Supplementary Materials: Probabilistic Numerical Methods for Partial Differential Equations and Bayesian Inverse Problems

Jon Cockayne[*] Chris J. Oates[†] T. J. Sullivan[‡]
Mark Girolami[§]

July 8, 2017

S1 Computational Details

In this supplement we provide details on the Markov chain Monte Carlo (MCMC) methods used for our experiments and on the optimisation methods used for experimental design in the main text.

S1.1 Infinite Dimensional MCMC

The goal here is to obtain samples from $\Pi_\theta^{\boldsymbol{y},h}$. Intractability of the normalisation constant Z_h motivates the use of MCMC techniques; these construct a measure-preserving Markov transition kernel over Θ that can be used to obtain approximately independent samples from $\Pi_\theta^{\boldsymbol{y},h}$. Crucially, MCMC requires knowledge of the Radon–Nikodým derivative only up to a multiplicative constant, avoiding the need to calculate Z_h.

For infinite-dimensional parameter inference problems we propose to use the preconditioned Crank–Nicolson (pCN) algorithm [5, 6]. This proceeds as follows: Assume that Θ is a Hilbert space and a Gaussian prior $\Pi_\theta = N(0, C)$ is assigned over Θ. Given potential function Φ_h, Algorithm 1 details the pCN method for constructing a Markov chain θ^i, $i = 1, \dots, I$, that targets the posterior $\Pi_\theta^{\boldsymbol{y},h}$:

Algorithm 1 (pCN Method).

```
• Pick $\theta^0 \in \Theta$
• For $i = 1 \dots I$:
    1. Draw $\xi^i \sim N(0, C)$
```

[*] University of Warwick (j.cockayne@warwick.ac.uk).

[†] Newcastle University and Alan Turing Institute (chris.oates@ncl.ac.uk).

[‡] Free University of Berlin and Zuse Institute Berlin (sullivan@zib.de).

[§] Imperial College and Alan Turing Institute (m.girolami@imperial.ac.uk).

2. Propose $\theta^* \leftarrow \sqrt{1-\lambda^2}\theta^i + \lambda\xi^i$

3. Compute

$$\alpha \leftarrow \min\left\{1, \frac{\exp(-\Phi_h(\boldsymbol{y},\theta^*))}{\exp(-\Phi_h(\boldsymbol{y},\theta^i))}\right\}$$

4. Set $\theta^{i+1} \leftarrow \theta^*$ *with probability* α*; otherwise set* $\theta^{i+1} \leftarrow \theta^i$

- *End.*

The parameter λ governs the scale of the proposal increments and should be tuned to achieve fastest mixing of the Markov chain.

S1.2 Pseudo-Marginal MCMC

Turning to the inverse problem, we now present an MCMC scheme for sampling the joint distribution of the solution vector $\boldsymbol{u}$, the latent states $\boldsymbol{z}$ and the parameter θ. Our primary interest is in θ, so a natural approach is to focus sampling effort on θ via Pseudo-Marginal MCMC [4]. For simplicity we restrict to finite dimensional $\Theta \subseteq \mathbb{R}^M$. This is to avoid technical issues associated with infinite-dimensional parameters in Pseudo-Marginal MCMC.

When the forward model is non-linear, the data-likelihood is an intractable integral

$$\pi(\boldsymbol{y}|\boldsymbol{g},\boldsymbol{b},\theta) = \int \pi(\boldsymbol{z}|\theta) \underbrace{\int \pi(\boldsymbol{y}|\boldsymbol{u})\pi(\boldsymbol{u}|\boldsymbol{z},\boldsymbol{g},\boldsymbol{b},\theta)\mathrm{d}\boldsymbol{u}\,\mathrm{d}\boldsymbol{z}}_{(*)} \tag{1}$$

where we note that the interior integral $(*)$ is available in closed-form as before in Section 3.2.2. An improper uniform measure $\pi(\boldsymbol{z}|\theta) = 1$ was taken; note that the impropriety of $\pi(\boldsymbol{z}|\theta)$ implies that Eq. 1 is only defined up to a multiplicative constant.

To construct an estimate to Eq. 1, an importance density $r(\boldsymbol{z}|\boldsymbol{y},\theta)$ was constructed, defined in Sec. S1.2.1, and we rewrite Eq. 1 as

$$\pi(\boldsymbol{y}|\boldsymbol{g},\boldsymbol{b},\theta) = \int \frac{1}{r(\boldsymbol{z}|\boldsymbol{y},\theta)} \int \pi(\boldsymbol{y}|\boldsymbol{u})\pi(\boldsymbol{u}|\boldsymbol{z},\boldsymbol{g},\boldsymbol{b},\theta)\mathrm{d}\boldsymbol{u}\; r(\boldsymbol{z}|\boldsymbol{y},\theta)\mathrm{d}\boldsymbol{z}.$$

Then an explicit, almost-surely positive, unbiased estimator $\hat{\pi}(\boldsymbol{y}|\boldsymbol{g},\boldsymbol{b},\theta)$ for the likelihood $\pi(\boldsymbol{y}|\boldsymbol{g},\boldsymbol{b},\theta)$ can be constructed as follows:

1. $\boldsymbol{z}^* \sim r(\boldsymbol{z}|\boldsymbol{y},\theta)$
2. $\hat{\pi}(\boldsymbol{y}|\boldsymbol{g},\boldsymbol{b},\theta) \leftarrow \frac{1}{r(\boldsymbol{z}^*|\boldsymbol{y},\theta)} \int \pi(\boldsymbol{y}|\boldsymbol{u})\pi(\boldsymbol{u}|\boldsymbol{z}^*,\boldsymbol{g},\boldsymbol{b},\theta)\mathrm{d}\boldsymbol{u}$

The Pseudo-Marginal approach constructs a Markov chain θ^i, $i = 1,\dots,I$, as follows: Specify a position-dependent proposal density $q(\theta^*|\boldsymbol{y},\theta)$, for example a random walk. Then:

Algorithm 2 (Pseudo-Marginal MCMC)**.**

```
Pick $\theta^0 \in \Theta$ and simulate $\hat{\pi}^0$, an unbiased estimate for $\pi(\boldsymbol{y}|\boldsymbol{g}, \boldsymbol{b}, \theta^0)$.
For $i = 1, \ldots, I$:
    1. Propose $\theta^* \sim q(\theta^*|\boldsymbol{y}, \theta^i)$
    2. Simulate $\hat{\pi}^*$, an unbiased estimate for $\pi(\boldsymbol{y}|\boldsymbol{g}, \boldsymbol{b}, \theta^i)$
    3. Compute
       $$\alpha \leftarrow \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta^i)} \cdot \frac{\hat{\pi}^*}{\hat{\pi}^i} \cdot \frac{q(\theta^i|\boldsymbol{y}, \theta^*)}{q(\theta^*|\boldsymbol{y}, \theta^i)}\right\}$$
    4. Set $\theta^{i+1} \leftarrow \theta^*$, $\hat{\pi}^{i+1} \leftarrow \hat{\pi}^*$ with probability $\alpha$; else set $\theta^{i+1} \leftarrow \theta^i$, $\hat{\pi}^{i+1} \leftarrow \hat{\pi}^i$
End.
```

S1.2.1 Multiple Solutions

The performance of Algorithm 2 depends on how close $r(\boldsymbol{z}|\boldsymbol{y}, \theta)$ can be made to the latent variable posterior $\pi(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{g}, \boldsymbol{b}, \theta)$. Adaptive proposals were employed that automatically take into account the varying nature of the parameter θ. This construction is complicated by the fact that non-linear PDEs are not guaranteed a unique solution, leading to multiple values of $\boldsymbol{z}$ which are each consistent with some solution of the PDE. Here details are provided for the choice of importance density.

A known, fixed number of solutions are assumed to exist for the non-linear forward problem and it is further assumed that these each vary smoothly with θ. The strategy is to augment the MCMC procedure with an additional parameter, i, describing the solution index. A joint proposal over (θ, i) then operates in a Metropolis-within-Gibbs sampler. The proposal distribution for i was taken to be uniform for simplicity. To be specific, consider a semi-linear PDE of the form $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$ as in Sec. 5.2.2. For fixed (θ, i), an approximation $\hat{\boldsymbol{u}}_i$ to the solution $\boldsymbol{u}_i$ of the PDE is obtained. We take $r(\boldsymbol{z}|\boldsymbol{y}, \theta, i) = N(\mathcal{A}_2^{-1}\hat{\boldsymbol{u}}_i, \mathcal{C})$ for an appropriate covariance $\mathcal{C}$. We emphasise that these choices affect only the mixing properties of the MCMC, not the posterior distributions that are the central focus of this work. In addition, the approximation $\hat{u}_i$ is not required to a high degree of accuracy as it is used only to construct the importance distribution $r(\boldsymbol{z}|\boldsymbol{y}, \theta, i)$. In fact, for computational efficiency a crude solution is preferable as this lowers the overall cost of the algorithm. To obtain these approximation, the recent "deflation" approach of [7] was applied.

The total computational cost of this method is equivalent to a single application of the deflation technique, followed by the cost of sampling from the importance distribution to obtain an unbiased estimate of the likelihood. This latter cost is minimal, as the matrices required to compute $\boldsymbol{u}|\theta, \boldsymbol{z}$ are predominantly independent of $\boldsymbol{z}$; as a result, using many samples from $\boldsymbol{z}$ to approximate the data-likelihood is computationally inexpensive.

S1.3 Experimental Design

In this paper A-optimal experimental designs were pursued, to limit scope. In this case the solution to Eq. 14 is generally unavailable and numerical minimisation is required. A hybrid strategy, that combines the Approximate Coordinate Exchange (ACE) algorithm of [12] with Bayesian Optimisation [11] performed well at this task. Here for convenience we suppress the subscript on the set of design points X_0.

A complication here is that due to the dependence of both $\mathcal{A}$ and $\mathcal{B}$ on θ, it is natural that X_0 should also depend on θ. This idea is not pursued in this work; instead a design is constructed for a single candidate $\hat{\theta}$, chosen heuristically based upon the problem. For the Allen–Cahn application this was $\hat{\theta} = \theta^\dagger$.

In addition to the loss function $L(\mathbf{\Sigma}(\hat{\theta}, X))$, viewed here as a function of X, a *deletion function* $d(X, M)$ is also required. Suppose that $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\}$ and let $X_{-j} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \dots, \boldsymbol{x}_m\}$; that is, X_j is the set X without the point $\boldsymbol{x}_j$. Then the deletion function $d(X, M)$ returns the ordered list of M points for which $L(\mathbf{\Sigma})\hat{\theta}, X_{-j}))$ is minimal. That is, the deletion function returns the indices of those coordinates which, when deleted, have the least impact on the loss of the set X.

The full implementation of the experimental design procedure is given in Algorithm 3.

Algorithm 3 (Experimental Design)**.**

```
Fix an initial design $X^0$.
For $i = 1, \dots, I$:
    1. Set $X^i = X^{i-1}$
    2. Compute $C_i = d(X^i, M)$.
    3. For $j \in C_i$:
        a) Find
               $\boldsymbol{x}_j^* = \arg\min_{\boldsymbol{x}} L(\mathbf{\Sigma}(\hat{\theta}, X^i_{-j} \cup \{\boldsymbol{x}\}))$
           with a numerical optimisation method.
        b) Set $X^i = X^i_{-j} \cup \{\boldsymbol{x}_j^*\}$, preserving the ordering of the set.
Return $X^I$
```

Here M is a parameter controlling how many points are considered at each iteration. Because moving a point generally changes all values of the deletion function, setting $M = 1$ is most desirable. However, computing the deletion function is expensive, and so setting $M > 1$ can reduce the amount of computational effort required to achieve convergence.

Related work on experimental design in inverse problems is more classical, in the sense that numerical error is assumed to be negligible and instead one aims to select the locations X of sensors, that will be used to obtain the data $\boldsymbol{y}$, in order to minimise expected posterior uncertainty over the parameter θ. In this context, recent work includes a series

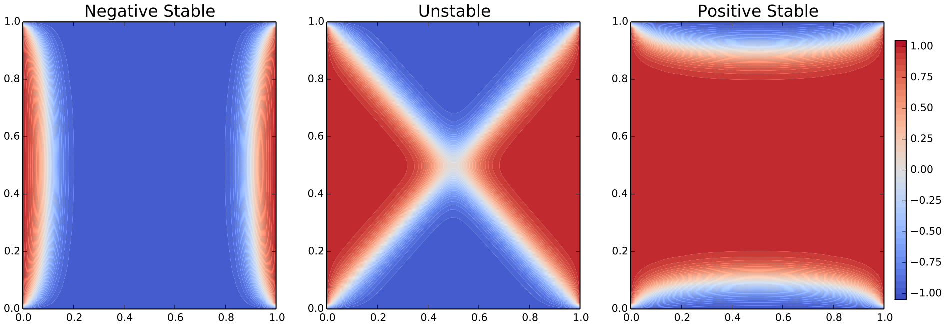

Figure 1: Application to Allen–Cahn: Three distinct solutions of the PDE, here shown at the parameter value $\theta^{\dagger} = 0.04$.

of papers by [2, 1, 3]. In addition, recent work by [8] applies these techniques to Gaussian process regression using the integrated variance criterion to determine optimal sensor locations; this is equivalent to the A-optimal approach which we pursued. The method used to attain the optimal design is to minimise a Monte-Carlo estimate of the objective function using gradient-based optimisers, rather than ACE.

From the meshless methods literature, [10] and [9] both consider the use of greedy algorithms to select locations X in order to minimise a criterion relating to numerical error in the forward problem. These papers differ to ours in several respects. First, the context is asymmetric collocation, rather than symmetric collocation. Second, the formulation is not probabilistic and does not have the associated interpretation as a problem in experimental design. Third, inverse problems are not considered.

To conclude this section, note that the direct approach of minimising uncertainty over the parameter θ appears to be challenging in this framework. The present proposal, to minimise uncertainty over the solution vector $\boldsymbol{u}$ at each value of the parameter θ, provides a practical approach that acts as a proxy for uncertainty in the parameter.

S1.3.1 Experimental Design and MCMC

Note that the ACE approach elegantly could be adapted to interlace with MCMC: For one iteration of the MCMC the difference between θ^{i+1} and θ^{i} will usually be small. It is therefore reasonable to expect that an optimal design $X_0^*(\theta^i)$ for θ^i will be a good approximation to the optimal design $X_0^*(\theta^{i+1})$ for θ^{i+1}. The ACE algorithm, at iteration i of the MCMC, could take a numerical approximation of $X_0^*(\theta^i)$ as a starting point and perform a local search in design space in order to locate an approximate $X_0^*(\theta^{i+1})$. This approach avoids the need to repeatedly solve challenging multi-variate optimisation problems *de novo* at each iteration of the MCMC.

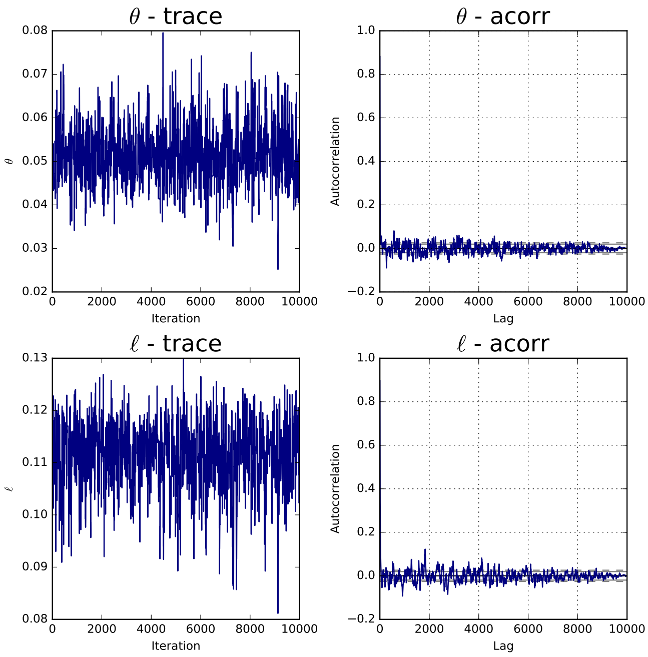

Figure 2: Application to Allen–Cahn: MCMC trace and autocorrelation plots for the unknown parameter θ and the kernel length scale ℓ, based on a probabilistic meshless method with $m_{\mathcal{A}} = 20$ design points.

References

[1] Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas. A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0-sparsification. *SIAM J. Sci. Comput.*, 36(5):A2122–A2148, 2014. ISSN 1064-8275. doi: 10.1137/130933381.

[2] Alen Alexanderian, Philip J. Gloor, and Omar Ghattas. On Bayesian A- and D-optimal experimental designs in infinite dimensions. *Bayesian Anal.*, 11(3):671–695, 2016. ISSN 1936-0975. doi: 10.1214/15-BA969.

[3] Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas. A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems. *SIAM J. Sci. Comput.*, 38(1):A243–A272, 2016. ISSN 1064-8275. doi: 10.1137/140992564.

[4] Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725, 2009. ISSN 0090-5364. doi: 10.1214/07-AOS574.

[5] Simon L. Cotter, Gareth O. Roberts, Andrew M. Stuart, and David. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statist. Sci.*, 28(3):424–446, 2013. ISSN 0883-4237. doi: 10.1214/13-STS421.

[6] Masoumeh Dashti and Andrew M. Stuart. The Bayesian approach to inverse problems, 2016.

[7] Patrick E. Farrell, Ásgeir Birkisson, and Simon W. Funke. Deflation techniques for finding distinct solutions of nonlinear partial differential equations. *SIAM J. Sci. Comput.*, 37(4):A2026–A2045, 2015. ISSN 1064-8275. doi: 10.1137/140984798.

[8] Alex a. Gorodetsky and Youssef M. Marzouk. Mercer kernels and integrated variance experimental design: connections between Gaussian process regression and polynomial approximation. *SIAM J. Sci. Comput.*, xx:1–32, 2015.

[9] Leevan Ling and Robert Schaback. Stable and convergent unsymmetric meshless collocation methods. *SIAM J. Numer. Anal.*, 46(3):1097–1115, 2008. ISSN 0036-1429. doi: 10.1137/06067300X.

[10] Leevan Ling, Roland Opfer, and Robert Schaback. Results on meshless collocation techniques. *Eng. Anal. Boundary Elements*, 30(4):247–253, 2006. doi: 10.1016/j.enganabound.2005.08.008.

[11] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(2):117–129, 1978.

[12] Antony Overstall, David Woods, and Ben Parker. Bayesian optimal design for ordinary differential equation models. arXiv:1509.04099v2, 2015.