

SCALABLE CONTROL VARIATES FOR MONTE CARLO METHODS VIA STOCHASTIC OPTIMIZATION

Shijing Si
Duke University
shijing.si@duke.edu

Chris. J. Oates
Newcastle University
chris.oates@ncl.ac.uk

Andrew B. Duncan
Imperial College London
a.duncan@imperial.ac.uk

Lawrence Carin
Duke University
lcarin@duke.edu

François-Xavier Briol
University College London
f.briol@ucl.ac.uk

June 16, 2020

ABSTRACT

Control variates are a well-established tool to reduce the variance of Monte Carlo estimators. However, for large-scale problems including high-dimensional and large-sample settings, their advantages can be outweighed by a substantial computational cost. This paper considers control variates based on Stein operators, presenting a framework that encompasses and generalizes existing approaches that use polynomials, kernels and neural networks. A learning strategy based on minimising a variational objective through stochastic optimization is proposed, leading to scalable and effective control variates. Our results are both empirical, based on a range of test functions and problems in Bayesian inference, and theoretical, based on an analysis of the variance reduction that can be achieved.

1 Introduction

This paper focuses on the approximation of the integral of an arbitrary function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with respect to a distribution Π , denoted $\Pi[f] := \int f d\Pi$. It will be assumed that Π admits a smooth and everywhere positive Lebesgue density π such that the gradient of $\log \pi$ can be pointwise evaluated. This situation is typical in Bayesian statistics, where Π represents a posterior distribution and, to circumvent this intractability, Markov chain Monte Carlo (MCMC) methods are used. Nevertheless, the ergodic average of MCMC output converges at a slow rate proportional to $n^{-1/2}$ and, for finite chain length n , there can be considerable stochasticity associated with the MCMC output.

Control variates are a variance reduction technique for Monte Carlo methods, including MCMC. Given a test function f , the general approach is to identify another function, g , such that the variance of the estimator with f replaced by $f - g$ is smaller than that of the original estimator, and such that $\Pi[g] = 0$, so the value of the integral is unchanged. Such a g is called a *control variate* (CV). CVs are widely-used in statistics and machine learning, including for the simulation of Markov processes [29, 41], stochastic optimization [65], stochastic gradient MCMC [3], reinforcement learning [26, 27, 34], variational inference [46, 52, 53] and Bayesian evidence evaluation [43, 44].

Given a test function f , the problem of selecting an appropriate CV is non-trivial and a variety of approaches have been proposed. Our discussion focuses only on the setting where π is provided only up to an unknown normalization constant; i.e., the setting where MCMC is typically used. The most widely-used approach to selection of a CV is based on $g = \nabla \log \pi$ and simple (e.g., linear) transformations thereof [2, 20, 40, 47]; note that under weak tail conditions on π , the CV property $\Pi[g] = 0$ is assured. Recently several authors have proposed the use of more complicated or even non-parametric transformations, such as based on high order polynomials [60], kernels [42, 43, 6] and neural networks (NNs) [26, 64, 34]. These new approaches have been shown empirically – and theoretically, in the case of kernels [6, 42] – to provide substantial reduction in variance for MCMC.

These recent developments are closely related to Stein’s method [14, 57, 61], a tool used in probability theory to quantify how well one distribution Π' approximates another distribution Π . Recall that, given a collection of functions g for which $\Pi[g] = 0$ is satisfied, Stein’s method uses $\sup_g \Pi'[g]$ as a means of quantifying the difference between Π and Π' . As a byproduct, researchers in this field have constructed a large range of functions g that can be considered as candidates for a CV method. Although Stein’s method has recently been applied to a variety of problems including MCMC convergence assessment [22, 23, 24], goodness-of-fit testing [17, 36, 66], variational inference [52, 53], estimators for models with intractable likelihoods [5, 38] and the approximation of complex posterior distributions [15, 16, 36, 35, 37, 55], a unified account of how Stein’s method can be exploited for the construction of CVs, encompassing existing polynomial, kernel and NN transformations, has yet to appear. The organization and contributions of the paper are as follows:

- The literature to-date on polynomial, kernel, and NN-based CVs is reviewed in Sec. 2.
- A general framework for the construction of CVs is presented in Sec. 3, rooted in Stein’s method, and an efficient learning strategy based on stochastic optimization is proposed.
- A theoretical analysis of the variance reduction is provided in Sec. 4.
- An empirical assessment is provided in Sec. 5 and covers a range of synthetic test problems, as well as problems arising in the Bayesian inferential context. These experiments highlight the flexibility of our framework and the efficiency of the stochastic optimization procedure that we have proposed for learning of a suitable CV.

2 Literature Review

We survey existing CV methods that are compatible with MCMC output, this being our application area or principal interest. However, elements of our discussion will be applicable to other situations where CVs are used. In what follows, it is assumed that some states $(x_i)_{i=1}^n \subset \mathbb{R}^d$ have already been obtained (for example, as a realization of a Markov chain with target Π) and our goal is to construct a low-variance estimator for $\Pi[f]$ by learning a CV based on a subset of size $m \leq n$.

Several CV approaches have been proposed. One approach is to use a Taylor expansion of the test function f [46, 65], or perhaps a polynomial approximation to f learned from regression [32]. Unfortunately, this will only be a feasible approach when integrating against simple probability distributions Π for which polynomials can be exactly integrated, such as a Gaussian. CVs may also be directly available through problem-specific knowledge [e.g., for certain Markov processes; 29, 41], but this is rarely the case in general. Alternatively, CVs can sometimes be built using known properties of the method used for obtaining samples; see [1, 7, 9, 12, 18, 28] for CVs that are developed with a particular MCMC method in mind. See also [30] for CVs specialized to quasi Monte Carlo. An obvious drawback to the methods above is that they impose strong restrictions on the methods that one may use to obtain the $(x_i)_{i=1}^m$.

An arguably more general framework, and our focus in this paper, is to first curate a rich set \mathcal{G} of candidate CVs, and then to employ a learning procedure to select the optimal CV $g \in \mathcal{G}$. This should be done according to a suitable optimality criterion based on f and the given set $(x_i)_{i=1}^m$. The methodological challenges are therefore twofold; first, we must construct \mathcal{G} and second, we must provide a procedure to select a suitable CV from this set. The construction of a candidate set \mathcal{G} has been approached by several authors using a variety of regression-based techniques:

- Motivated by physical considerations, [2] proposed CVs of the form $g = Hu$, based on the Schrödinger-type Hamiltonian $H = -0.5\Delta + 0.5(\sqrt{\pi})^{-1}\Delta\sqrt{\pi}$, where Δ is the Laplacian and u is a polynomial of fixed degree. See also [20, 40, 47].
- An approach called *control functionals* (CFs) was proposed in [43], where the set \mathcal{G} consisted of functions of the form $g = \nabla \cdot u + u \cdot \nabla \log \pi$, where $\nabla \cdot$ denotes the divergence operator, ∇ denotes the gradient operator and $u : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is constrained to belong to a suitable Hilbert space of vector fields on \mathbb{R}^d . See also [6, 42, 59] for the connection with Stein’s method.
- In more recent work, [64] extended the CF approach to the case where a NN is used to provide a parametric family of candidates for the vector field u . The set of all such functions g generated using a fixed architecture of NN is taken as \mathcal{G} . See also [34, 62].

Thus, several related options are available for constructing a suitable candidate set \mathcal{G} . However, where existing literature diverges markedly is in the procedure used to select a suitable CV from this set:

- For approaches based on polynomials, [2] proposed to select polynomial coefficients θ in order to minimize the sum-of-squares error $\sum_{i=1}^m (f(x_i) - g_\theta(x_i))^2$, which can be done through solving a linear system. Here g_θ is used to emphasize the dependence on coefficients θ of the polynomial. For even moderate degree polynomials, the combinatorial explosion in the number of coefficients as d grows necessitates regularized estimation of θ ; suitable regularizers are evaluated in [50, 60].
- For the CF approaches, regularized estimation is essential since the Hilbert space of vector fields is infinite dimensional. Here, [43] proposed to select g as a minimal norm element of the Hilbert space for which the interpolation equations $f(x_i) = c + g(x_i)$ are satisfied for all $i = 1, \dots, m$ and some $c \in \mathbb{R}$. A major drawback of this approach is the $O(m^3)$ computational cost.
- The approach based on NN also exploited a sum-of-squares error, but in [64] the authors proposed to include an additional regularizer term $\lambda \sum_{i=1}^m g_\theta(x_i)^2$, for some pre-specified constant λ , to avoid over-fitting of the NN. Optimization over θ , the parameters of the NN that enter into g_θ , was performed using stochastic gradient descent.

It is therefore apparent that, in existing literature, the construction of the candidate set \mathcal{G} is intimately tied to the approach used to select a suitable element from it. This makes it difficult to draw meaningful conclusions about which CVs are most suitable for a given task; from a theoretical perspective, existing analyses make assumptions that are mutually incompatible and, from a practical perspective, the different techniques and software involved in implementing existing methods precludes a straight-forward empirical comparison. Our attention therefore turns next to the construction of a general framework that can be used to learn a wide range of CVs, including polynomial, kernel and NN, under a single set of theoretical assumptions and algorithmic parameters, enabling a genuinely systematic assessment of CV methods to be performed.

3 Methods

We present a general framework for the construction of CVs: In Section 3.1 the construction of a candidate set \mathcal{G} is achieved using Stein operators, which unifies the CVs proposed in existing contributions such as [2, 43, 64] and covers simultaneously the case of polynomials, kernels and NNs. Then, in Section 3.2, we present an approach to selection of a suitable element $g \in \mathcal{G}$, based on a variational formulation and performing stochastic optimization on an appropriate objective functional.

3.1 Construction of a Class of Control Variates \mathcal{G}

The construction of non-trivial functions $g : \mathbb{R}^d \rightarrow \mathbb{R}$ with the property $\Pi[g] = 0$ is not straight-forward in the setting where MCMC would be used, since for general f the integral $\Pi[f]$ cannot be exactly computed. Stein’s method [61] offers a solution to this problem in the case where the gradient of $\log \pi$ can be evaluated pointwise, which we describe next. A *Stein characterization* of a distribution Π consists of a pair $(\mathcal{U}, \mathcal{L})$, where \mathcal{U} is a set of functions whose domain is \mathbb{R}^d and \mathcal{L} is an operator, such that $\Pi'[\mathcal{L}u] = 0 \forall u \in \mathcal{U}$ if and only if the distributions Π' and Π are equal. In this case \mathcal{U} is called a *Stein class* and \mathcal{L} is called a *Stein operator*¹. Clearly, if one can identify a Stein characterization for Π , then one could take $\mathcal{G} = \mathcal{L}\mathcal{U} = \{ \mathcal{L}u : u \in \mathcal{U} \}$ as a set of candidates CVs.

The literature on Stein’s method provides general approaches to identify a Stein characterization [14, 57, 61]. In the *generator approach*, \mathcal{L} is taken to be the infinitesimal generator of an Markov process which is ergodic with respect to Π [4]. For example, if \mathcal{L} is the infinitesimal generator of an overdamped Langevin diffusion then one obtains the *Langevin* Stein operator, which acts on vector fields u on \mathbb{R}^d as $\mathcal{L}_L u = \nabla \log \pi \cdot u + \nabla \cdot u$. This recovers the operator used in the CF approach of [43], as well as the operator used in the NN approach of [64]. Alternatively, we could construct an operator that acts on *scalar*-valued functions by replacing the vector field u with the potential ∇u in the previous operator, leading to the scalar-valued Langevin (SL) Stein operator $\mathcal{L}_{SL} u = \Delta u + \nabla u \cdot \nabla \log \pi$. This recovers the operator used with polynomials in [2, 40]. Trivially, a scalar multiple of a Stein operator is a Stein operator, and one may combine Stein characterizations $(\mathcal{U}_i, \mathcal{L}_i)$ linearly as $\mathcal{L}u = \mathcal{L}_1 u_1 + \mathcal{L}_2 u_2$, $u \in \mathcal{U}_1 \times \mathcal{U}_2$, so that considerable flexibility can be achieved. We will see in Section 5 that this can lead to scalable and flexible classes of CVs.

¹To simplify presentation in the paper, we always assume \mathcal{U} is a *maximal* set of functions for which $\mathcal{L}u$ is well-defined and $\Pi[\mathcal{L}u] = 0$.

3.2 Selection of a Suitable Control Variate $g \in \mathcal{G}$

Once a set \mathcal{G} of candidate CVs has been constructed, we must consider how to select a suitable element $g \in \mathcal{G}$ (or equivalently $u \in \mathcal{U}$) that leads to improved performance of the Monte Carlo estimator when f is replaced by $f - g$. In general this will depend on the specific details of the Monte Carlo method; for example, in MCMC one would select g to minimize asymptotic variance [7, 18], while in quasi Monte Carlo one would minimize the Hardy-Krause variation [30]. The situation simplifies considerably when \mathcal{G} contains an element g^* such that $f - g^*$ is constant. This optimal function $g^* = \mathcal{L}u^*$, if it exists, is given by the solution of *Stein's equation*: $\mathcal{L}u^*(x) = f(x) - \Pi[f]$. This paper proposes to directly approximate a solution of this equation (a linear partial differential equation) by casting it in a variational form and solving over a subset $\mathcal{V} \subseteq \mathcal{U}$. The variational characterization that we use is that $J(u^*) = 0$, where

$$J(u) := \|f - \mathcal{L}u - \Pi[f]\|_{L^2(\Pi)}^2 = \text{Var}_{\Pi}[f - \mathcal{L}u]$$

with $L^2(\Pi)$ being the space of square-integrable functions with respect to Π . Equivalently, this can be seen as the variance of $f - \mathcal{L}u$. In the spirit of empirical risk minimization, we propose to minimize an empirical approximation of this functional, computed based on samples $x_i, i = 1, \dots, m$ that are drawn either exactly or approximately from Π . There are two natural approximations that could be considered. The first is based on the *variance* representation

$$J(u) = \text{Var}_{\Pi}[f - \mathcal{L}u] \approx \frac{2}{m(m-1)} \sum_{i>j} (f(x_i) - \mathcal{L}u(x_i) - f(x_j) + \mathcal{L}u(x_j))^2 =: J_m^V(u), \quad (1)$$

providing an approximation of J at cost $O(m^2)$, used in [8]. The second is based on the *least-squares* representation

$$J(u) = \min_{c \in \mathbb{R}} \|f - \mathcal{L}u - c\|_{L^2(\Pi)}^2 \approx \min_{c \in \mathbb{R}} J_m^{\text{LS}}(c, u), \quad (2)$$

$$J_m^{\text{LS}}(c, u) := \frac{1}{m} \sum_{i=1}^m (f(x_i) - \mathcal{L}u(x_i) - c)^2,$$

providing an approximation of J at cost $O(m)$, used in [2, 40, 42, 43]. These approximations will be unbiased when the x_i are independent draws from Π , but this will not necessarily hold for the MCMC case. To approximately solve this variational formulation we consider a parametric subset $\mathcal{V} \subseteq \mathcal{U}$, where elements of \mathcal{V} can be written as v_{θ} for some parameter $\theta \in \mathbb{R}^p$. Depending on the specific nature of the functions g_{θ} , it can occur that the optimization problem is under-constrained, e.g., when $p > m$. Therefore, following [43, 60, 64], we also allow for the possibility of additional regularization at the level of θ . Thus we aim to minimize objectives of the form $\tilde{J}_m^V(\theta) + \lambda_m \Omega(\theta)$ and $\tilde{J}_m^{\text{LS}}(c, \theta) + \lambda_m \Omega(\theta)$ over $c \in \mathbb{R}$ and $\theta \in \mathbb{R}^p$, where $\tilde{J}_m^V(\theta) := J_m^V(v_{\theta})$, $\tilde{J}_m^{\text{LS}}(c, \theta) := J_m^{\text{LS}}(c, v_{\theta})$, $\lambda_m > 0$ and $\Omega(\theta)$ is a regularization term to be specified. To reduce notational overhead, for the least-squares case we let $\theta_0 := c$ and simply write $\tilde{J}_m^{\text{LS}}(\theta)$ where $\theta \in \mathbb{R}^{p+1}$.

To perform the minimization, we propose to use stochastic gradient descent (SGD) [26]. Thus, to minimize a functional $F(\theta)$, we iterate through $\theta^{(t+1)} = \theta^{(t)} - \alpha_t \widehat{\nabla F}(\theta^{(t)})$, where the *learning rate* α_t decreases as $t \rightarrow \infty$ and $\widehat{\nabla F}$ is an unbiased approximation to F . In our experiments, $\widehat{\nabla F}$ is constructed using a randomly chosen subset from $(x_i)_{i=1}^m$, with this subset being re-sampled at each step of SGD (*i.e.*, mini-batch SGD) [49, 67].

This framework is compatible with any parametric function class and has the potential to provide significant speed-ups, relative to existing methods, due to the efficiency of SGD. For example, taking \mathcal{V} to be the polynomials of degree at most k in each variable recovers the same class as [2, 40, 47, 60], but with a parameter optimization strategy based on SGD as opposed to exact least squares. For SGD with t iterations and mini-batches of size b , our computational cost will be of order $\mathcal{O}(d^k bt)$, whereas exact least squares must solve a linear system of size $\mathcal{O}(d^k)$, leading to a cost of $\mathcal{O}(d^{3k} + md^k)$. Similarly, taking \mathcal{V} to be a linear space spanned by m translates of a kernel recovers the CF method of [43]. SGD has computational cost of $\mathcal{O}(mdbt)$, whereas CFs requires $\mathcal{O}(m^3 + m^2 d)$ due to the need to invert an m -dimensional matrix. Significant reduction in computational cost can also be obtained for ensembles: a combination of polynomial and kernel basis functions, as considered in [59], would cost $\mathcal{O}((md + d^k)bt)$ compared to the $\mathcal{O}(m^3 + d^{3k} + m^2 + md^k)$ cost when the linear system is exactly solved. Furthermore, any hyper-parameters, such as kernel parameters, can be incorporated into the minimization procedure with SGD, so that nested computational loops are avoided.

Some of these speed-ups are illustrated on a toy example in Fig. 1. Even for this moderately-sized problem, the use of SGD provides significant speed-ups for polynomial and kernel CVs. Additional experiments with values of $m = 5000$ in Appendix D.1 show that larger speed-ups can be obtained for large scale problems.

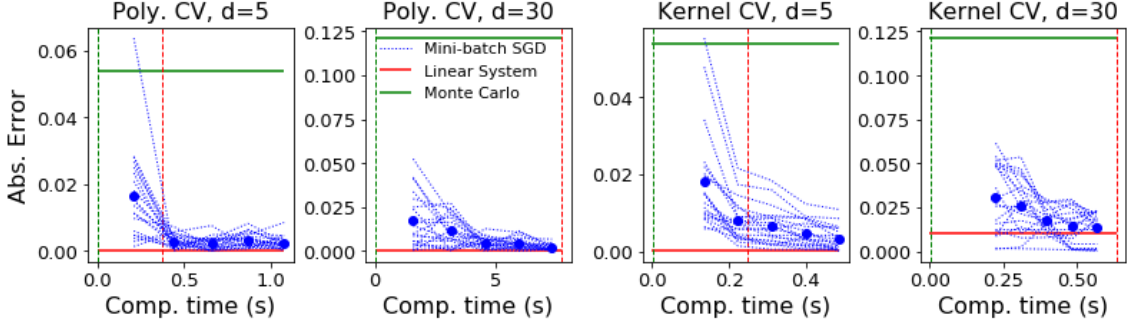


Figure 1: *Scalable Control Variates in High Dimensions*. Here we consider the toy problem of integrating $f(x) = x_1 + \dots + x_d$ against $\mathcal{N}(0, I_{d \times d})$. The total sample size is $n = 1000$ and $m = 500$ of these were used as the training set. Here 20 realizations (blue dashed lines) are shown and blue dots represent the mean absolute error. The red lines represent the performance and computational cost of solving the corresponding linear system exactly, our benchmark. Similarly, the green lines represent the Monte Carlo estimator with no control variate used.

4 Theoretical Assessment

Here we provide theoretical guarantees for our CVs trained using SGD. All proofs are in Appendix C.

The first question is whether it is possible to obtain *zero-variance* CVs, i.e. can we find a $u \in \mathcal{U}$ such that $J(u) = \text{Var}_{\Pi}[f - \mathcal{L}u] = 0$. The answer is “yes” under regularity conditions on Π and \mathcal{L} , and whenever \mathcal{V} is large enough. In particular, a fixed parametric class may not be large enough, but we can consider a nested sequence of sets $\mathcal{V}_1 \subseteq \mathcal{V}_2 \subseteq \dots$ such that $\cup_{p \in \mathbb{N}} \mathcal{V}_p$ is dense in \mathcal{U} . For example, \mathcal{V}_p could be polynomials of a certain degree p , or NNs with p hidden units.

Proposition 1. *Let \mathcal{U} be a normed space and $\mathcal{L} : \mathcal{U} \rightarrow L^2(\Pi)$ be a bounded linear operator. Consider a sequence of nested sets $\mathcal{V}_1 \subseteq \mathcal{V}_2 \subseteq \dots$ such that $\cup_{p \in \mathbb{N}} \mathcal{V}_p$ is dense in \mathcal{U} . If $\exists u \in \mathcal{U}$ that solves the Stein equation $\mathcal{L}u = f - \Pi[f]$, then $\lim_{p \rightarrow \infty} \inf_{v \in \mathcal{V}_p} J(v) = 0$.*

Of course, the existence of a solution to the Stein equation needs to be verified. The result below provides regularity conditions for the existence of a solution when using \mathcal{L}_{SL} , the Stein operator used in our experiments. Denote the Sobolev space $W^{k,p}(\Pi)$ of functions whose weak derivatives of order k are in $L^p(\Pi)$ and the Sobolev space $W_{\text{loc}}^{k,p}$ of functions whose p -th power weak derivatives of order k are locally integrable; these are formally defined in Appendix B. For a vector-valued function $h : \mathbb{R}^d \rightarrow \mathbb{R}^p$ we let $\|h\|_{L^p(\Pi)} := (\sum_{i=1}^d \|h_i\|_{L^p(\Pi)}^2)^{1/2}$.

Proposition 2. *Consider the vector space $\mathcal{U} = W^{2,2}(\Pi) \cap W^{1,4}(\Pi)$ equipped with norm $\|u\|_{\mathcal{U}} := \max(\|u\|_{W^{1,4}(\Pi)}, \|u\|_{W^{2,2}(\Pi)})$. Then $\mathcal{L}_{\text{SL}} : \mathcal{U} \rightarrow L^2(\Pi)$ is a bounded linear operator with $\|\mathcal{L}_{\text{SL}}\|_{\mathcal{U} \rightarrow L^2(\Pi)} \leq 2\sqrt{(\|\nabla \log \pi\|_{L^4(\Pi)}^2 + 1)}$. Furthermore, suppose that the distribution Π satisfies (i) $\int \|x\|_2^K d\Pi(x) < \infty$ for some $K > 8$, and that (ii) $(\nabla \log \pi)(x) \cdot (x/\|x\|_2) \leq -r\|x\|_2^\alpha$ for some $\alpha > -1$, $r > 0$, and all $\|x\|_2 > M$ for some $M > 0$. Suppose additionally that (iii) $|f(x)| \leq C_1 + C_2\|x\|_2^\beta$ for some $C_1, C_2 \geq 1$ and $\beta < K/4 - 2$. Then, $\exists u \in \mathcal{U}$ that solves the Stein equation $\mathcal{L}_{\text{SL}}u = f - \Pi[f]$.*

The fact that the space \mathcal{U} in Proposition 2 is separable ensures that suitable approximating sets \mathcal{V}_p can be constructed. For example, if $\{u_i\}_{i=1}^\infty$ is a spanning set for \mathcal{U} then we may set $\mathcal{V}_p = \text{span}(u_1, \dots, u_p)$, in which case $\cup_{p \in \mathbb{N}} \mathcal{V}_p$ is dense in \mathcal{U} so the result of Proposition 2 holds.

Notice that a solution to the Stein equation will not be unique, since one can introduce an additive constant. This motivates, in practice, the use of an additional regularizer $\Omega(\theta)$ to ensure uniqueness of the minimum of $\theta \mapsto J(v_\theta)$. To provide sufficient conditions for global convergence of SGD, we must also impose additional structure on the parametric class \mathcal{G} . The following result establishes convergence over a fixed set \mathcal{V}_p which is linear, both in the idealized scenario where we may directly sample from Π and in the practical scenario where we approximate Π with MCMC. Let $\sigma_{\min}(M)$ and $\sigma_{\max}(M)$ denote the minimum and maximum singular values of a matrix M .

Theorem 1. Let $\mathcal{L} : \mathcal{U} \rightarrow L^2(\Gamma)$ be a bounded linear operator for some distribution Γ on \mathbb{R}^d and let $\tilde{J}(\theta) := \|f - \theta_0 - \mathcal{L}v_\theta\|_{L^2(\Gamma)}^2$ for $\theta \in \mathbb{R}^{p+1}$. Assume that $\exists u \in \mathcal{U}$ solving the Stein equation $\mathcal{L}u = f - \Pi[f]$. Furthermore, assume:

- **Model:** Any $v \in \mathcal{V}_p$ can be expressed as $v_\theta = \sum_{i=1}^p \theta_i u_i$ where $u_1, \dots, u_p \in \mathcal{U}$. Furthermore, letting $\psi_0 := 1$ and $\psi_i := \mathcal{L}u_i$, we assume that the $\{\psi_i\}_{i=0}^p$ are linearly independent in $L^2(\Gamma)$.
- **Optimizer:** The random variables $x_i^{(t)}$ are distributed according to Γ , such that $x_i^{(s)}$ and $x_j^{(t)}$ are independent whenever $s \neq t$. Let $\theta^{(t)}$ denote the t -th iteration of SGD, with stochastic gradient at step t based on batch $(x_i^{(t)})_{i=1}^b$. Suppose the learning rate $(\alpha_t)_{t \in \mathbb{N}}$ satisfies:

$$\alpha_t = \frac{\beta}{\gamma+t}, \quad \beta > \frac{1}{2\sigma_{\min}(M)}, \quad \gamma > 0, \quad \alpha_1 \leq \frac{\sigma_{\min}(M^2)}{2\sigma_{\max}(M)(\sigma_{\max}(\mathbb{E}[(M_b^{(1)})^2]) + \sigma_{\min}(M^2))}$$

where $M_{i,j} := \Gamma[\psi_i \psi_j]$, $[M_b^{(t)}]_{i,j} := \frac{1}{b} \sum_{k=1}^b \psi_i(x_k^{(t)}) \psi_j(x_k^{(t)})$.

Then, there exists $\nu \geq 0$ such that $\mathbb{E}[\tilde{J}(\theta^{(t)})] \leq \frac{\nu}{\gamma+t} + \|\mathcal{L}\|_{\mathcal{U} \rightarrow L^2(\Gamma)}^2 \inf_{v \in \mathcal{V}_p} \|u - v\|_{\mathcal{U}}^2$.

The result is in expectation with respect to the law of $x_i^{(t)}$, and guarantees that the CVs trained with SGD will converge to the optimal CV of the form $g = \mathcal{L}v$, $v \in \mathcal{V}_p$. The second term is an upper bound on $\inf_{v \in \mathcal{V}_p} J(v)$, which will be zero when the assumptions of Proposition 1 or 2 hold. The case $\Gamma = \Pi$ corresponds to minimisation of $J(v)$ over $v \in \mathcal{V}_p$ using SGD with exact sampling from Π , while the case $\Gamma = \frac{1}{m} \sum_{i=1}^m \delta(x_i)$ corresponds to minimisation of the empirical risk $\tilde{J}(\theta) = \tilde{J}_m^{LS}(\theta)$ using SGD with mini-batches drawn from the (fixed) training dataset $(x_i)_{i=1}^m$. For the later case, the theorem is presented for \tilde{J}_m^{LS} , but a similar proof technique could be used for \tilde{J}_m^V .

The result does not apply to NNs; indeed, SGD is not expected to converge to the global minimum of \tilde{J} when a NN is employed since \tilde{J} will be non-convex. Bounds on the error incurred could however be obtained for alternative algorithms including stochastic gradient Langevin dynamics [13, 51, 68].

5 Empirical Assessment

Here we assess our method on both synthetic problems and on problems arising in a Bayesian statistical context. Our aim is twofold; (i) to assess whether our learning procedure provides a speed-up compared to existing approaches, and (ii) to gain insight into which class of CV may be most appropriate for a given context. The Stein operator \mathcal{L}_{SL} was used for all experiments. For the polynomial and kernel CVs, the regularizer $\Omega(\theta) = \|\theta\|_2^2$ was used, while for NN CVs the regularizer $\Omega(\theta) = \sum_{i=1}^m g_\theta(x_i)^2$ was used, following [64]. The regularization strength parameter λ was tuned by cross-validation. For some datasets, we employed two ensemble CVs: a sum of kernel and a polynomial (i.e., kernel + polynomial); and a sum including two kernels with different hyperparameters and a polynomial (i.e., multiple kernels + polynomial). Implementation details and further experiments are provided in Appendix D.

Genz Test Functions: The Genz functions are a standard benchmark used to evaluate a numerical integration method [21]. These functions f exhibit discontinuities and sharp peaks, but nevertheless they can be exactly integrated. The purpose of this first experiment is simply to assess whether *any* variance reduction can be achieved using our general framework in challenging and pathological situations.² Results are shown in Table 1 for polynomial-based and kernel-based CVs, as well as an ensemble of both. The CVs are trained using SGD on the least-squares objective functional with batch size $b = 8$ for 25 epochs. For each f , the mean absolute error (MAE) of polynomial CVs is always the largest while the linear combination of kernel and polynomial consistently performs the best. This is likely due to the increased flexibility of the CV. In all cases a substantial reduction in MAE was achieved, compared to MC. Full details and an extensive range of additional experiments are provided in Appendix D.2.

Integrating Gaussian Processes: To automatically generate test problems, we modelled f as a Gaussian process (GP) and sampled $(\Pi[f], f(x_1), \dots, f(x_n))$ from its Gaussian marginal; here the GP was centred and

²We emphasize that MC can be evaluated at negligible cost and we are not advocating that our methods should be preferred for this task.

Integrand f	MC	Poly. CV	Ker. CV	Poly.+Ker. CV
Continuous	2.77e-03	3.21e-03	3.28e-04	1.85e-04
Corner Peak	5.76e-03	1.07e-03	9.27e-06	6.05e-06
Discontinuous	2.04e-02	1.32e-02	3.91e-03	2.65e-03
Gaussian Peak	1.47e-03	1.40e-03	1.24e-05	1.05e-05
Oscillatory	4.17e-03	1.06e-03	4.63e-06	3.90e-06
Product Peak	1.37e-03	1.32e-03	2.12e-05	2.52e-06
Time (sec.)	7.10e-02	4.30e+00	2.60e+00	5.70e+00

Table 1: Mean absolute error (based on 20 repetitions) for polynomial-based CV, kernel-based CV and an ensemble of these, for the Genz benchmark [21]. We took $n = 1000$, $m = 500$ and $d = 1$. The training time presented is for 25 epochs, averaged over repetitions for all integrands.

a squared-exponential covariance function was used, and the distribution Π was taken to be an L -component Gaussian mixture model. In this way infinitely many problem instances can be generated, of a similar nature to those arising in computer experiments [31] and Bayesian numerical methods [11, 45]. We compared CVs based on polynomials, kernels, and NNs (three-layer ResNet with ReLU activation with 50 neurons per layer). Results are presented in Fig. 2, with implementational details in Appendix D.3. The left-most panel presents the performance of each CV for minimising either \tilde{J}_m^V or \tilde{J}_m^{LS} in $d = 1$. Polynomials are not flexible enough for such complex integrands, but kernels and NNs can achieve substantial reduction in error. However, we found that the “effective” time requires to implement a NN, including initialization of SGD and selecting an appropriate learning rate, meant that NN were not time-competitive with the other methods considered. The center-left panel studies the impact of d on the performance of each method. The performance of polynomial and kernel CVs degrades rapidly with d , but this is not the case for NNs. In both panels, \tilde{J}_m^{LS} leads to improved results compared to \tilde{J}_m^V . The centre-right and right panels report computational times of linear system and mini-batch SGD as d and m grows. These two panels verify that mini-batch SGD has linear time complexity as n or d is increased, whilst exact solution of linear systems leads to exponential computational costs for polynomial and kernel CVs.

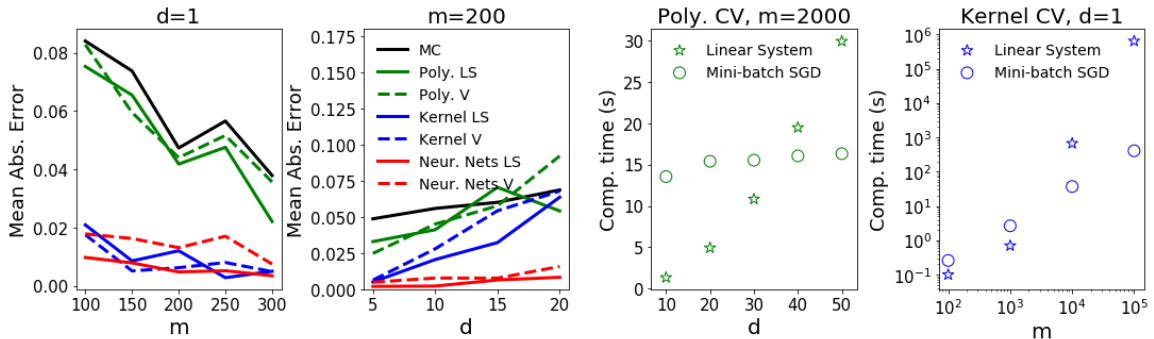


Figure 2: Integrating Gaussian Processes. Left and centre-left: The mean absolute error (based on 20 repetitions) of the CV estimators as a function of the training set size m and dimension d . Centre-right and right: Compute times for polynomial and kernel CVs as a function of m and d .

Parameter Inference for Ordinary Differential Equations: Here we consider the problem of inference for parameters $\alpha, \beta, \gamma, \delta$ of the Lotka–Volterra equations $\dot{x} = \alpha x - \beta xy$, $\dot{y} = \delta xy - \gamma y$, a popular ecological model for competing populations [39, 63]. Our experimental set up is identical to that used in [55]. Our task is to compute posterior means of these dynamic parameters based on datasets of size n arising as a subsample from Metropolis-adjusted Langevin algorithm output [56]; the full MCMC output provided the ground truth. Half of the sample was used to train CVs ($m = \frac{n}{2}$) and a batch size of $b = 8$ was used over 25 epochs in SGD based on the least-squares objective functional. Fig. 3 displays the performance of different CVs under sizes of training dataset. In each case the standard MC estimate is outperformed, with ensemble of multiple kernels with a polynomial or the NN performing uniformly best. Due to the computational cost of training NNs as shown in the rightmost panel, we found the ensemble to be preferable. The ensemble also leads to a convex objective which is easier to minimize.

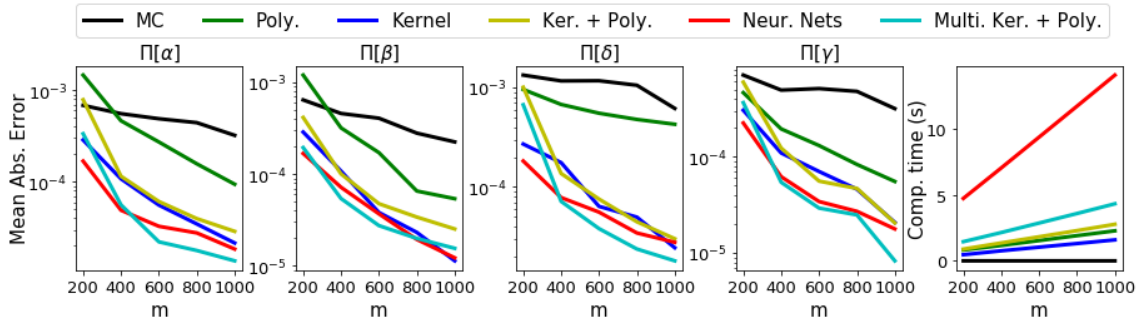


Figure 3: *Parameter Inference for Ordinary Differential Equations.* Each panel except the rightmost presents the mean absolute error (based on 20 repetitions) for approximation of posterior expectations of model parameters using MCMC output. The rightmost panel presents the computing time of training these CVs. Here “MC” represents the benchmark where no CV is used.

High-dimensional Bayesian Logistic Regression on Sonar Data: In this final example, we consider Bayesian logistic regression applied to sonar data from [19, 25], as considered in [60]. The parameter is the 61-dimensional regression coefficient β and contains information about the energy frequencies being reflected from either a metal cylinder ($y = 1$) or a rock ($y = 0$). MCMC was used to sample from the posterior of β as described in [60] and the full output constitutes our ground truth. Our task is to approximate the posterior probability that an unlabeled data point z corresponds to a metal cylinder, rather than rock, based on a subset of size m from the MCMC output. Thus $f(\beta) = (1 + \exp(-z^\top \beta))^{-1}$.

In these experiments, all MCMC samples were included in the training dataset (i.e. $m = n$). Batch sizes of $b = 8$ were used over 25 epochs in SGD and J_m^{LS} was used.

Fig. 4 compares the performance of different CV methods. The two ensemble CVs and the NNs perform significantly better than other CVs. When $m < 1000$, the NNs yield the smallest mean absolute errors, followed by the CV with multiple kernels and a polynomial.

When $m \geq 1000$, the ensemble CV surpasses NNs. One possible explanation is that for all values of m we used the same multi-layer perceptron (MLP) with 6 layers and 20 nodes in each of them. Therefore, the NNs size (capacity) remains the same while the training data size m increases. Further growing the depth of NN could lead to an improved performance. Furthermore, the results for polynomials and kernels demonstrate that our general framework based on SGD can achieve comparable MAE with exactly solving the linear systems, but with a fraction of the associated computational overhead. The compute time of NN in Fig. 4 does not capture the time required to manually calibrate SGD, so that the “effective” compute time is much higher than reported.

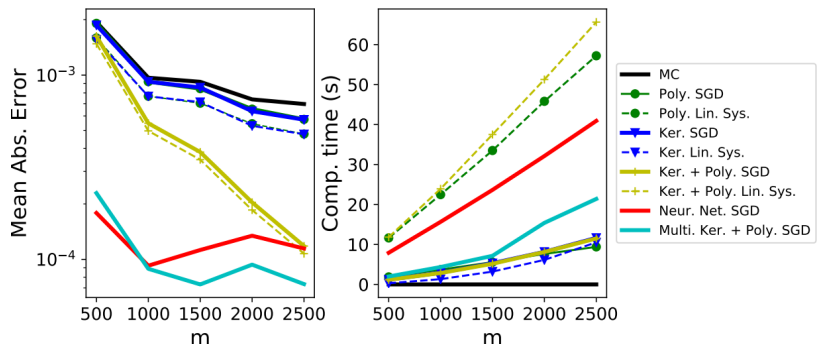


Figure 4: *Sonar Dataset.* The mean absolute error (left) and compute times (right), as a function of the size m of the training set; based on 20 repetitions.

6 Conclusion

This paper proposes a general framework for developing CVs using Stein operators and SGD. It was demonstrated that (i) the proposed training scheme leads to speed-ups compared to existing CV methods; (ii) novel CV methods (e.g., ensemble methods) can be easily developed; (iii) theoretical analysis can be performed in quite a general setting that simultaneously encompasses multiple CV methods. Further research could explore the use of other Stein classes and operators. In terms of Stein classes, one could consider the use of wavelets, which are known for their good performance for multi-scale function approximation, or other NN architectures which could provide further gains in high dimensions. Stein operators are not unique and one could explore

parameterized operators [33] and include these parameters in the optimization scheme. Finally, one could construct novel CVs on other spaces, such as general smooth manifolds or countable spaces [6].

Acknowledgment

The authors would like to thank Ricardo Henao and Charline Le Lan for helpful discussions, and Wilson Chen, Marina Riabiz and Leah South for sharing MCMC samples from the model of atmospheric pollutants, the predator-prey model and the logistic regression model respectively. CJO, ABD, FXB were also supported by the Lloyd’s Register Foundation Programme on Data-Centric Engineering and the Alan Turing Institute under the EPSRC grant [EP/N510129/1]. FXB was supported by an Amazon Research Award on “Transfer Learning for Numerical Integration in Expensive Machine Learning Systems”.

References

- [1] S. Andradóttir, D. P. Heyman, and T. J. Ott. Variance reduction through smoothing and control variates for Markov chain simulations. *ACM Transactions on Modeling and Computer Simulation*, 3(3):167–189, 1993.
- [2] R. Assaraf and M. Caffarel. Zero-variance principle for Monte Carlo algorithms. *Physical Review Letters*, 83(23):4682, 1999.
- [3] J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29:599–615, 2019.
- [4] A. D. Barbour. Stein’s method and Poisson process convergence. *Journal of Applied Probability*, 25:175–184, 1988.
- [5] A. Barp, F.-X. Briol, A. B. Duncan, M. Girolami, and L. Mackey. Minimum Stein discrepancy estimators. In *Neural Information Processing Systems*, pages 12964–12976, 2019.
- [6] A. Barp, C. J. Oates, E. Porcu, and M. Girolami. A Riemannian-Stein kernel method. *arXiv:1810.04946*, 2018.
- [7] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov. Variance reduction for Markov chains with application to MCMC. *Statistics and Computing*, 30:973–997, 2020.
- [8] D. Belomestny, L. Iosipoi, and N. Zhivotovskiy. Variance reduction via empirical variance minimization: convergence and complexity. *Doklady Mathematics*, 98:494–497, 2018.
- [9] D. Belomestny, E. Moulines, N. Shagadatov, and M. Urusov. Variance reduction for MCMC methods via martingale representations. *arXiv:1903.0737*, 2019.
- [10] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
- [11] F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? (with discussion). *Statistical Science*, 34(1):1–22, 2019.
- [12] N. Brosse, A. Durmus, S. Meyn, and E. Moulines. Diffusion approximations and control variates for MCMC. *arXiv:1808.01665*, 2018.
- [13] N. H. Chau, E. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. *arXiv:1905.13142*, 2019.
- [14] L. H. Y. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein’s Method*. Springer, 2010.
- [15] W. Y. Chen, A. Barp, F.-X. Briol, J. Gorham, M. Girolami, L. Mackey, and C. J. Oates. Stein point Markov chain Monte Carlo. In *International Conference on Machine Learning, PMLR 97*, pages 1011–1021, 2019.
- [16] W. Y. Chen, L. Mackey, J. Gorham, F.-X. Briol, and C. J. Oates. Stein points. In *Proceedings of the International Conference on Machine Learning, PMLR 80:843-852*, 2018.
- [17] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. *International Conference on Machine Learning*, 48:2606–2615, 2016.

- [18] P. Dellaportas and I. Kontoyiannis. Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(1):133–161, 2012.
- [19] D. Dua and C. Graff. UCI machine learning repository. 2017.
- [20] N. Friel, A. Mira, and C. J. Oates. Exploiting multi-core architectures for reduced-variance estimation with intractable likelihoods. *Bayesian Analysis*, 11(1):215–245, 2014.
- [21] A. Genz. Testing multidimensional integration routines. In *Proceedings of the International Conference on Tools, Methods and Languages for Scientific and Engineering Computation*, pages 81–94, 1984.
- [22] J. Gorham, A. Duncan, L. Mackey, and S. Vollmer. Measuring sample quality with diffusions. *Annals of Applied Probability*, 29(5):2884–2928, 2019.
- [23] J. Gorham and L. Mackey. Measuring sample quality with Stein’s method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.
- [24] J. Gorham and L. Mackey. Measuring sample quality with kernels. In *Proceedings of the International Conference on Machine Learning*, pages 1292–1301, 2017.
- [25] R. P. Gorman and T. J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural networks*, 1(1):75–89, 1988.
- [26] W. Grathwohl, D. Choi, Y. Wu, G. Roeder, and D. Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018.
- [27] E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.
- [28] H. Hammer and H. Tjelmeland. Control variates for the Metropolis-Hastings algorithm. *Scandinavian Journal of Statistics*, 35(3):400–414, 2008.
- [29] S. G. Henderson and P. W. Glynn. Approximating martingales for variance reduction in Markov process simulation. *Mathematics of Operations Research*, 27(2):253–271, 2002.
- [30] F. J. Hickernell, C. Lemieux, and A. B. Owen. Control variates for quasi-Monte Carlo. *Statistical Science*, 20(1):1–31, 2005.
- [31] M. C. Kennedy and A. O. Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(3):425–464, 2001.
- [32] R. Leluc, F. Portier, and J. Segers. Control variate selection for Monte Carlo integration. *arXiv:1906.10920*, 2019.
- [33] C. Ley and Y. Swan. Parametric Stein operators and variance bounds. *Brazilian Journal of Probability and Statistics*, 30(2), 2016.
- [34] H. Liu, Y. Feng, Y. Mao, D. Zhou, J. Peng, and Q. Liu. Action-dependent control variates for policy optimization via Stein’s identity. In *International Conference on Learning Representation*, 2018.
- [35] Q. Liu and J. D. Lee. Black-box importance sampling. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 952–961, 2017.
- [36] Q. Liu, J. D. Lee, and M. I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In *International Conference on Machine Learning*, pages 276–284, 2016.
- [37] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, 2016.
- [38] S. Liu, T. Kanamori, W. Jitkrittum, and Y. Chen. Fisher efficient inference of intractable models. In *Neural Information Processing Systems*, pages 8793–8803, 2019.
- [39] A. Lotka. Principles of physical biology. *Baltimore: Waverly*, 1925.
- [40] A. Mira, R. Solgi, and D. Imparato. Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.
- [41] N. J. Newton. Variance reduction for simulated diffusions. *SIAM Journal on Applied Mathematics*, 54(6):1780–1805, 1994.

- [42] C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami. Convergence rates for a class of estimators based on Stein’s identity. *Bernoulli*, 25(2):1141–1159, 2019.
- [43] C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society B: Statistical Methodology*, 79(3):695–718, 2017.
- [44] C. J. Oates, T. Papamarkou, and M. Girolami. The controlled thermodynamic integral for Bayesian model comparison. *Journal of the American Statistical Association*, 2016.
- [45] A. O’Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991.
- [46] J. Paisley, D. Blei, and M. Jordan. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, 2012.
- [47] T. Papamarkou, A. Mira, and M. Girolami. Zero variance differential geometric Markov chain Monte Carlo algorithms. *Bayesian Analysis*, 9(1):97–128, 2014.
- [48] E. Pardoux and A. Y. Vertennikov. On the Poisson equation and diffusion approximation. I. *Annals of Probability*, 29(3):1061–1085, 2001.
- [49] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
- [50] F. Portier and J. Segers. Monte Carlo integration with a growing number of control variates. *Journal of Applied Probability*, 56(4):1168–1186, 2019.
- [51] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a non-asymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory, PMLR 65*, pages 1674–1703, 2017.
- [52] R. Ranganath, J. Altsosaar, D. Tran, and D. M. Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.
- [53] R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.
- [54] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [55] M. Riabiz, W. Chen, J. Cockayne, P. Swietach, S. A. Niederer, L. Mackey, and C. J. Oates. Optimal thinning of MCMC output. *arXiv:2005.03952*, 2020.
- [56] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996.
- [57] N. Ross. Fundamentals of Stein’s method. *Probability Surveys*, 8:210–293, 2011.
- [58] D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*. Cambridge University Press, 2003.
- [59] L. F. South, T. Karvonen, C. Nemeth, M. Girolami, and C. J. Oates. Semi-exact control functionals from Sard’s method. *arXiv:2002.00033*, 2020.
- [60] L. F. South, C. J. Oates, A. Mira, and C. Drovandi. Regularised zero-variance control variates for high-dimensional variance reduction. *arXiv:1811.05073*, 2019.
- [61] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability*, pages 583–602. University of California Press, 1972.
- [62] G. Tucker, A. Mnih, C. J. Maddison, J. Lawson, and J. Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pages 2627–2636, 2017.
- [63] V. Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560, 1926.
- [64] R. Wan, M. Zhong, H. Xiong, and Z. Zhu. Neural control variates for monte carlo variance reduction. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 533–547, 2019.

- [65] C. Wang, X. Chen, A. Smola, and E. P. Xing. Variance reduction for stochastic gradient optimization. In *Advances in Neural Information Processing Systems*, pages 181–189, 2013.
- [66] J. Yang, Q. Liu, V. Rao, and J. Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570, 2018.
- [67] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning*, page 116, 2004.
- [68] Y. Zhang, O. D. Akyildiz, T. Damoulas, and S. Sabanis. Nonasymptotic estimates for stochastic gradient Langevin dynamics under local conditions in nonconvex optimization. *arXiv:1910.03008*, 2019.

Appendices

The following document contains the appendices for the paper *Scalable Control Variates for Monte Carlo Methods via Stochastic Optimization*. In Appendix A, we review existing methodology for CVs based on polynomials and kernels. Appendix B recalls mathematical definitions that are required for the proofs in Appendix C. Finally, Appendix D contains a detailed exposition of the experimental setup in the paper for reproducibility, and provides additional results.

A Additional Background on Control Variates

Let $\{x_i\}_{i=1}^n$ be a set containing approximate samples from Π . The classic approach to CVs is based on data-splitting, such that a CV g is constructed based on a subset of the samples $\{x_i\}_{i=1}^m$, then a Monte Carlo estimator based on $f - g$ is evaluated using the remainder of the samples, $\{x_i\}_{i=m+1}^n$. Thus $\Pi[f]$ is approximated using the CV estimator

$$\frac{1}{(n-m)} \sum_{i=m+1}^n (f(x_i) - g(x_i))$$

where $g(\cdot) = g(\cdot; x_1, \dots, x_m)$. If the x_i are independent samples from Π then such a CV estimator is unbiased. It is also common practice to use the same set $\{x_i\}_{i=1}^n$ for both the construction of g and evaluation of the Monte Carlo estimator; in this case the estimator is biased in general but may enjoy superior mean square error.

In this section we recall existing approaches to constructing CVs, providing references to existing literature where appropriate.

A.1 Control Variates based on Polynomials

As pointed out in the main text, the polynomial CVs of [2, 40, 47, 60] are based on \mathcal{L}_{SL} and take the form:

$$g_\theta(x) = (\mathcal{L}_{\text{SL}} v_\theta)(x) = \Delta_x v_\theta(x) + \nabla_x v_\theta(x) \cdot \nabla_x \log \pi(x),$$

where $v_\theta(x)$ is a polynomial of order $k \in \mathbb{N}$. For first order polynomials (i.e. $k = 1$ and $p = d$), we have $v_\theta(x) = \sum_{i=1}^d \theta_i x_i$ where $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$, and the CV estimator is of the form: $g_\theta(x) = \theta \cdot \nabla_x \log \pi(x)$. Note that the constant term is not included, since this is in the null space of \mathcal{L}_{SL} . More generally, for an arbitrary polynomial of order k ,

$$v_\theta(x) = \sum_{j=1}^p \theta_j x_1^{\alpha_{j1}} \dots x_d^{\alpha_{jd}},$$

for some $\theta = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p$ and where the rows of the matrix $\alpha \in \mathbb{Z}^{p \times d}$ are multi-indices containing polynomial coefficients such that the polynomial has total degree $k \geq 1$: $1 \leq \sum_{l=1}^d \alpha_{jl} \leq p$. The total number of polynomials satisfying this condition is $p = \binom{d+k}{d} - 1$. The CVs based on such polynomials take the form $g_\theta(x) = \theta \cdot b(x)$, where the vector $b(x) = (b_1(x), \dots, b_p(x))$ has components:

$$b_j(x) = \left[\sum_{l=1}^d \max(0, \alpha_{jl}) x_l^{\alpha_{jl}-1} \frac{\partial \log \pi}{\partial x_l} \prod_{z=1, z \neq l}^d x_z^{\alpha_{jz}} \right] \\ + \left[\max(0, \alpha_{jl}(\alpha_{jl} - 1)) x_l^{\alpha_{jl}-2} \prod_{z=1, z \neq l}^d x_z^{\alpha_{jz}} \right]$$

for $j = 1, \dots, p$; see Appendix A of [60]. The value of θ which minimizes the least-squares objective \hat{J}_m^{LS} is given by $\theta_m^* = \hat{V}_m^{-1} \hat{C}_m$ with:

$$\hat{V}_m = \frac{1}{(m-1)} \sum_{i=1}^m (b(x_i) - \frac{1}{m} \sum_{i=1}^m b(x_i)) (b(x_i) - \frac{1}{m} \sum_{i=1}^m b(x_i))^\top, \\ \hat{C}_m = \frac{1}{(m-1)} \sum_{i=1}^m (f(x_i) - \frac{1}{m} \sum_{i=1}^m f(x_i)) (b(x_i) - \frac{1}{m} \sum_{i=1}^m b(x_i))^\top.$$

The size m of the training dataset is required to be sufficiently large to ensure that the matrix \hat{V}_m is non-singular, otherwise additional regularisation is required [60]. Exact solution of this linear system for θ_m^* requires a computational cost of $O(p^3)$, which can be prohibitive since p increases rapidly with both d and k .

A.2 Control Functionals: Control Variates based on Reproducing Kernels

Control functionals are CVs constructed using a nonparametric kernel-based interpolant. Let $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a symmetric positive definite kernel with corresponding reproducing kernel Hilbert space \mathcal{H}_k . [43] noted that, under some regularity conditions, the kernels

$$k_0(x, y) := \nabla_x \cdot \nabla_y k(x, y) + \nabla_x k(x, y) \cdot \nabla_y \log \pi(y) \\ + \nabla_y k(x, y) \cdot \nabla_x \log \pi(x) + k(x, y) \nabla_x \log \pi(x) \cdot \nabla_y \log \pi(y), \quad (3)$$

and $k_+(x, y) := k_0(x, y) + \sigma^2$ for $\sigma > 0$ are also reproducing kernels with corresponding RKHS respectively denoted \mathcal{H}_{k_0} and \mathcal{H}_{k_+} . More specifically, \mathcal{H}_{k_+} is just \mathcal{H}_{k_0} with the addition of constant functions. The RKHS \mathcal{H}_{k_+} can be used to approximate the integrand f as follows:

$$\tilde{f}_\sigma \in \arg \min \{ \|h\|_{\mathcal{H}_+} \text{ s.t. } h \in \mathcal{H}_+, h(x_i) = f(x_i), i = 1, \dots, m \}.$$

Under regularity conditions this provides a unique approximation of the form (see e.g. Proposition 1 in [11]):

$$\tilde{f}_\sigma(x) = k_+(x, X) k_+(X, X)^{-1} f(X),$$

where we have used the matrix notation $[k_+(x, X)]_i = k_+(x, x_i)$, $[f(X)]_i = f(x_i)$ and $[k_+(X, X)]_{i,j} = k_+(x_i, x_j)$ for $i, j \in \{1, \dots, m\}$. The integral of this approximation can be obtained in closed form:

$$\Pi[\tilde{f}_\sigma] = \sigma^2 \mathbf{1}^\top k_+(X, X)^{-1} f(X),$$

where $\mathbf{1}$ is the vector $(1, \dots, 1)^\top$. Finally, the control functional is therefore given by $g_\sigma(x) = \tilde{f}_\sigma(x) - \Pi[\tilde{f}_\sigma]$, which takes the form:

$$g_\sigma(x) := (k_+(x, X) - \sigma^2 \mathbf{1}^\top) k_+(X, X)^{-1} f(X).$$

To remove the dependence on the regularization due to σ , we let $\sigma \rightarrow \infty$ and get the CV [43]:

$$g(x) = k_0(x, X) k_0(X, X)^{-1} \left[f(X) - \left(\frac{\mathbf{1}^\top k_0(X, X)^{-1} f(X)}{\mathbf{1}^\top k_0(X, X)^{-1} \mathbf{1}} \right) \mathbf{1} \right].$$

Properties of control functional estimators have been detailed in [6, 42, 43].

A.3 Control Variates Based on Ensembles of Kernels and Polynomials

In our experiments, we also employed a linear combination (or *ensemble*) of CVs, one based on a kernel and the other on a polynomial. Given a training set $X = \{x_i\}_{i=1}^m$, the CV is given by:

$$g_\theta(x) = \Delta_x \Phi_{\bar{\theta}}(x) + \nabla_x \Phi_{\bar{\theta}}(x) \cdot \nabla_x \log \pi(x) + \bar{\theta} \cdot k_0(x, X) \\ = \tilde{\theta}^\top b(x) + \bar{\theta} \cdot k_0(x, X),$$

where $\theta = (\tilde{\theta}, \bar{\theta})$, $\tilde{\theta} = (\theta_1, \dots, \theta_p)^\top$, $\bar{\theta} = (\bar{\theta}_1, \dots, \bar{\theta}_m)^\top$, $\Phi_{\bar{\theta}}(x)$ is some polynomial of order $k \in \mathbb{N}$.

This form of CV was proposed in [59] under the *semi-exact control functionals*, where the authors derived a closed-form expression for the parameter vector θ under the requirements that: (i) $g_\theta(x_i) = f(x_i)$ for $i = 1, \dots, m$ and (ii) $g_\theta = f$ whenever f belongs to a user-specified finite-dimensional vector space spanned by b_1, \dots, b_p . Requirement (ii) is an *exactness* condition, which motivated the name *semi-exact*. Let

$$B = \begin{bmatrix} 1 & b_1(x_1) & \cdots & b_p(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & b_1(x_m) & \cdots & b_p(x_m) \end{bmatrix}.$$

Then, under regularity conditions, [59] showed that $\tilde{\theta}$ and $\bar{\theta}$ can be found by solving:

$$\begin{bmatrix} k_0(X, X) & B \\ B^\top & \mathbf{0}_{p \times p} \end{bmatrix} \begin{bmatrix} \tilde{\theta} \\ \bar{\theta} \end{bmatrix} = \begin{bmatrix} f(X) \\ \mathbf{0}_{p \times 1} \end{bmatrix},$$

where $\mathbf{0}_{p \times 1}$ is a $p \times 1$ column vector of zeros and $\mathbf{0}_{p \times p}$ is a $p \times p$ matrix of zeros. For the experiments in this paper we do *not* enforce exactness constraints for mini-batch SGD algorithm; as shown in Fig. 4, the performance of ensemble CVs (a kernel with a polynomial) trained by mini-batch SGD and solving linear system is comparable.

B Mathematical Definitions

Before moving on to the proofs of the main results in the paper (presented in Appendix C), we recall basic definitions from functional analysis.

Let X and Y be two normed real vector spaces. A function $f : X \rightarrow Y$ is called *Lipschitz* continuous if there exists a constant L such that, $\forall x, x' \in X$

$$\|f(x) - f(x')\|_Y \leq L\|x - x'\|_X.$$

The smallest such $L \geq 0$ is called the *Lipschitz constant* of f .

The norm of a bounded linear operator $\mathcal{L} : X \rightarrow Y$ is given by:

$$\|\mathcal{L}\|_{X \rightarrow Y} := \inf \{c \geq 0 : \|\mathcal{L}x\| \leq c\|y\| \ \forall x \in X\}.$$

For $1 \leq p < \infty$ we denote

$$\begin{aligned} L^p(\Pi) &:= \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} \text{ measurable} \mid \|f\|_{L^p(\Pi)} := \left(\int_{\mathbb{R}^d} |f(x)|^p \Pi(dx) \right)^{\frac{1}{p}} < \infty \right\}. \\ L_{\text{loc}}^p &:= \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} \text{ measurable} \mid \left(\int_K |f(x)|^p dx \right)^{\frac{1}{p}} < \infty, \ \forall \text{compact } K \subset \mathbb{R}^d \right\}. \end{aligned}$$

As usual, $L^p(\Pi)$ can be interpreted as a normed space via identification of functions that agree Π -almost everywhere on \mathbb{R}^d . Using this definition, we can now also define weighted Sobolev spaces of integer smoothness. These are the space defined as:

$$\begin{aligned} W^{k,p}(\Pi) &:= \left\{ f \in L^p(\Pi) \mid D^\alpha f \in L^p(\Pi) \ \forall |\alpha| \leq k \right\} \\ W_{\text{loc}}^{k,p} &:= \left\{ f \in L_{\text{loc}}^p \mid D^\alpha f \in L_{\text{loc}}^p \ \forall |\alpha| \leq k \right\} \end{aligned}$$

In this definition, $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ is a multi-index and D^α denotes the weak derivative of order α , i.e. $D^\alpha f := \partial^{|\alpha|} f / \partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}$. Recall that $W^{k,p}(\Pi)$ can be interpreted as a normed space with norm

$$\|u\|_{W^{k,p}(\Pi)} := \left(\sum_{i=0}^k \sum_{\alpha: |\alpha|=i} \int |D^\alpha u(x)|^p d\Pi(x) \right)^{\frac{1}{p}},$$

again via identification of functions whose derivatives up to order $|\alpha| \leq k$ agree Π -almost everywhere on \mathbb{R}^d .

C Proofs of Theoretical Results

In this section we prove the theoretical results presented in the main text.

C.1 Proof of Proposition 1

Proof. Let $u \in \mathcal{U}$ solve the Stein equation $\mathcal{L}u = f - \Pi[f]$. Since \mathcal{L} is a bounded linear operator between normed spaces,

$$J(v) = \|f - \Pi[f] - \mathcal{L}v\|_{L^2(\Pi)}^2 = \|\mathcal{L}[u] - \mathcal{L}[v]\|_{L^2(\Pi)}^2 \leq \|\mathcal{L}\|_{\mathcal{U} \rightarrow L^2(\Pi)}^2 \|u - v\|_{\mathcal{U}}^2$$

where $\|\mathcal{L}\|_{\mathcal{U} \rightarrow L^2(\Pi)} < \infty$. Fix $\epsilon > 0$. Since $u \in \mathcal{U}$ and $\cup_{p \in \mathbb{N}} \mathcal{V}_p$ is dense in \mathcal{U} , there exists $v \in \cup_{p \in \mathbb{N}} \mathcal{V}_p$ such that $\|u - v\|_{\mathcal{U}} < \epsilon$. In particular, there exists $q \in \mathbb{N}$ such that $v \in \mathcal{V}_q$. Moreover, since $\mathcal{V}_q \subseteq \mathcal{V}_p$ for all $q \leq p$, the function $p \mapsto \inf_{v \in \mathcal{V}_p} J(v)$ is non-increasing. Thus

$$0 \leq \lim_{p \rightarrow \infty} \inf_{v \in \mathcal{V}_p} J(v) \leq \inf_{v \in \mathcal{V}_q} J(v) \leq \|\mathcal{L}\|_{\mathcal{U} \rightarrow L^2(\Pi)}^2 \inf_{v \in \mathcal{V}_q} \|u - v\|_{\mathcal{U}}^2 \leq \|\mathcal{L}\|_{\mathcal{U} \rightarrow L^2(\Pi)}^2 \epsilon^2.$$

Since $\epsilon > 0$ was arbitrary, the right hand side can be made arbitrarily small, proving the result. \square

C.2 Proof of Proposition 2

Proof. First we will show that \mathcal{L}_{SL} is a bounded linear operator from $\mathcal{U} = W^{2,2}(\Pi) \cap W^{1,4}(\Pi)$ to $L^2(\Pi)$. To this end:

$$\|\mathcal{L}_{\text{SL}}[u] - \mathcal{L}_{\text{SL}}[v]\|_{L^2(\Pi)}^2 = \|\nabla \log \pi \cdot \nabla(u-v) + \nabla \cdot \nabla(u-v)\|_{L^2(\Pi)}^2 \quad (4)$$

$$\leq 2 \left[\|\nabla \log \pi \cdot \nabla(u-v)\|_{L^2(\Pi)}^2 + \|\nabla \cdot \nabla(u-v)\|_{L^2(\Pi)}^2 \right] \quad (5)$$

$$\leq 2 \left[\|\nabla \log \pi\|_{L^4(\Pi)}^2 \|\nabla(u-v)\|_{L^4(\Pi)}^2 + \|u-v\|_{W^{2,2}(\Pi)}^2 \right] \quad (6)$$

$$\leq 2 \left(\|\nabla \log \pi\|_{L^4(\Pi)}^2 + 1 \right) \left(\|u-v\|_{W^{1,4}(\Pi)}^2 + \|u-v\|_{W^{2,2}(\Pi)}^2 \right), \quad (7)$$

$$\leq 4 \left(\|\nabla \log \pi\|_{L^4(\Pi)}^2 + 1 \right) \max(\|u-v\|_{W^{1,4}(\Pi)}, \|u-v\|_{W^{2,2}(\Pi)})^2$$

Eq. 4 follows by definition of the Stein operator, Eq. 5 follows from the fact that $(a+b)^2 \leq 2(a^2+b^2)$. Eq. 6 follows from the vector-valued Hölder inequality together with the definition of $\|\cdot\|_{W^{2,2}(\Pi)}$. Eq. 7 follows from the definition of $\|\cdot\|_{W^{1,4}(\Pi)}$.

Thus \mathcal{L}_{SL} is a bounded linear operator as claimed, and moreover $\|\mathcal{L}_{\text{SL}}\|_{\mathcal{U} \rightarrow L^2(\Pi)} \leq 2\sqrt{\|\nabla \log \pi\|_{L^4(\Pi)}^2 + 1}$.

The second task is to establish that there exists a solution to the Stein equation $\mathcal{L}_{\text{SL}}u = f - \Pi[f]$. For this we leverage [48, Theorem 1] which states that, if conditions (ii), (iii) hold, there exists a solution u to the Stein equation which is continuous and belongs to $W_{\text{loc}}^{2,q}$ for all $q > 1$. Moreover, $\forall m > \beta + 2$ there exists C_m such that $|u(x)| + |\nabla u(x)| \leq C_m(1 + |x|^m)$ for all $x \in \mathbb{R}^d$. By assumption (i) it follows that $u \in W^{1,4}(\Pi)$. Moreover, since π was assumed to be smooth (recall, this was assumed at the outset in Section 1), standard regularity results imply that u is smooth and so, is a classical solution. We can therefore write

$$|\Delta u(x)| \leq |f(x)| + |\Pi(f)| + |\nabla \log \pi(x) \cdot \nabla u(x)|, \quad x \in \mathbb{R}^d,$$

so that $\|\Delta u\|_{L^2(\Pi)} \leq 2\|f\|_{L^2(\Pi)} + \|\nabla \log \pi\|_{L^4(\Pi)}\|u\|_{W^{1,4}(\Pi)} < \infty$. It follows that $u \in W^{2,2}(\Pi) \cap W^{1,4}(\Pi)$, as claimed. \square

C.3 Proof of Theorem 1

The following elementary lemma will be required:

Lemma 1. *Let A and B be positive semi-definite matrices of equal dimension, such that $\sigma_{\min}(A) \geq \sigma_{\max}(B)$. Then $A - B$ is also a positive semi-definite matrix.*

Proof. Let A and B be $d \times d$ dimensional. For any $x \in \mathbb{R}^d$ we have that

$$\begin{aligned} x^\top (A - B)x &= x^\top Ax - x^\top Bx \geq \sigma_{\min}(A)\|x\|_2^2 - \sigma_{\max}(B)\|x\|_2^2 \\ &= (\sigma_{\min}(A) - \sigma_{\max}(B))\|x\|_2^2 \geq 0. \end{aligned}$$

\square

The implication of our linearity assumption on the model is that the parametrized objective function is a quadratic function in $\theta \in \mathbb{R}^{p+1}$; which simplifies the analysis of SGD.

Proof of Theorem 1. From linearity of \mathcal{L} , the objective function that we aim to minimize is

$$\tilde{J}(\theta) = \|f - \theta_0 - \sum_{i=1}^p \theta_i \mathcal{L}u_i\|_{L^2(\Gamma)}^2 = \|f - \sum_{i=0}^p \theta_i \psi_i\|_{L^2(\Gamma)}^2,$$

and we have $\psi_0 = 1$ and $\psi_i = \mathcal{L}u_i$, $i = 1, \dots, p$. This can be re-expressed in matrix notation as

$$\tilde{J}(\theta) = \theta^\top M \theta - 2a^\top \theta + \Gamma[f^2], \quad (8)$$

where $M_{i,j} = \Gamma[\psi_i \psi_j]$ and $a_i := \Gamma[f \psi_i]$. Our two cases of interest are $\Gamma = \Pi$ and $\Gamma = \frac{1}{m} \sum_{i=1}^m \delta(x_i)$ for a fixed set $\{x_i\}_{i=1}^m \subset \mathbb{R}^d$.

Our aim is to verify the preconditions of Theorem 4.7 in [10]. If these are satisfied then we may conclude that, under the assumptions on the learning rate in the statement of Proposition 1, for some constant $\nu \geq 0$,

$$\mathbb{E} \left[\tilde{J}(\theta^{(t)}) \right] \leq \frac{\nu}{\gamma+t} + \inf_{\theta \in \mathbb{R}^{p+1}} \tilde{J}(\theta).$$

In particular, since we have assumed that $\exists u \in \mathcal{U}$ that solves the Stein equation $\mathcal{L}u = f - \Pi[f]$ and that $\mathcal{L} : \mathcal{U} \rightarrow L^2(\Gamma)$ is a bounded linear operator, the same argument used in the proof of Proposition 1 shows that $\tilde{J}(\theta) \leq \|\mathcal{L}\|_{\mathcal{U} \rightarrow L^2(\Gamma)}^2 \|u - v_\theta\|_{\mathcal{U}}^2$, so that

$$\mathbb{E} \left[\tilde{J}(\theta^{(t)}) \right] \leq \frac{\nu}{\gamma+t} + \|\mathcal{L}\|_{\mathcal{U} \rightarrow L^2(\Gamma)}^2 \inf_{v \in \mathcal{V}_p} \|u - v\|_{\mathcal{U}}^2, \quad (9)$$

as claimed.

Theorem 4.7 in [10] requires that $\tilde{J}(\theta)$ is continuously differentiable with $\nabla \tilde{J}$ being Lipschitz. This is satisfied in our context, with Lipschitz constant $2\sigma_{\max}(M)$. The two remaining conditions that we must verify in order to apply Theorem 4.7 of [10] are; (i) the strong convexity property

$$\tilde{J}(\theta) - \tilde{J}(\vartheta) \geq \langle \nabla \tilde{J}(\vartheta), \theta - \vartheta \rangle + \frac{l}{2} \|\theta - \vartheta\|_2^2$$

for some $l > 0$, and (ii) the bound

$$\mathbb{E}[\|\nabla \tilde{J}_b(\theta)\|_2^2] \leq C_1 + C_2 \|\nabla \tilde{J}(\theta)\|_2^2 \quad (10)$$

for some constants C_1, C_2 where $\nabla \tilde{J}_b$ is a stochastic estimate of $\nabla \tilde{J}$ based on b samples from Γ . See the discussion of (4.9) in [10] for why establishing (10) is a sufficient condition for Theorem 4.7.

First we verify condition (i); that the optimization problem is strongly convex in $\theta \in \mathbb{R}^{p+1}$. From direct computation with (8) we obtain that $\tilde{J}(\theta)$ is strongly convex if and only if $(\theta - \vartheta)^\top M(\theta - \vartheta) \geq \frac{c}{2} \|\theta - \vartheta\|_2^2$. Since M is positive semi-definite and the ψ_i were assumed to be linearly independent in $L^2(\Gamma)$, the matrix M is non-singular and $\sigma_{\min}(M) > 0$. Thus \tilde{J} is strongly convex with strong convexity constant $c = 2\sigma_{\min}(M)$.

It remains only to verify condition (ii). Let $\psi(x) := (\psi_0, \psi_1(x), \dots, \psi_p(x))^\top$. Recall that, in the t th step of SGD, the gradient $\nabla \tilde{J}$ is unbiasedly estimated with

$$\begin{aligned} \nabla \tilde{J}_b(\theta) &:= \nabla \left[\frac{1}{b} \sum_{i=1}^b \left(f(x_i^{(t)}) - \psi(x_i^{(t)})^\top \theta \right)^2 \right] \\ &= -\frac{2}{b} \sum_{i=1}^b \left(f(x_i^{(t)}) - \psi(x_i^{(t)})^\top \theta \right) \psi(x_i^{(t)}), \end{aligned}$$

where the $x_i^{(t)}$ are independent samples from Γ . Let f be the vector with entries $f_i := f(x_i^{(t)})$, let a_b be the vector with entries $a_{b,j} := \frac{1}{b} \sum_{i=1}^b f(x_i^{(t)}) \psi_j(x_i^{(t)})$, so that $\mathbb{E}[a_b] = a$, and $\Psi_{i,j} := \psi_j(x_i^{(t)})$. Thus

$$\begin{aligned} \frac{1}{4} \|\nabla \tilde{J}_b(\theta)\|_2^2 &= \frac{1}{b^2} \|(f - \Psi\theta)^\top \Psi\|_2^2 = \frac{1}{b^2} (f - \Psi\theta)^\top \Psi \Psi^\top (f - \Psi\theta) \\ &= \frac{1}{b^2} (\theta^\top \Psi^\top \Psi \Psi^\top \Psi \theta - 2f^\top \Psi \Psi^\top \Psi \theta + f^\top \Psi \Psi^\top f) \\ &= \theta^\top M_b^2 \theta - 2a_b^\top M_b \theta + a_b^\top a_b \end{aligned}$$

where $M_b = \frac{1}{b} \Psi^\top \Psi$ satisfies $\mathbb{E}[M_b] = M$. Similarly,

$$\frac{1}{4} \|\nabla \tilde{J}(\theta)\|_2^2 = \theta^\top M^2 \theta - 2a^\top M \theta + a^\top a.$$

Since (10) is equivalent to non-negativity of

$$\begin{aligned} &\frac{1}{4} \left(C_1 + C_2 \|\nabla \tilde{J}(\theta)\|_2^2 - \mathbb{E}[\|\nabla \tilde{J}_b(\theta)\|_2^2] \right) \\ &= \theta^\top (C_2 M^2 - \mathbb{E}[M_b^2]) \theta - 2(C_2 a^\top M - a_b^\top M_b) \theta + \left(\frac{1}{4} C_1 + C_2 a^\top a - a_b^\top a_b \right), \quad (11) \end{aligned}$$

using Lemma 1 we choose to set $C_2 = \sigma_{\max}(\mathbb{E}[M_b^2]) / \sigma_{\min}(M^2)$ to ensure that the matrix $C_2 M^2 - \mathbb{E}[M_b^2]$ is semi-positive definite. Given this choice of C_2 , it is possible to take C_1 large enough to guarantee that the expression in Equation 11 is non-negative $\forall \theta \in \mathbb{R}^{p+1}$. This verifies (ii). From Theorem 4.7 in [10], the result follows under the stated assumptions on the learning rate α_t . \square

D Numerical Experiments

Here we discuss further implementation details for each of the examples in the paper, and also provide additional numerical experiments to complement the results in the main text.

For all experiments in this paper, the specific parametric forms of CVs that were considered were as follows:

1. The second order polynomial class was used for polynomial CVs, i.e.,

$$\Phi_\theta(x) = \frac{1}{2}x^\top Ax + b^\top x,$$

where $A \in \mathbb{R}^{d \times d}$ is a symmetric matrix, $b \in \mathbb{R}^d$, and $\theta = (A, b)$.

2. For the kernel CVs, the kernel $k_0(x, x')$ in Eq. 3 was used, and we followed [43] in taking

$$k(x, x') = (1 + \alpha_1 \|x\|_2^2 + \alpha_1 \|x'\|_2^2)^{-1} \exp\left\{-(2\alpha_2^2)^{-1} \|x - x'\|_2^2\right\} \quad (12)$$

for hyper-parameters $\alpha_1, \alpha_2 > 0$ to be specified.

3. The NN CVs were fully connected layers. For the Gaussian process realization experiment, the NN had 2 layers and each layer had 50 hidden nodes. For other experiments, the NN had 6 layers and each layer had 20 hidden nodes. The ReLU activation function was used for all neurons except the output neuron, where the identity function was employed.
4. The ensemble CV of polynomial and kernel was the sum of a degree 2 polynomial CV and a kernel interpolant CV. In the case of multiple kernels, the same base kernel was used, but with different choices of hyperparameters.

The hyper-parameters α_1 and α_2 in the kernel and ensemble CVs were selected via 5-fold cross-validation. In all experiments, the reported computing timings do not include the hyper-parameter tuning time. This is still fair to compare various CVs because all CVs and training methods—SGD and exact solution—necessitate tuning hyper-parameters.

The remainder of this section is devoted to reporting details of the experiments that were reported in the main text. In Section D.1 we describe the illustrative experiment from the main text and also provide additional experiments, not reported in the main text. Section D.2 contains details for the Genz test function experiment and reports additional results, not contained in the main text. Section D.3 contains details for the Gaussian Process experiment. In Section D.4 we report an additional experiment that considers posterior inference for a model of atmospheric pollution, not contained in the main text. In Section D.5 we present the ensemble CV of two kernels and a polynomial used in the last two experiments of this paper: ordinary differential equations and *sonar* dataset.

D.1 Numerical Integration of Polynomials

We start by comparing a range of CVs on polynomial integrands which are integrated against a Gaussian distribution. This is a good benchmark problem since the integrals can be computed in closed-form, and the performance of each method can hence be studied precisely.

Implementation Details

Consider an integrand which is a sum of p polynomials:

$$f(x) = \sum_{j=0}^p \prod_{i=1}^d \alpha_{ji} x_i^{\beta_{ji}}$$

where $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, $\alpha \in \mathbb{R}^{p \times d}$ & $\beta \in \mathbb{N}^{p \times d}$ (for both matrices, rows correspond to a polynomial, and each column to a dimension of the space). We can easily compute the integral of such a polynomial against a Gaussian distribution $\mathcal{N}(0, \Sigma)$ where $\Sigma = \sigma^2 I_{d \times d}$ using well-known Gaussian identities. In particular, denoting π the pdf of this Gaussian distribution, we have:

$$\Pi[f] = \int_{\mathbb{R}^d} f(x) \pi(x) dx = \sum_{j=0}^p \prod_{i=1}^d \alpha_{ji} \delta_{\{\beta_{ji} \in \{0, 2, 4, \dots\}\}} \sigma^{\beta_{ji}} (\beta_{ji} - 1)!!$$

where $\delta_{\{\beta_{ji} \in \{0, 2, 4, \dots\}\}}$ is an indicator function taking value 1 when β_{ji} is pair and 0 otherwise. Also, $x!!$ denotes the double factorial (also called semi-factorial), which is the product of all integers from 1 to x that have the same parity (odd or even) as x . We can therefore use integration of polynomials against a Gaussian distribution as a test-bed for various MC or CV methods.

In this case, we have that $\nabla_x \log \pi(x) = -x^2/\sigma^2$, and so $\mathcal{L}_{\text{SL}} u = \Delta u + \nabla u \cdot \nabla \log \pi$ will itself also be a polynomial whenever u is a polynomial.

Additional Experiments

We start by integrating $f(x) = \sum_{j=1}^d (1 - x_j)$ against a standard Gaussian: $\mathcal{N}(0, I_{d \times d})$. This integrand is particularly well suited for the polynomial-based CVs of [40, 47] since, in this case, the integrand is itself in the class of functions of the form $\mathcal{L}_{\text{SL}} u_\theta$. We use $n = 10^4$ design points obtained by sampling IID from a $\mathcal{N}(0, I_{d \times d})$. A 90/10 split is used for approximation data and MC data. We compare the polynomial-based CVs of degree two trained by solving the least-squares through a linear system, and these same CVs trained by SGD. The experiments are presented in Fig. 1 (top row).

The performance of the SGD CVs is usually worse initially, but approaches that of the linear system solution as t grows. This holds regardless of the initialization of SGD (see the blue lines). Such results are not surprising since the objective function is convex. In low dimensions, the computational cost associated with solving the linear system is low and there is hence not much point to using SGD. The main advantage of the SGD approach can be observed for large d , in which case we obtain a performance close, if not equal, to that of solving the linear system, but at a small fraction of the cost. This advantage of SGD also increases with d . Note that early stopping of SGD could provide good performance with a further reduction of the computational cost.

We also provide additional experiments using kernel-based CVs in Fig. 5 (bottom row). Similar conclusions can be made from these plots. Firstly, in low dimension, there is not much point using the SGD approach over solving the exact least squares problem, but as the dimensionality of the problem grows, the SGD methodology can provide close-to-optimal performance at a fraction of the cost. Secondly, we once again have that early stopping of the SGD procedure could provide further significant computational gains.

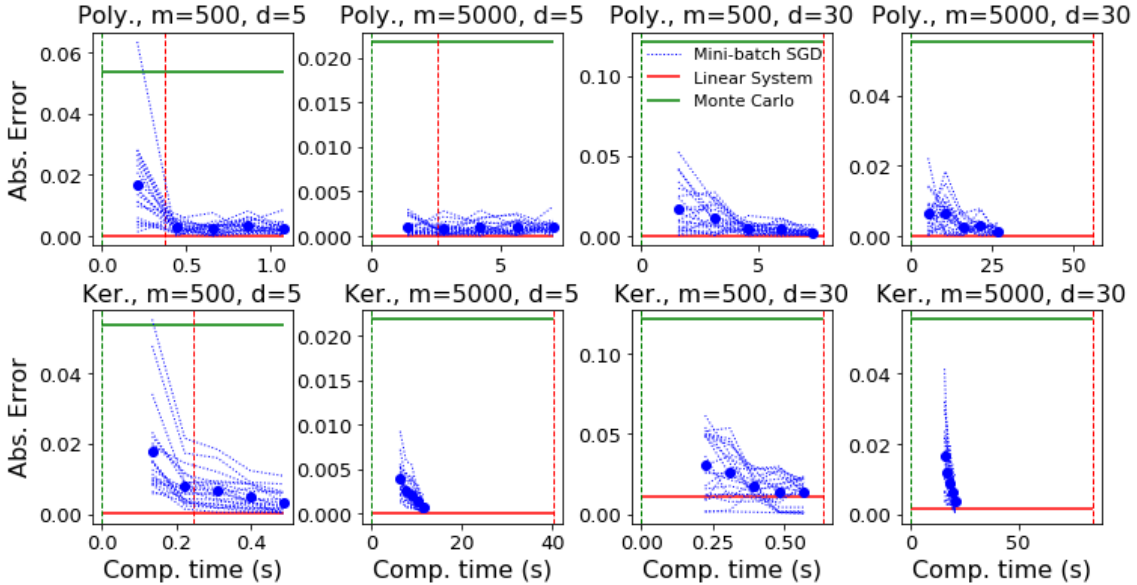


Figure 5: Performance of the polynomial-based CVs (top row) and kernel-based CVs (bottom row) on polynomial integrands. We compare CVs obtained by solving linear systems with the CVs trained using SGD. The y -axis gives the mean absolute error over 20 different datasets, whilst the dotted line gives the cost of solving the linear system. Here 20 realizations (blue dashed lines) are shown and blue dots represent the mean absolute error. The red lines represent the performance and computational cost of solving the corresponding linear system exactly, our benchmark. Similarly, the green lines represent the Monte Carlo estimator with no control variate used.

D.2 Genz Test Functions

A popular set of synthetic problems for numerical integration are the Genz test functions introduced in [21]. These functions, which can all be integrated analytically, were selected to test several difficult scenarios for

numerical integration tools based on functional approximation, such as sharp peaks and discontinuities. They are usually defined on $[0, 1]^d$, but can easily be transformed to be defined on the whole of \mathbb{R}^d , as we discuss next.

Implementation Details

We consider such a transformation here to keep the setting as close to possible to that of the polynomials. Let $h : [0, 1]^d \rightarrow \mathbb{R}$ be such a test function. Then, using a change of variables, we get:

$$\int_{[0,1]^d} h(y)dy = \int_{\mathbb{R}^d} h(\Phi(x))\phi(x)dx$$

where $\Phi(x)$ is a d -dimensional vector given by $\Phi(x) = (\Phi(x_1), \dots, \Phi(x_d))$ where Φ is the cumulative distribution function of a standard Gaussian distribution and ϕ is the corresponding probability density function. We therefore have integration problems of the form $\Pi[f] = \int_{\mathbb{R}^d} f(x)\pi(x)dx$, where $f(x) = h(\Phi(x))$, Π is a standard Gaussian, and h is any of the classical Genz functions [21], as described in the Table 2 below. See <https://www.sfu.ca/~ssurjano/integration.html> for implementations of these functions in R or MATLAB.

Genz Function	Integrand	Integral
Continuous	$\exp\left(-\sum_{i=1}^d a_i x_i - u_i \right)$	$\prod_{i=1}^d a_i^{-1} (2 - \exp(a_i(u_i - 1)) - \exp(-a_i u_i))$
Corner Peak	$\left(1 + \sum_{i=1}^d a_i x_i\right)^{-d-1}$	$\sum_{k=0}^d \sum_{\substack{I \subseteq \{1, \dots, d\}, \\ I =k}} (-1)^{k+d} \left(1 + \sum_{i=1}^d a_i - \sum_{j \in I} a_j\right)^{-1} \left(d! \prod_{i=1}^d a_i\right)^{-1}$
Discontinuous	$\begin{cases} 0, & \text{if } x_i > u_i \text{ for any } i \\ \exp\left(\sum_{i=1}^d a_i x_i\right), & \text{else} \end{cases}$	$\prod_{i=1}^d a_i^{-1} (\exp(a_i \min(1, u_i)) - 1)$
Gaussian Peak	$\exp\left(-\sum_{i=1}^d a_i^2 (x_i - u_i)^2\right)$	$\left(\frac{\sqrt{\pi}}{2}\right)^d \left(\prod_{i=1}^d a_i^{-1}\right) \times \left(\prod_{i=1}^d \text{Erf}(a_i(1 - u_i)) - \text{Erf}(-a_i u_i)\right)$
Oscillatory	$\cos\left(2\pi u_1 + \sum_{i=1}^d a_i x_i\right)$	$\sum_{k=0}^d \sum_{\substack{I \subseteq \{1, \dots, d\}, \\ I =k}} \frac{(-1)^k}{\prod_{i=1}^d a_i} g\left(2\pi u_1 + \sum_{i=1}^d a_i - \sum_{j \in I} a_j\right)$ where $g(x) = \begin{cases} \sin(x) & \text{if } \text{mod}(d, 4) = 1 \\ -\cos(x) & \text{if } \text{mod}(d, 4) = 2 \\ -\sin(x) & \text{if } \text{mod}(d, 4) = 3 \\ \cos(x) & \text{if } \text{mod}(d, 4) = 0 \end{cases}$
Product Peak	$\prod_{i=1}^d (a_i^{-2} + (x_i - u_i)^2)^{-1}$	$\prod_{i=1}^d a_i (\arctan((1 - u_i)a_i) - \arctan(-u_i a_i))$

Table 2: *Genz Test Functions*: This table contains 6 test functions defined on $[0, 1]^d$, as well as their corresponding integrals against a uniform distribution. The parameter vectors $a = (a_1, \dots, a_d) \in \mathbb{R}_{>0}^d$ and $u = (u_1, \dots, u_d) \in [0, 1]^d$ can be changed to adapt the difficulty of the integration problem. Their default values are $a = (5, \dots, 5)$ and $u = (0.5, \dots, 0.5)$.

Additional Experiments

The main numerical results are presented in the main text, but we now highlight additional results.

Firstly, results (in $d = 1$) are provided in Table 3. These results focus on kernel-based CVs trained with the least-squares objective either by solving the linear system (as per [43]), or with SGD with either 2 or 5 epochs. We split the data and assign 50% to constructing the CV and 50% for the estimator. In all experiments, the CVs provide significant improvement over a MC estimator. Overall, the linear system approach tends to outperform SGD, but SGD can obtain significant variance reduction at a fraction of the computational cost. Further results are presented in Table 4 in Appendix D.2, which demonstrates that the same conclusion holds for higher-dimensional integrands ($d = 5, 10, 15, 20$), and that the 50/50 split may be suboptimal. Indeed, it is found that assigning a greater proportion of the data on the construction of the CV may be preferable, but that this will generally increase computational cost.

Integrand	MC	Linear Sys.	SGD 2 Epoc.	SGD 5 Epoc.
Continuous	2.77e-03	3.04e-04	3.45e-04	3.28e-04
Corner Peak	5.76e-03	7.07e-06	1.69e-05	9.27e-06
Discontinuous	2.04e-02	2.39e-03	6.30e-03	3.91e-03
Gaussian Peak	1.47e-03	8.84e-06	1.10e-04	1.24e-05
Oscillatory	4.17e-03	3.68e-06	1.22e-05	4.63e-06
Product Peak	1.37e-03	1.79e-05	1.48e-04	2.12e-05
Time (sec.)	7.10e-02	5.68e-01	1.90e-01	4.50e-01

Table 3: *Performance of kernel CVs for the Genz Functions.* We take $n = 1000, m = 500$. The time presented is an average over repetitions for all six functions (the difference across integrand was negligible).

Secondly, Table 4 provides additional experiments in the case of the Genz peak function. The table demonstrates that the observation that SGD can provide results close to those of LS at a fraction of the price is still true regardless of the dimension.

Dim.	MC	Linear Sys.	SGD 2 Epoc.	SGD 5 Epoc.
5	2.85e-03	5.81e-04	6.51e-04	5.84e-04
10	2.29e-03	1.98e-04	2.79e-04	2.70e-04
15	2.10e-03	4.93e-04	1.22e-03	6.14e-03
20	1.73e-03	5.13e-04	6.35e-04	5.90e-04
Time (secs.)	7.00e-02	7.60e-01	3.30e-01	5.95e-01

Table 4: The mean absolute errors and computing times of kernel CVs on Genz product peak function (with parameters $a = (1.0, \dots, 1.0)$ and $u = (0.5, \dots, 0.5)$ of the same dimension as the integrand) of four dimensions: 5, 10, 15 and 20. The total sample size is 1000.

Secondly, in Table 5, we provide a comparison of the kernel-based CVs for different splits of the data (for solving Stein’s equation and MC estimation). We consider four cases: a 50/50 split (i.e. 50% of the data is used for solving Stein’s equation, and 50% for MC estimation), a 70/30 split, a 90/10 split and a 100/0 split.

The computational cost and mean absolute error (MAE) both depend on the number of data points allocated to each task. The larger we make the proportion of data points allocated to solving Stein’s equation, the more expensive the estimator becomes, but this usually comes with an increase in accuracy. Assuming that the number of data points is fixed to n , the user is therefore able to choose this split according to the computational power available.

D.3 Integrating Gaussian Processes

We are integrating realizations of a Gaussian process (GP) [54] with mean function $m(x) = 0$ and kernel function

$$c(x, y; \lambda, \sigma) = \lambda^2 \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right) = \lambda^2 (2\pi\sigma^2)^{\frac{d}{2}} \phi(x|y, \sigma^2, I_{d \times d}).$$

The integral is taken with respect to a mixture of Gaussian distributions with probability density function: $\pi(x) = \sum_{i=1}^L \rho_i \phi(x|\mu_i, \Sigma_i)$, where $\rho = (\rho_1, \dots, \rho_L) \in [0, 1]^d$ is a vector of mixture weights satisfying

Integrand	50/50	70/30	90/10	100/0
Continuous	3.28e-04	4.82e-04	7.40e-04	3.80e-04
Corner Peak	9.27e-06	1.57e-05	3.26e-05	1.02e-05
Discontinuous	3.91e-03	7.29e-03	3.2e-03	3.51e-03
Gaussian Peak	1.24e-05	2.18e-05	2.78e-05	2.05e-04
Oscillatory	4.63e-06	8.47e-06	1.67e-05	3.25e-05
Product Peak	2.12e-05	2.20e-05	3.03e-05	3.63e-05
Time (secs.)	4.70e-01	5.79e-01	6.89e-01	7.40e-01

Table 5: The mean absolute errors of kernel CV methods, as a function of the train/test data split. The sample size fixed at 1000, and four train/test splits are used: 50/50, 70/30, 90/10, and 100/0. The computing times for 5 epochs of mini-batch SGD training are shown in the bottom row.

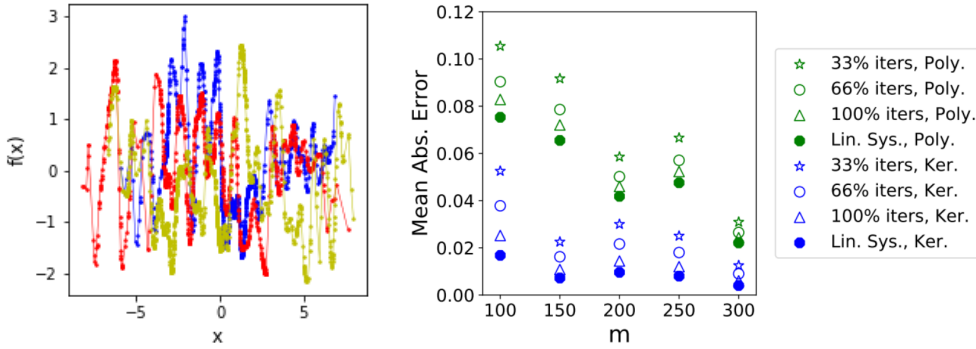


Figure 6: *Left*: Three realizations from a Gaussian Process. *Right*: Mean absolute errors of kernel and polynomial CVs evaluated after 33%, 66% and 100% SGD training.

$\sum_{l=1}^L \rho_l = 1$. For our problems the mean vectors μ_1, \dots, μ_L are generated at random from a zero-mean Gaussian distribution with covariance $3I_{d \times d}$, and random covariance matrices $\Sigma_1, \dots, \Sigma_L$ are obtained by taking a matrix A_l with entries uniformly random on $[0, 1)$, then setting the covariance $\Sigma_l = A_l^\top A_l$. The mixture weights are also simulated. We simulate the unweighted mixture weights from a uniform distribution between 0 and 1, and then we normalize them to have mixture weights. For this mixture of Gaussians, the score function is given by:

$$\nabla_x \log \pi(x) = \frac{\sum_{l=1}^L \rho_l \nabla_x \phi(x|\mu_l, \Sigma_l)}{\sum_{l=1}^L \rho_l \phi(x|\mu_l, \Sigma_l)} = \frac{\sum_{l=1}^L \rho_l \phi(x|\mu_l, \Sigma_l) \Sigma_l^{-1} (x - \mu_l)}{\sum_{l=1}^L \rho_l \phi(x|\mu_l, \Sigma_l)}.$$

We note that the integral of the mean function is $\Pi[m] = 0$, and the integrated covariance function is given by:

$$\Pi[c(x, \cdot)] = \lambda^2 (\sqrt{2\pi}\sigma)^d \sum_{l=1}^L \rho_l \phi(x|\mu_l, \Sigma_l + \sigma^2 I_{d \times d}).$$

Finally, the integral of the covariance function with respect to the both variables is:

$$\Pi\Pi[c] = \lambda^2 (\sqrt{2\pi}\sigma)^d \sum_{l,m=1}^L \rho_l \rho_m \phi(\mu_l|\mu_m, \Sigma_l + \Sigma_m + \sigma^2 I_{d \times d}).$$

These identities allow us to easily simulate a draw from a Gaussian process and its integral jointly. Indeed, under a Gaussian process model, the vector $(f(x_1), f(x_2), \dots, f(x_n), \Pi[f])$ is jointly distributed as a multivariate Gaussian distribution with mean $(m(x_1), m(x_2), \dots, m(x_n), \Pi[m])$ and covariance:

$$\begin{bmatrix} c(x_1, x_1) & c(x_1, x_2) & \dots & c(x_1, x_n) & \Pi[c(x_1, x)] \\ c(x_2, x_1) & c(x_2, x_2) & \dots & c(x_2, x_n) & \Pi[c(x_2, x)] \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c(x_n, x_1) & c(x_n, x_2) & \dots & c(x_n, x_n) & \Pi[c(x_n, x)] \\ \Pi[c(x, x_1)] & \Pi[c(x, x_2)] & \dots & \Pi[c(x, x_n)] & \Pi\Pi[c] \end{bmatrix}.$$

This procedure therefore allows us to create a wide range of examples of varying complexity, by changing the dimension d of the domain, the number L of mixture components and the parameters λ and σ of the GP covariance function.

In this experiment, all of the samples x_i are included in the training dataset (i.e. $m = n$). We investigate the performance of two objective functions: \tilde{J}_m^V and \tilde{J}_m^{LS} . The main results of this experiment are displayed in Fig. 2. When implementing SGD algorithm, the batch size is 8 and the number of training epochs is 10 for kernel CV, and 25 for other CVs.

D.4 Posterior Inference for a Model of Atmospheric Pollutant Detection

We start with the problem of computing posterior expectations for some Bayesian inference problem linked to the LIDAR (light detection and ranging) experiment, which considers the reflection of light emitted by some laser to detect pollutants in the atmosphere [58]. This dataset consists of $m = 221$ observations of the distance travelled before the light is reflected (denoted $\{l_i\}_{i=1}^m$), and of the log-ratios of received light from two laser sources (denoted $\{r_i\}_{i=1}^m$).

Following [16], we consider regression with a mean-zero GP model: $r_j = g(l_j) + \epsilon_j$, where $\epsilon_j \sim \mathcal{N}(0, \alpha)$, $\alpha = 0.04$. The kernel $c(l, l') = \lambda_1^2 \exp(-\lambda_2^2 \|l - l'\|_2^2 / 2)$ was parameterized with $\phi_1 = \log \lambda_1$ and $\phi_2 = \log \lambda_2$, and a Cauchy prior was placed on $\phi = (\phi_1, \phi_2)$. Denote by Π the posterior measure over ϕ given the observed data.

We are interested in computing posterior moments $\Pi[\phi_1]$, $\Pi[\phi_2]$, $\Pi[\phi_1^2]$ and $\Pi[\phi_2^2]$, as well as the marginal log-likelihood $\Pi[p(r|l, \phi)]$. The posterior marginal likelihood of the data is defined as the integral of the likelihood $p(y|X, \phi)$ with respect to the posterior on the parameters. This likelihood can be expressed in log-form as below:

$$\log p(y|X, \phi) = -\frac{1}{2} y^\top (C_\phi + \alpha I_{n \times n})^{-1} y - \frac{1}{2} \log |C_\phi + \alpha I_{n \times n}| - \frac{n}{2} \log 2\pi,$$

where C_ϕ denotes the $n \times n$ matrix with entries $(C_\phi)_{ij} = c(x_i, x_j)$, where we have made the dependence explicit on the hyper-parameters ϕ of the covariance function.

To do so, we use an adaptive MCMC algorithm to sample from Π . Results are presented in Fig. 7 for CVs based on polynomials, kernels and NNs all trained with SGD. When implementing SGD algorithm, the batch size is 8 and the number of training epochs is 10 for kernel CV, and 25 for other CVs. We notice that the performance can vary significantly based on the Stein space and operator. This clearly demonstrates the advantages of our general methods which can be easily adapted to the problem at hand.

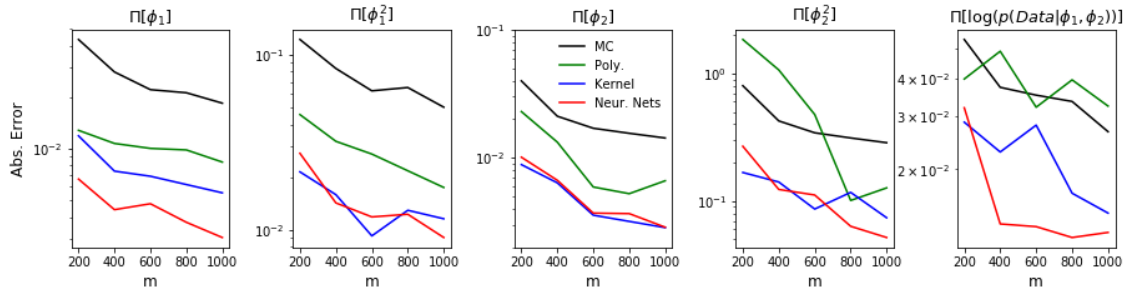


Figure 7: *Parameter Inference for Model of Atmospheric Pollutants.* We compute the first two posterior moments of the kernel parameters, as well as the marginal likelihood.

D.5 Parameter Inference for Ordinary Differential Equations

In this experiment and the *Sonar Dataset* experiment, we employ an ensemble CV of multiple kernels with a polynomial. Similar to ensemble CV of a kernel and a polynomial, the ensemble CV of multiple kernels and a polynomial employs the sum of a polynomial CV and two kernel interpolant. We employ two kernels and a polynomial. The two kernels are of the same form as $k_0(x, x')$ in Eq. 3, but they have different

hyper-parameters. When implementing this ensemble CV, the kernel in Eq. 12 is used and it is defined by two hyper-parameters α_1 and α_2 . For two kernels, we chose values of α_2 by the median heuristic, setting one to

$$\ell = \sqrt{\frac{1}{2} \text{Median}\{\|x_i - x_j\|_2^2 : 1 \leq i < j \leq m\}},$$

and the other kernel to $\sqrt{2}\ell$. The other hyper-parameter α_1 is the same for two kernels, and it is tuned via 5-fold cross validation over a grid $\{1.0e+6, 1.0e+5, 1.0e+4, 1.0e+3, 1.0e+2, 1.0e+1, 1.0, 1.0e-1, 1.0e-2\}$. For this ensemble CV of multiple kernels and a polynomial, we do not derive the exact solution of parameters, but rather we implement our SGD framework to learn this ensemble CV on a training dataset.