi,N}_U + E^{\xi,N}_U\right) \tag{S20}$$

where $(E^{\varphi,N}_U)_{N\in\mathbb{N}}$ and $(E^{\xi,N}_U)_{N\in\mathbb{N}}$ are uniformly integrable sequences of non-negative random variables with finite mean that converge in distribution to a limit with support on $\mathbb{R}_+$. Assumption A6 is necessary to quantify the Monte Carlo error involved when employing approximate projections and can be deduced for example using the central limit theorem in Theorem 1. The third and fourth terms of (S19) concern the mis-specification error of the chosen function classes. In particular, we have

$$\rho(F(\varphi),\psi^*)+\rho(F(\xi),\psi^*)=\rho(\tilde{\varphi},\varphi^*)+\rho(\tilde{\xi},\xi^*) \tag{S21}$$

where $\tilde{\varphi}$ and $\tilde{\xi}$ denote idealized ADP approximations of the optimal policies φ^* and ξ^* w.r.t. $\mathbb{Q}^{\varphi}$ and $\mathbb{Q}^{\xi}$ respectively. In the well-specified case, by consistency of least squares, the errors $e(\varphi):=\rho(\tilde{\varphi},\varphi^*)$ and $e(\xi):=\rho(\tilde{\xi},\xi^*)$ would be equal to zero.

Combining (S19), (S20) and (S21) gives

$$\rho(F_U^N(\varphi),F_U^N(\xi))\leq N^{-1/2}\left(E_U^{\varphi,N}+E_U^{\xi,N}\right)+e(\varphi)+e(\xi).$$

Therefore if $\rho(\varphi,\xi)\geq 1$, we have

$$\rho(F_U^N(\varphi),F_U^N(\xi))\leq L_U^N\rho(\varphi,\xi) \tag{S22}$$

with

$$L_U^N=N^{-1/2}\left(E_U^{\varphi,N}+E_U^{\xi,N}\right)+e(\varphi)+e(\xi). \tag{S23}$$

If Ψ is compact, which may be imposed by truncating our least squares estimators, the expectation of (S23) is bounded by

$$\begin{aligned}\mathbb{E}\left[L_U^N\right]&=N^{-1/2}\left\{\mathbb{E}\left[E_U^{\varphi,N}\right]+\mathbb{E}\left[E_U^{\xi,N}\right]\right\}+e(\varphi)+e(\xi)\\&\leq 2N^{-1/2}\sup_{\varphi\in\Psi}\mathbb{E}\left[E_U^{\varphi,N}\right]+2\sup_{\varphi\in\Psi}e(\varphi).\end{aligned}$$

In the well-specified case, we have $\sup_{\varphi\in\Psi}e(\varphi)=0$ so $\mathbb{E}[L_U^N]<1$ when the number of particles N is sufficiently large. In the mis-specified case, we also require that the mis-specification error $\sup_{\varphi\in\Psi}e(\varphi)$ be sufficiently small. If $\varphi=\xi$, (S22) also holds since $\rho(F_U^N(\varphi),F_U^N(\xi))=0$. Although the above arguments explain why one can expect contraction for policies φ and ξ that are distant, it does not capture the case $\rho(\varphi,\xi)\in(0,1)$, corresponding to when these policies are close.

The following example illustrates contraction in a simple setting with mis-specification.

Example 1. *Let the state space be $\mathsf{X} = [0,1]$, equipped with its Borel σ-algebra $\mathcal{X} = \mathfrak{B}([0,1])$. For simplicity, we consider a single time step and an initial distribution μ that is given by the uniform distribution on X. The potential function of interest is $G_0(x_0) = \exp(-x_0^2)$.*

The function class we specify for the ADP algorithm is $\mathsf{F} = \{\varphi(x) = ax : a \in \mathbb{R}\}$. Given a current policy $\psi_0(x_0) = \exp(-a_0 x_0)$, the ψ-twisted SMC method (Algorithm 1) will sample N independent samples $X_0^n \sim \mu^\psi$ for $n \in [1:N]$. In terms of independent uniform random variables $(U_0^n)_{n\in[1:N]}$, these samples can be generated using

$$X_0^n = -{a_0}^{-1} \log\left(1 - U_0^n\{1 - \exp(-a_0)\}\right). \tag{S24}$$

In this setting, ADP (Algorithm 2) would consider the following least squares problem

$$\begin{aligned} \alpha_0 &= \arg\min_{\varphi\in\mathsf{F}} \sum_{n=1}^N \left(\varphi(X_0^n) + \log G_0(X_0^n) - \log\psi_0(X_0^n)\right)^2 \\ &= \arg\min_{\alpha\in\mathbb{R}} \sum_{n=1}^N \left(\alpha X_0^n - \{(X_0^n)^2 - a_0 X_0^n\}\right)^2 \\ &= \frac{\sum_{n=1}^N (X_0^n)^3}{\sum_{n=1}^N (X_0^n)^2} - a_0. \end{aligned}$$

Therefore, the ADP algorithm can be represented as the iterated random function $F_U^N(a_0) = a_0 + \alpha_0$. By considering $a_0 = 0$, we see that Assumption A4 is satisfied since

$$\mathbb{E}\left[F_U^N(0)\right] = \mathbb{E}\left[\frac{\sum_{n=1}^N (U_0^n)^3}{\sum_{n=1}^N (U_0^n)^2}\right] < \infty.$$

As the number of particles $N \to \infty$,

$$F_U^N(a_0) \to \frac{\int_0^1 x_0^3 \exp(-a_0 x_0)\,\mathrm{d}x_0}{\int_0^1 x_0^2 \exp(-a_0 x_0)\,\mathrm{d}x_0} =: F(a_0). \tag{S25}$$

The limiting function F corresponds to the idealized ADP algorithm with exact projections. We note that Assumption A6 holds since the above convergence rate is $O(N^{-1/2})$ by the central limit theorem. In the left panel of Figure S1, we illustrate the distribution of the iterates $a_0^{(i)} = F_U^N(a_0^{(i-1)})$ with initialization $a_0^{(0)} = 0$ for different number of particles. This plot shows how *the estimates produced by the ADP algorithm concentrate around the fixed point iteration defined by F.*

We now turn our attention to Assumption A5. The derivative of F_U^N with respect to a_0 is

$$\frac{\mathrm{d}}{\mathrm{d}a_0}F_U^N(a_0) = \frac{\sum_{n=1}^N 3(X_0^n)^2 d_0^n}{\sum_{n=1}^N (X_0^n)^2} - \frac{\sum_{n=1}^N (X_0^n)^3 \cdot \sum_{n=1}^N 2X_0^n d_0^n}{\left\{\sum_{n=1}^N (X_0^n)^2\right\}^2}$$

where

$$d_0^n = \frac{U_0^n \exp(-a_0)}{a_0(1-U_0^n\{1-\exp(-a_0)\})} + {a_0}^{-2}\log\left(1-U_0^n\{1-\exp(-a_0)\}\right)$$

denotes the derivative of (S24) with respect to a_0. As $N\to\infty$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}a_0}F_U^N(a_0) \to\ & \frac{\int_0^1 3x_0^2 d(a_0,x_0)\exp(-a_0x_0)\,\mathrm{d}x_0}{\int_0^1 x_0^2\exp(-a_0x_0)\,\mathrm{d}x_0} \\ & - \frac{\int_0^1 x_0^3\exp(-a_0x_0)\,\mathrm{d}x_0 \cdot \int_0^1 2x_0 d(a_0,x_0)\exp(-a_0x_0)\,\mathrm{d}x_0}{\left\{\int_0^1 x_0^2\exp(-a_0x_0)\,\mathrm{d}x_0\right\}^2}\end{aligned} \tag{S26}$$

where

$$d(a_0,X_0^n) = a_0^{-1}\left\{\frac{\exp(-a_0)}{(1-\exp(-a_0))}(\exp(a_0X_0^n)-1) - X_0^n\right\} = d_0^n.$$

It is apparent from the right panel of Figure S1 that the idealized ADP algorithm with exact projection is a contraction. Moreover, for this particular example, the ADP algorithm is also a contraction (on average) even with a small number of particles.

5. Connection to Kullback-Leibler control. The Kullback-Leibler (KL) divergence from $\nu\in\mathcal{P}(\mathsf{E})$ to $\mu\in\mathcal{P}(\mathsf{E})$ is defined as $\mathrm{KL}(\mu|\nu) = \int_{\mathsf{E}}\log(\mathrm{d}\mu/\mathrm{d}\nu)(x)\mu(\mathrm{d}x)$ if the integral is finite and $\mu\ll\nu$, and $\mathrm{KL}(\mu|\nu)=\infty$ otherwise. The intent of this section is to show that ϕ^* defined in (19) is the optimal policy of an associated KL optimal control problem [15, 10]. Making this connection allows us to leverage existing methodology and analysis developed in the approximate dynamic programming literature [2, 16] in Sections 4.2 and 5.1 respectively.

Suppose that the current policy is $\psi\in\Psi$ and consider the following optimal control problem

$$\inf_{\phi\in\Phi}\mathrm{KL}\left((\mathbb{Q}^\psi)^\phi|\mathbb{P}\right) = \inf_{\phi\in\Phi}\mathbb{E}_{(\mathbb{Q}^\psi)^\phi}\left[C(X_{0:T})\right] \tag{S27}$$

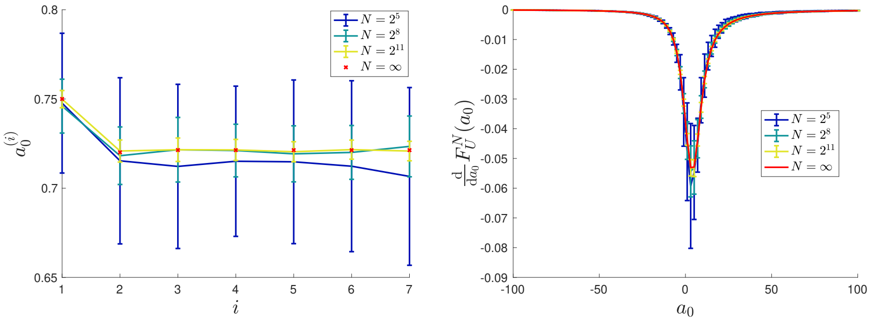

FIG S1. *Error bars illustrating the mean (± one standard deviation) of the iterates (*left*) and the derivative of the iterated random function (*right*) considered in Example 1. Each colour corresponds to a specific number of particles N in the ADP algorithm; the $N = \infty$ case corresponds to the expressions in (S25) and (S26).*

where the set of admissible policies for the control problem is

$$\Phi := \left\{ \phi \in \Psi : \mathrm{KL}\left((\mathbb{Q}^{\psi})^{\phi} | \mathbb{Q}^{\psi} \right) < \infty \right\}$$

and the cost functional $C : \mathsf{X}^{T+1} \to \mathbb{R}$ can be written as

$$C(x_{0:T}) := \log \frac{\mathrm{d}(\mathbb{Q}^{\psi})^{\phi}}{\mathrm{d}\mathbb{Q}^{\psi}}(x_{0:T}) - \log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}^{\psi}}(x_{0:T}). \tag{S28}$$

Using properties of KL divergence, it follows from Property 1 of Proposition 4.1 that ϕ^* defined in (19) solves the optimal control problem (S27). Rewriting (S28) gives

$$\mathbb{E}_{(\mathbb{Q}^{\psi})^{\phi}}\left[C(X_{0:T})\right] = \mathrm{KL}\left((\mu^{\psi})^{\phi} | \mu^{\psi} \right) + \sum_{t=1}^{T} \mathbb{E}_{(\mathbb{Q}^{\psi})^{\phi}}\left[\mathrm{KL}\left((M_t^{\psi})^{\phi} | M_t^{\psi})(X_{t-1}) \right) \right]$$
$$- \mathbb{E}_{(\mu^{\psi})^{\phi}}\left[\log G_0^{\psi}(X_0) \right] - \sum_{t=1}^{T} \mathbb{E}_{(\mathbb{Q}^{\psi})^{\phi}}\left[\log G_t^{\psi}(X_{t-1}, X_t) \right] + \log Z.$$

We shall henceforth redefine the cost functional (S28) to remove the intractable constant $\log Z$ that does not affect the minimizer of (S27).

Given a policy $\phi \in \Phi$, the corresponding value functions $V^{\phi} = (V_t^{\phi})_{t \in [0:T]}$ of the control problem are given by the expected cost-to-go from a fixed time and state (see for example [2, Section 2.1])

$$V_0^{\phi}(x_0) := \mathrm{KL}\left((M_1^{\psi})^{\phi}|M_1^{\psi})(x_0)\right) + \sum_{s=1}^{T-1} \mathbb{E}^{0,x_0}_{(\mathbb{Q}^{\psi})^{\phi}}\left[\mathrm{KL}\left((M_{s+1}^{\psi})^{\phi}|M_{s+1}^{\psi})(X_s)\right)\right]$$

(S29)

$$- \log G_0^{\psi}(x_0) - \sum_{s=1}^{T} \mathbb{E}^{0,x_0}_{(\mathbb{Q}^{\psi})^{\phi}}\left[\log G_s^{\psi}(X_{s-1}, X_s)\right],$$

$$V_t^{\phi}(x_{t-1}, x_t) := \mathrm{KL}\left((M_{t+1}^{\psi})^{\phi}|M_{t+1}^{\psi})(x_t)\right) + \sum_{s=t+1}^{T-1} \mathbb{E}^{t,x_t}_{(\mathbb{Q}^{\psi})^{\phi}}\left[\mathrm{KL}\left((M_{s+1}^{\psi})^{\phi}|M_{s+1}^{\psi})(X_s)\right)\right]$$

$$- \log G_t^{\psi}(x_{t-1}, x_t) - \sum_{s=t+1}^{T} \mathbb{E}^{t,x_t}_{(\mathbb{Q}^{\psi})^{\phi}}\left[\log G_s^{\psi}(X_{s-1}, X_s)\right], \quad t \in [1 : T-1],$$

$$V_T^{\phi}(x_{T-1}, x_T) := -\log G_T^{\psi}(x_{T-1}, x_T).$$

In this notation, the total value of policy ϕ is given by

$$v(\phi) := (\mu^{\psi})^{\phi}(V_0^{\phi}) + \mathrm{KL}\left((\mu^{\psi})^{\phi}|\mu^{\psi}\right) = \mathrm{KL}\left((\mathbb{Q}^{\psi})^{\phi}|\mathbb{P}\right) - \log Z.$$

We now define the optimal value v^* and optimal value functions $V^* = (V_t^*)_{t\in[0:T]}$ w.r.t. $\mathbb{Q}^{\psi}$ by taking the infimum over the set Φ

(S30)
$$\begin{aligned} v^* &:= \inf_{\phi} v(\phi), \\ V_0^*(x_0) &:= \inf_{\phi_s, s\in[1:T]} V_0^{\phi}(x_0), \\ V_t^*(x_{t-1}, x_t) &:= \inf_{\phi_s, s\in[t+1:T]} V_t^{\phi}(x_{t-1}, x_t), \quad t \in [1 : T-1], \\ V_T^*(x_{T-1}, x_T) &:= -\log G_T^{\psi}(x_{T-1}, x_T), \end{aligned}$$

and denote the minimizer (if it exists) as $\phi^* = (\phi_t)_{t\in[0:T]}$. We stress the dependence of both V^* and ϕ^* on the current policy $\psi \in \Psi$ as it is omitted notationally. These minimization problems can be solved using a backward dynamic programming approach. From (S29) and (S30), we have the dynamic programming recursion

(S31)

$$\begin{aligned}
&V_T^*(x_{T-1},x_T) = -\log G_T^{\psi}(x_{T-1},x_T),\\
&V_t^*(x_{t-1},x_t) = -\log G_t^{\psi}(x_{t-1},x_t) + \inf_{\phi_{t+1}} \Big\{ (M_{t+1}^{\psi})^{\phi}(V_{t+1}^*)(x_t)\\
&\qquad\qquad + \mathrm{KL}\left((M_{t+1}^{\psi})^{\phi} | M_{t+1}^{\psi} \right)(x_t) \Big\}, \quad t \in [1:T-1],\\
&V_0^*(x_0) = -\log G_0^{\psi}(x_0) + \inf_{\phi_1} \left\{ (M_1^{\psi})^{\phi}(V_1^*)(x_0) + \mathrm{KL}\left((M_1^{\psi})^{\phi} | M_1^{\psi} \right)(x_0) \right\},\\
&v^* = \inf_{\phi_0} \left\{ (\mu^{\psi})^{\phi}(V_0^*) + \mathrm{KL}\left((\mu^{\psi})^{\phi} | \mu^{\psi} \right) \right\}.
\end{aligned}$$

The above is commonly referred to as the discrete time Bellman recursion.

Owing to the use of KL costs, the minimizations in (S31) are tractable: assuming that the current policy $\psi \in \Psi$ satisfies $\mathrm{KL}(\mathbb{P}|\mathbb{Q}^{\psi}) < \infty$, applying [3, Proposition 2.3] gives

(S32)

$$\begin{aligned}
&V_T^*(x_{T-1},x_T) = -\log G_T^{\psi}(x_{T-1},x_T),\\
&V_t^*(x_{t-1},x_t) = -\log G_t^{\psi}(x_{t-1},x_t) - \log M_{t+1}^{\psi}(e^{-V_{t+1}^*})(x_t), \quad t \in [1:T-1],\\
&V_0^*(x_0) = -\log G_0^{\psi}(x_0) - \log M_1^{\psi}(e^{-V_1^*})(x_0),\\
&v^* = -\log \mu^{\psi}(e^{-V_0^*}) = -\log Z,
\end{aligned}$$

with infimum attained at $\phi_t^* = e^{-V_t^*}$ for $t \in [0:T]$. Observe that the optimal value functions are simply logarithmic transformations of the optimal policy, and the dynamic programming recursion (S32) corresponds to (19) in logarithmic scale. The optimal value is $v^* = -\log Z$ as we have adjusted the cost functional (S28). Lastly, the finite KL condition guarantees existence of a unique minimizer ϕ^* that lies in Φ. It should be clear from Proposition 4.1 that working with the subset $\Phi \subset \Psi$ is not necessary, i.e. such a condition is only required when we formulate ϕ^* as the optimal policy of a Kullback-Leibler control problem.

6. A non-linear multimodal state space model.

We consider a popular toy non-linear state space model [7, 12] which corresponds to working on $(\mathsf{X}, \mathcal{X}) = (\mathbb{R}, \mathfrak{B}(\mathbb{R})), \mathsf{Y} = \mathbb{R}$ and having

$$\begin{aligned} \nu(\mathrm{d}x_0) &= \mathcal{N}(x_0; 0, 5)\mathrm{d}x_0, \\ f_t(x_{t-1}, \mathrm{d}x_t) &= \mathcal{N}\left(x_t; \frac{1}{2}x_{t-1} + \frac{25x_{t-1}}{1+x_{t-1}^2} + 8\cos(1.2t), \sigma_f^2\right)\mathrm{d}x_t, \\ g_t(x_t, y_t) &= \mathcal{N}\left(y_t; \frac{1}{20}x_t^2, \sigma_g^2\right), \end{aligned} \tag{S33}$$

for $t \in [1:T]$, where $\theta = (\sigma_f^2, \sigma_g^2) \in \mathbb{R}_+ \times \mathbb{R}_+$. We will employ the BPF as uncontrolled SMC, i.e. set $\mu = \nu$ and $M_t = f_t$ for $t \in [1:T]$. As the smoothing distribution (5) is highly multimodal, owing to the uncertainty of the sign of the latent process, this example is commonly used as a benchmark to assess the performance of SMC methods. Moreover, we observe from Figure S2 that this problem also induces complex multimodal optimal policies.

6.1. *Approximate dynamic programming.* To approximate these policies, we rely on the following flexible function classes

$$\mathsf{F}_t = \left\{ \varphi(x_t) = -\log\left(\sum_{m=1}^{M} \alpha_{t,m} \exp\left(-\beta_t(x_t - \xi_{t,m})^2\right)\right) \right. \\ \left. : (\alpha_{t,m}, \beta_t, \xi_{t,m}) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R},\, m \in [1:M] \right\}$$

for all $t \in [0:T]$, which corresponds to a radial basis function (RBF) approximation of the optimal policy in the natural scale. With this choice of function classes, the approximate projections (31) can be implemented using non-linear least squares.

Given the output of a twisted SMC method based on the current policy, we adopt the following approach which is computationally more efficient. Firstly, we fix β_t as a pre-specified bandwidth factor $\tau \in \mathbb{R}_+$ multiplied by the sample standard deviation of particles $(X_t^n)_{n\in[1:N]}$ at time $t \in [0:T]$. Instead of performing the above logarithmic projections to learn the associated value functions, we fit the RBF approximation directly at the natural scale with $\xi_{t,n} = X_t^n$ for $n \in [1:N]$, as this can be efficiently implemented [13, p. 161] as a linear least squares problem with non-negativity constraints in $(\alpha_{t,n})_{n\in[1:N]}$. We note that care has to be taken to ensure that these computations are numerically stable. We then sort the estimated weights $(\alpha_{t,n})_{n\in[1:N]}$ and keep as knots $(\xi_{t,m})_{m\in[1:M]}$ particles with the M largest weights, as this avoids having to retain components with low weights. This selection procedure allows us to adaptively focus our computational effort on approximating the optimal policy at appropriate regions of the state space.

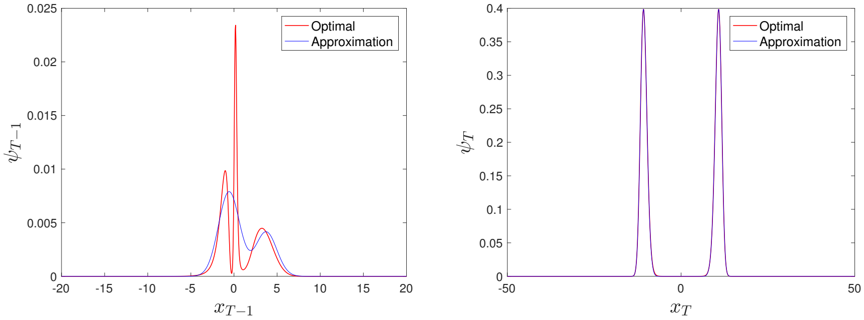

FIG S2. *Optimal policy of non-linear multimodal state space model (S33) at terminal times. The problem setting corresponds to $T = 100, \sigma_f^2 = 10, \sigma_g^2 = 1$ and the algorithmic settings of cSMC is $I = 1, N = 512, M = 16$.*

Writing $(\alpha_{t,m}^{i+1})_{m\in[1:M]}$ as the weights, β_t^{i+1} as the bandwidth and $(\xi_{t,m}^{i+1})_{m\in[1:M]}$ as the knots estimated by cSMC at iteration $i \in [0 : I-1]$ for $t \in [0 : T]$, the policy $\psi^{(i)} = (\psi_t^{(i)})_{t\in[0:T]}$ at iteration $i \in [1 : I]$ has the form

$$\psi_t^{(i)}(x_t) = \sum_{m\in[1:M]^i} \alpha_{t,m}^{(i)} \exp\left(-\beta_t^{(i)}(x_t - \xi_{t,m}^{(i)})^2\right), \quad t \in [0 : T], \tag{S34}$$

where $m = (m_j)_{j\in[1:i]} \in [1 : M]^i$ is a multi-index, $\beta_t^{(i)} := \sum_{j=1}^i \beta_t^j$, $\xi_{t,m}^{(i)} := \sum_{j=1}^i \beta_t^j \xi_{t,m_j}^j / \beta_t^{(i)}$ and

$$\alpha_{t,m}^{(i)} := \prod_{j=1}^i \alpha_{t,m_j}^j \exp\left(-\sum_{j=1}^i \beta_t^j (\xi_{t,m_j}^j)^2 + \beta_t^{(i)}(\xi_{t,m}^{(i)})^2\right).$$

It follows that under policy (S34), the initial distribution $\mu^{\psi^{(i)}}$ is a mixture of Gaussian distributions, Markov transition kernels $(M_t^{\psi^{(i)}})_{t\in[1:T]}$ are given by mixtures of Gaussian transition kernels and evaluation of the twisted potentials $(G_t^{\psi^{(i)}})_{t\in[0:T]}$ defined in (13) is tractable; exact expressions are given in Section 9.1 of the Supplementary Material. Figure S2 shows that such a parameterization is flexible enough to provide an adequate approximation of the optimal policy.

6.2. *Comparison of algorithmic performance.* We investigate the use of cSMC when the observation noise is small, i.e. high signal-to-noise ratio, since this is the regime where BPF exhibits poor performance. To do so, we fix $\sigma_f^2 = 10$ and simulate three sets of observations $y_{0:T} \in \mathsf{Y}^{T+1}$ of

			Observation noise		
			$\sigma_g^2 = 0.1$	$\sigma_g^2 = 0.5$	$\sigma_g^2 = 1$
Algorithm	BPF	N	2252	4710	6553
		ESS%	15.07%	30.24%	39.12%
		$\log Z$	-281.6185 ± 1.0054	-262.9861 ± 0.6037	-250.6369 ± 0.2845
		RVAR	1.27×10^{-5}	5.27×10^{-6}	1.29×10^{-6}
	cSMC	I	1	1	1
		N	512	512	512
		M	16	16	16
		τ	0.5	0.4	0.3
		ESS%	82.35%	92.51%	94.66%
		$\log Z$	-281.1483 ± 0.2295	-262.7223 ± 0.2425	-250.6949 ± 0.1439
		RVAR	6.67×10^{-7} (**19.1**)	8.52×10^{-7} (**6.18**)	3.29×10^{-7} (**3.91**)

Table S1

Non-linear multimodal state space model (S33): algorithmic settings and performance of BPF and cSMC for each observation noise considered. Notationally, N refers to the number of particles, I is the number of iterations taken by cSMC, M denotes the number of components and τ the bandwidth factor used in the ADP approximation. Results were obtained using 100 *independent repetitions each of method. The shorthand* ESS% *denotes the percentage of effective sample size averaged over time and repetitions,* $\log Z$ *refers to the estimation of the normalizing constant in logarithmic scale (± a standard deviation),* RVAR *is the sample relative variance of these estimates over the repetitions. Shown in bold is the gain that cSMC offers relative to BPF.*

length $T + 1 = 100$ according to (S33) as σ_g^2 takes values in $\{0.1, 0.5, 1\}$. We use $N = 512$ particles in cSMC and $I = 1$ iteration as preliminary runs indicate that policy refinement under the parameterization (S34) provides little improvement, especially when additional computing time is taken into account. The number of particles in BPF is then chosen to match computational time. The number of components M and bandwidth factor τ were tuned using preliminary runs. These algorithmic settings and the results obtained in 100 independent repetitions of each method are summarized in Table S1. As expected, although the performance gains over BPF diminish as the observation noise increases, it can be substantial when σ_g^2 is small.

7. Linear quadratic Gaussian control. This section considers a Gaussian static model (Section 2.4) which will allow us to draw connections to concepts from the linear quadratic Gaussian (LQG) control literature [1]. Consider $\mu(\mathrm{d}x_0) = \mathcal{N}(x_0; \mu_0, \Sigma_0)\mathrm{d}x_0$ on $(\mathsf{X}, \mathcal{X}) = (\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$ and $\ell(x, y) = \exp(-(y - x)^T R^{-1}(y - x)/2)$ for some $y \in \mathsf{Y} = \mathbb{R}^d$ and symmetric positive definite $R \in \mathbb{R}^{d\times d}$. By conjugacy, the models (6) are Gaussian and for $t \in [0 : T]$ we have $\eta_t(\mathrm{d}x_t) = \mathcal{N}(x_t; \mu_t, \Sigma_t)\mathrm{d}x_t$ with

$$\mu_t := \Sigma_t(\Sigma_0^{-1}\mu_0 + \lambda_t R^{-1}y), \quad \Sigma_t := (\Sigma_0^{-1} + \lambda_t R^{-1})^{-1}$$

and

$$Z_t = \det(\Sigma_0)^{-1/2} \det(\Sigma_t)^{1/2} \exp\left(-\frac{1}{2} \left\{ \mu_0^T \Sigma_0^{-1} \mu_0 + \lambda_t y^T R^{-1} y - \mu_t^T \Sigma_t^{-1} \mu_t \right\} \right).$$

7.1. *Riccati equation.* We now show that the backward recursion (19) with ψ initialized as a policy of constant one functions can be performed exactly to obtain analytic expressions of the optimal policy w.r.t. $\mathbb{Q}$. First note that under the choice of forward and backward Markov transition kernels specified in Section 7 with pre-conditioner $\Gamma = I_d$, the potentials (7) have the form

$$\text{(S35)} \qquad -\log G_t(x_{t-1}, x_t) = x_t^T \tilde{A}_t x_t + x_t^T \tilde{b}_t + \tilde{c}_t + x_{t-1}^T \tilde{D}_t x_{t-1} + x_{t-1}^T \tilde{e}_t,$$

where

(S36)

$$\tilde{A}_t := \frac{1}{8} h \Sigma_t^{-2}, \quad \tilde{b}_t := -\frac{1}{4} h \Sigma_t^{-2} \mu_t, \quad \tilde{c}_t := \frac{1}{2}(\lambda_t - \lambda_{t-1}) y^T R^{-1} y,$$

$$\tilde{D}_t := -\frac{1}{8} h \Sigma_t^{-2} + \frac{1}{2}(\lambda_t - \lambda_{t-1}) R^{-1}, \quad \tilde{e}_t := -(\lambda_t - \lambda_{t-1}) R^{-1} y + \frac{1}{4} h \Sigma_t^{-2} \mu_t,$$

for $t \in [1:T]$. For sufficiently small step size, observe that dropping $O(h)$ terms in (S36) gives $\log G_t(x_{t-1}, x_t) \approx (\lambda_t - \lambda_{t-1}) \log \ell(x_{t-1}, y)$ which, as expected, recovers the AIS potentials (8). For notational convenience, we set $(\tilde{A}_0, \tilde{b}_0, \tilde{c}_0, \tilde{D}_0, \tilde{e}_0)$ as the zero matrix or vector of the appropriate size and write the mean of the Euler-Maruyama move as $x_{t-1} + h \nabla \log \eta_t(x_{t-1})/2 = P_t x_{t-1} + q_t$ with $P_t := I_d - h\Sigma_t^{-1}/2$ and $q_t := h \Sigma_t^{-1} \mu_t / 2$.

PROPOSITION 3. *The optimal policy $\psi^* = (\psi_t^*)_{t \in [0:T]}$ w.r.t. $\mathbb{Q}$ is given by*

(S37)

$$-\log \psi_0^*(x_0) = x_0^T A_0^* x_0 + x_0^T b_0^* + c_0^*,$$

$$-\log \psi_t^*(x_{t-1}, x_t) = x_t^T A_t^* x_t + x_t^T b_t^* + c_t^* + x_{t-1}^T D_t^* x_{t-1} + x_{t-1}^T e_t^*, \quad t \in [1:T],$$

where the coefficients $(A_t^*, b_t^*, c_t^*, D_t^*, e_t^*)_{t \in [0:T]}$ are determined by the back-

ward recursion

$$
\begin{aligned}
\text{(S38)}\qquad A_t^* &= \tilde{A}_t + \frac{1}{2}h^{-1}P_{t+1}(I_d - K_{t+1}^*)P_{t+1} + D_{t+1}^*,\\
b_t^* &= \tilde{b}_t + P_{t+1}K_{t+1}^* b_{t+1}^* + e_{t+1}^* + \frac{1}{2}P_{t+1}(I_d - K_{t+1}^*)\Sigma_{t+1}^{-1}\mu_{t+1},\\
c_t^* &= \tilde{c}_t + c_{t+1}^* - \frac{1}{2}\log\det(K_{t+1}^*) + \frac{1}{2}h^{-1}q_{t+1}^T q_{t+1}\\
&\quad - \frac{1}{2}h^{-1}(q_{t+1} - hb_{t+1}^*)^T K_{t+1}^*(q_{t+1} - hb_{t+1}^*),\\
D_t^* &= \tilde{D}_t,\\
e_t^* &= \tilde{e}_t,
\end{aligned}
$$

for $t \in [T-1:0]$, *with* $K_t^* := (I_d + 2hA_t^*)^{-1}, t \in [1:T]$ *and initialization at* $(A_T^*, b_T^*, c_T^*, D_T^*, e_T^*) = (\tilde{A}_T, \tilde{b}_T, \tilde{c}_T, \tilde{D}_T, \tilde{e}_T)$.

Proof. We proceed by induction. Clearly, (S37) holds for $t = T$ since $\psi_T^* = G_T$. Assume that (S37) holds for time $t+1$. The recursion (19) can be written as

$$
-\log\psi_t^*(x_{t-1}, x_t) = -\log G_t(x_{t-1}, x_t) - \log M_{t+1}(\psi_{t+1}^*)(x_t).
$$

Some manipulations yield

$$
\begin{aligned}
-\log M_{t+1}(\psi_{t+1}^*)(x_t) &= x_t^T\left(\frac{1}{2}h^{-1}P_{t+1}(I_d - K_{t+1}^*)P_{t+1} + D_{t+1}^*\right)x_t\\
&+ x_t^T\left(P_{t+1}K_{t+1}^* b_{t+1}^* + e_{t+1}^* + \frac{1}{2}P_{t+1}(I_d - K_{t+1}^*)\Sigma_{t+1}^{-1}\mu_{t+1}\right) - \frac{1}{2}\log\det(K_{t+1}^*)\\
&+ c_{t+1}^* + \frac{1}{2}h^{-1}\left\{q_{t+1}^T q_{t+1} - (q_{t+1} - hb_{t+1}^*)^T K_{t+1}^*(q_{t+1} - hb_{t+1}^*)\right\}.
\end{aligned}
$$

Adding this to (S35) establishes that $-\log\psi_t^*$ has the desired form (S37) and equating coefficients of the polynomial gives (S38). □

The backward recursion (S38) for the coefficients is analogous to the Riccati equation in the context of LQG control. To illustrate the behaviour of these coefficients, we set the prior as $\mu_0 = 0_d$, $\Sigma_0 = I_d$ and the likelihood as $y = (\xi, \dots, \xi)^T$ for some $\xi \in \mathbb{R}$ and $R_{i,j} = \delta_{i,j} + (1 - \delta_{i,j})\rho$ for $i, j \in [1:d]$ and some $\rho \in [-1, 1]$ (here $\delta_{i,j}$ denotes the Kronecker delta). The time evolution of these coefficients is plotted in the top row of Figure S3 for the problem setting $d = 2$, $\xi = 4$, $\rho = 0.8$. Noting that the optimal value of the

Kullback-Leibler control problem (S30) is

$$v^* = -\log Z = c_0^* + \frac{1}{2}\log\det(\Sigma_0) - \frac{1}{2}\log\det(K_0^*)$$
$$- \frac{1}{2}(\Sigma_0^{-1}\mu_0 - b_0^*)^T K_0^*(\Sigma_0^{-1}\mu_0 - b_0^*) + \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0$$

with $K_0^* := (\Sigma_0^{-1} + 2A_0^*)^{-1}$, the dominant contribution that the constant c_0^* has to v^* suggests that it is important to estimate the constants in (S37) to learn good policies. Moving from the bottom left to top left plot, observe that increasing the location parameter ξ from 1 to 4 increases the magnitude of $(b_t^*, e_t^*)_{t\in[0:T]}$ but leaves $(A_t^*, D_t^*)_{t\in[0:T]}$ unchanged. This behaviour is evident from the expressions of $(D_t^*, e_t^*)_{t\in[0:T]}$ and is unsuprising for $(A_t^*)_{t\in[0:T]}$ as the parameter ξ does not alter the 'structure' of the problem. The increase in the magnitude of $(b_t^*)_{t\in[0:T]}$ shows that the optimally controlled SMC method achieves the desired terminal distribution by initializing

$$\mu^{\psi^*}(\mathrm{d}x_0) = \mathcal{N}(x_0; K_0^*(\Sigma_0^{-1}\mu_0 - b_0^*), K_0^*)\mathrm{d}x_0 \tag{S39}$$

closer to the posterior distribution and taking larger drifts in

$$M_t^{\psi^*}(x_{t-1}, \mathrm{d}x_t) = \mathcal{N}\left(x_t; K_t^*(P_t x_{t-1} + q_t - hb_t^*), hK_t^*\right)\mathrm{d}x_t, \quad t \in [1:T]. \tag{S40}$$

Comparing the plots in the bottom row reveals that the off-diagonal elements of $(A_t^*, D_t^*)_{t\in[0:T]}$ vanish under independence. Therefore these terms should be taken into account for posterior distributions that are very correlated. Having obtained the optimal policy w.r.t. $\mathbb{Q}$ in a backward sweep, we may then simulate the optimally controlled SMC method in a forward pass. In Figure S4, we contrast the output of the uncontrolled SMC method with that of the optimally controlled.

7.2. *Approximate dynamic programming.* The ability to compute the optimal policy in this setting allows us to evaluate the effectiveness of ADP algorithm (31) under correct parameterization, i.e. select the function classes

$$\mathsf{F}_0 = \left\{\varphi(x_0) = x_0^T A_0 x_0 + x_0^T b_0 + c_0 : (A_0, b_0, c_0) \in \mathbb{S}_d \times \mathbb{R}^d \times \mathbb{R}\right\},$$
$$\mathsf{F}_t = \left\{\varphi(x_{t-1}, x_t) = x_t^T A_t x_t + x_t^T b_t + c_t + x_{t-1}^T D_t x_{t-1} + x_{t-1}^T e_t \right.$$
$$\left. : (A_t, b_t, c_t, D_t, e_t) \in \mathbb{S}_d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{S}_d \times \mathbb{R}^d\right\}, \quad t \in [1:T].$$

This choice corresponds to function classes of the form (38), hence we can use linear least squares to estimate the coefficients at each iteration of cSMC

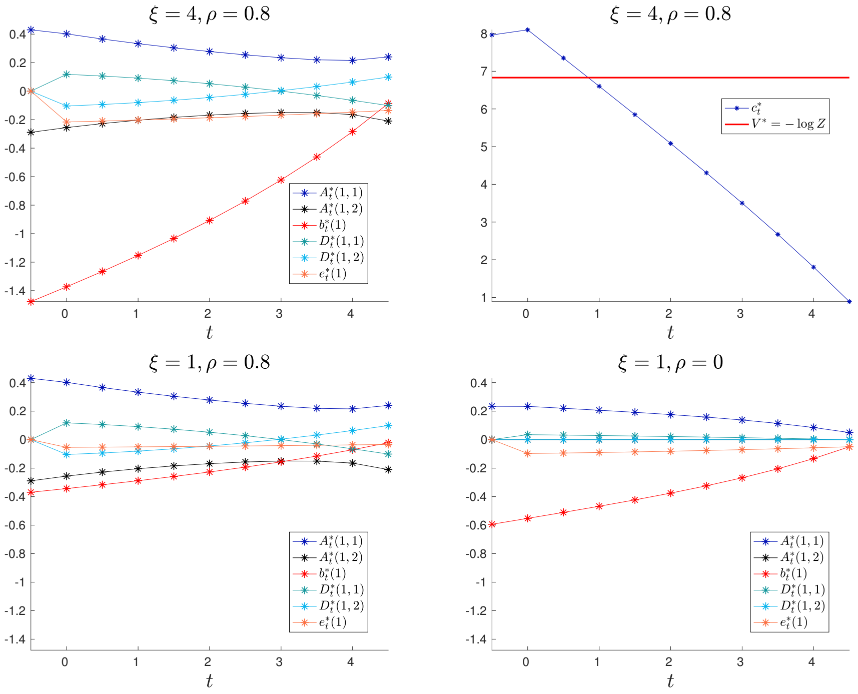

FIG S3. *Coefficients of the optimal policy w.r.t. $\mathbb{Q}$ in LQG control under various problem settings. The algorithmic settings of cSMC are $T = 10, h = 0.1, \lambda_t = t/T$. Note that all except the top right plot share the same axes.*

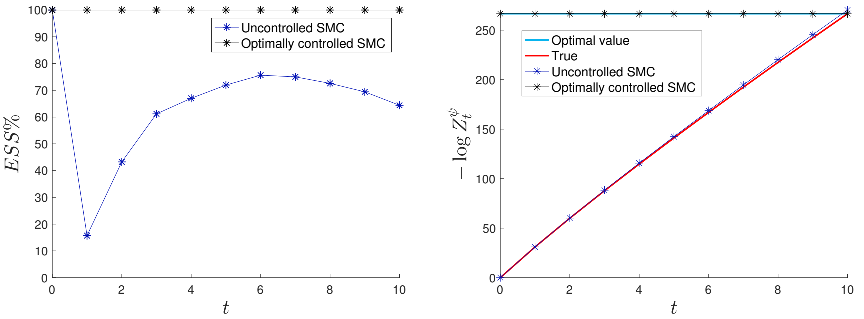

FIG S4. *Comparison of uncontrolled SMC and optimal LQG controlled SMC in terms of effective sample size* (left) *and normalizing constant estimation* (right)*. The problem setting considered here is $d = 32, \xi = 20, \rho = 0.8$ and the algorithmic settings of uncontrolled SMC are $N = 2048, T = 10, h = 0.1, \lambda_t = t/T$.*

– see (S4) and (S5). If $(A_t^{i+1}, b_t^{i+1}, c_t^{i+1}, D_t^{i+1}, e_t^{i+1})$ denote the coefficients estimated at iteration $i \in [0 : I-1]$ of Algorithm 3 in step 2(b), it follows that the policy at iteration $i \in [1 : I]$ is given by

$$\begin{aligned} -\log \psi_0^{(i)}(x_0) &= x_0^T A_0^{(i)} x_0 + x_0^T b_0^{(i)} + c_0^{(i)}, \\ -\log \psi_t^{(i)}(x_{t-1}, x_t) &= x_t^T A_t^{(i)} x_t + x_t^T b_t^{(i)} + c_t^{(i)} + x_{t-1}^T D_t^{(i)} x_{t-1} + x_{t-1}^T e_t^{(i)}, \end{aligned}$$

for $t \in [1 : T]$, where $A_t^{(i)} := \sum_{j=1}^i A_t^j, b_t^{(i)} := \sum_{j=1}^i b_t^j, c_t^{(i)} := \sum_{j=1}^i c_t^j, D_t^{(i)} := \sum_{j=1}^i D_t^j, e_t^{(i)} := \sum_{j=1}^i e_t^j$. Observe from (S39) and (S40) that we need to impose the following positive definite constraints

$$\Sigma_0^{-1} + 2A_0^{(i)} \succ 0, \quad I_d + 2hA_t^{(i)} \succ 0, \quad t \in [1 : T],$$

which can be done by projecting onto the set of real symmetric positive definite matrices [9]. In our numerical implementation, we find that these constraints are already satisfied when the step size h is sufficiently small. Although the computational complexity of this ADP procedure is $O(N)$, it scales quite costly in dimension d as computation of least squares estimators require inversion of $p \times p$ matrices where $p = d^2 + 3d + 1$. For problems with large d, it might be worth considering the use of iterative linear solvers which offer reduced complexity. We note that it is possible to avoid learning the x_{t-1} dependency in the policy $\psi_t^*(x_{t-1}, x_t), t \in [1 : T]$ and hence reduce computational complexity drastically; we do not exploit this observation here for simplicity of presentation but will do so for other applications.

Figure S5 displays the coefficients estimated by cSMC with $I = 2$ iterations. It is striking that with $N = 2048$ particles, we are able to accurately estimate, in a single ADP iteration, the true coefficients in dimension $d = 32$ (here $p = 1121$). That said, we typically need to increase N with d to prevent the Gram matrices (S4) from being ill-conditioned. Moreover, we find that it is unnecessary to perform policy refinement in this example, as the estimated policies converge immediately to an invariant distribution that is very concentrated around the optimal policy (S37), which is the fixed point of the idealized algorithm in Theorem 5.2 under correct parameterization. The performance of the resulting controlled SMC method is indistinguishable from that in Figure S4.

8. Bayesian logistic regression.

Consider a binary regression problem: each observation $y_m \in \{0, 1\}, m \in [1 : M]$ is modelled as an independent Bernoulli random variable with probability of success $\kappa(x^T X_m)$, where $\kappa(u) := (1 + \exp(-u))^{-1}$ for $u \in \mathbb{R}$ is the logistic link function, $x \in \mathsf{X} = \mathbb{R}^d$

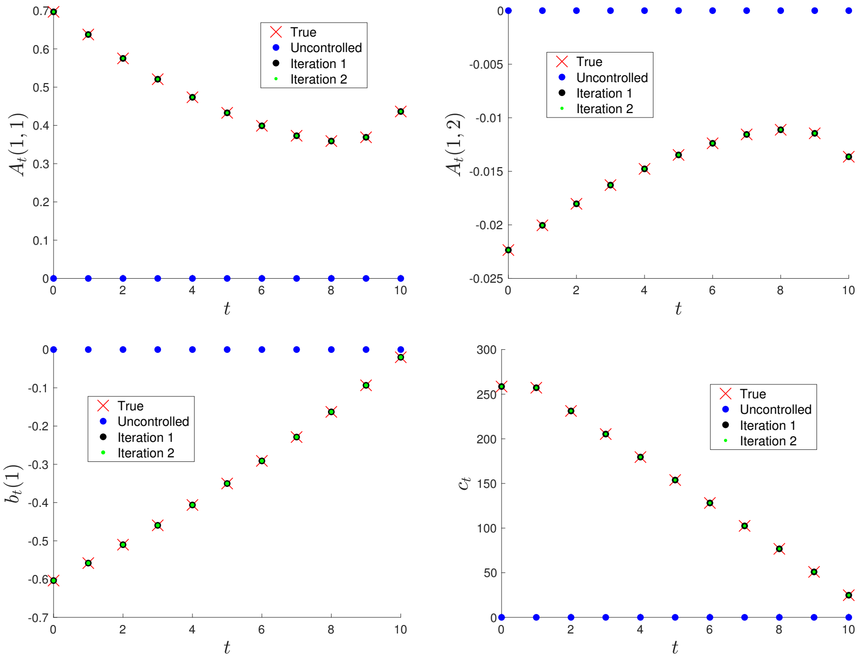

FIG S5. *Coefficients of the optimal policy w.r.t.* $\mathbb{Q}$ *in LQG control against estimates obtained using ADP algorithm. The problem setting is* $d = 32, \xi = 20, \rho = 0.8$ *and the algorithmic settings of uncontrolled SMC are* $N = 2048, T = 10, h = 0.1, \lambda_t = t/T$.

denotes the unknown regression coefficients and $X_m \in \mathbb{R}^d$ the $m \in [1:M]$ row of a model matrix $X \in \mathbb{R}^{M\times d}$. Hence the likelihood function and its gradient is given by

$$\ell(x,y) = \exp\left(y^T X x - \sum_{m=1}^{M} \log(1+\exp(x^T X_m))\right)$$

and

$$\nabla \log \ell(x,y) = X^T y - \sum_{m=1}^{M} (1+\exp(-x^T X_m))^{-1} X_m$$

where $y = (y_m)_{m\in[1:M]} \in \mathsf{Y} = \{0,1\}^M$ is a given dataset of interest. Following [8], we specify a Gaussian prior distribution $\mu(\mathrm{d}x_0) = \mathcal{N}(x_0;\mu_0,\Sigma_0)\mathrm{d}x_0$ on $(\mathsf{X},\mathcal{X}) = (\mathbb{R}^d, \mathfrak{B}(\mathbb{R}^d))$ of the form $\mu_0 = 0_d$ and $\Sigma_0 = \pi^2 M/(3d)(X^T X)^{-1}$.

8.1. *Approximate dynamic programming.* In view of Proposition 1 and the previous section on LQG control, we consider the function classes in (45). As before, coefficients $(A_t^{i+1}, b_t^{i+1}, c_t^{i+1})_{t\in[0:T]}$ at each iteration $i \in [0:I-1]$ can be estimated by linear least squares and the policy $\psi^{(i)} = (\psi_t^{(i)})_{t\in[0:T]}$ at iteration $i \in [1:I]$ has the form

$$\begin{aligned} -\log \psi_0^{(i)}(x_0) &= x_0^T A_0^{(i)} x_0 + x_0^T b_0^{(i)} + c_0^{(i)}, \\ -\log \psi_t^{(i)}(x_{t-1},x_t) &= x_t^T A_t^{(i)} x_t + x_t^T b_t^{(i)} + c_t^{(i)} - (\lambda_t - \lambda_{t-1}) \log \ell(x_{t-1},y), \end{aligned}$$

for $t \in [1:T]$, where $A_t^{(i)} := \sum_{j=1}^{i} A_t^j, b_t^{(i)} := \sum_{j=1}^{i} b_t^j, c_t^{(i)} := \sum_{j=1}^{i} c_t^j$ for $t \in [0:T]$. Assuming that the constraints $K_0^{(i)} := (\Sigma_0^{-1} + 2A_0^{(i)})^{-1} \succ 0$, $K_t^{(i)} := (\Gamma^{-1} + 2hA_t^{(i)})^{-1} \succ 0$, $t \in [1:T]$ are satisfied or imposed, then sampling from

$$\mu^{\psi^{(i)}}(\mathrm{d}x_0) = \mathcal{N}\left(x_0; K_0^{(i)}(\Sigma_0^{-1}\mu_0 - b_0^{(i)}), K_0^{(i)}\right)\mathrm{d}x_0$$

and

$$M_t^{\psi^{(i)}}(x_{t-1},\mathrm{d}x_t) = \mathcal{N}\left(x_t; K_t^{(i)}\{\Gamma^{-1} q_t(x_{t-1}) - h b_t^{(i)}\}, hK_t^{(i)}\right)\mathrm{d}x_t, \tag{S41}$$

with $q_t(x_{t-1}) := x_{t-1} + h\Gamma\nabla\log\eta_t(x_{t-1})/2$ for $t \in [1:T]$ is feasible and evaluation of the twisted potentials $(G_t^{\psi^{(i)}})_{t\in[0:T]}$ defined in (13) is tractable since

$$\begin{aligned} \mu(\psi_0^{(i)}) = {} & \det(\Sigma_0)^{-1/2} \det(K_0^{(i)})^{1/2} \exp\left(\frac{1}{2}(\Sigma_0^{-1}\mu_0 - b_0^{(i)})^T K_0^{(i)} (\Sigma_0^{-1}\mu_0 - b_0^{(i)})\right) \\ & \times \exp\left(-\frac{1}{2}\mu_0^T \Sigma_0^{-1}\mu_0 - c_0^{(i)}\right) \end{aligned}$$

and

$$
\begin{aligned}
M_t(\psi_t^{(i)})(x_{t-1}) = {} & \det(\Gamma)^{-1/2}\det(K_t^{(i)})^{1/2} \\
& \times \exp\left(\frac{1}{2}h^{-1}(\Gamma^{-1}q_t - hb_t^{(i)})^T K_t^{(i)}(\Gamma^{-1}q_t - hb_t^{(i)})(x_{t-1})\right) \\
& \times \exp\left(-\frac{1}{2}h^{-1}(q_t^T\Gamma^{-1}q_t)(x_{t-1}) - c_t^{(i)} + (\lambda_t - \lambda_{t-1})\log \ell(x_{t-1}, y)\right)
\end{aligned}
$$

for $t \in [1 : T]$. We note that imposing $A_t = 0$ and letting b_t depend on the argument x_{t-1} in (45) is related to the approach in [11, 14], as (S41) then corresponds to an Euler-Maruyama discretization of a controlled diffusion with an additive control $x_{t-1} \mapsto b_t^{(i)}(x_{t-1})$. For this application, we set the pre-conditioner as $\Gamma = I_d$ and we illustrate in Figure S6 that the parameterization (45) provides a good approximation of the optimal policy on a particular dataset concerning modeling of heart diseases.

8.2. *Comparison of algorithmic performance.* We now perform a comparison of algorithms on the analysis of three real datasets[1] with different characteristics, in the same manner as Section 7.2. We use $N = 1024$ number of particles in cSMC and select the number of iterations using preliminary runs – see Figure S6. The number of particles used in AIS is then chosen to match computational cost, measured in terms of run time. These algorithmic settings and the results obtained using 100 independent repetitions each of method are summarized in Table S2. Although AIS provides state-of-the-art results in complex scenarios for these models [6], the comparison shows that for all datasets considered, cSMC outperforms it and particularly so for the task of marginal likelihood estimation by several orders of magnitude.

9. Model specific expressions.

9.1. *Expressions for non-linear multimodal state space model.* For notational simplicity, we write $\mu_0 = 0, \sigma_0^2 = 5$ and $\mu_t(x_{t-1}) := x_{t-1}/2 + 25x_{t-1}/(1 + x_{t-1}^2) + 8\cos(1.2t)$ for $t \in [1 : T]$. Assume that the policy $\psi^{(i)} = (\psi_t^{(i)})_{t\in[0:T]}$ at iteration $i \in [1 : I]$ has the form (S34). The initial distribution is given by

$$
\mu^{\psi^{(i)}}(\mathrm{d}x_0) = \sum_{m\in[1:M]^i} A_{0,m}^{(i)}\mathcal{N}\left(x_0; \mu_{0,m}^{(i)}, (\sigma_0^{(i)})^2\right)\mathrm{d}x_0
$$

[1] Datasets were downloaded from the UCI machine learning repository and standardized before analysis.

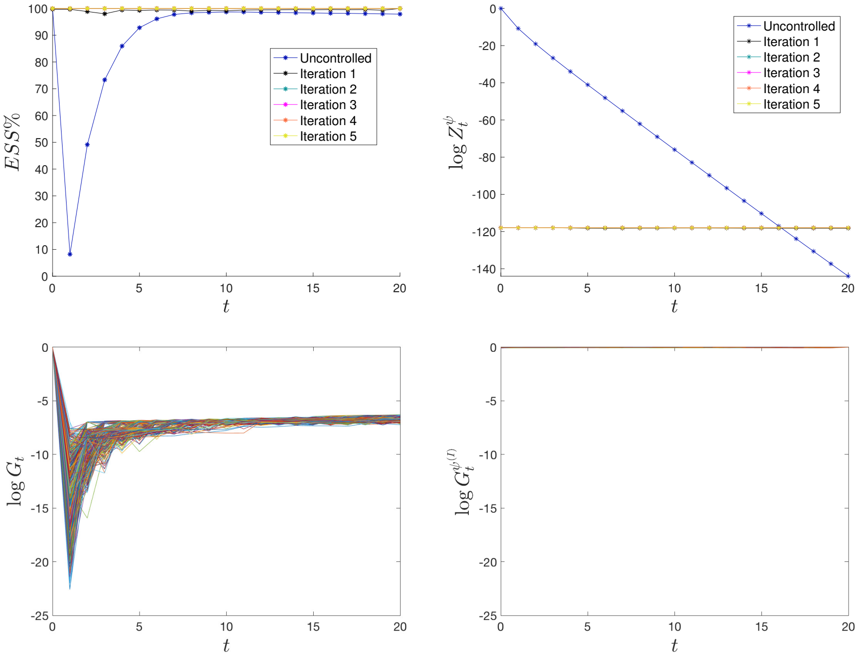

Fig S6. *Comparison of uncontrolled and controlled SMC methods in terms of effective sample size (*top left*), normalizing constant estimation (*top right*) and variance of particle weights (*bottom *row) when performing Bayesian logistic regression on the Heart disease dataset. The algorithmic settings of cSMC are* $I = 5, N = 1024, T = 20$, $h = 1 \times 10^{-4}$, $\lambda_t = t/T$.

			Dataset		
			Heart disease $(M = 270, d = 14)$	*Australian credit* $(M = 690, d = 15)$	*German credit* $(M = 1000, d = 25)$
Algorithm	AIS	N	1843	1843	2048
		h	5×10^{-2}	3×10^{-2}	1×10^{-2}
		ESS%	82.95%	79.75%	74.95%
		$\log Z$	-118.0198 ± 0.4383	-252.8699 ± 1.5128	-527.4392 ± 3.3088
		VAR	1.92×10^{-1}	2.29	10.95
		RMSE	4.40×10^{-1}	2.60	10.06
	cSMC	I	3	4	3
		N	1024	1024	1024
		h	1×10^{-4}	1×10^{-3}	5×10^{-4}
		ESS%	99.99%	99.95%	99.91%
		$\log Z$	-117.9638 ± 0.0117	-250.7504 ± 0.0101	-517.9299 ± 0.0092
		VAR	1.36×10^{-4} ($\mathbf{1.41 \times 10^{3}}$)	1.03×10^{-4} ($\mathbf{2.23 \times 10^{4}}$)	8.39×10^{-5} ($\mathbf{1.31 \times 10^{5}}$)
		RMSE	1.16×10^{-2} ($\mathbf{37.86}$)	1.02×10^{-2} ($\mathbf{2.55 \times 10^{2}}$)	9.11×10^{-3} ($\mathbf{1.10 \times 10^{3}}$)

TABLE S2

Algorithmic settings and performance of AIS and cSMC when performing Bayesian logistic regression for each dataset. Notationally, N refers to the number of particles, h the step size used in MALA for AIS and ULA for cSMC, and I is the number of iterations taken by cSMC. Both algorithms take $T = 20$ time steps for all datasets. Results were obtained using 100 *independent repetitions each of method. The shorthand* ESS% *denotes the percentage of effective sample size averaged over time and repetitions,* $\log Z$ *refers to the estimation of the normalizing constant in logarithmic scale (± a standard deviation),* VAR *is the sample variance of these estimates over the repetitions and* RMSE *the corresponding root mean squared error, which we computed by taking reference to an estimate obtained using many repetitions of a SMC method with a large number of particles. Shown in bold are the gains that cSMC offers relative to AIS.*

with

$$\mu_{0,m}^{(i)} := (\sigma_0^{(i)})^2 \left(2\beta_0^{(i)}\xi_{0,m}^{(i)} + \mu_0\sigma_0^{-2}\right), \quad (\sigma_0^{(i)})^2 := \left(2\beta_0^{(i)} + \sigma_0^{-2}\right)^{-1},$$

and

$$A_{0,m}^{(i)} := \frac{\alpha_{0,m}^{(i)} \exp\left(-\beta_0^{(i)}(\xi_{0,m}^{(i)})^2 + (\mu_{0,m}^{(i)})^2(\sigma_0^{(i)})^{-2}/2\right)}{\sum_{n\in[1:M]^i} \alpha_{0,n}^{(i)} \exp\left(-\beta_0^{(i)}(\xi_{0,n}^{(i)})^2 + (\mu_{0,n}^{(i)})^2(\sigma_0^{(i)})^{-2}/2\right)}.$$

For each $t \in [1:T]$, the Markov transition kernel

$$M_t^{\psi^{(i)}}(x_{t-1}, \mathrm{d}x_t) = \sum_{m\in[1:M]^i} A_{t,m}^{(i)}(x_{t-1})\mathcal{N}\left(x_t; \mu_{t,m}^{(i)}(x_{t-1}), (\sigma_t^{(i)})^2\right)\mathrm{d}x_t$$

with

$$\mu_{t,m}^{(i)}(x_{t-1}) := (\sigma_t^{(i)})^2 \left(2\beta_t^{(i)}\xi_{t,m}^{(i)} + \mu_t(x_{t-1})\sigma_f^{-2}\right), \quad (\sigma_t^{(i)})^2 := \left(2\beta_t^{(i)} + \sigma_f^{-2}\right)^{-1},$$

and

$$A_{t,m}^{(i)}(x_{t-1}) := \frac{\alpha_{t,m}^{(i)} \exp\left(-\beta_t^{(i)}(\xi_{t,m}^{(i)})^2 + \mu_{t,m}^{(i)}(x_{t-1})^2(\sigma_t^{(i)})^{-2}/2\right)}{\sum_{n\in[1:M]^i} \alpha_{t,n}^{(i)} \exp\left(-\beta_t^{(i)}(\xi_{t,n}^{(i)})^2 + \mu_{t,n}^{(i)}(x_{t-1})^2(\sigma_t^{(i)})^{-2}/2\right)}.$$

Evaluation of the twisted potentials $(G_t^{\psi^{(i)}})_{t\in[0:T]}$ defined in (13) is tractable since

$$\mu(\psi_0^{(i)}) = \frac{\sigma_0^{(i)}}{\sigma_0} \exp\left(-\frac{1}{2}\mu_0^2\sigma_0^{-2}\right) \sum_{m\in[1:M]^i} \alpha_{0,m}^{(i)} \exp\left(-\beta_0^{(i)}(\xi_{0,m}^{(i)})^2 + \frac{1}{2}(\mu_{0,m}^{(i)})^2(\sigma_0^{(i)})^{-2}\right)$$

and

$$\begin{aligned} M_t(\psi_t^{(i)})(x_{t-1}) = &\frac{\sigma_t^{(i)}}{\sigma_f} \exp\left(-\frac{1}{2}\mu_t(x_{t-1})^2\sigma_f^{-2}\right) \\ &\times \sum_{m\in[1:M]^i} \alpha_{t,m}^{(i)} \exp\left(-\beta_t^{(i)}(\xi_{t,m}^{(i)})^2 + \frac{1}{2}\mu_{t,m}^{(i)}(x_{t-1})^2(\sigma_t^{(i)})^{-2}\right) \end{aligned}$$

for $t \in [1:T]$.

9.2. *Expressions for Lorenz-96 model.* Suppose that the current policy is given by (44) and write

$$\tilde{A}_t^{(i)} := A_t^{(0)} + A_t^{(i)}, \quad \tilde{b}_t^{(i)} := b_t^{(0)} + b_t^{(i)}, \quad \tilde{c}_t^{(i)} := c_t^{(0)} + c_t^{(i)},$$

for $t \in [0:T]$, where $(A_t^{(0)}, b_t^{(0)}, c_t^{(0)})_{t\in[0:T]}$ are the coefficients corresponding to APF. If the constraints $K_0^{(i)} := (\sigma_f^{-2} I_d + 2\tilde{A}_0^{(i)})^{-1} \succ 0$, $K_t^{(i)} := (\sigma_f^{-2} h^{-1} I_d + 2\tilde{A}_t^{(i)})^{-1} \succ 0, t \in [1:T]$ are satisfied or imposed, then sampling from

$$\mu^{\psi^{(i)}}(\mathrm{d}x_0) = \mathcal{N}\left(x_0; -K_0^{(i)}\tilde{b}_0^{(i)}, K_0^{(i)}\right) \mathrm{d}x_0$$

and

$$M_t^{\psi^{(i)}}(x_{t-1}, \mathrm{d}x_t) = \mathcal{N}\left(K_t^{(i)}\{\sigma_f^{-2}h^{-1}q(x_{t-1}) - \tilde{b}_t^{(i)}\}, K_t^{(i)}\right) \mathrm{d}x_t, \quad t \in [1:T],$$

is feasible and evaluation of the twisted potentials $(G_t^{\psi^{(i)}})_{t\in[0:T]}$ defined in (13) is tractable since

$$\mu(\psi_0^{(i)}) = \sigma_f^{-d} \det(K_0^{(i)})^{1/2} \exp\left(\frac{1}{2}(\tilde{b}_0^{(i)})^T K_0^{(i)} \tilde{b}_0^{(i)} - \tilde{c}_0^{(i)}\right)$$

and

$$M_t(\psi_t^{(i)})(x_{t-1}) = \sigma_f^{-d} h^{-d/2} \det(K_t^{(i)})^{1/2} \exp\left(-\frac{1}{2}\sigma_f^{-2}h^{-1}(q^T q)(x_{t-1}) - \tilde{c}_t^{(i)}\right)$$
$$\times \exp\left(\frac{1}{2}(\sigma_f^{-2}h^{-1}q - \tilde{b}_t^{(i)})^T K_t^{(i)} (\sigma_f^{-2}h^{-1}q - \tilde{b}_t^{(i)})(x_{t-1})\right)$$

for $t \in [1:T]$.

9.3. *Expressions for neuroscience model.* Assume that the constraints $k_0^{(i)} := (1 + 2a_0^{(i)})^{-1} > 0$, $k_t^{(i)} := (\sigma^{-2} + 2a_t^{(i)})^{-1} > 0$, $t \in [1:T]$ are satisfied or imposed. Then the initial distribution

$$\mu^{\psi^{(i)}}(\mathrm{d}x_0) = \mathcal{N}\left(x_0; -k_0^{(i)}b_0^{(i)}, k_0^{(i)}\right) \mathrm{d}x_0$$

and the Markov transition kernels

$$M_t^{\psi^{(i)}}(x_{t-1}, \mathrm{d}x_t) = \mathcal{N}\left(x_t; k_t^{(i)}(\alpha\sigma^{-2}x_{t-1} - b_t^{(i)}), k_t^{(i)}\right) \mathrm{d}x_t$$

for $t \in [1:T]$. Moreover, the twisted potentials $(G_t^{\psi^{(i)}})_{t\in[0:T]}$ defined in (13) can be evaluated since

$$\mu(\psi_0^{(i)}) = (k_0^{(i)})^{1/2} \exp\left(\frac{1}{2}k_0^{(i)}(b_0^{(i)})^2 - c_0^{(i)}\right)$$

and

$$M_t(\psi_t^{(i)})(x_{t-1}) = (k_t^{(i)})^{1/2}\sigma^{-1}\exp\left(\frac{1}{2}k_t^{(i)}(\alpha\sigma^{-2}x_{t-1}-b_t^{(i)})^2-\frac{1}{2}\sigma^{-2}\alpha^2x_{t-1}^2-c_t^{(i)}\right)$$

for $t \in [1:T]$.

References.

[1] B. D. Anderson and J. B. Moore. *Optimal Control: Linear Quadratic Methods*. Dover Publications, 2007.

[2] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic Programming*. Athena Scientific, 1996.

[3] P. Dai Pra, L. Meneghini and W.J. Runggaldier. Connections between stochastic control and dynamic games. *Mathematics of Control, Signals and Systems*, 9(4):303–326, 1996.

[4] P. Del Moral. *Feynman-Kac Formulae*. Springer, 2004.

[5] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*. 41(1):45–76, 1999.

[6] N. Chopin and J. Ridgeway. Leave Pima Indians alone: binary regression as a benchmark for Bayesian computation. *Statistical Science*, 32(1):64-87, 2017.

[7] N. Gordon, J. Salmond and A. Smith. A novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEE Proceedings on Radar and Signal Processing*, 140:107–113, 1993.

[8] T. E. Hanson, A. J. Branscum and W. O. Johnson. Informative g-priors for logistic regression. *Bayesian Analysis*, 9(3):597–612, 2014.

[9] N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

[10] H. J. Kappen, V. Gómez and M. Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

[11] H. J. Kappen and H. C. Ruiz. Adaptive importance sampling for control and inference. *Journal of Statistical Physics*, 162(5):1244–1266, 2016.

[12] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

[13] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, 1974.

[14] H. C. Ruiz and H. J. Kappen. Particle smoothing for hidden diffusion processes: adaptive path integral smoother. *IEEE Transactions on Signal Processing*, 65(12):3191–3203, 2017.

[15] E. Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009.

[16] J. N. Tsitsiklis and B. Van Roy. Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks*, 12(4):694–703, 2001.

5 Nepal Park
Singapore 139408, Singapore
E-mail: heng@essec.edu

15 Broadway
Sydney, NSW 2007, Australia
E-mail: adrian.bishop@uts.edu.au

24-29 St Giles'
Oxford OX1 3LB, UK
E-mail: deligian@stats.ox.ac.uk; doucet@stats.ox.ac.uk