PASOA- PArticle baSed Bayesian Optimal Adaptive design

Jacopo Iollo [1 2 3] **Christophe Heinkelé** [3] **Pierre Alliez** [2] **Florence Forbes** [1]

Abstract

We propose a new procedure named PASOA, for Bayesian experimental design, that performs sequential design optimization by simultaneously providing accurate estimates of successive posterior distributions for parameter inference. The sequential design process is carried out via a contrastive estimation principle, using stochastic optimization and Sequential Monte Carlo (SMC) samplers to maximise the Expected Information Gain (EIG). As larger information gains are obtained for larger distances between successive posterior distributions, this EIG objective may worsen classical SMC performance. To handle this issue, tempering is proposed to have both a large information gain and an accurate SMC sampling, that we show is crucial for performance. This novel combination of stochastic optimization and tempered SMC allows to jointly handle design optimization and parameter inference. We provide a proof that the obtained optimal design estimators benefit from some consistency property. Numerical experiments confirm the potential of the approach, which outperforms other recent existing procedures.

1. Introduction

A design refers to some experimental conditions required to perform an experiment and get observations from the phenomenon under study. Assuming that such a design is characterized by some parameters denoted by $\boldsymbol{\xi}$, simple examples of $\boldsymbol{\xi}$ are coordinates in a 2D space or a frequency at which we wish to measure a quantity of interest. The overall goal of experimental design can be summarized as the acquisition of good quality data, which is of increasing importance in numerous scientific and industrial applications, see *e.g.* (Ryan et al., 2016; Rainforth et al., 2024). Quality can be defined from numerous view points, allocating resources for information gathering, improving precision and/or prediction or reducing experimental costs. In this work, we assume that the desired designs are continuous parameters $\boldsymbol{\xi} \in \mathcal{E}$ that maximize information on some parameters of interest $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$. In Bayesian Optimal Experimental Design (BOED), a Bayesian model is specified over $\boldsymbol{\theta}$ (Chaloner & Verdinelli, 1995; Sebastiani & Wynn, 2000; Amzal et al., 2006; Ryan et al., 2016; Rainforth et al., 2024). The criterion defining the optimality of a design may then depend on the application (Ryan et al., 2016; Kleinegesse & Gutmann, 2021) but the most used of them is the expected information gain (EIG). It targets designs that most reduce the entropy of the posterior distribution over $\boldsymbol{\theta}$, or equivalently that maximize the Kullback-Leibler divergence between the posterior and the prior, and this in expectation over all possible experimental outcomes, see eq. (1,2). The Bayesian formulation is particularly convenient in sequential contexts, where we wish to exploit the results of previous experiments to guide the design of future ones. Unfortunately, computing and optimizing the EIG sequentially requires sampling from an evolving sequence of intractable posterior distributions. We propose to use a sequential Monte-Carlo (SMC) approach for an efficient sequential sampling (Del Moral et al., 2006; Doucet & Lee, 2018; Naesseth et al., 2019; Chopin & Papaspiliopoulos, 2020; Dai et al., 2022) and in the spirit of some original and recent approaches (Huan & Marzouk, 2014; Kleinegesse & Gutmann, 2021; Foster et al., 2019; 2020; 2021; Blau et al., 2022), we adopt a stochastic optimization approach to optimize the EIG. The originality of our approach is to use SMC in conjunction with stochastic optimization to efficiently sample the relevant quantities and estimate the required noisy gradients. However, naive SMC is potentially problematic. In sequential EIG-based optimization, larger gains are obtained for larger distances between successive posterior distributions, which is a particularly bad scenario for SMC performance. A solution is to consider *distribution tempering or annealing* (Neal, 2001; Del Moral et al., 2006), which creates a path of intermediate closer distributions for a more accurate SMC sampling (Figures 1 and 2). We name our approach PASOA, for Particle baSed Bayesian Optimal Adaptive design, where

[1]Université Grenoble Alpes, Inria, CNRS, G-INP, France [2]Université Côte d'Azur, Inria, France [3]Cerema, Endsum-Strasbourg, France. Correspondence to: Jacopo Iollo <jacopo.iollo@inria.fr>, Florence Forbes <florence.forbes@inria.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

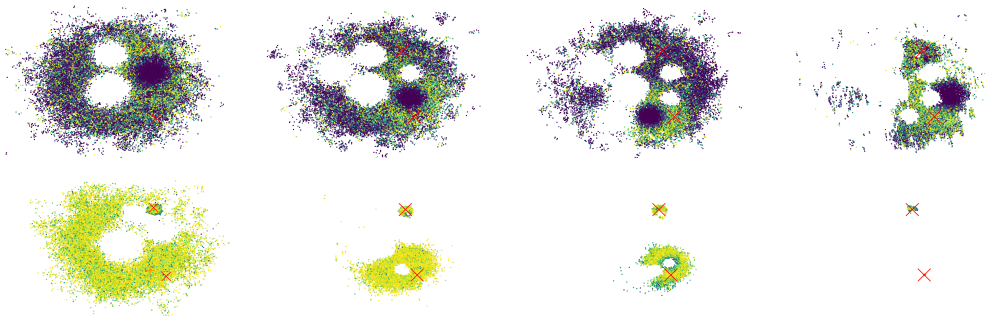

Figure 1. Source location example at steps 5, 7, 9, 11. Over design steps, particles concentrate faster to true sources (red crosses) with PASOA (2nd line) than with SMC (1st line). Lower particle weights in blue, higher in yellow.

particles refer to samples featuring an evolving Bayesian posterior information that adaptively guides the successive design estimation. To our knowledge, such a combination, of stochastic optimization and SMC or tempered SMC, has not been proposed before for BOED. It allows to benefit from both techniques' advantages. Stochastic optimization provides a more scalable way to handle sequential design optimization, while the principled SMC framework allows to efficiently perform an accurate parameter inference by reusing samples from past experiments. This high quality inference is exploited in turn via stochastic optimization to yield informative design estimators coming with good theoretical features. We show that these estimators are consistent and converge in probability to the optimal designs. On the numerical side, we verify their practical performance on benchmark design problems. We observe significantly higher information gains than previous methods, including sophisticated and reinforcement learning approaches (Blau et al., 2022; Foster et al., 2020), with these gains leading, in turn, to improved posterior inference and parameter estimation. We thus provide the first theoretical result of this kind in the BOED literature, while also highlighting tempering as a mean to greatly enhance the performance of SMC in BOED.

2. Related work

Simulations required for EIG optimization can be obtained via Markov Chain Monte Carlo (MCMC) algorithms, see *e.g.* Kleinegesse & Gutmann (2020) for a static design. In a sequential design context, the necessity to do so at each step is very simulation intensive (Amzal et al., 2006; Kleinegesse et al., 2020). Typically, Kleinegesse et al. (2020) extend Kleinegesse & Gutmann (2020) to a sequential setting but shows up to only $K = 4$ or 5 experiments. As an alternative, SMC samplers have been used in previous work, *e.g.* Drovandi et al. (2014; 2013) use SMC but only for finite-valued design parameters, avoiding the problematic optimization by reducing it to a finite number of comparisons. Another attempt (Kuck et al., 2006) proposes SMC to handle the optimisation part but restricts to static one-step design. In our work, we consider both the continuous design and sequential setting, which was referred to as an open question by Ryan et al. (2016). Recent ideas using stochastic optimization (Blau et al., 2022; Kleinegesse & Gutmann, 2021; Foster et al., 2020; 2021; 2019; Huan & Marzouk, 2014) have since then paved the way to efficient optimisation formulation of sequential BOED but without consideration of SMC techniques and somewhat neglecting posterior distribution inference. A solution explored by Foster et al. (2019; 2020) uses mean field variational approximations of the posterior distributions but with no guarantee on their quality, while Foster et al. (2021); Blau et al. (2022) bypass the need for such posterior estimation using a reinforcement learning formulation. In this work, we show that superior performance than these previous solutions can be obtained with SMC, from which we also derive new good features. We provide theoretical results and guarantees that have not been formulated before in *modern* BOED and we shed a new light on tempering by showing that it can be crucial in SMC-based sequential BOED, to face the contradiction between searching for the largest information gain and maintaining good sampling performance.

3. Sequential Bayesian Optimal Experimental Design (BOED)

The Bayesian framework is a unified way to account for prior information via a probability distribution $p(\boldsymbol{\theta})$, for uncertainties about the observations $\mathbf{y}$ through a distribution $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})$, and for a design criterion (also called utility function) $F(\boldsymbol{\xi}, \boldsymbol{\theta}, \mathbf{y})$ describing the experimental aims. The prior is assumed to be independent of $\boldsymbol{\xi}$ and $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})$ available in closed-form.

Expected Information Gain (EIG). There exists various utility functions F depending on the targeted task (Ryan et al., 2016; Kleinegesse & Gutmann, 2021). In this

work, we focus on parameter estimation and consider an information-based utility leading to the so-called expected information gain (EIG). The EIG, denoted by I, admits several equivalent expressions, see *e.g.* (Foster et al., 2019). It can be written as the expected loss in entropy when accounting for an observation $\mathbf{y}$ at $\boldsymbol{\xi}$ (eq. (1)) or as a mutual information or expected Kullback-Leibler divergence (eq. (2)). Using $p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\xi}) = p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\xi})p(\mathbf{y}|\boldsymbol{\xi}) = p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})p(\boldsymbol{\theta})$,

$$I(\boldsymbol{\xi}) = \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\xi})}[H(p(\boldsymbol{\theta})) - H(p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\xi})] \quad (1)$$

$$= \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\xi})}\left[\mathrm{KL}(p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\xi}), p(\boldsymbol{\theta}))\right] \quad (2)$$

where random variables are indicated with uppercase letters, $\mathbb{E}_p[\cdot]$ denotes the expectation with respect to p, KL the Kullback-Leibler divergence and $H(p(\boldsymbol{\theta})) = -\mathbb{E}_{p(\boldsymbol{\theta})}[\log p(\boldsymbol{\theta})]$ is the entropy of p. We thus look for $\boldsymbol{\xi}^*$ satisfying

$$\boldsymbol{\xi}^* \in \arg\max_{\boldsymbol{\xi}\in\mathbb{R}^d} I(\boldsymbol{\xi}) \ . \quad (3)$$

Before optimizing $I(\boldsymbol{\xi})$, evaluating $I(\boldsymbol{\xi})$ is often difficult due to the intractability of $p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\xi})$ and $p(\mathbf{y}|\boldsymbol{\xi})$.

Sequential design. Solving (3) is a *static* or *one-step* design problem. A single $\boldsymbol{\xi}$ or multiple $\{\boldsymbol{\xi}_1, \cdot, \boldsymbol{\xi}_K\}$ are selected prior to any observation, measurements $\{\mathbf{y}_1, \cdot, \mathbf{y}_K\}$ are made for these design parameters and the experiment is stopped. The prior $p(\boldsymbol{\theta})$ can be used to encode previous observations but in static design, the selected designs depend only on the model. In contrast, in sequential or iterated design, K experiments are planned sequentially to construct an adaptive strategy, meaning that for the kth experiment, the best $\boldsymbol{\xi}_k$ is selected taking into account the previous design parameters and associated observations $\boldsymbol{D}_{k-1} = \{(\mathbf{y}_1, \boldsymbol{\xi}_1), \cdot, (\mathbf{y}_{k-1}, \boldsymbol{\xi}_{k-1})\}$. Then, $\mathbf{y}_k$ is measured at $\boldsymbol{\xi}_k$ and $\boldsymbol{D}_k$ is updated into $\boldsymbol{D}_k = \boldsymbol{D}_{k-1} \cup (\mathbf{y}_k, \boldsymbol{\xi}_k)$. This approach is referred to as *greedy* or *myopic*, in that the next design $\boldsymbol{\xi}_k$ is chosen as if it was the last one. Non-myopic approaches exist, using for instance reinforcement learning principles (Blau et al., 2022; Foster et al., 2021) but with another layer of complexity and performance that are not always superior, see (Blau et al., 2022) or our results in Section 7. In this work, we limit ourselves to the greedy approach, replacing in (1) or (2) the prior $p(\boldsymbol{\theta})$ by our current belief on $\boldsymbol{\theta}$, namely $p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1}) = p(\boldsymbol{\theta}|\mathbf{y}_1, \boldsymbol{\xi}_1, \cdot, \mathbf{y}_{k-1}, \boldsymbol{\xi}_{k-1})$, and to solve iteratively for

$$\boldsymbol{\xi}_k^* \in \arg\max_{\boldsymbol{\xi}\in\mathbb{R}^d} I_k(\boldsymbol{\xi}), \quad (4)$$

$$\text{where } I_k(\boldsymbol{\xi}) = \mathbb{E}[H(p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})) - H(p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\xi}, \boldsymbol{D}_{k-1}))]$$

$$= \mathbb{E}\left[\mathrm{KL}(p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\xi}, \boldsymbol{D}_{k-1}), p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1}))\right] \quad (5)$$

and $\mathbb{E}$ is with respect to $p(\mathbf{y}|\boldsymbol{\xi}, \boldsymbol{D}_{k-1})$. Observations are assumed conditionally independent so that $p(\boldsymbol{\theta}|\boldsymbol{D}_k) \propto p(\boldsymbol{\theta}) \prod_{i=1}^k p(\mathbf{y}_i|\boldsymbol{\theta}, \boldsymbol{\xi}_i)$ which also leads to

$$p(\boldsymbol{\theta}|\boldsymbol{D}_k) \propto p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})\ p(\mathbf{y}_k|\boldsymbol{\theta}, \boldsymbol{\xi}_k) \ . \quad (6)$$

EIG contrastive bound optimization. Going back to the optimization in (3), we focus on the continuous design case and assume that quantities are differentiable when needed and that gradients are well defined. This is generally not a restrictive assumption except for discrete design spaces that may require specific treatments, see *e.g.* (Blau et al., 2022; Drovandi et al., 2014). A standard gradient ascent algorithm would consist, at iteration t, of updating $\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t + \gamma_t\ \nabla_{\boldsymbol{\xi}} I(\boldsymbol{\xi})_{|\boldsymbol{\xi}=\boldsymbol{\xi}_t}$ with a stepsize γ_t, but in practice both $I(\boldsymbol{\xi})$ and its gradient $\nabla_{\xi} I(\boldsymbol{\xi})$ are intractable. However, they can both be expressed as expectations, which naturally leads to consider stochastic approximation approaches (Borkar, 2008), among which the most popular is the Stochastic Gradient (SG) algorithm. In a BOED setting, if $\nabla_{\boldsymbol{\xi}} I(\boldsymbol{\xi})$ is expressed as an expectation $\nabla_{\boldsymbol{\xi}} I(\boldsymbol{\xi}) = \mathbb{E}[f(\boldsymbol{\xi}, \mathbf{X})]$ over a random variable $\mathbf{X}$, the SG iteration writes $\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t + \gamma_t\ f(\boldsymbol{\xi}_t, \mathbf{x}_t)$ with $\mathbf{x}_t$ a realisation of $\mathbf{X}$. This assumes that we can differentiate under the integral sign in (2). There exist different ways to differentiate, including the popular reparametrization trick, but SG for $I(\boldsymbol{\xi})$ remains difficult to perform due to the intractability of the integrand in (2). An alternative has been proposed by Foster et al. (2019; 2020) referred to as a variational approach. It consists of optimizing a tractable lower bound of $I(\boldsymbol{\xi})$ and computing $\boldsymbol{\xi}^*$ via an alternate maximization. We also consider such a bound, I_{PCE}, introduced by Foster et al. (2020) and named the Prior Contrastive Estimation (PCE) bound. It is based on contrastive samples from L additional variables $\boldsymbol{\theta}_\ell$, for $\ell = 1 : L$, distributed following the prior $p(\boldsymbol{\theta})$ as $\boldsymbol{\theta}$, which is rewritten as $\boldsymbol{\theta}_0$. I_{PCE} is defined as

$$I_{PCE}(\boldsymbol{\xi}) = \mathbb{E}_{p(y|\boldsymbol{\theta}_0, \boldsymbol{\xi}) \prod_{\ell=0}^L p(\boldsymbol{\theta}_\ell)}\left[F(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, \mathbf{Y})\right] \quad (7)$$

$$\text{with } F(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, \mathbf{y}) = \log \frac{p(\mathbf{y}|\boldsymbol{\theta}_0, \boldsymbol{\xi})}{\frac{1}{L+1}\sum_{\ell=0}^L p(\mathbf{y}|\boldsymbol{\theta}_\ell, \boldsymbol{\xi})}.$$

I_{PCE} is a lower bound $I(\boldsymbol{\xi}) \geq I_{PCE}(\boldsymbol{\xi})$ and the bound is tight when L tends to ∞ (see Foster et al. (2020) for a proof). It is tractable as all expressions $p(\mathbf{y}|\boldsymbol{\theta}_\ell, \boldsymbol{\xi})$ are tractable, and its gradient requires only the gradient $\nabla_{\boldsymbol{\xi}} p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})$. Stochastic approximation can be applied to maximize I_{PCE} as $\nabla_{\boldsymbol{\xi}} I_{PCE}(\boldsymbol{\xi})$ can be expressed as an expectation via reparametrization. More specifically, we assume that there exists a transformation $T_{\boldsymbol{\theta}_0}^{\boldsymbol{\xi}}$ such that $\mathbf{Y} = T_{\boldsymbol{\theta}_0}^{\boldsymbol{\xi}}(U)$ with $U \in \mathcal{U}$ independent of $\boldsymbol{\xi}$ and $\boldsymbol{\theta}_0$ and easy to simulate, *e.g.* U is a standard Gaussian. It follows under some mild conditions specified in the Appendix that

$$\nabla_{\boldsymbol{\xi}} I_{PCE}(\boldsymbol{\xi}) = \mathbb{E}_{p(u)\prod_{\ell=0}^L p(\boldsymbol{\theta}_\ell)}\left[\nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, T_{\boldsymbol{\theta}_0}^{\boldsymbol{\xi}}(U))\right] \quad (8)$$

A similar bound and its gradient can be derived for sequential design optimization (4). Using conditional independence (6), at each step k, $p(\boldsymbol{\theta})$ only needs to be

replaced by the current posterior $p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})$,

$$I^k_{PCE}(\boldsymbol{\xi}) = \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta}_0,\boldsymbol{\xi}))\prod_{\ell=0}^{L} p(\boldsymbol{\theta}_\ell|\boldsymbol{D}_{k-1})}[F(\boldsymbol{\xi},\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L,\mathbf{Y})] \quad (9)$$

$$\nabla_{\boldsymbol{\xi}} I^k_{PCE}(\boldsymbol{\xi}) = \mathbb{E}_{p(u)\prod_{\ell=0}^{L} p(\boldsymbol{\theta}_\ell|\boldsymbol{D}_{k-1})}\left[\nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi},\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L,T^{\boldsymbol{\xi}}_{\boldsymbol{\theta}_0}(U))\right].$$

The stochastic gradient algorithm is summarized in Algorithm 1, with the additional possibility to estimate gradients with minibatches of size N_t, in line 5.

Algorithm 1 SG with minibatches $(N_t)_{t=1:T}$ at step k to optimize I^k_{PCE} in (9)

1: Set T iterations, $\boldsymbol{\xi}_0$, stepsizes $(\gamma_t)_{t=1:T}$
2: **while** $t \leq T$ **do**
3: Sample $\boldsymbol{\theta}^i_{\ell,t} \sim p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})$, $\ell = 0:L, i = 1:N_t$
4: Sample $u^i_t \sim p(u)$, for $i = 1:N_t$
5: Set $\nabla_{t+1} = \frac{1}{N_t}\sum_{i=1}^{N_t} \nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi},\boldsymbol{\theta}^i_{0,t},\cdot,\boldsymbol{\theta}^i_{L,t},T^{\boldsymbol{\xi}}_{\boldsymbol{\theta}^i_{0,t}}(u^i_t))|_{\boldsymbol{\xi}=\boldsymbol{\xi}_t}$
6: Update $\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t + \gamma_t \nabla_{t+1}$
7: **return** $\boldsymbol{\xi}^*_k = \boldsymbol{\xi}_T$ or a Polyak averaging value

Optimizing the contrastive bound I^k_{PCE} requires then sampling $\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L$ from $p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})$ (Algorithm 1 line 3). This is more costly than sampling from the prior, as $p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})$ is only know up to a normalizing constant (6), but can be efficiently dealt with using a SMC approach as detailed in the next section.

4. Tempered Sequential Monte Carlo

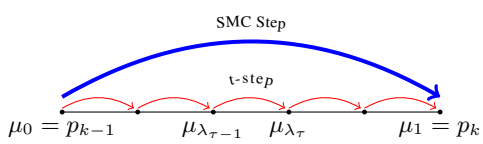

Figure 2. Tempered SMC, a SMC step from p_{k-1} to p_k (blue) performed in $\mathcal{T}$ intermediate tempering steps (red).

SMC samplers extend the idea of importance sampling by re-using samples from one distribution to another, and benefit from numerous theoretical results (Del Moral et al., 2006; Doucet & Lee, 2018; Naesseth et al., 2019; Chopin & Papaspiliopoulos, 2020; Dai et al., 2022). More specifically, denoting $\mathcal{P}(\boldsymbol{\Theta})$ the set of probability measures on $\boldsymbol{\Theta}$, the goal is to provide samples (called particles) from a sequence of probability distributions $\{p_k\}_{k=1:K}$ in $\mathcal{P}(\boldsymbol{\Theta})$. To simplify, we deal with probability densities assuming absolute continuity with respect to the Lebesgue measure but the setting is more general, *e.g.* (Chopin & Papaspiliopoulos, 2020). A standard MCMC approach would require to build an ergodic kernel M_k and to run

Algorithm 2 Adaptive tempered SMC at step k

1: Set $\tau = 0, \lambda_0 = 0, M, ESS_{min}, M_\lambda$, resample
2: Sample $\boldsymbol{\theta}^{1:M}_0 \sim p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1}) = \mu_{\lambda_0}(\boldsymbol{\theta})$
3: Set $w^i_0 = \frac{1}{M}$, $i = 1:M$ and $\mathbf{w}^{1:M}_0 = \{w^1_0,\cdot,w^M_0\}$
4: **while** $\lambda_\tau < 1$ **do**
5: Set $\tau = \tau + 1$
6: Set $\tilde{\boldsymbol{\theta}}^{1:M}_{\tau-1} = \text{resample}(\boldsymbol{\theta}^{1:M}_{\tau-1}, \mathbf{w}^{1:M}_{\tau-1})) \sim \mu_{\lambda_{\tau-1}}$
7: Sample $\boldsymbol{\theta}^i_\tau \sim M_{\lambda_{\tau-1}}(\tilde{\boldsymbol{\theta}}^i_{\tau-1},\cdot)$ for i=1:M,
8: Set $\boldsymbol{\theta}^{1:M}_\tau = \{\boldsymbol{\theta}^1_\tau,\cdot,\boldsymbol{\theta}^M_\tau\}$ Resampling and Markov kernel step
9: Solve for γ, $\frac{\left(\sum_{i=1}^M p(\mathbf{y}_k|\boldsymbol{\theta}^i_\tau,\boldsymbol{\xi}_k)^\gamma\right)^2}{\sum_{i=1}^M p(\mathbf{y}_k|\boldsymbol{\theta}^i_\tau,\boldsymbol{\xi}_k)^{2\gamma}} = ESS_{min}$
10: Set $\lambda_\tau = \lambda_{\tau-1} + \gamma$ Tempered step
11: Set $\tilde{w}^i_\tau = p(\mathbf{y}_k|\boldsymbol{\theta}^i_\tau,\boldsymbol{\xi}_k)^\gamma$, $w^i_\tau = \frac{\tilde{w}^i_\tau}{\sum_{j=1}^M \tilde{w}^j_\tau}$ for i=1:M
12: Set $\mathbf{w}^{1:M}_\tau = \{w^1_\tau,\cdots,w^M_\tau\}$ Compute new weights
13: **return** $\boldsymbol{\theta}^{1:M}_k = \boldsymbol{\theta}^{1:M}_\tau, \mathbf{w}^{1:M}_k = \mathbf{w}^{1:M}_\tau$ for a particle approximation $p^M_k = \sum_{i=1}^M w^i_k \delta_{\boldsymbol{\theta}^i_k}$ of $p(\cdot|\boldsymbol{D}_k) = \mu_1$

a Markov chain from scratch for each p_k. In contrast, SMC samplers provide the possibility to approximate p_k recycling samples from p_{k-1}. SMC samplers aim at propagating M particles $\boldsymbol{\theta}^{1:M}_{k-1} = \{\boldsymbol{\theta}^1_{k-1},\cdot,\boldsymbol{\theta}^M_{k-1}\}$ and their corresponding weights $\mathbf{w}^{1:M}_{k-1} = \{w^1_{k-1},\cdot,w^M_{k-1}\}$ in such a way that the empirical distribution p^M_k of the particles at time k converges to p_k in some sense: meaning that for all integrable functions ϕ, $\mathbb{E}_{p^M_k}[\phi(\boldsymbol{\theta})] = \sum_{i=1}^M w^i_k \phi(\boldsymbol{\theta}^i_k) \xrightarrow[M\to\infty]{} \mathbb{E}_{p_k}[\phi(\boldsymbol{\theta})]$. However, as showed by Agapiou et al. (2017), the number of particles M required for an accurate *particle approximation* p^M_k scales exponentially with the Kullback-Leibler distance between the proposal p_{k-1} and target p_k distributions. For $p_{k-1}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})$ and $p_k(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{D}_k)$, this is problematic as EIG optimization aims at increasing this distance, see (4,5). Moving from $p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})$ to $p(\boldsymbol{\theta}|\boldsymbol{D}_k)$ with just one SMC step might then yield poor results. A solution is to consider *tempering* (Neal, 2001; Del Moral et al., 2006; Syed et al., 2021) to move along a sequence of probability distributions interpolating between $p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})$ and $p(\boldsymbol{\theta}|\boldsymbol{D}_k)$. A tempering path is a sequence of the form μ_{λ_τ} with $0 = \lambda_0 < \lambda_\tau < \ldots < \lambda_{\mathcal{T}} = 1$ where $\mu_0 = p_{k-1}$ and $\mu_1 = p_k$ (Figure 2). Usually, only the initial and final distributions are imposed. Intermediate distributions μ_{λ_τ} are not of interest so that the λ_τ's can be chosen as desired. A popular approach is to use what is know as the geometric path: $\mu_{\lambda_\tau}(\boldsymbol{\theta}) \propto p_{k-1}(\boldsymbol{\theta})^{1-\lambda_\tau}\ p_k(\boldsymbol{\theta})^{\lambda_\tau}$, which using (6), takes the form $\mu_{\lambda_\tau}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})\ p(\mathbf{y}_k|\boldsymbol{\theta},\boldsymbol{\xi}_k)^{\lambda_\tau}$ or

equivalently $\mu_{\lambda_\tau}(\boldsymbol{\theta}) \propto \mu_{\lambda_{\tau-1}}(\boldsymbol{\theta})\, p(\mathbf{y}_k|\boldsymbol{\theta},\boldsymbol{\xi}_k)^{\lambda_\tau-\lambda_{\tau-1}}$. As setting the sequence λ_τ manually can be a challenging task with disappointing results, we follow the adaptive strategy proposed by Jasra et al. (2011). Given a user-set threshold ESS_{min} interpreted as an effective sample size, at iteration τ of the tempered SMC procedure, given a current set of particles $\boldsymbol{\theta}_\tau^{1:M}$, we set recursively $\lambda_\tau = \lambda_{\tau-1} + \gamma$ with γ the solution in $[0, 1-\lambda_{\tau-1}]$ of the equation $\frac{\left(\sum_{i=1}^M p\left(\mathbf{y}_k|\boldsymbol{\theta}_\tau^i,\boldsymbol{\xi}_k\right)^\gamma\right)^2}{\sum_{i=1}^M p\left(\mathbf{y}_k|\boldsymbol{\theta}_\tau^i,\boldsymbol{\xi}_k\right)^{2\gamma}} = ESS_{min}$. If γ is not in $[0, 1-\lambda_{\tau-1}]$, λ_τ is set to 1 and the tempering stops. This is a relatively simple task to solve with numerical root finding. This procedure guarantees that the SMC approximation error remains stable over iterations, see (Jasra et al., 2011) for details. It can be interpreted as a way to control the Chi-square pseudo-distance between the successive distributions, see (Chopin & Papaspiliopoulos, 2020) Proposition 17.2. Tempered SMC then requires like SMC an unbiased resampling scheme denoted by $resample(\boldsymbol{\theta}^{1:M}, \mathbf{w}^{1:M})$. Resampling is the action of drawing randomly from a weighted sample, so as to obtain an unweighted sample. Several unbiased resampling schemes are listed in Chapter 9 of Chopin & Papaspiliopoulos (2020) and studied in *e.g.* (Gerber et al., 2019; Crisan & Doucet, 2002). The most standard one is *multinomial* resampling which draws samples independently according to their weights. Tempered SMC also requires a family of Markov kernels $(M_\lambda)_\lambda$ so that $M_\lambda : \boldsymbol{\Theta} \to \mathcal{P}(\boldsymbol{\Theta})$ leaves μ_λ invariant. Tempering is illustrated in Figure 2 and in Algorithm 2.

5. Particle EIG contrastive bound approximation and optimization

Algorithms 1 and 2 can then be combined iteratively. We call the resulting algorithm PASOA. Algorithm 2 at step k provides a particle approximation of $p(\boldsymbol{\theta}|\boldsymbol{D}_k)$ used in Algorithm 1 line 3 at step $k+1$ to optimize the next EIG contrastive bound I_{PCE}^{k+1} and get $\boldsymbol{\xi}_{k+1}^*$. A new $\mathbf{y}_{k+1}$ is then measured at $\boldsymbol{\xi}_{k+1}^*$, $\boldsymbol{D}_{k+1}$ is set to $\boldsymbol{D}_k \cup \{\mathbf{y}_{k+1}, \boldsymbol{\xi}_{k+1}^*\}$ and Algorithm 2 is used again to get a particle approximation of $p(\boldsymbol{\theta}|\boldsymbol{D}_{k+1})$ etc. More specifically, the M weighted particles produced by Algorithm 2 are used to approximate the intractable $p(\boldsymbol{\theta}|\boldsymbol{D}_k)$ and to simplify the sampling of the $L+1$ contrastive variables in line 3 of Algorithm 1 at step $k+1$. In practice, line 3 can be performed in different ways. We propose the following one, which has good numerical and asymptotic properties (see Section 6). Due to the use of generally large numbers L (typically $L = 200$) of contrastive samples, expectations of interest, *e.g.* (7,8), are computed in a large dimensional space $\boldsymbol{\Theta}^{L+1}$ even if $\boldsymbol{\Theta}$ is of moderate dimension. A simple way to mitigate the dimension impact is to start from $M = N(L+1)$ particles using a SMC procedure on $\boldsymbol{\Theta}$ and partition them into $L+1$ disjoint subsets of N particles, denoted by $\boldsymbol{\theta}_{k,\ell}^{1:N}$, for $\ell = 0:L$, with their associated weights denoted by $W_{k,\ell}^{1:N}$ with $\sum_{i=1}^N W_{k,\ell}^i = 1$. Seeing particles and weights as random variables, if the $resampling$ procedure used in Algorithm 2 is so that the resampled values are independent conditionally on the previous particles, then the $L+1$ collections of random variables denoted by $\zeta_{k,\ell}^N = \{W_{k,\ell}^i, \boldsymbol{\theta}_{k,\ell}^i\}_{i=1:N}$ are independent and identically distributed (*i.i.d.*) conditionally on the previous particles. For instance, multinomial resampling preserves conditional independence. From the weighted particles produced by Algorithm 2, we then derive $L+1$ *i.i.d.* random probability measures $P_{k,\ell}^N = \sum_{i=1}^N W_{k,\ell}^i \delta_{\boldsymbol{\theta}_{k,\ell}^i}$ ($\delta_{\boldsymbol{\theta}}$ is the Dirac measure at $\boldsymbol{\theta}$) and the gradient estimate at each iteration t, Algorithm 1 line 5, is computed with $\theta_{0,t}^i, \ldots, \theta_{L,t}^i$ values sampled independently from realizations $p_{k,\ell}^N$ of the respective independent particle approximations $P_{k,\ell}^N$ for $i = 1 : N_t$ and $\ell = 0 : L$. We thus define $P_{k,\ell}^N$ on $\boldsymbol{\Theta}$ and use in our SG algorithm approximated distributions on $\boldsymbol{\Theta}^{L+1}$ which are product measures $\otimes_{\ell=0}^L P_{k,\ell}^N$. It follows a SG procedure, detailed in the Appendix (Algorithm 3), which produces an estimator $\boldsymbol{\xi}_{k+1,N}^*$ from the optimization of a particle approximation of I_{PCE}^{k+1} defined as

$$I_{PCE}^{k+1,N}(\boldsymbol{\xi}) = \mathbb{E}_{p(u)\prod_{\ell=0}^L p_{k,\ell}^N(\boldsymbol{\theta}_\ell)}\Big[F(\boldsymbol{\xi},\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L,T_{\boldsymbol{\theta}_0}^{\boldsymbol{\xi}}(U))\Big] \quad (10)$$

The advantage of such a product form particle approximation over standard SMC lies in the following observation. Standard SMC would consist in using $\prod_{\ell=0}^L p(\boldsymbol{\theta}_\ell) \approx P_k^N(\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L)$ where

$$P_k^N = \sum_{i=1}^N \left(\prod_{\ell=0}^L W_{k,\ell}^i\right) \delta_{(\boldsymbol{\theta}_{k,0}^i,\cdot,\boldsymbol{\theta}_{k,L}^i)} \,. \quad (11)$$

As an alternative, with the same M weights and particles, our *product form* approach is using $\prod_{\ell=0}^L p(\boldsymbol{\theta}_\ell) \approx \prod_{\ell=0}^L p_{k,\ell}^N(\boldsymbol{\theta}_\ell)$, which can be rewritten as a sum, *i.e.*

$$\prod_{\ell=0}^L p_{k,\ell}^N = \sum_{i_0=1}^N \cdot \sum_{i_L=1}^N \left(\prod_{\ell=0}^L W_{k,\ell}^{i_\ell}\right) \delta_{(\boldsymbol{\theta}_{k,0}^{i_0},\cdot,\boldsymbol{\theta}_{k,L}^{i_L})} \,. \quad (12)$$

There are two main differences between (11) and (12). First, (12) is more statistically efficient: the $\boldsymbol{\theta}_{k,\ell}^i$ are *i.i.d.* and so all permutations $(\boldsymbol{\theta}_{k,0}^{i_0}, \cdot, \boldsymbol{\theta}_{k,L}^{i_L})$ of these samples, for $1 \le i_0, \ldots, i_L \le N$, are identically distributed. Hence (12) averages over N^{L+1} tuples while its conventional counterpart (11) only averages over N tuples. This increase in tuple number leads to a decrease in estimators variance. With the same particles, product-form (12) makes the most out of every sample available. Second, in contrast to tuples in (11), those in (12) are not independent because the same $\boldsymbol{\theta}_{k,\ell}^i$ appears in many terms of the sum. As a consequence, using product form approximations requires to generalize

standard SMC results, which rely on the independence of the terms in the sum. In the next section, we take that into account to show the convergence of the quantities involved in PASOA.

6. Asymptotic properties

With the contrastive bound approach, to increase approximations quality, we can play with L the number of contrastive samples and N the number of particles. It has already been shown by Foster et al. (2020) that increasing L improves the quality of the I_{PCE} bound that becomes tight, *i.e.* to the limit equal to the EIG. We then study the behavior of our approximations when the number N of particles tends to ∞. The sequential design values produced can be seen as realizations of random estimators $\{\boldsymbol{\xi}^*_{k+1,N}\}_{N\geq 1}$ targeting points of maximum of random criterion functions $\{I^{k+1,N}_{PCE}(\boldsymbol{\xi})\}_{N\geq 1}$. We use here the same notation for random quantities and their realizations. We first establish the convergence of the criterion functions adapting standard SMC techniques in Proposition 1, and then the convergence of their points of maximum, *i.e.* the consistency of the $\{\boldsymbol{\xi}^*_{k+1,N}\}_{N\geq 1}$ to the optimal designs in Proposition 2.

Convergence of product particle approximations. With the use of contrastive samples, expectations are computed in $\boldsymbol{\Theta}^{L+1}$. Particles could then be considered as $(L+1)$-dimensional elements in this space so that standard SMC convergence results could apply (Chopin & Papaspiliopoulos, 2020), Chapter 11. However, working in a $(L+1)$-dimensional space is more challenging for SMC and not necessary for the type of convergence result we need. As importance sampling, SMC suffers from the weights' variance which scales unfavorably with the dimension of the problem (Naesseth et al., 2019; Chopin & Papaspiliopoulos, 2020). In contrast, in Section 5, we define $P^N_{k,\ell}$ on $\boldsymbol{\Theta}$ and use in our SG algorithm (Algorithm 3 in Appendix) approximated distributions on $\boldsymbol{\Theta}^{L+1}$ which are product measures $\otimes_{\ell=0}^{L} P^N_{k,\ell}$. Such products have been used in various settings but have been only recently studied in a more general way (Kuntz et al., 2022). However, theoretical results therein do not cover the use of product form estimators within SMC samplers. Other papers that consider *structured* SMC settings (Rebeschini & van Handel, 2015; Lindsten et al., 2017; Kuntz et al., 2024; Aitchison, 2019) differ from our work in key aspects further discussed in Appendix Section C.1.

Let $\mathcal{C}_b(\boldsymbol{\Theta}^{L+1})$ denote the set of functions $\phi : \boldsymbol{\Theta}^{L+1} \to \mathbb{R}$ that are measurable and bounded, and let $||\phi||_\infty$ denote the supremum norm $||\phi||_\infty = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^{L+1}} |\phi(\boldsymbol{\theta})|$. Using the notation and terminology of Chopin & Papaspiliopoulos (2020) and previous authors before them, we introduce the potential function $G_{k,\tau}$ which in our setting is equal to $G_{k,\tau}(\boldsymbol{\theta}) = p(\mathbf{y}_k|\boldsymbol{\theta},\boldsymbol{\xi}_k)^\gamma$ and used in Algorithm 2 to compute weights. We simplified the notation but γ may also depend on k and τ. More specifically, in standard tempering, γ is a preset constant but in Algorithm 2, we use an adaptive tempering where γ depends on previous particles. Although adaptive tempering schemes are often more challenging to study, *e.g.* (Salomone et al., 2018; Beskos et al., 2016; Del Moral et al., 2012), this is not problematic in our setting as $\gamma \leq 1$ and all we require is $G_{k,\tau}$ to be bounded.

For our result, the assumption that $p(\mathbf{y}_k|\boldsymbol{\theta},\boldsymbol{\xi}_k)$ is bounded and strictly positive is enough but more general results and cases are available in (Beskos et al., 2016; Jasra et al., 2011). The following proposition establishes the convergence in L_2-norm, which also implies the convergence in L_1-norm and in probability.

Following Chopin & Papaspiliopoulos (2020); Crisan & Doucet (2002), we can establish similarly L_p-norm ($p > 2$) and almost-sure convergence but convergence in probability is sufficient for consistency of the design estimators. We denote by $\zeta^N_{k,L}$ the collection of random variables $\zeta^N_{k,L} = \{W^i_{k,\ell}, \boldsymbol{\theta}^i_{k,\ell}\}_{i=1:N,\ell=0:L}$, and use the short notation $p_k^{\otimes L+1}[\phi]$ for $p_k^{\otimes L+1}[\phi] = \mathbb{E}_{\prod_{\ell=0}^{L} p(\boldsymbol{\theta}_\ell|\boldsymbol{D}_k)}[\phi(\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L)]$.

Proposition 1 (L_2 convergence) *Using Algorithm 2 with multinomial resampling and assuming that all potential functions $G_{k,\tau}$ are upper bounded, there exists a constant c_k such that, for all functions $\phi \in \mathcal{C}_b(\boldsymbol{\Theta}^{L+1})$,*

$$\mathbb{E}_{\zeta^N_{k,L}}\left[\left\{\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\prod_{\ell=0}^{L} W^{i_\ell}_{k,\ell}\ \phi(\boldsymbol{\theta}^{i_0}_{k,0},\cdot,\boldsymbol{\theta}^{i_L}_{k,L}) - p_k^{\otimes L+1}[\phi]\right\}^2\right] \leq c_k\ c_{N,L}\ \|\phi\|^2_\infty$$

with $c_{N,L} = 1-(1-\frac{1}{N})^{L+1} \approx \frac{L+1}{N}$ and where the expectation is taken over all the realizations of the random tempered SMC method, or equivalently on $\zeta^N_{k,L}$.

The proof is detailed in Appendix and extends the proof in Chapter 11 of Chopin & Papaspiliopoulos (2020) which is itself mainly based on (Crisan & Doucet, 2002). Note that the first term in the outer expectation above corresponds to the expectation of ϕ with respect to $\otimes_{\ell=0}^{L} P^N_{k,\ell}$, which is a random variable as the $P^N_{k,\ell}$ are random measures. For any $\boldsymbol{\xi} \in \mathcal{E}$, when ϕ is set to $\phi_{\boldsymbol{\xi}}(\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L) = f_{PCE}(\boldsymbol{\xi},\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L)$, with $f_{PCE}(\boldsymbol{\xi},\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L) = \mathbb{E}_{p(u)}\left[F(\boldsymbol{\xi},\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L,T^{\boldsymbol{\xi}}_{\boldsymbol{\theta}_0}(U))\right]$ assumed to be bounded on $\boldsymbol{\Theta}^{L+1}$ for all $\boldsymbol{\xi}$, then $p_k^{\otimes L+1}[\phi_{\boldsymbol{\xi}}] = I^{k+1}_{PCE}(\boldsymbol{\xi})$, the PCE lower bound (9) we seek to optimize at step $k+1$. Similarly, the expectation of $\phi_{\boldsymbol{\xi}}$ with respect to $\otimes_{\ell=0}^{L} P^N_{k,\ell}$ is $I^{k+1,N}_{PCE}(\boldsymbol{\xi})$ in (10). It follows from Proposition 1 that

$$\mathbb{E}_{\zeta^N_{k,L}}\left[\left\{I^{k+1,N}_{PCE}(\boldsymbol{\xi}) - I^{k+1}_{PCE}(\boldsymbol{\xi})\right\}^2\right] \leq c_k\ c_{N,L}\ \|\phi\|^2_\infty,$$

or in other words that for all $\boldsymbol{\xi} \in \mathcal{E}$, the sequence of random variables $\{I_{PCE}^{k+1,N}(\boldsymbol{\xi})\}_{N\geq 1}$ converges in L_2-norm to $I_{PCE}^{k+1}(\boldsymbol{\xi})$ when N tends to ∞. Using Chebyshev's inequality this also implies the convergence in probability, pointwise in $\boldsymbol{\xi}$, that is for all $\boldsymbol{\xi}$, for all $\epsilon > 0$,

$$\lim_{N\to\infty} p_{\zeta_{k,L}^N}\left(\left|I_{PCE}^{k+1,N}(\boldsymbol{\xi}) - I_{PCE}^{k+1}(\boldsymbol{\xi})\right| \geq \epsilon\right) = 0. \quad (13)$$

To further comment this result and give more insight on the constants involved, we note that interestingly, if the normalized potential functions are bounded by 1 (see Appendix Section C.2 for details), the constant denoted by c_k in Proposition 1 remains low and decreases when L increases. Otherwise, the constant may increase very fast with k and, even if k is relatively low in BOED, the bound may become uninformative for finite N. This is actually a well referenced behavior in SMC, see *e.g.* (Chopin & Papaspiliopoulos, 2020). In our work, we limited the presentation to *simple* conditions but with stronger assumptions, in particular on the Markov kernels, it is possible to limit the growing of c_k over steps k, see Section 11.4. in (Chopin & Papaspiliopoulos, 2020).

Consistency of the sequential design estimators. Under additional assumptions, specified in Proposition 2, it follows from Proposition 1 that the design parameters found at each step of our procedure tend to the ideal ones.

Assume $I_{PCE}^{k+1}(\boldsymbol{\xi})$ reaches its maximum in $\boldsymbol{\xi}_{k+1}^*$ and consider a sequence $\{\boldsymbol{\xi}_{k+1,N}^*\}_{N\geq 1}$ of estimators defined by

$$\boldsymbol{\xi}_{k+1,N}^* \in \arg\max_{\boldsymbol{\xi}\in\mathcal{E}} I_{PCE}^{k+1,N}(\boldsymbol{\xi}) \,. \quad (14)$$

Proposition 2 states that $\{\boldsymbol{\xi}_{k+1,N}^*\}_{N\geq 1}$ are consistent estimators of $\boldsymbol{\xi}_{k+1}^*$, *i.e.* that for all $\epsilon > 0$,

$$\lim_{N\to\infty} p_{\zeta_{k,L}^N}\left(\|\boldsymbol{\xi}_{k+1,N}^* - \boldsymbol{\xi}_{k+1}^*\| \geq \epsilon\right) = 0. \quad (15)$$

The assumptions in Proposition 2 are in a simplified form. More general situations could be handled but would require more complex developments while not changing the general idea. To simplify the presentation, our first assumption (A1) is that the design space $\mathcal{E}$ is compact. The second assumption (A2) is to ensure that the point of maximum $\boldsymbol{\xi}_{k+1}^*$ is isolated or well-separated and that only design values in the neighborhood of $\boldsymbol{\xi}_{k+1}^*$ reaches values close to $I_{PCE}^{k+1}(\boldsymbol{\xi}_{k+1}^*)$ (see Lemma 5 in Appendix). Assumption (A3) indicates that estimators $\boldsymbol{\xi}_{k+1,N}^*$ could be points of near maximum only. The formulation of (A3) uses additional variables ρ_N that account for the fact that estimators $\boldsymbol{\xi}_{k+1,N}^*$ are usually only approximations of the maxima in (14), as in practice the optimization task is solved numerically with a given precision that can be controlled in the sense that a bound on the approximation error can be provided. If we assume that $\boldsymbol{\xi}_{k+1,N}^*$ is an exact maximizer of $I_{PCE}^{k+1,N}(\boldsymbol{\xi})$, then (A3) is trivially satisfied with $\rho_N = 0$. In our setting, $\boldsymbol{\xi}_{k+1,N}^*$ is obtained by a stochastic gradient algorithm for which we could exhibit the ρ_N sequence but this is a particular case left to the interested reader. This point of view, the assumptions and the technique of proof we are using, are similar to the ones used for establishing asymptotic properties of M-estimators (van der Vaart, 1998; van der Vaart & Wellner, 1996). A proof is given in Appendix, using the pointwise convergence (13) from Proposition 1 and several intermediate results.

Proposition 2 (Consistency) *Assume*

A1 *$\mathcal{E} \in \mathbb{R}^d$ is a compact set.*

A2 *For all $\boldsymbol{\xi} \neq \boldsymbol{\xi}_{k+1}^*$, $I_{PCE}^{k+1}(\boldsymbol{\xi}) < I_{PCE}^{k+1}(\boldsymbol{\xi}_{k+1}^*)$*

A3 *There exists a sequence $\{\rho_N\}_{N\geq 1}$ of positive random variables and a sequence of random variables $\{\boldsymbol{\xi}_{k+1,N}^*\}_{N\geq 1}$ in $\mathcal{E}$ that satisfy*

$$\forall \epsilon > 0, \quad \lim_{N\to\infty} p_{\zeta_{k,L}^N}(\rho_N \geq \epsilon) = 0$$

$$\liminf_{N\to\infty} p_{\zeta_{k,L}^N}\left(I_{PCE}^{k+1,N}(\boldsymbol{\xi}_{k+1,N}^*) \geq I_{PCE}^{k+1,N}(\boldsymbol{\xi}_{k+1}^*) - \rho_N\right) = 1$$

Then the sequence $\{\boldsymbol{\xi}_{k+1,N}^*\}_{N\geq 1}$ is consistent, i.e. for all $\epsilon > 0$,

$$\lim_{N\to\infty} p_{\zeta_{k,L}^N}\left(\|\boldsymbol{\xi}_{k+1,N}^* - \boldsymbol{\xi}_{k+1}^*\| \geq \epsilon\right) = 0.$$

7. Numerical experiments

To benchmark our method in terms of information gained, we use the *sequential prior contrastive estimation* (SPCE) and *sequential nested Monte Carlo* (SNMC) bounds introduced in (Foster et al., 2021) and used in (Blau et al., 2022). For a number K of experiments leading to $\boldsymbol{\xi}_1, \cdot, \boldsymbol{\xi}_K$, and L contrastive variables, SPCE and SNMC are respectively lower and upper bounds for the total EIG, *i.e.* the expected information gained from the entire sequence $\boldsymbol{\xi}_1, \cdot, \boldsymbol{\xi}_K$ and they become tight when L tends to ∞. Their exact expressions are given in Appendix Sections D.1 and D.2. They have the advantage of using only samples from the prior $p(\boldsymbol{\theta})$ and not from the successive posterior distributions, thus not considering posterior approximation errors and making them convenient criteria to compare methods on design sequences only. Methods can be compared via their [SPCE, SNMC] intervals which contain the total EIG. However, one benefit of our approach is to also provide accurate posterior estimation for subsequent accurate parameter estimation. To assess this aspect, we use the L_2 Wasserstein distance between the weighted

particles yielded by our method and the true parameter $\boldsymbol{\theta}$. We consider two other recent approaches, namely a reinforcement learning-based approach RL-BOED from (Blau et al., 2022) and the *variational prior contrastive estimation* VPCE of Foster et al. (2020). RL-BOED is a non-myopic approach, while VPCE involves iteratively optimizing I_{PCE} in a myopic manner but estimating posterior distributions with variational inference. We also compare with a non tempered version of our approach and with a random baseline, where the observations $\{\mathbf{y}_1, \cdot, \mathbf{y}_K\}$ are simulated with designs generated randomly. For methods that do not provide posterior estimations or poor quality ones (RL-BOED, VPCE, Random), we compute Wasserstein distances on posterior samples obtained by using tempered SMC on their design and observation sequences. In contrast, naive SMC Wasserstein distances are computed on the SMC posterior samples to better assess the impact of tempering via the comparison with PASOA.

Source location. We consider the 2D location finding experiment used in (Foster et al., 2021; Blau et al., 2022). It consists of S hidden sources in $\mathbb{R}^2$ whose locations $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \cdot, \boldsymbol{\theta}_S\}$ are unknown. Each source emits a signal whose intensity attenuates according to the inverse-square law. The measured signal is the superposition of all these signals. The design problem is to choose where to make the measurements to best learn the source locations. If a measurement is performed at a point $\boldsymbol{\xi} \in \mathbb{R}^2$, the signal strength is $\mu(\boldsymbol{\theta}, \boldsymbol{\xi}) = b + \sum_{s=1}^{S} \frac{\alpha_s}{m + ||\boldsymbol{\theta}_s - \boldsymbol{\xi}||_2^2}$ where α_s, b and m are constants. A standard Gaussian prior is assumed for each $\boldsymbol{\theta}_s \sim \mathcal{N}(0, \boldsymbol{I})$ and the likelihood is assumed log-normal, $(\log \mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\xi}) \sim \mathcal{N}(\log \mu(\boldsymbol{\theta}, \boldsymbol{\xi}), \sigma)$ with standard deviation σ. We set $S = 2$, $\alpha_1 = \alpha_2 = 1$, $m = 10^{-4}$, $b = 10^{-1}$, $\sigma = 0.5$ and we plan $K = 30$ successive design optimisations. The Markov kernel is that of a Metropolis-Hasting scheme with a Gaussian proposal centered at the current particle with a variance set to the empirical variance of the particles. We use $L = 200$ contrastive variables for the I_{PCE}^k bound. Algorithm 2 is used to get $N = 100$ simulations $\boldsymbol{\theta}_\ell^{1:N}$ of each contrastive variable. The Adam algorithm (Kingma & Ba, 2015) is then used with standard hyperparameters to perform the stochastic gradient.

The whole experiment is repeated 100 times but drawing source locations at random each time. Figure 3, column 1, shows, with respect to k, the median and standard error for SPCE, SNMC and the L_2 Wasserstein distances between weighted particles and the true source locations.

A first observation is that design optimization leads to a significant improvement over the naive random baseline. For VPCE and RL-BOED we recover the results shown in Figure 1 of Blau et al. (2022). Our method leads to a significant improvement, both in terms of information gain and posterior estimation. It improves by 30% RL-BOED results on SPCE and provides much higher SNMC. The L_2 Wasserstein distance is two order of magnitude lower, suggesting the higher quality of our measurements.

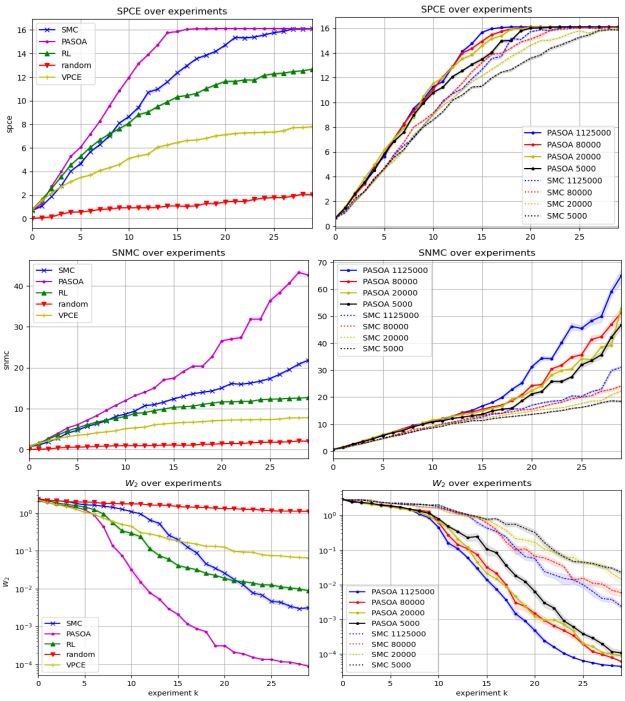

Figure 3. Source location. Column 1: median and standard error over 100 rollouts for SPCE (top), SNMC (middle) and L_2 Wasserstein distance (bottom) with respect to the number of experiments k. Column 2: impact of the number of particles (5K to 1M) on median SPCE, SNMC and Wasserstein distance for PASOA (plain) and SMC (dotted). Note the logarithmic scale.

Both plain SMC and PASOA outperform the other methods. Additional illustration of the benefit of tempering is visible in Figure 1 where PASOA better estimates the posterior concentrating around the true sources when SMC leads to more scattered particles and is less efficient in removing particles from non informative locations. Tempering allows to reduce the number of particles, outperforming SMC even with much fewer particles, *e.g.* $5e^3$ *vs* 10^6 for a lower computation time (Figure 3, column 2, and Table 1). The number of tempering steps tends to decrease with the number of experiments as more information is gathered (Appendix Figure 6). PASOA is also more robust to prior misspecification (Figure 7).

Constant Elasticity of Substitution (CES). In this other model (Blau et al., 2022; Foster et al., 2020), an agent compares two baskets of goods with 3 items each and gives a rating in $[0, 1]$. The design is a 6-dimensional vector representing quantities for each item in each basket. There are 3 parameters $\boldsymbol{\theta} = (\rho, \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3), u)$ in dimension 5, which have to be recovered from the

Table 1. Impact of the number of particles on average computation times for Source location on a V100 GPU and on average total tempering steps for PASOA.

Particles	50 000	100 000	320 000	1 000 000
SMC	25.1s $\pm$ 0.1s	30.1s $\pm$ 0.1s	61.7 $\pm$ 0.1s	130.1 $\pm$ 0.2s
PASOA	25.9s $\pm$ 0.1s	31.4s $\pm$ 0.1s	63.4 $\pm$ 0.1s	134.4 $\pm$ 0.2s
Tempering steps	120.2	120.0	116.1	114.2

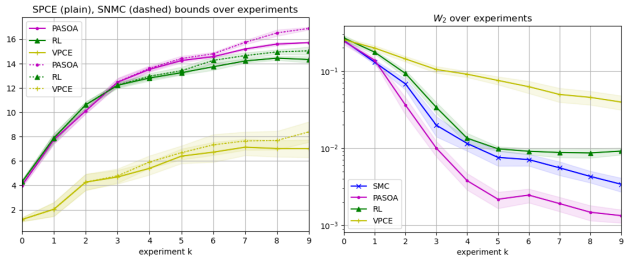

Figure 4. CES example. Median and standard error over 100 rollouts, with respect to the number of experiments k, for SPCE -plain , SNMC -dashed (left) and Wasserstein distance (right).

agent's ratings of different baskets. The design task is challenging since most basket configurations provide very little information. Figure 4 shows the obtained SPCE, SNMC and Wasserstein distance curves for $K = 10$ experiments. SMC-based methods outperform again RL-BOED and VPCE, both in terms of design sequences and Wasserstein distances. For PASOA, the [SPCE, SNMC] intervals (Figure 4 left) are distinctly above the other ones. For clarity sake, the SMC [SPCE, SNMC] intervals are not plotted but shown in Appendix Figure 8.

For a global view of the methods, Table 2 summarizes the main features and running times, with more explanations in Appendix Section D.3. Additional experimental results and implementation details are also given in the Appendix. The chosen models are benchmarks used in BOED. A real-word application would be a great addition but is out of the scope of this paper.

Table 2. Main features of the compared methods. Column 5 shows training times required for amortization.

Method	Posterior	Amortized	Non-myopic	Training Time
PASOA	✓	✗	✗	—
SMC	✓	✗	✗	—
RL-BOED (Blau et al., 2022)	✗	✓	✓	CES: $\sim$ 20h Sources: $\sim$ 10h
VPCE (Foster et al., 2020)	✓	✗	✗	—

8. Conclusion

We have introduced a new Bayesian sequential design optimization algorithm which also provides posterior distribution estimates for parameter inference. The procedure uses a tempering principle to handle the fact that maximizing information gain jeopardizes standard SMC samplers accuracy. Although greedy, our approach performs better than the long-sighted reinforcement learning approach of Blau et al. (2022). As already observed by Blau et al. (2022), a possible explanation is the use of posterior information in our optimisation process. Moreover, the lesser performance of VPCE, which uses suboptimal variational posterior approximations, suggests that the key to a good EIG optimization is an accurate posteriors estimation, which can be achieved via SMC and further improved with tempering. Accurately estimating posterior distributions seems more important and more critical than planning multiple steps ahead. In addition, when the design parameter is high dimensional, all methods are likely to suffer from the difficulty of searching in a high dimensional design space, but as a myopic approach PASOA is likely to suffer less than non myopic approaches for which the search space grows exponentially as the number of experiments increases. Nevertheless, investigating the possibility to extend our approach via a policy-based method would be interesting. Another difficulty may come from a high dimensional parameter. Augmenting the parameter dimension makes inference more difficult. If posterior inference fails, PASOA may loose its specific advantage and result in design of lesser quality. Methods that do not rely on these posterior estimation are less prone to this loss of design performance. However, the goal of experimental design is to provide informative data to infer and ultimately gain information on the parameter of interest. Even if other methods may then provide better SPCE values, this is useless if the posterior is impossible to estimate correctly. In practice, for PASOA, one might use a more sophisticated tuned Markov kernel and exploit the advantage of tempering that a good kernel allows to scale to higher dimensions, see Section 17.2.3 in (Chopin & Papaspiliopoulos, 2020). At last, investigating amortized simulation-based inference such as (Ivanova et al., 2021; Kleinegesse & Gutmann, 2021; Kleinegesse et al., 2020) to handle models which are only available through simulations would be an important next step for applications.

Acknowledgements

The authors thank the reviewers and area chair for their interesting and useful comments and the Inria Challenge project ROAD-AI for partial funding. This work was performed using HPC/AI resources from GENCI-IDRIS (Grant 2023-AD011014217R1). Pierre Alliez is supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency ANR-19-P3IA-0002.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pp. 405–431, 2017.

Aitchison, L. Tensor Monte Carlo: Particle Methods for the GPU era. In *Advances in Neural Information Processing Systems 2019*, pp. 7146–7155, 2019.

Amzal, B., Bois, F., Parent, E., and Robert, C. P. Bayesian-Optimal Design via Interacting Particle Systems. *Journal of the American Statistical Association*, 101(474):773–785, 2006.

Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Dedieu, A., Fantacci, C., Godwin, J., Jones, C., Hemsley, R., Hennigan, T., Hessel, M., Hou, S., Kapturowski, S., Keck, T., Kemaev, I., King, M., Kunesch, M., Martens, L., Merzic, H., Mikulik, V., Norman, T., Papamakarios, G., Quan, J., Ring, R., Ruiz, F., Sanchez, A., Schneider, R., Sezener, E., Spencer, S., Srinivasan, S., Stokowiec, W., Wang, L., Zhou, G., and Viola, F. The DeepMind JAX Ecosystem, 2020.

Benassi, R., Bect, J., and Vazquez, E. Bayesian optimization using sequential monte carlo. In *Learning and Intelligent Optimization. LION 2012. Lecture Notes in Computer Science*, volume 7219, pp. 339–342. Springer, Berlin, Heidelberg, 11 2011.

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

Blau, T., Bonilla, E. V., Chades, I., and Dezfouli, A. Optimizing sequential experimental design with deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 2107–2128. PMLR, 17–23 Jul 2022.

Borkar, V. *Stochastic approximation: a dynamical view point*. Cambridge University Press, 2008.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2020.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, August 1995.

Chopin, N. and Papaspiliopoulos, O. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

Crisan, D. and Doucet, A. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746, 2002.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Dai, C., Heng, J., Jacob, P. E., and Whiteley, N. An invitation to sequential Monte Carlo samplers. *Journal of the American Statistical Association*, 117(539):1587–1600, June 2022.

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

Del Moral, P., Doucet, A., and Jasra, A. On Adaptive Resampling Procedures for Sequential Monte Carlo Methods. *Bernoulli*, 18(1):252–278, 2012.

Doucet, A. and Lee, A. Sequential Monte Carlo methods. In *Handbook of Graphical Models: 165–189*. Ed. by M. Maathuis, M. Drton, S. L. Lauritzen, and M. Wainwrigth, 2018.

Drovandi, C., McGree, J. M., and Pettitt, A. N. A sequential Monte Carlo algorithm to incorporate model uncertainty in Bayesian sequential design. *Journal of Computational and Graphical Statistics*, 23(1):3–24, March 2014.

Drovandi, C. C., McGree, J., and Pettitt, A. N. Sequential Monte Carlo for Bayesian sequentially designed experiments for discrete data. *Computational Statistics & Data Analysis*, 57(1):320–335, 2013.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., and Goodman, N. Variational Bayesian Optimal Experimental Design. In *Advances in Neural Information Processing Systems*, pp. 14059–14070, 2019.

Foster, A., Jankowiak, M., O'Meara, M., Teh, Y. W., and Rainforth, T. A Unified Stochastic Gradient Approach to Designing Bayesian-Optimal Experiments. In *Proceedings of the 23rd International Conference in Artificial Intelligence and Statistics*, volume 108, pp. 2959–2969, 2020.

Foster, A., Ivanova, D. R., Malik, I., and Rainforth, T. Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 3384–3395. PMLR, July 2021.

Gerber, M., Chopin, N., and Whiteley, N. Negative association, ordering and convergence of resampling methods. *Annals of Statistics*, 47(4):2236–2260, August 2019.

Henrández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. Predictive Entropy Search for Efficient Global Optimization of Black-Box Functions. In *Advances in Neural Information Processing Systems*, volume 1, pp. 918–926, 2014.

Huan, X. and Marzouk, Y. Gradient-based stochastic optimization methods in Bayesian experimental design. *International Journal for Uncertainty Quantification*, 4 (6):479–510, 2014.

Ivanova, D. R., Foster, A., Kleinegesse, S., Gutmann, M. U., and Rainforth, T. Implicit Deep Adaptive Design: Policy-Based Experimental Design without Likelihoods. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25785–25798, 2021.

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

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

Kleinegesse, S. and Gutmann, M. U. Bayesian experimental design for implicit models by mutual information neural estimation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 493, pp. 5316–5326. PMLR, July 2020.

Kleinegesse, S. and Gutmann, M. U. Gradient-based Bayesian Experimental Design for Implicit Models using Mutual Information Lower Bounds. *ArXiv*, abs/2105.04379, 2021.

Kleinegesse, S., Drovandi, C., and Gutmann, M. U. Sequential Bayesian Experimental Design for Implicit Models via Mutual Information. *Bayesian Analysis*, 16: 773–802, July 2020.

Kuck, H., de Freitas, N., and Doucet, A. SMC Samplers for Bayesian Optimal Nonlinear Design. In *IEEE Nonlinear Statistical Signal Processing Workshop*, pp. 99–102, 2006.

Kuntz, J., Crucinio, F., and Johansen, A. Product-form estimators: exploiting independence to scale up Monte Carlo. *Statistics and Computing*, 32(12), 2022.

Kuntz, J., Crucinio, F., and Johansen, A. The divide-and-conquer sequential Monte Carlo algorithm: theoretical properties and limit theorems. *Annals of Applied Probability*, 34(1B):1469–1523, Feb. 2024.

Lao, J. and Louf, R. Blackjax: A sampling library for JAX. *Astrophysics Source Code Library*, 2022.

Lindsten, F., Johansen, A., Naesseth, C., Kirkpatrick, B., Schön, T., Aston, J., and Bouchard-Côté, A. Divide-and-Conquer With Sequential Monte Carlo. *Journal of Computational and Graphical Statistics*, 26(2):445 – 458, 2017.

Moffat, H., Hainy, M., Papanikolaou, N., and Drovandi, C. Sequential experimental design for predator-prey functional response experiments. *Journal of Royal Society Interface*, 17(166), 2020.

Naesseth, C. A., Lindsten, F., and Schön, T. B. Elements of Sequential Monte Carlo. *Foundations and Trends in Machine Learning*, 12(3):307–392, 2019.

Neal, R. Annealed importance sampling. *Statistics and Computing*, 11:125–139, April 2001.

Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F. Modern Bayesian Experimental Design. *Statistical Science*, 39(1):100–114, Feb. 2024.

Rebeschini, P. and van Handel, R. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866, Oct. 2015.

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N. A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, 84(1):128–154, April 2016.

Salomone, R., South, L., Drovandi, C. C., and Kroese, D. P. Unbiased and Consistent Nested Sampling via Sequential Monte Carlo. *ArXiv*, 2018.

Sebastiani, P. and Wynn, H. P. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.

Snoek, J., Larochelle, H., and Adams, R. P. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 2951–2959, 2012.

Syed, S., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. Non-reversible parallel tempering: a scalable highly parallel MCMC scheme. *Journal of the Royal Statistical Society (Series B)*, 84:321–350, 2021.

van der Vaart, A. W. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Process: With Applications to Statistics*. springer, 1996.

In this appendix, we provide proofs for the theoretical results in our paper and additional details on the implementation and numerical experiments.

A. Particle EIG contrastive bound optimization

The algorithm below summarizes the procedure developed for design optimization. It is the result of the combination of Algorithms 1 and 2 in the main body. It results from the replacement in the former of the current posterior $p(\boldsymbol{\theta}|\boldsymbol{D}_{k-1})$ by its particle approximation resulting from the adaptive tempering in Algorithm 2.

Algorithm 3 Particle EIG contrastive bound stochastic optimization at step $k+1$

Set T iterations, $\boldsymbol{\xi}_0$, stepsizes $(\gamma_t)_{t=1:T}$
Run tempered SMC Algorithm 2 (main body) at step k for $M = N(L+1)$, and set $L+1$ independent particle approximations $P^N_{k,\ell}$, for $\ell = 0:L$ (see main paper Section 5 for details)
while $t \leq T$ **do**
  Sample $\boldsymbol{\theta}^i_{\ell,t} \sim P^N_{k,\ell}$, for $\ell = 0:L, i = 1:N_t$
  Sample $u^i_t \sim p(u)$, for $i = 1:N_t$
  Set $\nabla_{t+1} = \frac{1}{N_t}\sum_{i=1}^{N_t} \nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi}, \boldsymbol{\theta}^i_{0t}, \cdot, \boldsymbol{\theta}^i_{Lt}, T^{\boldsymbol{\xi}}_{\boldsymbol{\theta}^i_{0t}}(u^i_t))|_{\boldsymbol{\xi}=\boldsymbol{\xi}_t}$
  Update $\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t + \gamma_t \nabla_{t+1}$
return $\boldsymbol{\xi}^*_{k+1,N} = \boldsymbol{\xi}_T$ or a Polyak averaging value

B. Differentiation under the integral sign

We briefly recall standard conditions under which it is possible to exchange differentiation and expectation operators. In our setting this implies conditions that are also useful for other results in Section C.3. Indeed to carry out a stochastic gradient algorithm (Algorithms 1 and 3), we assume that all quantities are well defined. In particular,

$$\nabla_{\boldsymbol{\xi}} I_{PCE}(\boldsymbol{\xi}) = \mathbb{E}_{p(u)\prod_{\ell=0}^{L} p(\boldsymbol{\theta}_\ell)}\left[\nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, T^{\boldsymbol{\xi}}_{\boldsymbol{\theta}_0}(U))\right]. \tag{16}$$

Denoting $p_u \otimes p_\theta^{\otimes L+1}$ the product probability distribution $p(u)\prod_{\ell=0}^{L} p(\boldsymbol{\theta}_\ell)$ on $\mathcal{U} \times \boldsymbol{\Theta}^{L+1}$, sufficient conditions for (16) are that the function F is differentiable in $\boldsymbol{\xi}$ for $p_u \otimes p_\theta^{\otimes L+1}$-almost all $(u, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L)$ and its gradient satisfies $\left|\nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, T^{\boldsymbol{\xi}}_{\boldsymbol{\theta}_0}(u))\right| \leq H(\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, u)$ where H has a bounded expectation. In particular this implies the continuity properties (ii) and (iii) required in Lemma 4 and 5 below.

In practice, it can also be useful to notice that the gradient required within the expectation (16) can be expressed using only the log-likelihood gradient. To simplify, in the expression below we denote $T^{\boldsymbol{\xi}}_{\boldsymbol{\theta}_0}(u)$ by y. Using the definition of F, it comes,

$$\nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, y) = \nabla_{\boldsymbol{\xi}} \log p(y|\boldsymbol{\xi}, \boldsymbol{\theta}_0) - \sum_{\ell=0}^{L} w_\ell(\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, \boldsymbol{\xi}, y)\, \nabla_{\boldsymbol{\xi}} \log p(y|\boldsymbol{\xi}, \boldsymbol{\theta}_\ell) \tag{17}$$

where the $w_\ell(\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, \boldsymbol{\xi}, y) = \frac{p(y|\boldsymbol{\xi}, \boldsymbol{\theta}_\ell)}{\sum_{\ell'=0}^{L} p(y|\boldsymbol{\xi}, \boldsymbol{\theta}_{\ell'})}$ sum to 1 and act as weights. Thus we only need to compute for all $\boldsymbol{\theta}_0, \boldsymbol{\theta}_\ell$, the gradient $\nabla_{\boldsymbol{\xi}} \log p(T^{\boldsymbol{\xi}}_{\boldsymbol{\theta}_0}(u)|\boldsymbol{\xi}, \boldsymbol{\theta}_\ell)$.

C. Proofs of main and intermediate results

We first specify the steps that lead to Proposition 1 in the main body. The overall goal is to show that products of particle approximations have similar convergence properties than single particle approximations.

C.1. Product form estimators

Product form estimators have been used in various places in the literature but only studied in a more formal and general way in some recent work (Kuntz et al., 2022). Their advantages are clearly highlighted in (Kuntz et al., 2022) but the

theoretical results therein do not cover the use of product form estimators within SMC samplers. In contrast, (Rebeschini & van Handel, 2015; Lindsten et al., 2017; Kuntz et al., 2024; Aitchison, 2019) consider SMC settings that may look similar to ours but differ in key aspects. Indeed, (Rebeschini & van Handel, 2015) proposed a so-called block particle filter consisting of a product of local smaller particle filters to approximate particle filters in high dimensional models. The last line of their Algorithm 2 clearly shows that the block particle filter is equivalent to running several independent particle filters on smaller dimensional spaces. In contrast, we run a single SMC from which we extract at each step a product form estimate of the target distribution. This allows to compute expectations and other inferential quantities in a more efficient way, using the advantages of product form estimators, notably their lower variance, as explained in (Kuntz et al., 2022). The Divide-and-Conquer SMC (DaC-SMC) methodology introduced in (Lindsten et al., 2017) and further studied in (Kuntz et al., 2024) targets a different objective, which is to generalize the sequential nature of standard SMC to a more general tree-structured execution flow of the sampler. Hierarchical decompositions of variables of interest into subsets are considered and product form estimators are used to recombined the splitted subsets of particles. Typical use cases include sets of variables exhibiting a graph dependence structure such as Markov random fields. In contrast, in our work, the execution flow remains sequential, corresponding to a tree that is a simple line, and we do not address complex parameter structures. At last, the Tensor Monte Carlo technique proposed in (Aitchison, 2019) shares similarity with DaC-SMC but focuses on computational aspects and does not cover as many theoretical properties. The work in (Aitchison, 2019) proposes to exploit dependence structures in variables and conditional independence properties to address the issue of computing the very large sums induced by product form estimators. This is not an issue we encounter as, in our work, the computation of very large sums is avoided via stochastic approximation.

The advantage of using a product of particle approximations justifies a new analysis of the SMC procedure. Interestingly, the product-form can also help to mitigate the growth of constant c_k in the L_2 bound in Proposition 1 (c'_τ in the proof below). Increasing L can allow to reduce c_k to an extent that can make the bound informative for finite sample N. See the remark at the end of Section C.2.

C.2. Convergence of product form particle approximations

The result is intuitive and generalizes standard SMC proofs but as mentioned in the previous section, we believe it is not covered in previous work. It requires additional care as illustrated in Lemma 1 below. We follow the presentation in Chapter 11 of Chopin & Papaspiliopoulos (2020) which is mainly based on the presentation in (Crisan & Doucet, 2002). As mentioned in (Chopin & Papaspiliopoulos, 2020) it is not the most general presentation but it has the advantage of using standard tools while presenting the general idea. For simplicity, we only present the L_2 convergence result, which is enough to prove Proposition 1.

To simplify the notation, we use τ as the time or step index and μ_τ, M_τ for , μ_{λ_τ}, M_{λ_τ} as in Algorithm 2, to denote the step and successive targets and Markov kernels, but this extends to k and p_k, as we repeat the same process sequentially. For a SMC procedure on $\boldsymbol{\Theta}$, G_τ denotes then the so-called potential function involved in the weights normalization and $\tilde{G}_\tau = G_\tau/(\mu_{\tau-1}M_{\tau-1}(G_\tau))$ its normalization. For example, in our setting in Section 6, $G_\tau(\boldsymbol{\theta}) = G_{k,\tau}(\boldsymbol{\theta}) = p(\mathbf{y}_k|\boldsymbol{\theta}, \boldsymbol{\xi}_k)^\gamma$. Recall that for a function ϕ on $\boldsymbol{\Theta}$, we then have

$$\mu_\tau[\phi] = \mathbb{E}_{\mu_\tau}[\phi] = \mu_{\tau-1}M_{\tau-1}[\phi\ \tilde{G}_\tau] = \int \tilde{G}_\tau(\boldsymbol{\theta})\phi(\boldsymbol{\theta}) \left\{ \int M_{\tau-1}(\boldsymbol{\theta}', \boldsymbol{\theta})\ \mu_{\tau-1}(\boldsymbol{\theta}')\ d\boldsymbol{\theta}' \right\} d\boldsymbol{\theta}. \tag{18}$$

As a particular case, we have

$$\mu_\tau(\boldsymbol{\theta}_\ell) = \mu_\tau[\delta_{\boldsymbol{\theta}_\ell}] = \tilde{G}_\tau(\boldsymbol{\theta}_\ell) \left\{ \int M_{\tau-1}(\boldsymbol{\theta}', \boldsymbol{\theta}_\ell)\ \mu_{\tau-1}(\boldsymbol{\theta}')\ d\boldsymbol{\theta}' \right\}\ d\boldsymbol{\theta} \tag{19}$$

$$\mu_{\tau-1}M_{\tau-1}[\tilde{G}_\tau] = \mu_\tau[1] = 1 \tag{20}$$

We generalize these definitions and results on the product space $\boldsymbol{\Theta}^{L+1}$ with ϕ being now a function of $L+1$ variables. For a function ϕ on $\boldsymbol{\Theta}^{L+1}$, we use notation $M_\tau\phi$ for the function

$$M_\tau\phi : (\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L) \to \int \cdot \int \phi(\boldsymbol{\theta}'_0, \cdot, \boldsymbol{\theta}'_L) \left(\prod_{\ell=0}^{L} M_\tau(\boldsymbol{\theta}_\ell, \boldsymbol{\theta}'_\ell)\ d\boldsymbol{\theta}'_\ell \right). \tag{21}$$

Similarly, we denote $\mu_\tau^{\otimes L+1}[\phi] = \mathbb{E}_{\mu_\tau^{\otimes L+1}}[\phi] = \int \cdot \int \phi(\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L) \left(\prod_{\ell=0}^L \mu_\tau(\boldsymbol{\theta}_\ell)\, d\boldsymbol{\theta}_\ell\right)$. Defining,

$$(\mu_{\tau-1}M_{\tau-1})^{\otimes L+1}[\phi] = \int \cdot \int \phi(\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L) \prod_{\ell=0}^L \left\{\int M_{\tau-1}(\boldsymbol{\theta}'_\ell, \boldsymbol{\theta}_\ell)\, \mu_{\tau-1}(\boldsymbol{\theta}'_\ell)\, d\boldsymbol{\theta}'_\ell\right\} d\boldsymbol{\theta}_\ell.$$

we can deduce from (18) (respectively via (19) and (20) that

$$\mu_\tau^{\otimes L+1}[\phi] = (\mu_{\tau-1}M_{\tau-1})^{\otimes L+1}\left[\phi \prod_{\ell=0}^L \tilde{G}_\tau(\boldsymbol{\theta}_\ell)\right], \tag{22}$$

$$(\mu_{\tau-1}M_{\tau-1})^{\otimes L+1}\left[\prod_{\ell=0}^L \tilde{G}_\tau(\boldsymbol{\theta}_\ell)\right] = \mu_\tau^{\otimes L+1}[1] = 1. \tag{23}$$

The strategy is to establish Proposition 1 by decomposing the error at step τ as a sum of contributions from each step (sampling, reweighting, resampling, etc.) of the current and previous steps. We first need a general result on variances of sum of random variables.

Lemma 1 *Let $\{X_\ell^i\}_{\ell=0:L, i=1:N}$ a collection of i.i.d. random variables on $\mathcal{X}$. For any function ϕ on $\mathcal{X}^{L+1}$ so that the variances exist we have, for all $(i_0, \cdot, i_L) \in [1:N]^{L+1}$,*

$$Var\left(\phi(X_0^{i_0}, \cdot, X_L^{i_L})\right) = Var\left(\phi(X_0^1, \cdot, X_L^1)\right)$$

and we denote this common variance by $Var(\phi_1)$. We can then bound the following variance

$$Var\left(\sum_{i_0=1}^N \cdot \sum_{i_L=1}^N \phi(X_0^{i_0}, \cdot, X_L^{i_L})\right) \leq N^{L+1}\left(N^{L+1} - (N-1)^{L+1}\right) Var(\phi_1)$$

Proof. Let us denote $I = [1:N]$, $J = I^{L+1}$, $j = (i_0, \cdot, i_L)$ for $i_\ell \in I$, and use the short notation $\phi_j = \phi(X_0^{i_0}, \cdot, X_L^{i_L})$. Then

$$\begin{aligned} Var\left(\sum_{i_0=1}^N \cdot \sum_{i_L=1}^N \phi(X_0^{i_0}, \cdot, X_L^{i_L})\right) &= Var(\sum_{j\in J} \phi_j) = \sum_{j\in J} Var(\phi_j) + \sum_{j\in J}\sum_{j'\neq j} Cov(\phi_{j'}, \phi_j) \\ &= N^{L+1} Var(\phi_1) + \sum_{j\in J}\sum_{j'\neq j} Cov(\phi_{j'}, \phi_j)\ . \end{aligned}$$

The second covariance term is a sum of $N^{2(L+1)} - N^{L+1}$ pairwise covariances among which a number are zero due to the independence of the $\{X_\ell^i\}_{\ell=0:L, i=1:N}$. To count them, we notice that for each $j = (i_0, \cdot, i_L) \in J$, $Cov(\phi_j, \phi'_j) = 0$ for every $j' \in J$ such that $j' = (i'_0, \cdot, i'_L)$ with $i'_\ell \in I \backslash \{i_\ell\}$. There are thus $(N-1)^{L+1}$ such j'. It follows that there are altogether $N^{L+1}(N-1)^{L+1}$ null covariance terms and $N^{L+1}(N^{L+1} - (N-1)^{L+1} - 1)$ non null ones. For each non null covariance, the Cauchy-Schwartz inequality implies that $Cov(\phi_{j'}, \phi_j) \leq \sqrt{Var(\phi_{j'})Var(\phi_j)} = Var(\phi_1)$, which concludes the proof. □

We then recall some notation and definitions specific to our product form particle approximation. The tempered SMC procedure in Algorithm 2 produces $M = N(L+1)$ particles denoted by $\boldsymbol{\theta}_\tau^{1:M}$ at each step indexed by τ. When partitioning these particles into $L+1$ disjoint subset we will denote by $\boldsymbol{\theta}_{\ell,\tau}^{1:N}$ the N particles in each subset with $\ell = 0 : L+1$. We denote by $\zeta_{k,L}^N$ the collection of random variables $\zeta_{k,L}^N = \{W_{k,\ell}^i, \boldsymbol{\theta}_{k,\ell}^i\}_{i=1:N, \ell=0:L}$, and use the short notation $p_k^{\otimes L+1}[\phi]$ for $p_k^{\otimes L+1}[\phi] = \mathbb{E}_{\prod_{\ell=0}^L p(\boldsymbol{\theta}_\ell|\boldsymbol{D}_k)}[\phi(\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L)]$. We write $\mathcal{F}_{\tau-1}$ for the σ-algebra $\mathcal{F}_{\tau-1} = \sigma\left(\boldsymbol{\theta}_{\tau-1}^{1:M}\right)$. Let $\mathcal{C}_b(\boldsymbol{\Theta}^{L+1})$ denote the set of functions $\phi : \boldsymbol{\Theta}^{L+1} \to \mathbb{R}$ that are measurable and bounded, and let $||\phi||_\infty$ denote the supremum norm $||\phi||_\infty = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}^{L+1}} |\phi(\boldsymbol{\theta})|$. The following lemma bounds the Monte Carlo error at step τ.

Lemma 2 (Monte Carlo error) *Using Algorithm 2 with a multinomial resampling scheme, for all functions $\phi \in \mathcal{C}_b(\boldsymbol{\Theta}^{L+1})$,*

$$\mathbb{E}_{\zeta^N_{k,L}}\left[\left\{\frac{1}{N^{L+1}}\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\phi(\boldsymbol{\theta}^{i_0}_{0,\tau},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau})-\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\left(\prod_{\ell=0}^L W^{i_\ell}_{\ell,\tau-1}\right)M_{\tau-1}\phi(\boldsymbol{\theta}^{i_0}_{0,\tau-1},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau-1})\right\}^2\right]\leq c_{N,L}\|\phi\|^2_\infty \quad (24)$$

with $c_{N,L}=1-\left(1-\frac{1}{N}\right)^{L+1}$ and where the expectation is taken over all the realizations of the random tempered SMC method, or equivalently on $\zeta^N_{k,L}$.

Proof. Multinomial resampling preserves conditional independence so that the particles $\boldsymbol{\theta}^{1:N}_{\ell,\tau}$ are *i.i.d.* conditionally on $\mathcal{F}_{\tau-1}$ and their conditional distribution is

$$\boldsymbol{\theta}^i_{\ell,\tau}|\mathcal{F}_{\tau-1}\sim\sum_{i=1}^N W^i_{\ell,\tau-1}M_{\tau-1}(\boldsymbol{\theta}^i_{\ell,\tau-1},\cdot).$$

Thus for all $i_0,\cdot,i_L$,

$$\begin{aligned}\mathbb{E}_{\zeta^N_{k,L}}\left[\phi(\boldsymbol{\theta}^{i_0}_{0,\tau},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau})\mid\mathcal{F}_{\tau-1}\right]&=\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\left(\prod_{\ell=0}^L W^{i_\ell}_{\ell,\tau-1}\right)M_{\tau-1}\phi(\boldsymbol{\theta}^{i_0}_{0,\tau-1},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau-1})\\&=\mathbb{E}_{\zeta^N_{k,L}}\left[\frac{1}{N^{L+1}}\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\phi(\boldsymbol{\theta}^{i_0}_{0,\tau},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau})\mid\mathcal{F}_{\tau-1}\right].\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbb{E}_{\zeta^N_{\tau,L}}\left[\left\{\frac{1}{N^{L+1}}\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\phi(\boldsymbol{\theta}^{i_0}_{0,\tau},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau})-\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\left(\prod_{\ell=0}^L W^{i_\ell}_{\ell,\tau-1}\right)M_{\tau-1}\phi(\boldsymbol{\theta}^{i_0}_{0,\tau-1},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau-1})\right\}^2\mid\mathcal{F}_{\tau-1}\right]\\&=Var\left[\frac{1}{N^{L+1}}\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\phi(\boldsymbol{\theta}^{i_0}_{0,\tau},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau})\mid\mathcal{F}_{\tau-1}\right]=\frac{1}{N^{2(L+1)}}Var\left[\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\phi(\boldsymbol{\theta}^{i_0}_{0,\tau},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau})\mid\mathcal{F}_{\tau-1}\right]\\&\leq c_{N,L}\,Var\left[\phi(\boldsymbol{\theta}^1_{0,\tau},\cdot,\boldsymbol{\theta}^1_{L,\tau})\mid\mathcal{F}_{\tau-1}\right]\\&\leq c_{N,L}\,\mathbb{E}_{\zeta^N_{\tau,L}}\left[\phi(\boldsymbol{\theta}^1_{0,\tau},\cdot,\boldsymbol{\theta}^1_{L,\tau})^2\mid\mathcal{F}_{\tau-1}\right]\leq c_{N,L}\,\|\phi\|^2_\infty\end{aligned}$$

The line before last above results from Lemma 1. Using the tower property, we get (24). □

The constant $c_{N,L}$ is equivalent to $\frac{L+1}{N}$ and tends to 0 when N tends to ∞. The next Lemma involves the potential functions G_τ. Note that in Section 6, G_τ corresponds to $G_{k,\tau}(\boldsymbol{\theta})=p(\mathbf{y}_k|\boldsymbol{\theta},\boldsymbol{\xi})^\gamma$, where γ usually also depends on τ and k as it is found adaptively. However all we need is that the potential and then the likelihood is upper bounded in $\boldsymbol{\theta}$.

Lemma 3 (Weights normalization error) *If G_τ is upper bounded, then for all functions $\phi\in\mathcal{C}_b(\boldsymbol{\Theta}^{L+1})$,*

$$\begin{aligned}&\mathbb{E}_{\zeta^N_{\tau,L}}\left[\left\{\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\left(\prod_{\ell=0}^L W^{i_\ell}_{\ell,\tau}\right)\phi(\boldsymbol{\theta}^{i_0}_{0,\tau},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau})-\frac{1}{N^{L+1}}\sum_{i_0=1}^N\cdot\sum_{i_L=1}^N\left(\prod_{\ell=0}^L\tilde{G}_\tau(\boldsymbol{\theta}^{i_\ell}_{\ell,\tau})\right)\phi(\boldsymbol{\theta}^{i_0}_{0,\tau},\cdot,\boldsymbol{\theta}^{i_L}_{L,\tau})\right\}^2\right]\\&\qquad\leq\|\phi\|^2_\infty\,\mathbb{E}_{\zeta^N_{k,L}}\left[\left\{\prod_{\ell=0}^L\left(\frac{1}{N}\sum_{i=1}^N\tilde{G}_\tau(\boldsymbol{\theta}^i_{\ell,\tau})\right)-1\right\}^2\right]\end{aligned}$$

where $\tilde{G}_\tau=G_\tau/(\mu_{\tau-1}M_{\tau-1}(G_\tau))$.

Proof. By definition

$$W^{i_\ell}_{\ell,\tau}=\frac{G_\tau(\boldsymbol{\theta}^{i_\ell}_{\ell,\tau})}{\sum_{i=1}^N G_\tau(\boldsymbol{\theta}^i_{\ell,\tau})}=\frac{\tilde{G}_\tau(\boldsymbol{\theta}^{i_\ell}_{\ell,\tau})}{\sum_{i=1}^N\tilde{G}_\tau(\boldsymbol{\theta}^i_{\ell,\tau})}$$

Thus the term in the square in the left hand side is

$$\left(\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\left(\prod_{\ell=0}^{L}W_{\ell,\tau}^{i_\ell}\right)\phi(\boldsymbol{\theta}_{0,\tau}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau}^{i_L})\right)\left(1-\frac{1}{N^{L+1}}\prod_{\ell=0}^{L}(\sum_{i=1}^{N}\tilde{G}_\tau(\boldsymbol{\theta}_{\ell,\tau}^{i}))\right)$$

and the first factor is bounded by $\|\phi\|_\infty$ since $\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\left(\prod_{\ell=0}^{L}W_{\ell,\tau}^{i_\ell}\right)=1$. □

The two Lemmas can now be combined to establish the proposition below.

Proposition 1 ($\mathbf{L}_2$ convergence) *Using Algorithm 2 with multinomial resampling and assuming that potential functions G_τ are upper bounded, there exists constants c_τ, c'_τ so that, for all functions $\phi\in\mathcal{C}_b(\boldsymbol{\Theta}^{L+1})$,*

$$\mathbb{E}_{\zeta_{k,L}^N}\left[\left\{\frac{1}{N^{L+1}}\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\phi(\boldsymbol{\theta}_{0,\tau}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau}^{i_L})-(\mu_{\tau-1}M_{\tau-1})^{\otimes L+1}[\phi]\right\}^2\right]\leq c_\tau\ c_{N,L}\ \|\phi\|_\infty^2 \tag{25}$$

$$\mathbb{E}_{\zeta_{k,L}^N}\left[\left\{\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\left(\prod_{\ell=0}^{L}W_{\ell,\tau}^{i_\ell}\right)\phi(\boldsymbol{\theta}_{0,\tau}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau}^{i_L})-\mu_\tau^{\otimes L+1}[\phi]\right\}^2\right]\leq c'_\tau\ c_{N,L}\ \|\phi\|_\infty^2 \tag{26}$$

with $c_{N,L}=1-(1-\frac{1}{N})^{L+1}$.

Proof The proof works by induction on τ. Replacing $\mu_{\tau-1}M_{\tau-1}$ by μ_0, at $\tau=0$, (25) holds with $c_0=1$. Assume (25) holds at step τ, we first show that (26) also holds at step τ. The left hand side in (26) can be decomposed into

$$\begin{aligned}&\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\left(\prod_{\ell=0}^{L}W_{\ell,\tau}^{i_\ell}\right)\phi(\boldsymbol{\theta}_{0,\tau}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau}^{i_L})-\mu_\tau^{\otimes L+1}[\phi]=\\&\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\left(\prod_{\ell=0}^{L}W_{\ell,\tau}^{i_\ell}\right)\phi(\boldsymbol{\theta}_{0,\tau}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau}^{i_L})-\frac{1}{N^{L+1}}\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\left(\prod_{\ell=0}^{L}\tilde{G}_\tau(\boldsymbol{\theta}_{\ell,\tau}^{i_\ell})\right)\phi(\boldsymbol{\theta}_{0,\tau}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau}^{i_L})\\&+\frac{1}{N^{L+1}}\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\left(\prod_{\ell=0}^{L}\tilde{G}_\tau(\boldsymbol{\theta}_{\ell,\tau}^{i_\ell})\right)\phi(\boldsymbol{\theta}_{0,\tau}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau}^{i_L})-\mu_\tau^{\otimes L+1}[\phi]\ .\end{aligned}$$

For all ϕ, when $\phi'(\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L)=\left(\prod_{\ell=0}^{L}\tilde{G}_\tau(\boldsymbol{\theta}_\ell)\right)\phi(\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L)$, we have $(\mu_{\tau-1}M_{\tau-1})^{\otimes L+1}[\phi']=\mu_\tau^{\otimes L+1}[\phi]$ by (22). The second term can then be bounded by $c_\tau c_{N,L}\|\phi'\|_\infty^2$, applying (25) with ϕ'. For the first term, we apply Lemma 3 and then (25) with $\phi(\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L)=\prod_{\ell=0}^{L}\tilde{G}_\tau(\boldsymbol{\theta}_\ell)$ since then $(\mu_{\tau-1}M_{\tau-1})^{\otimes L+1}[\phi]=\mu_\tau^{\otimes L+1}[1]=1$ by (23). Finally we get (26) with $c'_\tau=4c_\tau\|\tilde{G}_\tau\|_\infty^{2(L+1)}$ where we have used that $\mathbb{E}[(X+Y)^2]\leq 2(\mathbb{E}[X^2]+\mathbb{E}[Y^2])$. We then show that (26) at step $\tau-1$ implies (25) at step τ. Similarly, we have,

$$\begin{aligned}&\frac{1}{N^{L+1}}\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\phi(\boldsymbol{\theta}_{0,\tau}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau}^{i_L})-(\mu_{\tau-1}M_{\tau-1})^{\otimes L+1}[\phi]=\\&\frac{1}{N^{L+1}}\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\phi(\boldsymbol{\theta}_{0,\tau}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau}^{i_L})-\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\left(\prod_{\ell=0}^{L}W_{\ell,\tau-1}^{i_\ell}\right)M_{\tau-1}\phi(\boldsymbol{\theta}_{0,\tau-1}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau-1}^{i_L})\\&+\sum_{i_0=1}^{N}\cdot\sum_{i_L=1}^{N}\left(\prod_{\ell=0}^{L}W_{\ell,\tau-1}^{i_\ell}\right)M_{\tau-1}\phi(\boldsymbol{\theta}_{0,\tau-1}^{i_0},\cdot,\boldsymbol{\theta}_{L,\tau-1}^{i_L})-(\mu_{\tau-1}M_{\tau-1})^{\otimes L+1}[\phi]\ .\end{aligned}$$

The first term can be bounded using Lemma 2 and for the second term by applying (26) at $\tau-1$ to function $M_{\tau-1}\phi$ defined in (21). Then (25) holds with $c_\tau=2(1+c'_{\tau-1})$. □

Proposition 1 in the paper is just the result above but reduced to the presentation of (26). To give more insight on the constants involved in Proposition 1, we note that the constant denoted by c_k corresponds to c'_τ in the above proof. A well referenced behavior in SMC, see *e.g.* (Chopin & Papaspiliopoulos, 2020), is that this constant may increase very fast with k and become uninformative for finite N. Without considering tempering for simplicity, it comes that $c_0 = 1$ and $c_k = 4(2 + c_{k-1})\|\tilde{G}\|_\infty^{2(L+1)}$, where $\|\tilde{G}\|_\infty$ denotes the upper bound in $\boldsymbol{\theta}$ of $\tilde{G}_k(\boldsymbol{\theta}) \propto p(\mathbf{y}_k|\boldsymbol{\theta}, \boldsymbol{\xi}_k)$, which is assumed here to be independent on the specific value of $\mathbf{y}_k$ and $\boldsymbol{\xi}_k$. It follows that $c_k = 2\sum_{i=1}^{k}(4\|\tilde{G}\|_\infty^{2(L+1)})^i$. Interestingly, with our product form approximation, if $\|\tilde{G}\|_\infty < 1$, c_k remains low and decreases when L increases.

C.3. Consistency of the sequential design estimators

We specify the steps that lead to the proof of Proposition 2 in the main body. We first recall some notation and definitions. The sequential design values produced by Algorithm 3 above can be seen as realizations of random estimators $\{\boldsymbol{\xi}^*_{k+1,N}\}_{N\geq 1}$ targeting points of maximum of random criterion functions $\{I_{PCE}^{k+1,N}(\boldsymbol{\xi})\}_{N\geq 1}$,

$$I_{PCE}^{k+1,N}(\boldsymbol{\xi}) = \mathbb{E}_{p(u)\prod_{\ell=0}^{L} p_{k,\ell}^{N}(\boldsymbol{\theta}_\ell)}\left[F(\boldsymbol{\xi},\boldsymbol{\theta}_0,\cdot,\boldsymbol{\theta}_L,T_{\boldsymbol{\theta}_0}^{\boldsymbol{\xi}}(U))\right] .$$

A natural question is to study the limiting distributions of these random quantities when the number of particles N tends to infinity. The main property is the following Theorem 1, which results from a general result from M-estimator theory (van der Vaart, 1998; van der Vaart & Wellner, 1996). The two following lemmas are then useful to provide simpler sufficient conditions to satisfy the Theorem's assumptions and to establish our main result in Proposition 2 in the main body, also recalled below.

Theorem 1 (Theorem 5.7 in (van der Vaart, 1998)) *Assume*

(i) For all $\epsilon > 0$,

$$\lim_{N\to\infty} p_{\zeta_{k,L}^N}\left(\sup_{\boldsymbol{\xi}\in\mathcal{E}} |I_{PCE}^{k+1,N}(\boldsymbol{\xi}) - I_{PCE}^{k+1}(\boldsymbol{\xi})| \geq \epsilon\right) = 0. \tag{27}$$

(ii) For all $\epsilon > 0$,

$$\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}^*_{k+1}\|\geq\epsilon} I_{PCE}^{k+1}(\boldsymbol{\xi}) < I_{PCE}^{k+1}(\boldsymbol{\xi}^*_{k+1}) . \tag{28}$$

(iii) There exists a sequence of positive random variables $\{\rho_N\}_{N\geq 1}$ and a sequence of random variables $\{\boldsymbol{\xi}^*_{k+1,N}\}_{N\geq 1}$ in $\mathcal{E}$ that satisfy

$$\forall\epsilon > 0, \quad \lim_{N\to\infty} p_{\zeta_{k,L}^N}(\rho_N \geq \epsilon) = 0$$

$$\lim\inf_{N\to\infty} p_{\zeta_{k,L}^N}\left(I_{PCE}^{k+1,N}(\boldsymbol{\xi}^*_{k+1,N}) \geq I_{PCE}^{k+1,N}(\boldsymbol{\xi}^*_{k+1}) - \rho_N\right) = 1.$$

Then the sequence of estimators $\{\boldsymbol{\xi}^*_{k+1,N}\}_{N\geq 1}$ is consistent, i.e. for all $\epsilon > 0$,

$$\lim_{N\to\infty} p_{\zeta_{k,L}^N}\left(\|\boldsymbol{\xi}^*_{k+1,N} - \boldsymbol{\xi}^*_{k+1}\| \geq \epsilon\right) = 0.$$

Proof. The proof is a special case of Theorem 5.7 in (van der Vaart, 1998). We reproduce it using our notation. In all the following proofs, to simplify, we drop the $k+1$ notation, so that $I_{PCE}^{k+1}(\boldsymbol{\xi})$ and $I_{PCE}^{k+1,N}(\boldsymbol{\xi})$ are now simply denoted by $I_{PCE}(\boldsymbol{\xi})$ and $I_{PCE}^{N}(\boldsymbol{\xi})$ and their respective maximizers by $\{\boldsymbol{\xi}^*\}$ and $\{\boldsymbol{\xi}^*_N\}$. Since $\boldsymbol{\xi}^*$ maximizes $I_{PCE}(\boldsymbol{\xi})$, it comes that for all $N \geq 1$,

$$\begin{aligned} 0 &\leq I_{PCE}(\boldsymbol{\xi}^*) - I_{PCE}(\boldsymbol{\xi}^*_N) \\ &= (I_{PCE}(\boldsymbol{\xi}^*) - I_{PCE}^N(\boldsymbol{\xi}^*)) + (I_{PCE}^N(\boldsymbol{\xi}^*) - I_{PCE}^N(\boldsymbol{\xi}^*_N)) + (I_{PCE}^N(\boldsymbol{\xi}^*_N) - I_{PCE}(\boldsymbol{\xi}^*_N)) . \end{aligned}$$

The first and third terms in the sum are bounded by $\sup_{\boldsymbol{\xi}\in\mathcal{E}} |I_{PCE}(\boldsymbol{\xi}) - I^N_{PCE}(\boldsymbol{\xi})|$ while the second term can be bounded by $\rho_N + (I^N_{PCE}(\boldsymbol{\xi}^*) - I^N_{PCE}(\boldsymbol{\xi}^*_N) - \rho_N)\ \delta_{\{I^N_{PCE}(\boldsymbol{\xi}^*) - I^N_{PCE}(\boldsymbol{\xi}^*_N) > \rho_N\}}$ where $\delta_{\mathcal{A}}$ is the indicator function which is 1 if $\mathcal{A}$ is satisfied and 0 otherwise. It follows that

$$\begin{aligned} 0 &\leq I_{PCE}(\boldsymbol{\xi}^*) - I_{PCE}(\boldsymbol{\xi}^*_N) \\ &\leq 3\max\left(2\sup_{\boldsymbol{\xi}\in\mathcal{E}} |I_{PCE}(\boldsymbol{\xi}) - I^N_{PCE}(\boldsymbol{\xi})|, \rho_N, \left(I^N_{PCE}(\boldsymbol{\xi}^*) - I^N_{PCE}(\boldsymbol{\xi}^*_N) - \rho_N\right)\delta_{\{I^N_{PCE}(\boldsymbol{\xi}^*) - I^N_{PCE}(\boldsymbol{\xi}^*_N) > \rho_N\}}\right). \end{aligned}$$

Assumptions (i) and (iii) imply that all three terms in the max tend to 0 in probability so that for all $\eta > 0$,

$$\lim_{N\to\infty} p_{\zeta^N_{k,L}}(I_{PCE}(\boldsymbol{\xi}^*) - I_{PCE}(\boldsymbol{\xi}^*_N) \geq \eta) = 0. \tag{29}$$

Then, for all $\epsilon > 0$, using (ii), for all $\boldsymbol{\xi}$ satisfying $\|\boldsymbol{\xi} - \boldsymbol{\xi}^*\| \geq \epsilon$, there exists $\eta > 0$ so that $I_{PCE}(\boldsymbol{\xi}) \leq I_{PCE}(\boldsymbol{\xi}^*) - \eta$. This implies that $\{\|\boldsymbol{\xi}^*_N - \boldsymbol{\xi}^*\| \geq \epsilon\} \subset \{I_{PCE}(\boldsymbol{\xi}^*_N) \leq I_{PCE}(\boldsymbol{\xi}^*) - \eta\}$ and

$$p_{\zeta^N_{k,L}}(\|\boldsymbol{\xi}^*_N - \boldsymbol{\xi}^*\| \geq \epsilon) \leq p_{\zeta^N_{k,L}}(I_{PCE}(\boldsymbol{\xi}^*_N) \leq I_{PCE}(\boldsymbol{\xi}^*) - \eta) = p_{\zeta^N_{k,L}}(I_{PCE}(\boldsymbol{\xi}^*) - I_{PCE}(\boldsymbol{\xi}^*_N) \geq \eta).$$

The limit in N of this last term tends to 0 by (29), which concludes the proof. □

In general, the uniform convergence in (i) is the most difficult assumption to check but in our setting, when $\mathcal{E}$ is compact, it is easily derived from previous assumptions and the pointwise convergence in probability ((iv) below), which can be derived from the L_2 convergence in Proposition 1 (see comments in the main body). Recall the following definition,

$$f_{PCE}(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L) = \mathbb{E}_{p(u)}\left[F(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L, T^{\boldsymbol{\xi}}_{\boldsymbol{\theta}_0}(U))\right]$$

and the shortcut notation $p_k^{\otimes L+1} = \prod_{\ell=0}^{L} p(\boldsymbol{\theta}_\ell|\boldsymbol{D}_k)$.

Lemma 4 *Assume*

(i) $\mathcal{E} \in \mathbb{R}^d$ is a compact set.

(ii) $I^{k+1}_{PCE}(\boldsymbol{\xi})$ is a continuous function in $\boldsymbol{\xi}$.

(iii) $f_{PCE}(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L)$ is a continuous function in $\boldsymbol{\xi}$ for $p_k^{\otimes L+1}$-almost all $(\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L)$.

(iv) Convergence in probability pointwise: For all $\boldsymbol{\xi} \in \mathcal{E}$ and all $\epsilon > 0$,

$$\lim_{N\to\infty} p_{\zeta^N_{k,L}}\left(|I^{k+1,N}_{PCE}(\boldsymbol{\xi}) - I^{k+1}_{PCE}(\boldsymbol{\xi})| \geq \epsilon\right) = 0.$$

Then the uniform convergence in Theorem 1 (i) is satisfied, that is, for all $\epsilon > 0$,

$$\lim_{N\to\infty} p_{\zeta^N_{k,L}}\left(\sup_{\boldsymbol{\xi}\in\mathcal{E}} |I^{k+1,N}_{PCE}(\boldsymbol{\xi}) - I^{k+1}_{PCE}(\boldsymbol{\xi})| \geq \epsilon\right) = 0.$$

Proof. Continuous functions on a compact set are uniformly continuous. It follows from (ii) and (iii) that for all $\epsilon > 0$, there exists $\eta > 0$ so that for all $\boldsymbol{\xi}' \in \mathcal{E}$,

$$\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}'\|\leq\eta} |I_{PCE}(\boldsymbol{\xi}) - I_{PCE}(\boldsymbol{\xi}')| \leq \epsilon \tag{30}$$

and for $p_k^{\otimes L+1}$-almost all $(\boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L)$,

$$\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}'\|\leq\eta} |f_{PCE}(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L) - f_{PCE}(\boldsymbol{\xi}', \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L)| \leq \epsilon. \tag{31}$$

Let $\mathcal{B}(\boldsymbol{\xi}, \eta)$ be a ball centered at $\boldsymbol{\xi}$ with radius η. As $\mathcal{E}$ is compact, for all $\eta > 0$, it is possible to extract, from the cover set $\bigcup_{\boldsymbol{\xi}\in\mathcal{E}} \mathcal{B}(\boldsymbol{\xi}, \eta)$, a finite subcover $\bigcup_{b=1:B} \mathcal{B}(\boldsymbol{\xi}^{(b)}, \eta)$ so that $\mathcal{E} \subset \bigcup_{b=1:B} \mathcal{B}(\boldsymbol{\xi}^{(b)}, \eta)$ and

$$\sup_{\boldsymbol{\xi}\in\mathcal{E}} |I^N_{PCE}(\boldsymbol{\xi}) - I_{PCE}(\boldsymbol{\xi})| \quad = \quad \max_{b=1:B} \sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}^{(b)},\eta)} |I^N_{PCE}(\boldsymbol{\xi}) - I_{PCE}(\boldsymbol{\xi})| \,. \tag{32}$$

For all $b = 1 : B$, and all $\boldsymbol{\xi} \in \mathcal{B}(\boldsymbol{\xi}^{(b)}, \eta)$, we also have that

$$|I^N_{PCE}(\boldsymbol{\xi}) - I_{PCE}(\boldsymbol{\xi})| \leq |I^N_{PCE}(\boldsymbol{\xi}) - I^N_{PCE}(\boldsymbol{\xi}^{(b)})| + |I^N_{PCE}(\boldsymbol{\xi}^{(b)}) - I_{PCE}(\boldsymbol{\xi}^{(b)})| + |I_{PCE}(\boldsymbol{\xi}^{(b)}) - I_{PCE}(\boldsymbol{\xi})|$$

which implies

$$\begin{aligned} \sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}^{(b)},\eta)} |I^N_{PCE}(\boldsymbol{\xi}) - I_{PCE}(\boldsymbol{\xi})| \quad &\leq \quad \sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}^{(b)},\eta)} \left\{ |I^N_{PCE}(\boldsymbol{\xi}) - I^N_{PCE}(\boldsymbol{\xi}^{(b)})| \right\} \\ &+ \quad |I^N_{PCE}(\boldsymbol{\xi}^{(b)}) - I_{PCE}(\boldsymbol{\xi}^{(b)})| \\ &+ \quad \sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}^{(b)},\eta)} \left\{ |I_{PCE}(\boldsymbol{\xi}^{(b)}) - I_{PCE}(\boldsymbol{\xi})| \right\} \end{aligned} \tag{33}$$

For the first term in the right-hand side,

$$\sup_{\boldsymbol{\xi}\in\mathcal{B}(\boldsymbol{\xi}^{(b)},\eta)} \left\{ |I^N_{PCE}(\boldsymbol{\xi}) - I^N_{PCE}(\boldsymbol{\xi}^{(b)})| \right\} \leq \mathbb{E}_{p_k^{N\otimes L+1}} \left[\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}^{(b)}\|\leq\eta} |f_{PCE}(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L) - f_{PCE}(\boldsymbol{\xi}^{(b)}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L)| \right]$$

By Proposition 1 (26), the right-hand side above tends in L_2-norm and then in probability when N tends to ∞ to

$$\mathbb{E}_{p_k^{\otimes L+1}} \left[\sup_{\|\boldsymbol{\xi}-\boldsymbol{\xi}^{(b)}\|\leq\eta} |f_{PCE}(\boldsymbol{\xi}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L) - f_{PCE}(\boldsymbol{\xi}^{(b)}, \boldsymbol{\theta}_0, \cdot, \boldsymbol{\theta}_L)| \right] ,$$

which is smaller than ϵ by (31). The second term in the right-hand side of (33) tends in probability to 0 by (iv) for all $\boldsymbol{\xi}^{(b)}$. Finally, using (30), the third term in (33) is smaller than ϵ. To conclude, (32) implies the uniform convergence (27). □

The following result gives simpler sufficient conditions for Assumption (ii) in Theorem 1 to hold.

Lemma 5 *Assume*

(i) $\mathcal{E} \in \mathbb{R}^d$ *is a compact set.*

(ii) $I^{k+1}_{PCE}(\boldsymbol{\xi})$ *is a continuous function in* $\boldsymbol{\xi}$ *.*

(iii) For all $\boldsymbol{\xi} \neq \boldsymbol{\xi}^*_{k+1}$, $I^{k+1}_{PCE}(\boldsymbol{\xi}) < I^{k+1}_{PCE}(\boldsymbol{\xi}^*_{k+1})$ *.*

Then Assumption (ii) *of Theorem 1 is satisfied.*

Proof. A continuous function reaches its maximum on a compact set. It follows that $I^{k+1}_{PCE}(\boldsymbol{\xi})$ reaches its maximum on the compact subset $\mathcal{E}\backslash\mathcal{B}(\boldsymbol{\xi}^*_{k+1}, \epsilon)$. Let $\boldsymbol{\xi}_\epsilon$ denote a value at which this maximum is reached. Then (iii) implies that $I^{k+1}_{PCE}(\boldsymbol{\xi}_\epsilon) < I^{k+1}_{PCE}(\boldsymbol{\xi}^*_{k+1})$, which leads to (28). □

Using Lemmas 4 and 5, we can now replace in Theorem 1 Assumptions (i) and (ii) by simpler conditions. It follows the Proposition 2 presented in the paper and recalled below.

Proposition 2 *Assume*

A1 $\mathcal{E} \in \mathbb{R}^d$ *is a compact set.*

A2 *For all* $\boldsymbol{\xi} \neq \boldsymbol{\xi}^*_{k+1}$, $I^{k+1}_{PCE}(\boldsymbol{\xi}) < I^{k+1}_{PCE}(\boldsymbol{\xi}^*_{k+1})$

A3 *There exists a sequence of positive random variables* $\{\rho_N\}_{N\geq 1}$ *and a sequence of random variables* $\{\boldsymbol{\xi}^*_{k+1,N}\}_{N\geq 1}$ *in* $\mathcal{E}$ *that satisfy*

$$\forall \epsilon > 0, \quad \lim_{N\to\infty} p_{\zeta^N_{k,L}}(\rho_N \geq \epsilon) = 0$$

$$\lim\inf_{N\to\infty} p_{\zeta^N_{k,L}}\Big(I^{k+1,N}_{PCE}(\boldsymbol{\xi}^*_{k+1,N}) \geq I^{k+1,N}_{PCE}(\boldsymbol{\xi}^*_{k+1}) - \rho_N\Big) = 1.$$

Then the sequence of estimators $\{\boldsymbol{\xi}^*_{k+1,N}\}_{N\geq 1}$ *is consistent, i.e. for all* $\epsilon > 0$,

$$\lim_{N\to\infty} p_{\zeta^N_{k,L}}\big(\|\boldsymbol{\xi}^*_{k+1,N} - \boldsymbol{\xi}^*_{k+1}\| \geq \epsilon\big) = 0.$$

Proof. With $\mathcal{E}$ compact, we can use Lemma 4. The continuity of $I^{k+1}_{PCE}(\boldsymbol{\xi})$ and f_{PCE} has been already assumed earlier as specified in Section B and Lemma 4 (iv) is a consequence of Proposition 1. It follows the uniform convergence property (i) in Theorem 1. Then (A1-2) and Lemma 5 imply (ii) in Theorem 1. With (A3), Theorem 1 leads to the result. Note that if we assume that $\boldsymbol{\xi}^*_{k+1,N}$ is an exact maximizer of $I^{k+1,N}_{PCE}(\boldsymbol{\xi})$ then (A3) is trivially satisfied with $\rho_N = 0$.

D. Numerical experiments

D.1. Sequential prior contrastive estimation (SPCE) criterion

We specify the SPCE introduced by Foster et al. (2021) and used in our experiments and those of Blau et al. (2022) to assess the design sequence quality in our comparison. For a number K of experiments, $\boldsymbol{D}_K = \{(\mathbf{y}_1, \boldsymbol{\xi}_1), \cdot, (\mathbf{y}_K, \boldsymbol{\xi}_K)\}$ and L contrastive variables, SPCE is defined as

$$SPCE(\boldsymbol{\xi}_1, \cdot, \boldsymbol{\xi}_K) = \mathbb{E}_{\prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\xi}_k, \boldsymbol{\theta}_0) \prod_{\ell=0}^{L} p(\boldsymbol{\theta}_\ell)} \left[\log \frac{\prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\theta}_0, \boldsymbol{\xi}_k)}{\frac{1}{L+1}\sum_{\ell=0}^{L}\prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\theta}_\ell, \boldsymbol{\xi}_k)} \right]. \tag{34}$$

SPCE is a lower bound of the total EIG which is the expected information gained from the entire sequence of design parameters $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$ and it becomes tight when L tends to ∞. In addition, SPCE has the advantage to use only samples from the prior $p(\boldsymbol{\theta})$ and not from the successive posterior distributions. It makes it a fair criterion to compare methods on design sequences only. Considering a true parameter value denoted by $\boldsymbol{\theta}^*$, given a sequence of design values $\{\boldsymbol{\xi}_k\}_{k=1:K}$, observations $\{\mathbf{y}_k\}_{k=1:K}$ are simulated using $p(\mathbf{y}|\boldsymbol{\theta}^*, \boldsymbol{\xi}_k)$ respectively. Therefore, for a given $\boldsymbol{D}_k$, the corresponding SPCE is estimated numerically by sampling $\boldsymbol{\theta}_1, \cdot, \boldsymbol{\theta}_L$ from the prior,

$$SPCE(\boldsymbol{D}_K) = \frac{1}{N}\sum_{i=1}^{N} \left\{ \log \frac{\prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\theta}^*, \boldsymbol{\xi}_k)}{\frac{1}{L+1}\left(\prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\theta}^*, \boldsymbol{\xi}_k) + \sum_{\ell=1}^{L}\prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\theta}^i_\ell, \boldsymbol{\xi}_k)\right)} \right\}.$$

As shown in (Foster et al., 2021) (Appendix A), SPCE increases with L to reach the total EIG $I(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K)$ when $L \to \infty$ at a rate $\mathcal{O}(L^{-1})$ of convergence. More specifically, it is shown in (Foster et al., 2021) that

$$0 \leq I(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K) - SPCE(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K) \leq \frac{C}{L+1} \tag{35}$$

where $C = \mathbb{E}_{p(\boldsymbol{D}_K)p(\boldsymbol{\theta}|\boldsymbol{D}_K)}\left[\frac{p(\boldsymbol{D}_K|\boldsymbol{\theta})}{p(\boldsymbol{D}_K)}\right] - 1$ with the notation $p(\boldsymbol{D}_K|\boldsymbol{\theta}) = \prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\theta}, \boldsymbol{\xi}_k)$.

It is also shown in (Foster et al., 2021) that for a given L, SPCE is bounded by $\log(L+1)$ while the upper bound SNMC below is potentially unbounded. As in (Blau et al., 2022), if we use $L = 10^7$ to compute SPCE and SNMC, the bound is $\log(L+1) = 16.12$ for SPCE. In practice this does not impact the numerical methods comparison as the intervals [SPCE, SNMC] containing the total EIG remain clearly distinct.

D.2. Sequential nested Monte Carlo (SNMC) criterion

Similarly, an upper bound on the total EIG, with similar features, has also been introduced by Foster et al. (2021). Its expression is very similar to that of SPCE, varying only through the sum in the denominator,

$$SNMC(\boldsymbol{\xi}_1, \cdot, \boldsymbol{\xi}_K) = \mathbb{E}_{\prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\xi}_k, \boldsymbol{\theta}_0) \prod_{\ell=0}^{L} p(\boldsymbol{\theta}_\ell)} \left[\log \frac{\prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\theta}_0, \boldsymbol{\xi}_k)}{\frac{1}{L} \sum_{\ell=1}^{L} \prod_{k=1}^{K} p(\mathbf{y}_k|\boldsymbol{\theta}_\ell, \boldsymbol{\xi}_k)} \right].$$

D.3. Implementation details

For VPCE (Foster et al., 2020) and RL-BOED (Blau et al., 2022), we use the code available at github.com/csiro-mlai/RL-BOED, using the settings recommended therein to reproduce the results in the respective papers. From the obtained sequences of observations and design values, we compute SPCE and SNMC as explained above and retrieve the same results as in their respective papers.

Our code is implemented in Jax (Bradbury et al., 2020) and available at github.com/iolloj/pasoa. Several packages are used through the repository. Namely, we used Optax (Babuschkin et al., 2020) to run Gradient Descents, the Sequential Monte Carlo part was heavily inspired and built using Kernels from BlackJax (Lao & Louf, 2022) and we used OTT (Cuturi et al., 2022) to compute Wasserstein distances.

Table 3 summarizes the main features and running times of the compared methods. The RL-BOED method has the advantage to be both non-myopic and amortized in the sense that a policy is learnt upfront and then used straightforwardly at each new experiment. It follows a much longer training time, which does not exist for the other methods. Note that in comparison the deployment times of all methods are neglible (see Table 1 in the paper). In contrast RL-BOED does not provide approximations for the posterior distributions.

Table 3. Main features and training times of the compared methods: the second column indicates whether a method also provides approximation of posterior distribution, the third if it is amortized and the fourth if it is non-myopic. The last column shows training times for the amortized method RL-BOED and a sequence of K experiments run on a single Nvidia V100 GPU, for the source finding and CES examples.

Method	Posterior	Amortized	Non-myopic	Training Time
PASOA	✓	✗	✗	—
SMC	✓	✗	✗	—
RL-BOED (Blau et al., 2022)	✗	✓	✓	CES: $\sim$ 20h Sources: $\sim$ 10h
VPCE (Foster et al., 2020)	✓	✗	✗	—

D.4. Hardware details

Our method can be run on a local machine and was tested on a Apple M1 Pro 16Gb chip. However, for a faster running time, each experiment was finally produced by running our method on a single Nvidia V100 GPU. One other advantage of tempering and of our PASOA method is that by reducing the number of needed particles for an accurate procedure, it lowers the hardware requirements for this method as it becomes feasible to run it on CPUs (see Table 1).

D.5. Checking the assumptions given in the theoretical results

Ideally, the models used in experiments should satisfy the assumptions appearing in our propositions. For the L_2 convergence result (Proposition 1), the conditions are easy to check. Proposition 1 requires that the potential functions $G_{k,\tau}$ are bounded. It is sufficient to check that the likelihood $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})$ as a function of $\boldsymbol{\theta}$ is bounded (main Section 6 before Proposition 1). For the source location model, the likelihood is log-normal and is bounded independently of $\boldsymbol{\theta}, \mathbf{y}$, and $\boldsymbol{\xi}$. For the CES example, the likelihood is a mixture given in equation (36), Section D.7 below where the last component is a logit-normal distribution.

Both p_0 and p_1 in equations (37) and (38) below are in $[0, 1]$. The only potentially problematic case may be when $\sigma_\eta \to 0$. In that case, $1 - p_0$ tends to 0 and for p_1 we can use the approximation below equation (38). It follows that the third term $(1 - p_0 - p_1)\ q(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\xi})$ in (36) remains bounded.

For the consistency result (Proposition 2), conditions (A1) and (A2) can be stronger than necessary. Note that condition (A3) is not directly related to the model but to the optimization procedure and could be ignored. The important *weaker* condition is (ii) in Theorem 1. Similarly to consistency results in M-estimator theory (see *e.g.* van der Vaart (1998)), in our work, we assume that $\boldsymbol{\xi}^*_{k+1}$ is a global and unique maximum of I^{k+1}_{PCE} ((A2)). Condition (ii) actually states that $\boldsymbol{\xi}^*_{k+1}$ is in addition well separated (see Figure 5.2 of van der Vaart (1998) for an illustration of this notion). Lemma 5 gives sufficient conditions for (ii), which results in (A1) and (A2) in Proposition 2. (A1) is that the design space is compact and (A2) states that $\boldsymbol{\xi}^*_{k+1}$ is a unique global maximum (not necessary well separated). (A1) is easy to check but (A2) is strong and not usually easy to check. Both can be relaxed with additional technicalities, see section 5.2.1 of van der Vaart (1998). For the CES model, the design space is compact. For the source example, it can be restricted to $[-X, X]^2$ without specific care, as *e.g.* in (Blau et al., 2022). For (A2), we have not found yet a general way to check this for the I_{PCE} bound. Note though, that this *unchecked* assumption is common practice as it would require more care to talk about consistency if the maximum was not unique and global.

D.6. Source location example

For the 2D location finding example used in (Foster et al., 2021; Blau et al., 2022) and tested in the paper, with 2 sources, $K = 30$ successive design optimisations, and 100 repetitions of the experiment, the number of gradient steps was set to 5000 and the ESS for the SMC procedure to 0.9. Also, in practice stratified resampling was preferred to multinomial resampling. The latter has the advantage to considerably simplify the proofs and this mismatch between theory and practice is very common in the SMC literature. Figure 3 in the paper shows the SPCE, SNMC and the L_2 Wasserstein distances between weighted particles and the true source locations, providing three quantitative assessment and comparison of methods. As an additional, visual assessment of the quality of the posterior approximation provided by our method, Figure 1 in the paper and Figure 5 below illustrate the evolution of the particles over the design steps, starting from a sample following the prior to a sample concentrating around the true source locations. In particular, the $k = 0$ step shows particles simulated according to the prior. In most use cases, plain SMC already gives better results than other reference methods. Figure 1 and Table 1 in the paper show that tempering allows to reduce the number of particles. In Figure 5 below, the source locations, indicated by red crosses in the plots, are chosen in a part of the space not well covered by the prior to illustrate the robustness of our approach to a potential prior misspecification. After some iterations PASOA is able to explore the parameter space to finally concentrate the posterior on the true source locations. In contrast, SMC may miss some of the sources when they are outside the prior mass or when there are too many of them. We suspect that prior misspecification is a typical very common feature that jeopardizes SMC performance while impacting much less PASOA. This is actually the same problem encountered with IS that tempering aims at solving. Similarly, Figure 7 shows, in terms of SPCE, SNMC and Wasserstein distance, that SMC is more robust with tempering. Figure 6 indicates the number of tempering steps taken on average. The median (over 100 rollouts) number of tempering steps varies from 14 to 2 and is globally decreasing, being under 5 after 15 experiments. The number of tempering steps reduces when the posterior concentrates and when adding new observations becomes less informative.

D.7. Constant Elasticity of Substitution example

This other model, used in (Blau et al., 2022; Foster et al., 2020), comes from behavioral economics. In this model, an agent compares two baskets of goods and gives a rating y on a sliding scale from 0 to 1. The goal is to design the two baskets of goods so as to infer the agent's utility function, which depends on some unknown parameters. The designs are 6-dimensional vectors $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ corresponding to the two baskets with 3 values each $\boldsymbol{\xi}_d = (\boldsymbol{\xi}_{d,1}, \boldsymbol{\xi}_{d,2}, \boldsymbol{\xi}_{d,3}) \in [0 : 100]^3$ for $d = 1, 2$, which represent quantities for 3 items in each basket. There are 3 parameters $\boldsymbol{\theta} = (\rho, \boldsymbol{\alpha}, u)$ in dimension 5, whose prior distributions are respectively $\rho \sim Beta(1, 1)$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3) \sim Dirichlet(1, 1, 1)$ and $\log u \sim \mathcal{N}(1, 3)$.

The model likelihood is given by the following model that uses a subjective utility function U and two hyperparameters

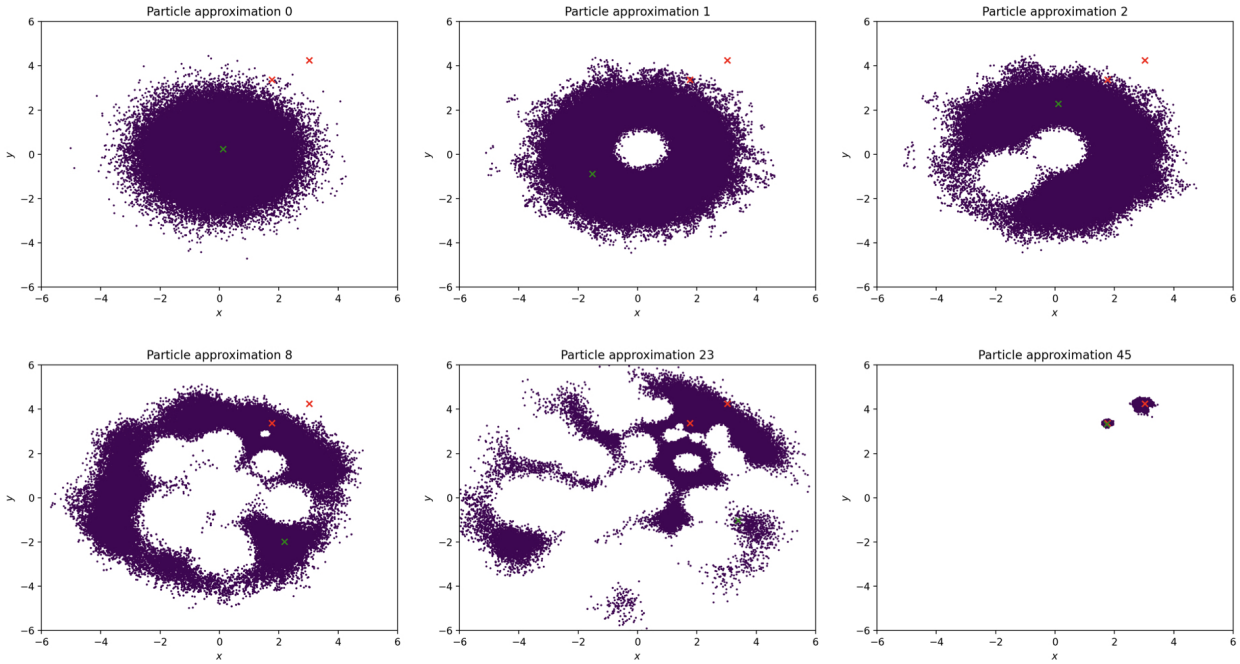

Figure 5. PASOA evolution of particles (in purple) over some selected steps k. Particles correspond initially to a sample from the prior $p(\boldsymbol{\theta})$ and progressively evolve to a sample of particles located around the initially unknown true source positions indicated by red crosses. Green crosses indicate the optimal measurement locations $\boldsymbol{\xi}_k^*$ obtained at each step k.

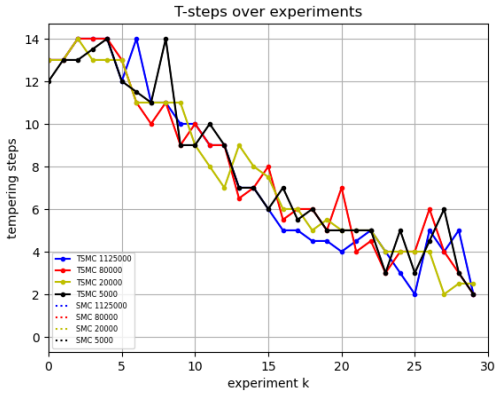

Figure 6. Source location example: median (over 100 rollouts) number of tempering steps over the number of experiments, with respect to the number of particles.

$\epsilon = 2^{-22}$ and $\tau = 0.005$,

$$
\begin{aligned}
y &= f(\eta, \epsilon) \\
\text{where } \eta &\sim \mathcal{N}(\mu_\eta, \sigma_\eta^2) \\
\text{with } \mu_\eta &= (U(\boldsymbol{\xi}_1) - U(\boldsymbol{\xi}_2))\ u \\
\sigma_\eta &= (1 + \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|)\ \tau\ u \\
\text{For } d = 1, 2, \quad U(\boldsymbol{\xi}_d) &= \left(\alpha_1 \boldsymbol{\xi}_{d,1}^\rho + \alpha_2 \boldsymbol{\xi}_{d,2}^\rho + \alpha_3 \boldsymbol{\xi}_{d,3}^\rho\right)^{1/\rho}
\end{aligned}
$$

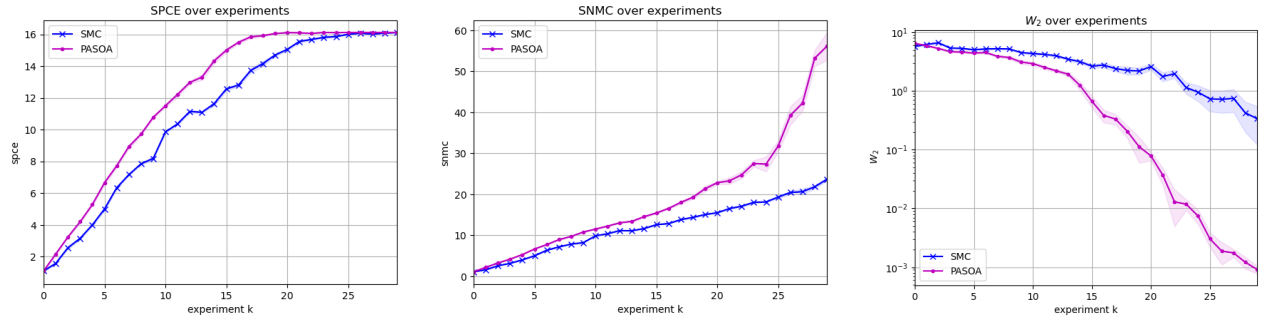

Figure 7. Source location example. Prior misspecification: median (over 100 rollouts) (a) SPCE, (b) SNMC and (c) Wasserstein distances for SMC (blue) and PASOA (red).

where $f(\eta, \epsilon)$ takes it values in $[\epsilon, 1-\epsilon]$ and is a censored sigmoid defined by

$$\begin{aligned} f(\eta, \epsilon) &= 1-\epsilon \quad \text{if } \eta \geq logit(1-\epsilon) \\ &= \epsilon \quad \text{if } \eta \leq logit(\epsilon) \\ &= (1+\exp(-\eta))^{-1} \quad \text{otherwise} \end{aligned}$$

with $logit(y) = \log(y/(1-y))$. In other words, y is a censored logit-normal distribution with parameters μ_η and σ_η. Its density is a mixture

$$p(y|\boldsymbol{\theta}, \boldsymbol{\xi}) = p_0 \delta_\epsilon(y) + p_1 \delta_{1-\epsilon}(y) + (1-p_0-p_1)\, q(y|\boldsymbol{\theta}, \boldsymbol{\xi}) \tag{36}$$

where $q(y|\boldsymbol{\theta}, \boldsymbol{\xi}) = \frac{1}{\sigma_\eta \sqrt{2\pi} y(1-y)} \exp(\frac{(logit(y)-\mu_\eta)^2}{2\sigma_\eta^2})$ is the density of a logit-normal distribution and p_0 and p_1 are defined by the following logit-normal CDF values

$$p_0 = q(y \leq \epsilon) = p(\eta \leq logit(\epsilon)) = F\left(\frac{logit(\epsilon) - \mu_\eta}{\sigma_\eta}\right) \tag{37}$$

$$p_1 = 1 - q(y \leq 1-\epsilon) = 1 - p(\eta \leq logit(1-\epsilon)) = 1 - F\left(\frac{logit(1-\epsilon) - \mu_\eta}{\sigma_\eta}\right) \tag{38}$$

with the last equalities involving the normal CDF values of variable η and the standard normal CDF F values. In practice, computing $\log p_0$ and $\log p_1$ may sometimes be numerically problematic when p_0 or p_1 become too small. Computing p_0 or p_1 involves computing lower and upper Gaussian tails. In this case, following (Foster et al., 2020), we use the following first order asymptotic approximation of the standard normal CDF, when x is large,

$$1 - F(x) \approx \frac{1}{x\sqrt{2\pi}} \exp(-x^2/2).$$

and when x is small (negative)

$$F(x) = 1 - F(-x) \approx \frac{1}{-x\sqrt{2\pi}} \exp(-x^2/2).$$

Thus, denoting f the pdf of the standard normal distribution, $\log p_0 \approx \log f(x) - log(-x)$ with $x = \frac{logit(\epsilon)-\mu_\eta}{\sigma_\eta}$ and $\log p_1 \approx \log f(x) - log(x)$ with $x = \frac{logit(1-\epsilon)-\mu_\eta}{\sigma_\eta}$ or to summarize both approximations when $|x|$ is large, $\log f(x) - \log(|x|)$.

Implementation details, if not otherwise specified, are the same as for the source location example. We plan $K = 10$ successive design optimisations and repeat the whole experiment 100 times for varying values of the true parameters, for all methods, PASOA, SMC, RL-BOED, VPCE and the random design baseline. This is overall a more challenging example as the objective function has many suboptimal local maxima, and the stochastic gradient procedure may be more sensitive to initialization. The number of gradient steps was set to 5000. For the SMC procedure, the ESS was set to 0.9, the Markov kernel is that of a random walk Metropolis-Hasting scheme with prior transformations mapping the parameters to $\mathbb{R}^4$. The transformations used are respectively, $u' = \log u$, $\rho' = logit(\rho) = \log \frac{\rho}{1-\rho}$, and $\alpha_1' = \log \frac{\alpha_1}{\alpha_3}$, $\alpha_2' = \log \frac{\alpha_2}{\alpha_3}$, with the inverse

transformations being $u = \exp u'$, $\rho = \frac{\exp \rho'}{1+\exp \rho'}$, $\alpha_1 = \frac{\exp \alpha_1'}{1+\exp \alpha_1' + \exp \alpha_2'}$, $\alpha_2 = \frac{\exp \alpha_2'}{1+\exp \alpha_1' + \exp \alpha_2'}$ and $\alpha_3 = \frac{1}{1+\exp \alpha_1' + \exp \alpha_2'}$. We use then $L = 100$ contrastive variables with each $N = 500$ simulations.

Figure 4 in the paper and Figures 8 and 9 below show, with respect to k, the median and standard error of the SPCE, SNMC and Wasserstein distances between weighted particles and the true parameters. We observe for all methods more variability in the repetitions for this example. In terms of total EIG, the difference with RL-BOED is not as large as in the source location example, but the difference remains large for Wasserstein distances. PASOA still shows better performance in terms of information gain as measured by SPCE and SNMC. In Figure 4 in the paper, we observe that in experiments 0-2, our approach temporarily loses its advantage over RL-BOED due to insufficiently refined particle approximations of the posteriors. However, this edge is regained in subsequent experiments as more information from the posteriors becomes available. Our better design sequences are also visible in the Wasserstein distance plot presented in Figure 4. Figure 8 left shows on the same plot the SPCE and SNMC curves. Without tempering, SMC gains an advantage only in the latter steps 7-9 in terms of information gained, while, in the Wasserstein distance plot presented in Figure 4, SMC shows better performance from the start. A possible explanation is that, as shown in Figure 9 below, RL-BOED performs better on parameter ρ at the expense of sacrificing precision on the others. Overall the Wasserstein distance for all parameters remains in favor of our methods but it may be that a better precision on ρ leads to a slightly higher information gain (Figure 8 left).

In our current tempering implementation, the Markov kernel is fixed to a standard Metropolis-Hastings scheme for all steps. It is out of the scope of this paper but possible directions for improvement include using more sophisticated kernels, such as Langevin or Hamiltonian Monte Carlo moves, as suggested in the *Tuning parameters* section p.1591 of Dai et al. (2022) and in references therein. More generally, a number of recommendations, as reviewed in (Dai et al., 2022), have been reported as efficient and could be investigated.

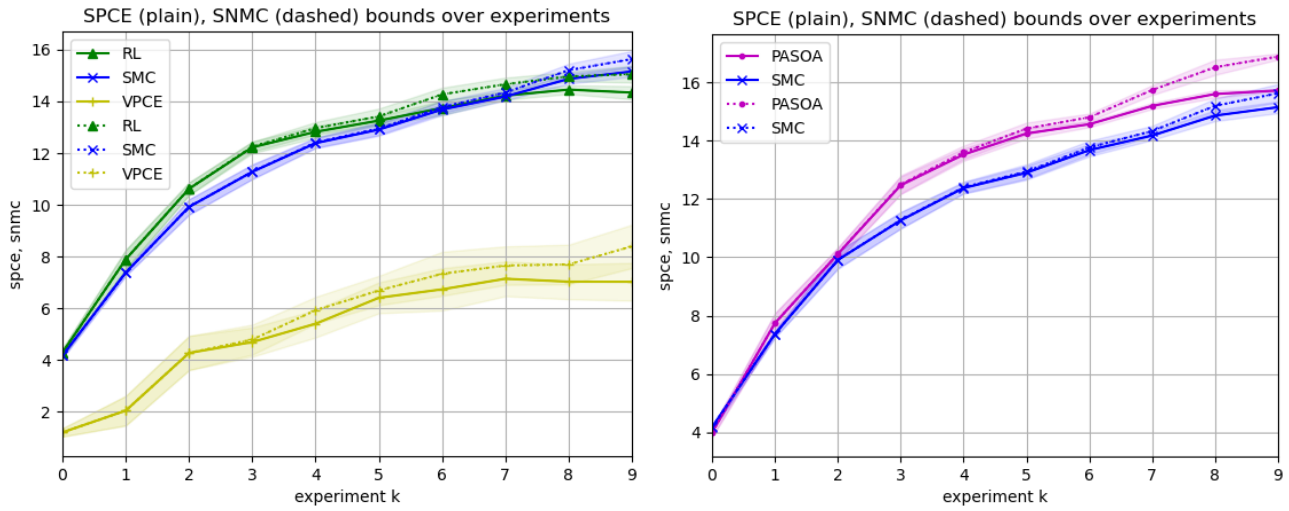

Figure 8. CES example. Median and standard error over 100 tollouts, with respect to the number of experiments k. The [SPCE, SNMC] intervals containing the totel EIG are plot, respectively with plain (SPCE lower bound) and dashed (SNMC upper bound) lines. Left: SMC (blue) vs RL-BOED (green) and VPCE (yellow). Right: SMC (blue) vs PASOA (red).

D.8. Non differentiable examples

When the model log-likelihood is not differentiable, either because the gradient is not available or difficult to compute, or because the design space is discrete, the stochastic gradient part of our method cannot be directly applied. However, we can still use the other parts by replacing the optimization step by either an exhaustive argmax, in the case of a finite design space, or by Bayesian optimization (Snoek et al., 2012; Henrández-Lobato et al., 2014; Benassi et al., 2011) which does not requires gradients. This can be seen as an advantage of myopic solutions, which allow such replacements to be easily performed. For each sequential optimization, the search space remains of reasonable size or dimension and does not increases exponentially. This is not the case for other policy-based approach, *e.g.* (Foster et al., 2021), which would involve a challenging high-dimensional Bayesian optimization of the policy parameters, or an exhaustive search in an exponentially increasing space with K.

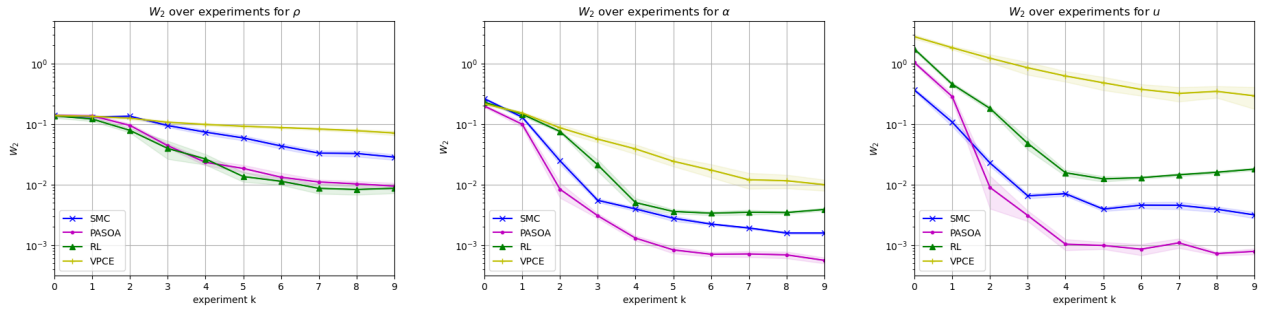

Figure 9. CES example. Median and standard error of Wasserstein distances for each parameter $(\rho, \boldsymbol{\alpha}, u)$ separately.

To illustrate this situation, another benchmark example used in (Blau et al., 2022; Moffat et al., 2020) is the Prey population example. The design is a discrete variable. Instead of using stochastic gradient descent to optimize the I_{PCE}^k bound at each step sequentially, we can compute it for every possible design and take the argmax. The same can be done to adapt VPCE to this discrete design space, while RL-BOED has the advantage to be applicable for both continuous and discrete spaces. Without the gradient part, our approach is similar to that of Moffat et al. (2020) but with an additional tempering, which was already reported to compare favorably to RL-BOED in Figure 4 of Blau et al. (2022).