# SURROGATE MODELS FOR STATISTICAL COMPUTATION AND ESTIMATION

**Supervisor**: Dr. Sam Power (sam.power@bristol.ac.uk)

## 1. Introduction

In specifying a statistical model, one implicitly agrees to the computational burden of computing the associated estimator. With the growth of computational resources over time, the statistician is emboldened to specify models of greater and greater complexity, seeking finer fidelity between model and reality. Where once a simple linear model might have sufficed to obtain a working understanding of some natural phenomenon, the ambitious modern statistician might now seek a finer description by turning to stochastic dynamical systems (diffusions, jump processes, ...), differential equations (ordinary, partial, ...), and beyond. Still, in accepting the conceptual capabilities of these advanced models, the same statistician must carefully face up to the computational expense which is incurred in calibrating these models to observed data.

A guiding principle in the numerical analysis of complex problems reminds us that even when high-fidelity computations are of primary interest, the use of reduced-cost, low-fidelity, approximate models can be a useful tool for accelerating computations. The classical multigrid literature explains how coarse discretisations of suitable PDE can be used to facilitate their solution on finer grids, the modern literature on Multilevel Monte Carlo demonstrates that the same principle remains powerful for stochastic computations, and various other intervening methods reflect this same principle: a well-chosen approximate model (or hierarchy of such models) is a valuable asset.

In the statistical world, there is a related but distinct literature concerning the design and application of approximate or 'surrogate' likelihoods for statistical models. While these approximations might take the form of coarsening some discretisation hyperparameter in the numerical-analytic style, they might equally be derived by neglecting some dependence structure in the 'ideal' model, or even by directly fitting some simplified functional form to the expensive 'full' likelihood, according to some appropriate regression procedure. Instead of only questioning the accuracy of the surrogate itself as a function, the more pertinent question thus becomes 'is my new estimator statistically adequate?', and 'how can I effectively trade off computational complexity and statistical accuracy?'.

## 2. Project Proposal

The plan for this project would be to explore and advance the use of surrogate likelihoods in a statistical context, with an eye towards developing computationally efficient approaches to estimation in the context of sampling, optimisation, and beyond. There are many possible directions to consider, each typically with some practical and some theoretical directions to explore. Some particular examples could be:

2.1. **Local v.s. Global Surrogates.** Many existing applications of surrogate models have a 'global' character to them: there is some expensive ideal model $\mathcal{M}$, some cheap surrogate model $\widehat{\mathcal{M}}$, and each of them is defined on the entirety of the parameter space. Theoretical results on the accuracy of surrogate-driven procedures often depend on the worst-case discrepancy between $\mathcal{M}$ and $\widehat{\mathcal{M}}$ across the entire parameter space, and it can be challenging to ensure that this is controlled well.

By contrast, it can often be the case that if we are interested in approximating the ideal model $\mathcal{M}$ well only in a neighbourhood of some given reference parameter $\theta_0$, then we can credibly construct a 'local' surrogate $\widehat{\mathcal{M}}_{\theta_0}$ which is a uniformly-good approximation of $\mathcal{M}$ only within this neighbourhood. It is natural to expect that judicious use of local surrogates should enable the construction of estimators which are both computationally efficient and rigorously justified to be statistically accurate; a project could be structured around demonstrating this possibility in both theory and practice.

2.2. **Adaptive Construction of Surrogates.** Classical surrogate likelihoods for a specific statistical model are often 'fixed-form' in character: an expert proposes an approximation, justifies its accuracy, and one then proceeds to use the approximation more-or-less as-is. More modern approaches go a step further, and seek to refine the approximation over time by e.g. regression approaches. By identifying and amending the deficiencies of the surrogate as appropriate, one hopes to eventually end up with a surrogate and associated statistical estimator which are of adequate quality, ideally improving upon the baseline 'fixed-form' method.

While such approaches are both intuitively appealing and widely-used, there are still several gaps in our understanding of their behaviour. At a practical level, one can ask how frequently and how ambitiously one should refine the surrogate: is it more efficient to expend a large effort 'learning the error' as an up-front cost, or is it more reliable to amortise this adaptation cost over several iterative 'rounds' of model-fitting? Several related theoretical questions arise naturally in the same context. A project could be structured around evaluating the merits these procedures through many of these lenses.

■

Various other topics could also be considered, e.g. effective use of hierarchies of surrogate models, theoretical aspects of Monte Carlo methods based on surrogate models, surrogate models in simulation-based inference, surrogate models for online problems, and more; feel free to email me if these options are of interest.

## 3. Supervisor

My research interests centre around the design and analysis of numerical algorithms, with applications to problems in modern statistics and machine learning. I am particularly interested in Monte Carlo methods, such as Markov Chain Monte Carlo and Sequential Monte Carlo, but retain broad interests in other classes of algorithm, particularly those which are applicable to large-scale problems. I enjoy trying to understand the convergence and complexity behaviour of these algorithms, with the hope that the theoretical insights gained in doing so can guide their practical implementations to be more automatic, robust, and efficient.