Bayesian Analysis (2023) **18**, Number 3, pp. 865–888

Regularized Zero-Variance Control Variates*

L. F. South[†,‡], C. J. Oates[§], A. Mira[¶], and C. Drovandi[‖]

Abstract. Zero-variance control variates (ZV-CV) are a post-processing method to reduce the variance of Monte Carlo estimators of expectations using the derivatives of the log target. Once the derivatives are available, the only additional computational effort lies in solving a linear regression problem. Significant variance reductions have been achieved with this method in low dimensional examples, but the number of covariates in the regression rapidly increases with the dimension of the target. In this paper, we present compelling empirical evidence that the use of penalized regression techniques in the selection of high-dimensional control variates provides performance gains over the classical least squares method. Another type of regularization based on using subsets of derivatives, or *a priori* regularization as we refer to it in this paper, is also proposed to reduce computational and storage requirements. Several examples showing the utility and limitations of regularized ZV-CV for Bayesian inference are given. The methods proposed in this paper are accessible through the R package ZVCV.

MSC2020 subject classifications: Primary 62-08; secondary 62F15.

Keywords: Bayesian inference, controlled thermodynamic integration – CTI, curse of dimensionality, Markov Chain Monte Carlo simulation, MCMC, Monte Carlo simulations, penalized regression, sequential Monte Carlo – SMC, Stein operator, variance reduction.

1 Introduction

Our focus in this paper is on calculating the expectation of a square integrable function $\varphi(\boldsymbol{\theta})$ with respect to a distribution with (Lebesgue) density $p(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$. Given independent and identically distributed (iid) samples $\{\boldsymbol{\theta}_i\}_{i=1}^N \overset{\text{iid}}{\sim} p(\boldsymbol{\theta})$, the standard

arXiv: 1811.05073

*LFS was supported by an Australian Research Training Program Stipend, by ACEMS and by the Engineering and Physical Sciences Research Council grant EP/S00159X/1. CJO was supported by the Lloyd's Register Foundation programme on data centric engineering at the Alan Turing Institute, UK. CD and CJO were supported by an Australian Research Council Discovery Project (DP200102101). AM was partially supported by the Swiss National Science Foundation grant 100018_200557.

[†]School of Mathematical Sciences, Queensland University of Technology, Australia, ll.south@qut.edu.au
ARC Centre of Excellence for Mathematical & Statistical Frontiers (ACEMS)
[‡]Department of Mathematics and Statistics, Lancaster University, United Kingdom
[§]School of Mathematics, Statistics and Physics, Newcastle University, UK
Alan Turing Institute, UK
[¶]Faculty of Economics, Università della Svizzera italiana, Switzerland
University of Insubria, Italy
[‖]School of Mathematical Sciences, Queensland University of Technology, Australia
ARC Centre of Excellence for Mathematical & Statistical Frontiers (ACEMS)

 https://doi.org/10.1214/22-BA1328

Monte Carlo estimator,

$$\widehat{\mathbb{E}_p[\varphi(\boldsymbol{\theta})]} = \frac{1}{N}\sum_{i=1}^{N}\varphi(\boldsymbol{\theta}_i), \tag{1.1}$$

is an unbiased estimator of $\mathbb{E}_p[\varphi(\boldsymbol{\theta})] = \int_\Theta \varphi(\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ and its variance is $\mathcal{O}(1/N)$. Reducing the variance of this estimator by increasing N is often infeasible due to the cost of sampling from $p(\boldsymbol{\theta})$ and potentially the cost of evaluating $\varphi(\boldsymbol{\theta})$. If the samples are not iid then the functional form of the estimator is the same and the methods described in this work still apply.

Recent control variate methods have focused on reducing the variance of (1.1) using the derivatives of the log target, $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$, or some unbiased estimator of this quantity. Zero-variance control variates (ZV-CV) (Assaraf and Caffarel, 1999; Mira et al., 2013) and control functionals (CF) (Oates et al., 2017) are two such methods. ZV-CV amounts to solving a linear regression problem and CF is a non-parametric alternative. These methods can be used as post-processing procedures after N samples, not necessarily iid, from p have been produced along with evaluations of $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\theta})$ for each of the samples. Often $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$ is already available because derivative-based methods like Metropolis adjusted Langevin algorithm (MALA) (Roberts and Stramer, 2002; Girolami and Calderhead, 2011) or Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Girolami and Calderhead, 2011) have been used in the sampling algorithm.

The parametric approximation in ZV-CV is based on a polynomial in $\boldsymbol{\theta}$, so the number of coefficients to estimate rapidly increases both with the polynomial order and with d. As a result of restricting to a low polynomial order, ZV-CV tends to offer less substantial improvements than CF for challenging low-dimensional applications. This is not surprising given the good statistical properties of CF which have been described in Oates et al. (2019); Barp et al. (2022). However, CF has an $\mathcal{O}(N^3)$ computational cost, compared to ZV-CV with has computational cost of $\mathcal{O}(N)$, and it also suffers from the curse of dimensionality with respect to d due to the use of non-parametric methods. Some results in Oates et al. (2017), shown mainly in the appendices, suggest that the performance of CF compared to ZV-CV may deteriorate in higher dimensions.

One aim of this work is to develop derivative-based control variate methods which are inexpensive, effective and capable of handling higher dimensions than existing derivative-based methods. The novel methods that we introduce are referred to as *regularized* ZV-CV and they are based on two types of regularization: penalization methods for linear regression and what we refer to as *a priori* regularization. Penalized ZV-CV allows higher order polynomials to be used than could be considered with ordinary least squares. This method is motivated by showing that $\mathcal{L}_2$ penalized ZV-CV is equivalent to CF with a second-order differential operator and finite-dimensional polynomial kernel. *A priori* ZV-CV is most beneficial when $N < d$. Empirical results in Section 4 suggest that significant variance reductions can be achieved with *a priori* ZV-CV when $N < d$ or with penalized ZV-CV when the polynomial order is pushed beyond the limits of what standard ZV-CV can handle. We have developed an R package, ZVCV (South, 2018), which implements several derivative-based variance reduction techniques including standard ZV-CV, CF and regularized ZV-CV.

An important application area for ZV-CV and regularized ZV-CV is Bayesian inference, where Monte Carlo integration is commonly used. The use of ZV-CV and CF to improve posterior expectations based on samples from Markov chain Monte Carlo (MCMC) (Metropolis et al., 1953) is well established (see e.g. Mira et al. (2013); Papamarkou et al. (2014); Oates et al. (2017)). ZV-CV and CF have also been applied to the power posterior (Friel and Pettitt, 2008) estimator of the normalizing constant in an MCMC setting by Oates et al. (2016) and Oates et al. (2017), where they refer to this method as controlled thermodynamic integration (CTI). In this paper we go beyond existing literature and describe how regularized ZV-CV fits naturally into the context of sequential Monte Carlo (SMC) samplers (Del Moral et al., 2006; Chopin, 2002). In doing so, we provide a setting where adaptive methods can easily be applied to the CTI estimator. A novel reduced-variance normalizing constant estimator using the standard SMC identity is also proposed.

An introduction to derivative-based Monte Carlo variance reduction methods is provided in Section 2. The main methodological contributions in terms of developing regularized ZV-CV methods can be found in Section 3. Section 4 contains a simulation study comparing methods and estimators on the novel application to SMC. A final discussion of limitations and possible future work is given in Section 5.

2 Control Variates Based on Stein Operators

In this section, we recall previous work on control variate methods. The classical framework for control variates (Ripley, 1987; Hammersley and Handscomb, 1964) is to determine an auxiliary function $\tilde{\varphi}(\boldsymbol{\theta}) = \varphi(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$ such that $\mathbb{E}_p[\tilde{\varphi}(\boldsymbol{\theta})] = \mathbb{E}_p[\varphi(\boldsymbol{\theta})]$ and $\mathbb{V}_p[\tilde{\varphi}(\boldsymbol{\theta})] < \mathbb{V}_p[\varphi(\boldsymbol{\theta})]$, where $\mathbb{V}_p$ denotes the variance with respect to $p(\boldsymbol{\theta})$. Estimator (1.1) can then be replaced with the unbiased, reduced variance estimator,

$$\widehat{\mathbb{E}_p[\varphi(\boldsymbol{\theta})]} = \frac{1}{N}\sum_{i=1}^{N}\left[\varphi(\boldsymbol{\theta}_i) + h(\boldsymbol{\theta}_i)\right]. \tag{2.1}$$

A control variate which has been considered in (Assaraf and Caffarel, 1999; Mira et al., 2013; Barp et al., 2022) is

$$\begin{aligned} h_g(\boldsymbol{\theta}) = \mathcal{L}g(\boldsymbol{\theta}) &= \frac{\nabla_{\boldsymbol{\theta}} \cdot (p(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}))}{p(\boldsymbol{\theta})} \\ &= \Delta_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}), \end{aligned} \tag{2.2}$$

where $\mathcal{L}$ is a second-order Langevin Stein operator (Stein, 1972; Gorham and Mackey, 2015) depending on p, $\Delta_{\boldsymbol{\theta}}$ is the Laplacian operator represented in coordinates as $\sum_{j=1}^{d} \nabla^2_{\boldsymbol{\theta}[j]}$ on $\mathbb{R}^d$, $\cdot$ is the dot product operator such that $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^\top \mathbf{b}$ and $g : \Theta \to \mathbb{R}$ is a twice continuously differentiable function to be specified.

Langevin Stein operators are helpful in generating control variates for two reasons. Firstly, they can be applied in Bayesian inference because they do not require the

normalizing constant of p. Furthermore, by definition a Stein operator $\mathcal{L}$ depending on $p(\boldsymbol{\theta})$ satisfies $\mathbb{E}_p[\mathcal{L}g(\boldsymbol{\theta})] = 0$ for all functions $g(\boldsymbol{\theta})$ in a set called a Stein function class (see Section 2.2) and therefore $\mathbb{E}_p[\tilde{\varphi}(\boldsymbol{\theta})] = \mathbb{E}_p[\varphi(\boldsymbol{\theta})]$ under mild conditions. Typically it is also a requirement that Stein operators $\mathcal{L}$ fully characterize p but this is not necessary for Stein-based control variates.

What remains is to choose g. The function g for which $\tilde{\varphi}(\boldsymbol{\theta})$ is constant, and thus zero variance is achieved, is generally intractable. In practice, g is restricted to some function class $\mathcal{G}$ and is estimated based on samples targeting p.

2.1 Choice of Function g

Variance reduction is effected through judicious choice of g. Once a function class $\mathcal{G}$ has been selected, the function $g \in \mathcal{G}$ is estimated by performing a regression task. As described in Barp et al. (2022), a generalization of several existing regression methods for this problem is

$$(\hat{c}, \hat{g}) \in \underset{\substack{c\in\mathbb{R}\\ g\in\mathcal{G}}}{\arg\min} \frac{1}{N}\sum_{i=1}^{N} [\varphi(\boldsymbol{\theta}_i) - c + \mathcal{L}g(\boldsymbol{\theta}_i)]^2 + \lambda \mathrm{PEN}(g), \tag{2.3}$$

where $\mathrm{PEN}(g)$ is a penalty function to be specified and $\lambda \geq 0$. This amounts to a penalized least squares approach to estimating $\varphi(\boldsymbol{\theta})$ using the functional form $c - \mathcal{L}g(\boldsymbol{\theta})$. This perspective on the optimization problem encompasses ZV-CV, CF and neural control variates (Zhu et al., 2019) as special cases. Further details on how ZV-CV and CF fit into this framework are given below. The main developments in this paper are based on considering alternative penalty functions.

Two recent contributions in control variates have optimization functions which do not fit into this framework, though the developments in penalty functions that are proposed in this paper could still be considered in these alternative frameworks. Belomestny et al. (2017) consider empirical variance minimization since minimizing a square error objective function may not be optimal. Brosse et al. (2019) consider an alternative optimization problem which is motivated by minimizing the asymptotic variance of a Langevin diffusion, which may be more suitable when samples have been obtained using MCMC with multivariate normal random walk or MALA proposals.

Control Functionals

CF (Oates et al., 2017; Barp et al., 2022) is based on choosing $\mathcal{G} \equiv \mathcal{H}$ where $\mathcal{H}$ is a user-specified Hilbert space of twice differentiable functionals on Θ. The penalty term $\mathrm{PEN}(g)$ considered in CF is $\mathrm{PEN}(g) = \|g\|^2_{\mathcal{H}}$, where $\|\cdot\|_{\mathcal{H}}$ is the norm associated with the Hilbert space $\mathcal{H}$. The existence of a solution pair $(\hat{c}, \hat{g}) \in \mathbb{R} \times \mathcal{H}$, together with an explicit algorithm for its computation, was obtained in that work under the assumption that the Hilbert space $\mathcal{H}$ admits a reproducing kernel (see Berlinet and Thomas-Agnan (2011) for background). This method leads to estimators with super-root-N convergence under conditions described in Oates et al. (2019) and Barp et al. (2022). However, the

cost associated with computation of $\hat{g}$ is $O(N^3)$, due to the need to invert a dense kernel matrix, and moreover this matrix is typically not well-conditioned. For applications that involve MCMC and SMC, typically N will be at least 10^3 and thus (in the absence of further approximations) the algorithm of Oates et al. (2017); Barp et al. (2022) can become impractical.

Zero-Variance Control Variates

ZV-CV (Assaraf and Caffarel, 1999; Mira et al., 2013) amounts to using $\mathcal{G}$ as the class of Qth order polynomial functions in $\boldsymbol{\theta}$, and $\lambda = 0$. The polynomials $P(\boldsymbol{\theta})$ that we consider have total degree $Q \in \mathbb{Z}_{\geq 0}$, meaning that the maximum sum of exponents is Q and the monomial basis is $\theta[1]^{\alpha_1} \cdots \theta[d]^{\alpha_d}$ where $\sum_{j=1}^{d} \alpha_j \leq Q$ and $\alpha \in \mathbb{Z}_{\geq 0}^{d}$. Substituting $g(\boldsymbol{\theta}) = P(\boldsymbol{\theta}) = \sum_{j=1}^{J} \beta_j P_j(\boldsymbol{\theta})$ into (2.2), where $P_j(\boldsymbol{\theta})$ is the jth monomial in the polynomial and $\boldsymbol{\beta} \in \mathbb{R}^J$ is the vector of polynomial coefficients, gives

$$\begin{aligned} h_P(\boldsymbol{\theta}) &= \mathcal{L}P(\boldsymbol{\theta}) \\ &= \sum_{j=1}^{J} \beta_j \mathcal{L}P_j(\boldsymbol{\theta}) \\ &= \boldsymbol{\beta}^\top \boldsymbol{x}(\boldsymbol{\theta}). \end{aligned}$$

The jth zerovariate (covariate in the regression), $x_j(\boldsymbol{\theta}) = \mathcal{L}P_j(\boldsymbol{\theta})$, is a term containing $\boldsymbol{\theta}$ and $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$. Its exact form is given in Appendix A of the Online Resources (South et al., 2022). For a Qth order polynomial when the dimension of $\boldsymbol{\theta}$ is d, the constant $J = \binom{d+Q}{d} - 1$ is the number of regression parameters, excluding the intercept which is in the null space of the Stein operator.

The standard approach in the literature for choosing $\boldsymbol{\beta}$ is to perform ordinary least squares (OLS) (Glasserman, 2003). This is equivalent to choosing $\lambda = 0$ in (2.3). The computational cost of ZV-CV is $\mathcal{O}(J^3 + NJ^2)$, which scales better with N than CF which has computational cost $\mathcal{O}(N^3)$, where often $J \ll N$. Unlike in CF, regularization methods have not previously been used in connection with ZV-CV.

Common practice is to default to $Q = 2$ in ZV-CV. Mira et al. (2013) consider $Q = 1$ to at most $Q = 3$ and find that $Q = 2$ is sufficient to achieve orders of magnitude variance reduction in most of their examples. Papamarkou et al. (2014) consider $Q \leq 2$, pointing out that "first and second degree polynomials suffice to attain considerable variance reduction." Low polynomial orders are also typically used in applications, for example Baker et al. (2019) use $Q = 1$ and Oates et al. (2016) use $Q \leq 2$. Oates et al. (2017) compare CF with ZV-CV using $Q = 2$ in most examples.

It has previously been proposed to increase the number of control variates as the sample size increases (see e.g. Portier and Segers (2019) and the appendices of Oates et al. (2017)). This approach can be motivated by the Stone-Weierstrass theorem (Stone, 1948), which states that polynomial functions can be used to uniformly approximate, to an arbitrary level of precision, continuous functions on closed intervals. However,

the increased number of coefficients in higher order polynomials may not be feasible or efficient to estimate with OLS.

We demonstrate in Section 4 that the common practice of defaulting to $Q = 2$ with OLS is often sub-optimal. Regularization approaches are proposed in Section 3 to enable higher degree polynomials to be employed whilst avoiding over-fitting of the regression model used to estimate the optimal coefficients of the polynomials.

2.2 Unbiasedness

Suppose that $g(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are sufficiently regular so that $\log p(\boldsymbol{\theta})$ has continuous first order derivatives and $g(\boldsymbol{\theta})$ has continuous first and second order derivatives. Also suppose that, if g is to be estimated, then the samples used in estimating $g(\boldsymbol{\theta})$ are independent of those used in evaluating (2.1). If $\Theta \neq \mathbb{R}^d$, then we require that Θ is compact and has piecewise smooth boundary $\partial\Theta$. Under these conditions, estimator (2.1) with samples $\{\boldsymbol{\theta}_i\}_{i=1}^N \overset{\text{iid}}{\sim} p(\boldsymbol{\theta})$ is unbiased if

$$\oint_{\partial\Theta} p(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) \cdot \boldsymbol{n}(\boldsymbol{\theta}) S(\mathrm{d}\boldsymbol{\theta}) = 0, \tag{2.4}$$

where $\oint_{\partial\Theta}$ is a surface integral over $\partial\Theta$, $\boldsymbol{n}(\boldsymbol{\theta})$ is the unit vector orthogonal to $\boldsymbol{\theta}$ at the boundary $\partial\Theta$ and $S(\mathrm{d}\boldsymbol{\theta})$ is the surface element at $\boldsymbol{\theta} \in \partial\Theta$. When $\Theta = \mathbb{R}^d$ is unbounded, condition (2.4) becomes a tail condition which is satisfied if $\int_{\Gamma_r} p(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) \cdot \boldsymbol{n}(\boldsymbol{\theta}) S(\mathrm{d}\boldsymbol{\theta}) \to 0$ as $r \to \infty$ where $\Gamma_r \in \mathbb{R}^d$ is a sphere centred at the origin with radius r and $\boldsymbol{n}$ is the unit vector orthogonal to $\boldsymbol{\theta}$ at Γ_r. This requirement is given in Equation 9 of Mira et al. (2013) and Assumption 2 of Oates et al. (2017) and it is a direct result of applying the divergence theorem to $\mathbb{E}_p[\mathcal{L} g(\boldsymbol{\theta})] = 0$. In the ZV-CV context for $\Theta = \mathbb{R}^d$, a sufficient condition for (2.4) is that the tails of p decay faster than polynomially (Appendix B of Oates et al., 2016).

The unbiased estimator which uses independent samples for estimation of $g(\boldsymbol{\theta})$ and evaluation of (2.1) is referred to as the "split" estimator. In practice, the so-called "combined" estimator which uses the full set of N samples for both estimation of $g(\boldsymbol{\theta})$ and evaluation of (2.1) can have lower mean square error than the split estimator but is no longer unbiased. If MCMC methods are employed then bias is unavoidable and the combined estimator is likely to be preferred.

2.3 Parameterization

An additional consideration when performing either ZV-CV or CF is the adopted parameterization. Any deterministic, invertible transformation of the random variables $\boldsymbol{\psi} = f(\boldsymbol{\theta})$ can be used so one can estimate

$$\widehat{\mathbb{E}_{p_{\boldsymbol{\theta}}}[\varphi(\boldsymbol{\theta})]} = \frac{1}{N} \sum_{i=1}^{N} \Big(\varphi(f^{-1}(\boldsymbol{\psi}_i)) + \Delta_{\boldsymbol{\psi}} g(\boldsymbol{\psi}_i) + \nabla_{\boldsymbol{\psi}} g(\boldsymbol{\psi}_i) \cdot \nabla_{\boldsymbol{\psi}} \log p_{\boldsymbol{\psi}}(\boldsymbol{\psi}_i) \Big), \tag{2.5}$$

instead of (2.1), where $p_{\boldsymbol{\theta}} \equiv p$ is the probability density function for $\boldsymbol{\theta}$, $p_{\boldsymbol{\psi}}$ is the probability density function for $\boldsymbol{\psi}$ obtained through a change of measure and $\{\boldsymbol{\psi}_i\}_{i=1}^N \sim p_{\boldsymbol{\psi}}$. For simplicity, the $\boldsymbol{\theta}$ parameterization is used in notation throughout the paper. The best parametrization to adopt for any given application is an open problem. If the original parameterization does not satisfy boundary condition (2.4), one could consider a reparameterization such that the boundary condition is satisfied.

3 Regularized Zero-Variance Control Variates

The aim of this section is to develop methods which are computationally less demanding than CF and offer improved statistical efficiency over standard ZV-CV. We describe two types of regularization: regularization through penalized regression and *a priori* regularization. The latter is primarily for cases where not all derivatives of the log target are available or when $N \ll d$. Combinations of the two regularization ideas are also possible. Methods to choose between control variates are described in Section 3.3.

3.1 Regularization Through Penalized Regression

As mentioned earlier, the number of regression parameters in ZV-CV grows rapidly with the order Q of the polynomial and with the dimension d of $\boldsymbol{\theta}$. Therefore, the polynomial order that could be considered is limited by the number of samples required to ensure existence of a unique solution to the OLS problem, eliminating the potential reduction that could be achieved using higher order polynomials. In this section, we propose to use penalized regression techniques to help overcome this problem.

In most contexts, using penalized regression reduces variance at the cost of introducing bias. Recall that the conditions for unbiasedness in Section 2.2 do not depend on the mechanism for estimating $\boldsymbol{\beta}$, as long as the samples used in estimating $\boldsymbol{\beta}$ are independent of those used in evaluating (2.1). Thus, the use of penalized regression methods does not introduce bias into ZV-CV.

The regularization methods introduced in Sections 3.1 and 3.1 involve a penalty function on $\boldsymbol{\beta}$ so we use standardization for stability and to be able to employ a single λ. The regression problem becomes:

$$(\hat{c}, \hat{\boldsymbol{\beta}}_s) \in \underset{\substack{c\in\mathbb{R} \\ \boldsymbol{\beta}_s\in\mathbb{R}^J}}{\arg\min} \frac{1}{N}\sum_{i=1}^{N} \left[\varphi_s(\boldsymbol{\theta}_i) - c + \boldsymbol{\beta}_s^\top \boldsymbol{x}_s(\boldsymbol{\theta}_i)\right]^2 + \lambda \mathrm{PEN}(\boldsymbol{\beta}_s), \tag{3.1}$$

where the subscript s is in reference to the response and predictors being standardized by their sample mean and standard deviation. Specifically, using the notation $\bar{a} = \frac{1}{N}\sum_{i=1}^N a_i$ and $\sigma_a = \sqrt{\sum_{i=1}^N (a_i - \bar{a})^2/(N-1)}$, we have that $\varphi_s(\boldsymbol{\theta}_i) = (\varphi(\boldsymbol{\theta}_i) - \overline{\varphi})/\sigma_\varphi$, $\boldsymbol{x}_s[j](\boldsymbol{\theta}_i) = (\boldsymbol{x}[j](\boldsymbol{\theta}_i) - \overline{\boldsymbol{x}[j]})/\sigma_{\boldsymbol{x}_j}$ for $j = 1, \ldots, J$ and $\boldsymbol{\beta}_s$ represents the coefficients on this standardized scale. The estimated coefficients on the original scale are $\hat{\beta}[j] = \hat{\beta}_s[j]\frac{\sigma_\varphi}{\sigma_{x[j]}}$.

The parameter λ is chosen to minimize the k-fold cross-validation mean square error.

$\mathcal{L}_2$ Penalization: PEN$(g) = \|\boldsymbol{\beta}_s\|_2^2$

The first type of penalization that we consider is Tikhonov regularization (Tikhonov et al., 2013), or ridge regression as it is known when applied in regression (Hoerl and Kennard, 1970). This involves using a squared $\mathcal{L}_2$ penalty, PEN$(g) = \|\boldsymbol{\beta}_s\|_2^2$. Ridge regression mitigates overfitting and allows for estimation when the regression problem is ill-posed due to a small number of observations. Closed form solutions for $\hat{c}$ and $\hat{\boldsymbol{\beta}}$ are available, leading to the same computational cost as OLS of $\mathcal{O}(J^3 + NJ^2)$. The use of $\mathcal{L}_2$ penalized ZV-CV can also be motivated using the results of Belkin et al. (2019), who argued that using $J \geq N$ with an interpolation-based approach (i.e. CF or regularized ZV-CV) can lead to better mean square loss compared to restricting to $J \leq N$ (i.e. standard ZV-CV) in situations where there is no reason to pre-suppose the first J basis functions are also the most useful. The latter condition may be satisfied when φ is too complex to be well-approximated using control variates based on low order polynomials.

To motivate this particular form of penalization, we now consider interpreting this method as a computationally efficient variant of CF. To facilitate a comparison with the approach of Barp et al. (2022), we consider a particular instance of CF with a reproducing kernel Hilbert space $\mathcal{H}$ that is carefully selected to lead to an algorithm with lower computational cost. Namely, we select a polynomial kernel

$$k(\boldsymbol{\theta}, \boldsymbol{\theta'}) = \sum_{j=1}^{J} P_j(\boldsymbol{\theta}) P_j(\boldsymbol{\theta'}),$$

where $P_j(\boldsymbol{\theta})$ denotes the jth of all J monomial terms in $\boldsymbol{\theta}$ up to order Q. For such a kernel, a well-defined Hilbert space $\mathcal{H} = \text{span}\{P_j\}_{j=1,\ldots,J}$ is reproduced and we have an explicit expression for the Hilbert norm

$$\left\| \sum_{j=1}^{J} \beta_j P_j \right\|_{\mathcal{H}} = \left(\sum_{j=1}^{J} \beta_j^2 \right)^{1/2},$$

which reveals the method of Barp et al. (2022) as an $\mathcal{L}_2$-penalized regression method. As such, the optimization problem in ZV-CV with PEN$(g) = \|\boldsymbol{\beta}_s\|_2^2$ is equivalent to the optimization problem in CF (without the standardization of the response and predictors) and it can be solved as a least-squares problem with complexity $\mathcal{O}(J^3 + NJ^2)$. The first main contribution of our work is to propose a more practical alternative to the method of Barp et al. (2022), which we recall has $\mathcal{O}(N^3)$ computational cost, by using such a finite-dimensional polynomial kernel. Our results in this direction are empirical (only) and we explore the properties of this method for various values of Q in Section 4.

Tikhonov regularization has been applied implicitly in the context of CF but, to the best of our knowledge, this is the first time that general penalized regression methods have been proposed in the context of ZV-CV. Results in Section 4 demonstrate that the new estimators can offer substantial variance reduction in practice when the number of samples is small relative to the number of coefficients being estimated.

$\mathcal{L}_1$-Penalization: PEN$(g) = \|\boldsymbol{\beta}_s\|_1$

The principal aim in the design of a control variate h is to accurately predict the value that the function φ takes at an input $\boldsymbol{\theta}^*$ not included in the training dataset $\{(\boldsymbol{\theta}_i, \varphi(\boldsymbol{\theta}_i))\}_{i=1}^N$. It is well-understood that $\mathcal{L}_1$-regularization can outperform $\mathcal{L}_2$-regularization in the predictive context when the function φ can be well-approximated by a relatively sparse linear combination of predictors. In our case, the unstandardized predictors are the functions in the set $\{1\} \cup \{\mathcal{L}P_j\}_{j=1,\ldots,J}$. Given that low-order polynomial approximation can often work well for integrands φ of interest, it seems plausible that $\mathcal{L}_1$-regularization could offer an improvement over the $\mathcal{L}_2$-regularization used in Oates et al. (2017); Barp et al. (2022). Investigating this question is the second main contribution of our work.

In the context of ZV-CV, $\mathcal{L}_1$-penalization can be interpreted as using the least absolute shrinkage and selection operator (LASSO, Tibshirani (1996)). LASSO introduces an $\mathcal{L}_1$ penalty $\text{PEN}(g) = \|\boldsymbol{\beta}_s\|_1$ where $||\boldsymbol{\beta}_s||_1 = \sum_j |\beta_s[j]|$. The effect of the penalty is that some coefficients are estimated to be exactly zero.

3.2 A priori Regularization

As an alternative to penalized regression methods, in this section we consider restricting the function g to vary only in a lower-dimensional subspace of the domain $\Theta \subseteq \mathbb{R}^d$. More specifically, a subset of parameters $S \subseteq \{1, \ldots, d\}$ is selected prior to estimation and the function g is defined, in a slight abuse of notation, as $g(\boldsymbol{\theta}) = P(\boldsymbol{\theta}[S])$. The log target derivatives, $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$, only appear in the control variates (2.2) through the dot product $\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$. Therefore if $j \notin S$ then the derivative $\nabla_{\boldsymbol{\theta}[j]} \log p(\boldsymbol{\theta})$ is not required. We refer to this approach as *a priori* regularization.

A priori regularization makes ZV-CV feasible when some derivatives cannot be used, for example due to intractability, numerical instability, computational expense or storage constraints. An example of where some derivatives may be difficult to obtain is in Bayesian inference for ordinary differential equation (ODE) models. Evaluating $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta})$ requires the sensitivities of the ODE to be computed, which involves augmenting the system of ODEs with additional equations. If some additional equations render the system stiff, then more costly implicit numerical solvers need to be used and in such cases it would be useful to avoid including sensitivities corresponding to the difficult elements of $\boldsymbol{\theta}$. It may also be infeasible to use the $\mathcal{O}(d)$ storage required to run standard ZV-CV. Storing a subset of the parameters and derivatives for use in *a priori* regularization may, however, be achievable. Another benefit of *a priori* ZV-CV is that it reduces the number of coefficients to estimate, making estimation feasible when $N \ll d$. Zhuo et al. (2018) consider similar ideas to *a priori* ZV-CV in the context of Stein variational gradient descent, where they use the conditional independence in $p(\boldsymbol{\theta})$ for probabilistic graphical models to separate high dimensional inference problems into a series of lower dimensional problems.

The downside of using *a priori* ZV-CV is that the potential for variance reduction is reduced, except for under both conditions (a) $\boldsymbol{\theta}[S]$ is independent of $\boldsymbol{\theta}[\bar{S}]$ according

to $p(\boldsymbol{\theta})$, where $\bar{S} = \{1, \ldots, d\} \setminus S$, and (b) $\varphi(\boldsymbol{\theta}) = \varphi(\boldsymbol{\theta}[S])$. Outside of this situation, restricting the polynomial to $g(\boldsymbol{\theta}) = P(\boldsymbol{\theta}[S])$ will give varying levels of performance depending on the subset that is selected. Intuitively, one may wish to choose the subset of variables so that $\boldsymbol{\theta}[S]$ and/or $\nabla_{\boldsymbol{\theta}[S]} \log p(\boldsymbol{\theta})$ have high correlations with $\varphi(\boldsymbol{\theta})$. In practice, this is easiest to do when there is *a priori* knowledge and therefore not all derivatives need to be calculated and stored. Given (b), it is suspected that this method will be more useful for individual parameter expectations than for expectations of functions of multiple parameters.

Estimators using this approach are unbiased under the same conditions as ZV-CV and penalized ZV-CV. This method is also applicable to CF, though nonlinear approximation may be more difficult in this non-parametric setting.

3.3 Automatic Selection of Control Variates

The performance of regularized ZV-CV depends upon the polynomial order, the penalization type and on S. We demonstrate in Section 4 that the common practice of defaulting to $Q = 2$ with OLS is often sub-optimal and also that the optimal control variate depends on a variety of factors including N and $p(\boldsymbol{\theta})$. It has previously been proposed to increase the number of control variates as the sample size increases (see e.g. Portier and Segers (2019) and the appendices of Oates et al. (2017)). However, in these existing works the mechanism whereby the complexity of the control variate was increased was not data-dependent.

To choose between control variates in this work, we use 2-fold cross-validation so that our selection is data-dependent. For each combination of penalization type and S, we start with polynomial order $Q = 1$ and we continue to increase the polynomial order until the average cross-validation error is larger for $Q+1$ than for Q. The combination of regularization method and polynomial order which gives the minimum cross-validation error is selected and we perform estimation using that method on the full set of samples. The cross-validation error that we use here is the sums of square residuals in the hold-out set, averaged across the two folds.

4 Empirical Assessment

In this section, we perform comparisons of regularized ZV-CV to ZV-CV and CF on Bayesian inference examples. In Bayesian statistics, the posterior distribution of the parameters $\boldsymbol{\theta}$ of a statistical model given observed data $\boldsymbol{y}$ is

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\ell(\boldsymbol{y}|\boldsymbol{\theta})p_0(\boldsymbol{\theta})}{Z},$$

where the function $\ell(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood function, $p_0(\boldsymbol{\theta})$ encorporates prior information and Z is a normalizing constant. Interest is in estimating posterior expectations $\int_\Theta \varphi(\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}$ and the normalizing constant or so-called "evidence" $Z = \int_\Theta \ell(\boldsymbol{y}|\boldsymbol{\theta})p_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ for Bayesian model choice. Posterior expectations and Z are typically analytically intractable and challenging to estimate due to the potentially high dimensional integration required.

ZV-CV and CF have both been applied in the context of estimating posterior expectations, for example by Mira et al. (2013); Papamarkou et al. (2014); Friel et al. (2016); Oates et al. (2017); Baker et al. (2019). Oates et al. (2016) and Oates et al. (2017) have also applied ZV-CV and CF, respectively, to a thermodynamic integration (Gelman and Meng, 1998; Ogata, 1989; Friel and Pettitt, 2008) estimator for the evidence, calling the resulting method controlled thermodynamic integration (CTI). The thermodynamic integration estimator gives the log evidence as the sum of multiple expectations with respect to p_t where $p_t = \ell(\boldsymbol{y}|\boldsymbol{\theta})^t p_0(\boldsymbol{\theta})/Z_t$ and t is referred to as the inverse temperature. Oates et al. (2017) use population Monte Carlo (Jasra et al., 2007) to obtain the samples from p_t for $t = 0, \ldots, T$ and they consider specifically $t_j = (j/T)^5$. A total of $2(T + 1)$ expectations are involved, with ZV-CV applied to the estimator for each expectation.

We propose to use sequential Monte Carlo (SMC, Del Moral et al. (2006)) with the tuning method of Salomone et al. (2018) for sampling, rather than the standard choices of MCMC or population MCMC. The benefit of this approach is that the samples are roughly independent which can be preferable over the high autocorrelation that can be seen in MCMC samples. The standard SMC evidence estimator is the product of T expectations, so we consider improving this estimator using ZV-CV and CF. Further details about implementation in SMC and the advantages of this approach are given in Appendix B of the Online Resources. From the perspective of comparing variance reduction methods, the application of ZV-CV to posterior expectations and to multiple evidence estimators means that ZV-CV can be compared on a variety of functions $\varphi(\boldsymbol{\theta})$ and distributions $p(\boldsymbol{\theta})$.

We perform an empirical comparison of the following methods using examples of varying complexity:

- **vanilla**: Monte Carlo integration without control variates.
- **ZV**$_Q$: ZV-CV with OLS and order Q polynomial.
- l-**ZV**$_Q$: ZV-CV with LASSO and order Q polynomial.
- r-**ZV**$_Q$: ZV-CV with ridge regression and order Q polynomial.
- **sub**$_k$-: This prefix indicates *a priori* ZV-CV with a subset of size k. Applications are limited to $d > 1$ dimensions. We only apply **sub**$_k$- ideas to posterior expectations since $\varphi(\boldsymbol{\theta})$ is a function of a single parameter and a potentially reasonable subset may be known *a priori*.
- **crossval**: Control variate selection using 2-fold cross-validation. This method chooses between ZV$_\mathrm{Q}$, l-ZV$_\mathrm{Q}$, r-ZV$_\mathrm{Q}$ and **sub**$_k$- where applicable.
- **CF**: Control functionals with a second-order Stein operator, a Gaussian kernel $k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp(-\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2/\sigma^2)$ and selection of σ^2 using 5-fold cross-validation with the generous 15-value grid $10^{\boldsymbol{\kappa}}$ where $\kappa_i = -3 + 0.5i$ for $i = 0, \ldots, 14$.

Methods written with the prefix sub$_k$- or the name l-ZV$_Q$, r-ZV$_Q$ or crossval are novel for all Q and k. The main purposes of these comparisons are to investigate the perfor-

mance of higher order polynomials, the utility of penalized regression and the ability to achieve variance reduction using a subset of derivatives. The purpose of the comparisons to CF is not necessarily to outperform CF, as CF can be infeasible to apply in its basic form for large N, but to benchmark the performance of these novel methods against CF. A variety of sample sizes, integrands $\varphi(\boldsymbol{\theta})$ and target distributions p are used for fair comparisons. We focus on sample sizes that are typical of SMC, ranging from $N = 10$ to $N = 10{,}000$ but we note that larger sample sizes can be accommodated by the regularized ZV-CV methods which have a computational complexity of $\mathcal{O}(N)$.

Estimators are compared on the basis of mean square error (MSE), where the gold standard of estimation is carefully chosen for each example. The main quantity of interest reported in this section is $\widehat{\mathsf{MSE}_p[\text{vanilla}]}/\widehat{\mathsf{MSE}_p[\cdot]}$, the MSE of the vanilla Monte Carlo estimator estimated from 100 independent SMC runs divided by the estimated MSE for the method in question. This quantity is referred to as statistical efficiency and it is reported for each fixed N. Values above one are preferred.

Control variate methods are most valuable when the sampling algorithm is expensive, for example due to the cost of evaluating the likelihood, or when evaluation of the function $\varphi(\boldsymbol{\theta})$ is costly. The overall efficiency, as measured by

$$\frac{\widehat{\mathsf{MSE}_p[\text{vanilla}]} \times \widehat{\mathsf{time}[\text{vanilla}]}}{\widehat{\mathsf{MSE}_p[\cdot]} \times \widehat{\mathsf{time}[\cdot]}},$$

is also considered for these examples. Here $\widehat{\mathsf{time}[\cdot]}$ is the average time across the 100 runs to compute the estimator in question, including the time spent running the SMC sampler. We note that the run time is subject to the efficiency of the code and here (penalized) ZV-CV is based on the R package glmnet (Friedman et al., 2010), cross-validated ZV-CV is written as a loop in R and CF is implemented in C++. Nevertheless, our proposed methods offer improved overall efficiency in several of the applications considered. The computational benefits of our approach will improve with increasing model complexity in terms of likelihood calculations, since the overhead associated with penalized regression will become relatively negligible.

Two examples are described in detail in this section. Appendices E, F and G also include results for a 61-dimensional logistic regression example, a one-dimensional ODE example which motivates higher order polynomials and a challenging nine-dimensional ODE model, respectively.

In terms of bias, boundary condition (2.4) is satisfied using the specified parameterizations for all examples considered in this paper. This can be verified through the sufficient condition that the tails of p decay faster than polynomially and $\Theta = \mathbb{R}^d$ (Appendix B of Oates et al. (2016)). However, the estimators are generically biased due to the use of SMC, as they would be with MCMC. All results are based on combined estimators as opposed to split estimators, so all pairs $\{\boldsymbol{\theta}_i, \varphi(\boldsymbol{\theta}_i)\}_{i=1}^N$ are used to build $\tilde{\varphi}$ and also to estimate $\mathbb{E}_p[\tilde{\varphi}(\boldsymbol{\theta})]$.

4.1 Recapture Example

This 11-dimensional example demonstrates that reduced variance estimators can be obtained with the use of higher order polynomials and regularization.

Marzolin (1988) collected data on the capture and recapture of the bird species *Cinclus cinclus* over six years. Like Brooks et al. (2000), Nott et al. (2018) and South et al. (2019b), we use a Bayesian approach to estimate the parameters of a Cormack-Jolly-Seber model (Lebreton et al., 1992) for the capture and recapture of this species. The parameters of the Cormack-Jolly-Seber model used here are the probability of survival from year i to $i+1$, ϕ_i, and the probability of being captured in year k, p_k, where $i = 1, \ldots, 6$ and $k = 2, \ldots, 7$. Denote the number of birds released in year i as D_i and the number of animals caught in year k out of the number released in year i as y_{ik}. It is simple to show that the number released in year i that are never caught is $d_i = D_i - \sum_{k=i+1}^{7} y_{ik}$ and the probability of a bird being released in year i and never being caught is $\chi_i = 1 - \sum_{k=i+1}^{7} \phi_i p_k \prod_{m=i+1}^{k-1} \phi_m(1-p_m)$. The likelihood is given by

$$\ell(\boldsymbol{y}|\boldsymbol{\theta}) \propto \prod_{i=1}^{6} \chi_i^{d_i} \prod_{k=i+1}^{7} \left[\phi_i p_k \prod_{m=i+1}^{k-1} \phi_m(1-p_m) \right]^{y_{ik}},$$

where $\boldsymbol{\theta} = (\phi_1, \ldots, \phi_5, p_2, \ldots, p_6, \phi_6 p_7)$. Following South et al. (2019b), the parameters ϕ_6 and p_7 are multiplied together due to a parameter identifiability issue.

The prior is $\boldsymbol{\theta}[j] \sim \mathcal{U}(0,1)$ for $j = 1, \ldots, 11$. To satisfy the boundary condition (2.4) and to improve the efficiency of MCMC proposals, all parameters are transformed to the real line using $\boldsymbol{\psi}[j] = \log(\boldsymbol{\theta}[j]/(1-\boldsymbol{\theta}[j]))$ so the prior density for $\boldsymbol{\psi}[j]$ is $\exp(\boldsymbol{\psi}[j])/(1+\exp(\boldsymbol{\psi}[j]))^2$, for $j = 1, \ldots, 11$.

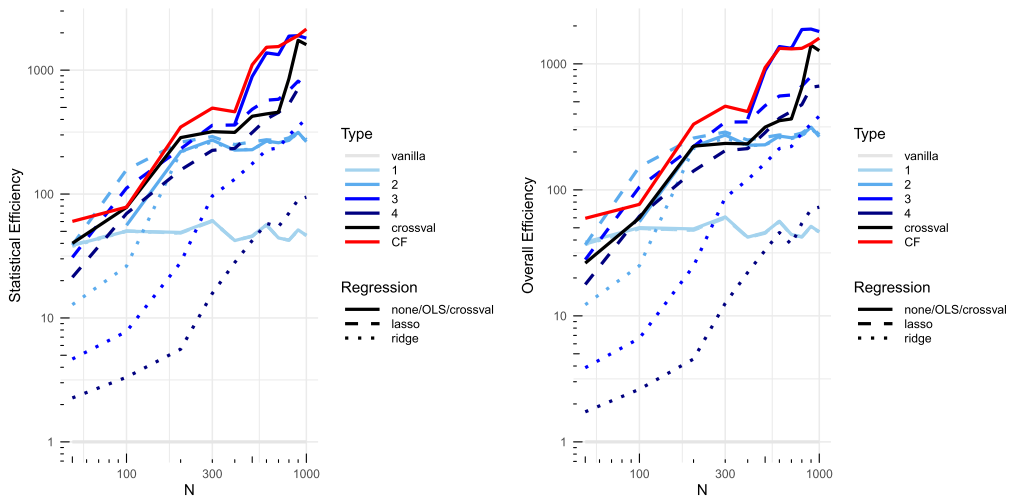
The gold standard of evidence estimation for this example is the mean evidence estimate for l-ZV$_1$ at $N = 5000$. The posterior expectation gold standard is the average posterior mean for ZV$_4$ at $N = 5000$.

Posterior Expectations

The average statistical efficiency and overall efficiency across parameters is shown in Figure 1, excluding *a priori* regularization results for simplicity. Higher order polynomials become more efficient as N increases and the use of penalized regression means that higher order polynomials can be considered for smaller N. LASSO regression is preferable over ridge regression for this example.

Using *a priori* ZV-CV with $S = j$, where j is the index of the current parameter of interest, sub$_1$-ZV$_1$ is on average roughly 10 times more efficient than vanilla Monte Carlo integration.

Cross-validation generally gives similar results to CF and to the best performing fixed method. More details about the selected control variates can be found in Appendix C of the Online Resources.



(a) Statistical efficiency

(b) Overall efficiency

Figure 1: Recapture example: (a) statistical efficiency averaged over 11 parameters and (b) overall efficiency averaged over 11 parameters.

Evidence Estimation

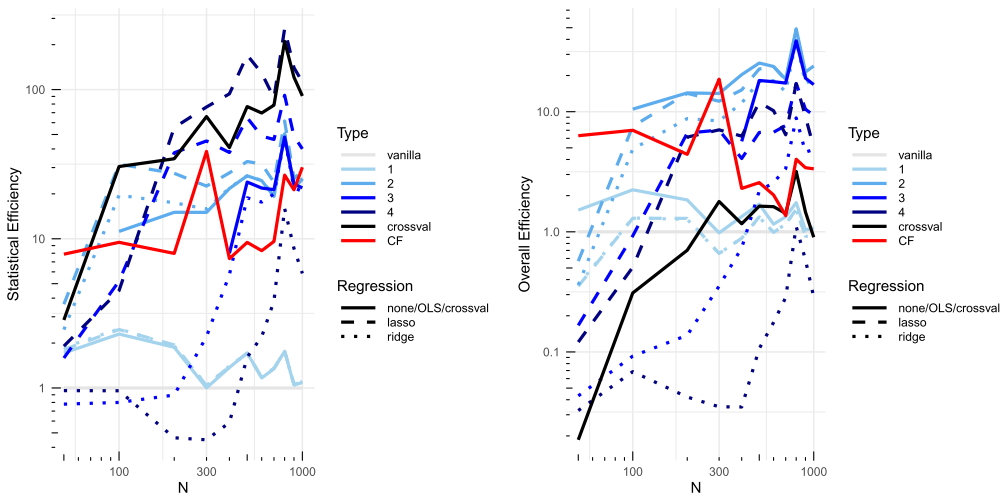
Regularized ZV-CV and automatic control variates give improved statistical efficiency over ZV-CV and CF for the range of N that are considered here, as seen in Figure 2. However, there is less improvement in terms of overall efficiency due to the fact that multiple expectations are required for evidence estimation. This puts the more computationally intensive methods including higher order polynomials, cross-validation and CF at a significant disadvantage. We note that this example was selected to allow for extensive comparisons and the cost of post-processing would have less impact under more expensive likelihood functions.

The selected control variates for $N = 50$ and $N = 1000$ can be found in Appendix C of the Online Resources.

4.2 Log-Gaussian Cox Point Process Example

We now consider an example where the dimension can be adjusted. The log-Gaussian Cox point process example of Møller et al. (1998) consists of locations of 126 Scots pine saplings in a 10×10 m^2 plot. The plot can be discretized into $n \times n$ grid cells, so that the dimension $d = n^2$ of the problem can be varied. Here we consider $n = 4$, $n = 8$ and $n = 16$ so that we have Bayesian inference problems of size $d = 16$, $d = 64$ and $d = 256$.

The model specifications, including code for the log likelihood, log prior and their gradients, match that of Heng and Jacob (2019). After normalizing the plot to fit onto a



(a) Statistical Efficiency CTI

(b) Overall Efficiency CTI

(c) Statistical Efficiency SMC

(d) Overall Efficiency SMC

Figure 2: Recapture example: (a) statistical efficiency for the CTI estimator, (b) overall efficiency for the CTI estimator, (c) statistical efficiency for the SMC estimator and (d) overall efficiency for the SMC estimator.

unit square, the number of points at grid cell (i, j) for $i, j = 1, \ldots, n$ is denoted $y_{i,j}$. It is assumed that the $y_{i,j}$ are conditionally independent and Poisson distributed with mean $\frac{\Lambda_{i,j}}{n^2}$. The prior is $\Lambda_{i,j} = \exp(\theta_{i,j})$ where $\theta_{i,j}$ has a Gaussian process prior with mean

Approach	Q	Stat. Efficiency	Comp. Efficiency	Overall Efficiency
Vanilla	NA	1.0×10^0	1.0000	1.0×10^0
CF	NA	4.4×10^1	0.9787	4.3×10^1
ZV_1	1	2.0×10^1	0.9994	2.0×10^1
l-ZV_1	1	2.2×10^1	0.9722	2.2×10^1
r-ZV_1	1	2.0×10^1	0.9719	1.9×10^1
ZV_2	2	–	–	–
l-ZV_2	2	1.2×10^3	0.9315	1.1×10^3
r-ZV_2	2	2.1×10^1	0.8435	1.8×10^1
ZV_3	3	–	–	–
l-ZV_3	3	5.0×10^2	0.8330	4.2×10^2
r-ZV_3	3	1.7×10^1	0.6675	1.1×10^1
ZV_4	4	–	–	–
l-ZV_4	4	3.2×10^2	0.6151	2.0×10^2
r-ZV_4	4	1.4×10^1	0.3857	5.3×10^0
sub_1-ZV_1	1	2.4×10^1	0.9996	2.4×10^1
sub_1-ZV_2	2	5.0×10^3	0.9996	5.0×10^3
sub_1-ZV_3	3	3.5×10^6	0.9996	3.5×10^6
sub_1-ZV_4	4	6.9×10^7	0.9995	6.9×10^7
crossval	NA	2.8×10^8	0.2125	5.5×10^7

Table 1: 16-dimensional Cox example: efficiency for marginal posterior expectations when $N = 100$, averaged over results for all 16 parameters. A "–"' indicates that the population size $N = 100$ is insufficient for standard ZV-CV. We refer the reader to the beginning of Section 4 for acronym definitions.

μ and covariance function $\Sigma_{(i,j),(i',j')} = s^2 \exp\left[-\sqrt{(i-i')^2 + (j-j')^2}/(n\beta)\right]$, where $\beta = 1/33$, $s^2 = 1.91$ and $\mu = \log(126) - s^2/2$.

Our goal is to estimate the posterior means for the parameters $\theta_{i,j}$ for $i, j = 1, \ldots, n$, and we do so using SMC runs with $N = 100$ particles. We do not consider evidence estimation due to lack of a reliable gold standard. Due to memory and time constraints, the maximum polynomial order is constrained so that the maximum number of covariates in ZV-CV is 5000. The gold standards in this example are the average posterior expectations across many independent unbiased Riemann-manifold HMC runs (Heng and Jacob, 2019) with unbiased control variates as described in South et al. (2019a). Details of the gold standard are available in Appendix D.

Tables 1, 2 and 3 show the mean relative statistical, computational and overall efficiency for posterior expectations in dimensions $d = 16$, $d = 64$ and $d = 256$, respectively, when $N = 100$. In all three settings, the existing methods (vanilla MC, ZV-CV with OLS and CF) are outperformed by the novel approaches of *a priori* ZV-CV, cross-validation and LASSO with a higher order polynomial than OLS could handle. The best performing novel method has an overall efficiency which is better than the best performing existing method by a factor of over 1,600,000 for $d = 16$, over 1,800 for $d = 64$ and over 25 for $d = 256$. Results showing the competitive performance of the novel methods for

Approach	Q	Stat. Efficiency	Comp. Efficiency	Overall Efficiency
Vanilla	NA	1.0	1.0000	1.0
CF	NA	5.9	0.9953	5.9
ZV_1	1	4.5	0.9999	4.5
l-ZV_1	1	12.0	0.9931	11.9
r-ZV_1	1	5.2	0.9924	5.2
ZV_2	2	–	–	–
l-ZV_2	2	17.2	0.9574	16.5
r-ZV_2	2	5.2	0.9071	4.7
sub_1-ZV_1	1	13.8	0.9999	13.8
sub_1-ZV_2	2	2505.3	0.9999	2505.1
sub_1-ZV_3	3	7181.6	0.9999	7181.0
sub_1-ZV_4	4	11081.3	0.9999	11080.4
crossval	NA	7271.9	0.4631	3369.3

Table 2: 64-dimensional Cox example: efficiency for marginal posterior expectations when $N = 100$, averaged over results for all 64 parameters. A "–"' indicates that the population size $N = 100$ is insufficient for standard ZV-CV. We refer the reader to the beginning of Section 4 for acronym definitions.

Approach	Q	Stat. Efficiency	Comp. Efficiency	Overall Efficiency
Vanilla	NA	1.0	1.0000	1.0
CF	NA	2.0	0.9990	2.0
ZV_1	1	–	–	–
l-ZV_1	1	21.2	0.9989	21.2
r-ZV_1	1	2.1	0.9980	2.1
sub_1-ZV_1	1	21.1	1.0000	21.1
sub_1-ZV_2	2	52.3	1.0000	52.3
sub_1-ZV_3	3	53.4	1.0000	53.4
sub_1-ZV_4	4	42.3	1.0000	42.3
crossval	NA	32.3	0.9855	31.8

Table 3: 256-dimensional Cox example: efficiency for marginal posterior expectations when $N = 100$, averaged over results for all 256 parameters. A "–"' indicates that the population size $N = 100$ is insufficient for standard ZV-CV. We refer the reader to the beginning of Section 4 for acronym definitions.

$N = 1{,}000$ and $N = 10{,}000$ are given in Appendix D. Like the results for $N = 100$, *a priori* ZV-CV, cross-validation and LASSO outperform existing alternatives in the majority of settings.

5 Discussion

In this paper, we introduced two types of regularized ZV-CV: regularization through penalized regression and regularization by selecting a subset of parameters to include in the regression model. Higher order polynomial basis functions have the potential to

outperform the commonly used polynomial with $Q = 2$ as N – the number of Monte Carlo, MCMC or SMC simulations – increases. Our penalized ZV-CV ensures that the resulting functional approximation problem remains well-defined when N is less than the number of control variate coefficients (J) while performing similarly to standard ZV-CV when $N > J$. For the examples considered here, we found that LASSO generally resulted in better performance than ridge regression. *A priori* ZV-CV led to significant improvements over vanilla Monte Carlo for posterior expectations, with little computational overhead.

One of the main applications of the proposed methods is in models where the dimension, d, is too high for standard variance reduction techniques to be efficient. Empirical evidence suggests that using ZV-CV and penalized ZV-CV, where Q is increased with N, offers better statistical performance than CF in high dimensions. However, the computational cost of (penalized) ZV-CV is $\mathcal{O}(N\binom{d+Q}{d}^2 + \binom{d+Q}{d}^3)$, which may prohibit the application of these methods with large Q in high dimensions. This explosion in complexity for large Q and d is a disadvantage relative to CF when the sample size is comparable or less than the dimension, though the complexity is similar when $N \approx d$ and $Q = 1$ in (penalized) ZV-CV. One could consider speeding up these algorithms by using partial LASSO searches (e.g. Efron et al., 2004; Fan and Lv, 2008) or by using approximate solvers as proposed in Si et al. (2022). Alternatively, in very large dimensions, the *a priori* ZV-CV approach can be used to obtain variance reductions with a complexity that is $\mathcal{O}(N|S|^2 + |S|^3)$ where $1 \leq |S| \leq d$. This *a priori* approach also offers benefits when not all derivatives are available, when $N \ll d$, or when information about the relationships between the integrand and parameters is known (for example when $p(\boldsymbol{\theta})$ has a directed acyclic graph factorization).

Leluc et al. (2019) provide additional theoretical support for LASSO-based control variate selection. The work of Leluc et al. (2019), which was publicly available after the pre-print of our paper (South et al., 2018), gives concentration inequalities for the integration error with LASSO-based control variates and also shows that the correct control variates are selected with high probability. The theoretical results are based on bounded control variates, which do not apply in ZV-CV and CF when p has unbounded support. Leluc et al. (2019) find empirically that a methodological adjustment of performing OLS for estimation once the control variates have been selected via LASSO is helpful in reducing the variance of the estimator. We point out that this modification is necessary to obtain the zero-variance property of ZV-CV. The optimal coefficients required to obtain zero-variance estimators cannot be obtained directly from penalized regression methods like LASSO and ridge regression with non-zero λ.

We have proposed the consideration of different penalty functions in the optimization problem for control variates, but we focus specifically on LASSO and ridge regression. Some other potentially useful regularization methods for the situation where $N < \binom{d+Q}{d}$ are elastic net (Zou and Hastie, 2005) and partial least squares (PLS, Wold (1975)). Elastic net is a compromise between LASSO and ridge regression which uses two tuning parameters. PLS is based on choosing the $k < \binom{d+Q}{d} - 1$ independent linear combinations of covariates that explain the maximum variance in the response, where k is chosen through cross-validation. Active subspaces (Constantine, 2015) are a more recent dimension-reduction technique which use the derivatives of the function of interest

to find the linear combinations of covariates that are best at predicting the function. It would be of interest in future research to compare our LASSO and ridge regression ZV-CV methods with these alternatives.

The concept of regularization by selecting a subset of parameters is referred to as nonlinear approximation in approximation theory and applied mathematics (DeVore, 1998), and there is some theoretical evidence to suggest that this can outperform linear approximation (e.g. penalized regression which is described in Section 3.1). Selecting a particular subset of monomials which are used in a polynomial interpolant is also the same idea as in sparse grid algorithms for numerical integration (Smolyak, 1963). These methods are known to work well in high dimensions and could be useful alternatives for selecting the subset of monomials in ZV-CV.

Stein-based control variates using neural networks have recently appeared in the literature (Zhu et al., 2019). Zhu et al. (2019) added details of penalization methods to their approach, where the control variates cannot be fitted exactly and stochastic optimization is required. Penalization methods are simpler and more stable in the linear regression context but in future research it would be of interest to compare to neural control variates with regularization. This alternative approach is likely to outperform ZV-CV in some applications, such as when $\varphi(\boldsymbol{\theta})$ is multi-modal.

Finding the optimal parameterization for a given application is a challenging open problem. Choosing the parameterization is a trade-off between making p_ϕ simpler and making $\varphi(f^{-1}(\phi))$ simpler. Another potential benefit of reparameterizing for ZV-CV is that there is the potential to enforce more sparsity in the predictors for improved performance in $\mathcal{L}_1$ penalization.

Derivatives are available in closed form or can be unbiasedly estimated for a large class of problems. ZV-CV has been applied in big data settings in the context of post-processing after stochastic gradient MCMC (Baker et al., 2019) and for models with intractable likelihoods (Friel et al., 2016). Regularized ZV-CV also applies in these settings. Regularized ZV-CV could also be used in exact approximate settings where a particle filtering estimate of the likelihood is used (see for example Dahlin et al. (2015) and Nemeth et al. (2016)). However, derivative-based methods are most appealing when the derivative of the log target can be obtained with little additional cost relative to the likelihood itself. An interesting avenue for future research may be to consider automatic differentiation.

Supplementary Material

Supplementary Material for Regularized Zero-Variance Control Variates
(DOI: 10.1214/22-BA1328SUPP; .pdf). This document provides three additional applications, further simulation results for the examples in the paper and a more detailed description of how to implement ZV-CV methods in SMC.

References

Assaraf, R. and Caffarel, M. (1999). "Zero-Variance Principle for Monte Carlo Algorithms." *Physical Review Letters*, 83(23): 4682–4685. 866, 867, 869

Baker, J., Fearnhead, P., Fox, E. B., and Nemeth, C. (2019). "Control Variates for Stochastic Gradient MCMC." *Statistics and Computing*, 29(3): 599–615. MR3969063. doi: https://doi.org/10.1007/s11222-018-9826-2. 869, 875, 883

Barp, A., Oates, C. J., Porcu, E., and Girolami, M. (2022). "A Riemann-Stein Kernel Method." *Bernoulli*, 28(4): 2181–2208. *arXiv preprint* 1810.04946. MR4474540. doi: https://doi.org/10.3150/21-bej1415. 866, 867, 868, 869, 872, 873

Belkin, M., Hsu, D., and Xu, J. (2019). "Two models of double descent for weak features." *arXiv preprint* arXiv:1903.07571. MR4186534. doi: https://doi.org/10.1137/20M1336072. 872

Belomestny, D., Iosipoi, L., and Zhivotovskiy, N. (2017). "Variance reduction via empirical variance minimization: convergence and complexity." *arXiv preprint* arXiv:1712.04667. 868

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media. MR2239907. doi: https://doi.org/10.1007/978-1-4419-9096-9. 868

Brooks, S. P., Catchpole, E. A., and Morgan, B. J. T. (2000). "Bayesian animal survival estimation." *Statistical Science*, 15(4): 357–376. MR1847773. doi: https://doi.org/10.1214/ss/1009213003. 877

Brosse, N., Durmus, A., Meyn, S., Éric Moulines, and Radhakrishnan, A. (2019). "Diffusion approximations and control variates for MCMC." *arXiv preprint* arXiv:1808.01665. 868

Chopin, N. (2002). "A sequential particle filter method for static models." *Biometrika*, 89(3): 539–552. MR1929161. doi: https://doi.org/10.1093/biomet/89.3.539. 867

Constantine, P. G. (2015). *Active subspaces: Emerging ideas for dimension reduction in parameter studies*, volume 2. Society for Industrial and Applied Mathematics. MR3486165. doi: https://doi.org/10.1137/1.9781611973860. 882

Dahlin, J., Lindsten, F., and Schon, T. B. (2015). "Particle Metropolis-Hastings using gradient and Hessian information." *Statistics and Computing*, 25: 81–92. MR3304908. doi: https://doi.org/10.1007/s11222-014-9510-0. 883

Del Moral, P., Doucet, A., and Jasra, A. (2006). "Sequential Monte Carlo samplers." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68: 411–436. MR2278333. doi: https://doi.org/10.1111/j.1467-9868.2006.00553.x. 867, 875

DeVore, R. A. (1998). "Nonlinear approximation." *Acta numerica*, 7: 51–150. MR1689432. doi: https://doi.org/10.1017/S0962492900002816. 883

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). "Hybrid Monte Carlo." *Physical Letters B*, 195(2). MR3960671. doi: https://doi.org/10.1016/0370-2693(87)91197-x. 866

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression." *The Annals of statistics*, 32(2): 407–499. MR2060166. doi: https://doi.org/10.1214/009053604000000067. 882

Fan, J. and Lv, J. (2008). "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911. MR2530322. doi: https://doi.org/10.1111/j.1467-9868.2008.00674.x. 882

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, 33(1): 1–22. URL https://www.jstatsoft.org/v33/i01/. 876

Friel, N., Mira, A., and Oates, C. J. (2016). "Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods." *Bayesian Analysis*, 11(1): 215–245. MR3447097. doi: https://doi.org/10.1214/15-BA948. 875, 883

Friel, N. and Pettitt, A. N. (2008). "Marginal likelihood estimation via power posteriors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3): 589–607. MR2420416. doi: https://doi.org/10.1111/j.1467-9868.2007.00650.x. 867, 875

Gelman, A. and Meng, X.-L. (1998). "Simulating normalizing constants: from importance sampling to bridge sampling to path sampling." *Statistical Science*, 13(2): 163–185. MR1647507. doi: https://doi.org/10.1214/ss/1028905934. 875

Girolami, M. and Calderhead, B. (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214. MR2814492. doi: https://doi.org/10.1111/j.1467-9868.2010.00765.x. 866

Glasserman, P. (2003). *Monte Carlo methods in financial engineering*, chapter 4, 185–279. Springer Science & Business Media. MR1999614. 869

Gorham, J. and Mackey, L. (2015). "Measuring sample quality with Stein's method." In *Proceedings of the 28th Conference on Neural Information Processing Systems*, volume 28, 226–234. MR4257226. 867

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*. Chapman & Hall. MR0223065. 867

Heng, J. and Jacob, P. (2019). "Unbiased Hamiltonian Monte Carlo with couplings." *Biometrika*, 106(2): 287–302. MR3949304. doi: https://doi.org/10.1093/biomet/asy074. 878, 880

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1): 55–67. MR0611894. doi: https://doi.org/10.1080/01966324.1981.10737061. 872

Jasra, A., Stephens, D. A., and Holmes, C. C. (2007). "On population-based simulation for static inference." *Statistics and Computing*, 17(3): 263–279. MR2405807. doi: https://doi.org/10.1007/s11222-007-9028-9. 875

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). "Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies." *Ecological Monographs*, 61(1): 67–118. 877

Leluc, R., Portier, F., and Segers, J. (2019). "Control variate selection for Monte Carlo integration." *arXiv preprint* arXiv:1906.10920. MR4277333. doi: https://doi.org/10.1007/s11222-021-10011-z. 882

Marzolin, G. (1988). "Polygynie du cincle plongeur (cinclus cinclus) dans le côtes de Loraine." *Oiseau et la Revue Francaise d'Ornithologie*, 58(4): 277–286. 877

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equations of state calculations by fast computing machines." *Journal of Chemical Physics*, 12(6): 1087–1092. 867

Mira, A., Solgi, R., and Imparato, D. (2013). "Zero variance Markov chain Monte Carlo for Bayesian estimators." *Statistics and Computing*, 23(5): 653–662. MR3094805. doi: https://doi.org/10.1007/s11222-012-9344-6. 866, 867, 869, 870, 875

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). "Log Gaussian Cox processes." *Scandinavian Journal of Statistics*, 25(3): 451–482. MR1650019. doi: https://doi.org/10.1111/1467-9469.00115. 878

Nemeth, C., Fearnhead, P., and Mihaylova, L. (2016). "Particle approximations of the score and observed information matrix for parameter estimation in state-space models with linear computational cost." *Journal of Computational and Graphical Statistics*, 25(4): 1138–1157. MR3572033. doi: https://doi.org/10.1080/10618600.2015.1093492. 883

Nott, D. J., Drovandi, C. C., Mengersen, K., and Evans, M. (2018). "Approximation of Bayesian predictive p-values with regression ABC." *Bayesian Analysis*, 13(1): 59–83. MR3737943. doi: https://doi.org/10.1214/16-BA1033. 877

Oates, C. J., Cockayne, J., Briol, F. X., and Girolami, M. (2019). "Convergence rates for a class of estimators based on Stein's method." *Bernoulli*, 25(2): 1141–1159. MR3920368. doi: https://doi.org/10.3150/17-bej1016. 866, 868

Oates, C. J., Girolami, M., and Chopin, N. (2017). "Control functionals for Monte Carlo integration." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 695–718. MR3641403. doi: https://doi.org/10.1111/rssb.12185. 866, 867, 868, 869, 870, 873, 874, 875

Oates, C. J., Papamarkou, T., and Girolami, M. (2016). "The controlled thermodynamic integral for Bayesian model evidence evaluation." *Journal of the American Statistical Association*, 111(514): 634–645. MR3538693. doi: https://doi.org/10.1080/01621459.2015.1021006. 867, 869, 870, 875, 876

Ogata, Y. (1989). "A Monte Carlo method for high dimensional integration." *Nu-*

merical Mathematics, 55(2): 137–157. MR0987382. doi: https://doi.org/10.1007/BF01406511. 875

Papamarkou, T., Mira, A., and Girolami, M. (2014). "Zero variance differential geometric Markov chain Monte Carlo algorithms." *Bayesian Analysis*, 9(1): 97–128. MR3188301. doi: https://doi.org/10.1214/13-BA848. 867, 869, 875

Portier, F. and Segers, J. (2019). "Monte Carlo integration with a growing number of control variates." *Journal of Applied Probability*, 56(4): 1168–1186. MR4041454. doi: https://doi.org/10.1017/jpr.2019.78. 869, 874

Ripley, B. (1987). *Stochastic Simulation*. John Wiley & Sons. MR0875224. doi: https://doi.org/10.1002/9780470316726. 867

Roberts, G. O. and Stramer, O. (2002). "Langevin diffusions and Metropolis-Hastings algorithms." *Methodology and Computing in Applied Probability*, 4(4): 337–357. MR2002247. doi: https://doi.org/10.1023/A:1023562417138. 866

Salomone, R., South, L. F., Drovandi, C. C., and Kroese, D. P. (2018). "Unbiased and Consistent Nested Sampling via sequential Monte Carlo." *arXiv preprint* arXiv:1805.03924. 875

Si, S., Oates, C., Duncan, A. B., Carin, L., and Briol, F.-X. (2022). "Scalable control variates for Monte Carlo methods via stochastic optimization." In *Proceedings of the 14th International Conference on Monte Carlo and Quasi Monte Carlo Methods in Scientific Computing*. MR4461055. doi: https://doi.org/10.1007/978-3-030-98319-2_10. 882

Smolyak, S. A. (1963). "Quadrature and interpolation formulas for tensor products of certain classes of functions." In *Doklady Akademii Nauk*, volume 148, 1042–1045. Russian Academy of Sciences. MR0147825. 883

South, L. F. (2018). *ZVCV: Zero-Variance Control Variates*. R package version 1.1.0. URL https://cran.r-project.org/web/packages/ZVCV/index.html 866

South, L. F., Nemeth, C., and Oates, C. J. (2019a). "Discussion of "Unbiased Markov chain Monte Carlo with couplings" by Pierre E. Jacob, John O'Leary and Yves F. Atchadé." *arXiv preprint* arXiv:1912.10496. MR4112777. doi: https://doi.org/10.1111/rssb.12336. 880

South, L. F., Oates, C. J., Mira, A., and Drovandi, C. (2018). "Regularised Zero-Variance Control Variates for High-Dimensional Variance Reduction." *arXiv preprint* arXiv:1811.05073. 882

South, L. F., Oates, C. J., Mira, A., and Drovandi, C. (2022). "Supplementary Material for Regularized Zero-Variance Control Variates." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1328SUPP. 869

South, L. F., Pettitt, A. N., and Drovandi, C. C. (2019b). "Sequential Monte Carlo Samplers with independent MCMC proposals." *Bayesian Analysis*, 14(3): 753–776. MR3960770. doi: https://doi.org/10.1214/18-BA1129. 877

Stein (1972). "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables." In Cam, M. L., Neyman, J., and Scott, E. L. (eds.), *Proc. 6th Berkeley Symp. Mathematical Statistics and Probability*, volume 2, 583–602. Berkeley: University of California Press. MR0402873. 867

Stone, M. H. (1948). "The generalized Weierstrass approximation theorem." *Mathematics Magazine*, 21(5): 237–254. MR0027121. doi: https://doi.org/10.2307/3029750. 869

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1): 267–288. MR1379242. 873

Tikhonov, A. N., Goncharsky, A., Stepanov, V. V., and Yagola, A. G. (2013). "Numerical methods for the solution of ill-posed problems." *Springer Science & Business Media*. MR1126915. 872

Wold, H. (1975). "Soft modeling by latent variables; the Non-linear Iterative Partial Least Squares Approach." In Gani, J. (ed.), *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett*, volume 12, 117–142. London: Academic Press. MR0431486. doi: https://doi.org/10.1017/s0021900200047604. 882

Zhu, Z., Wan, R., and Zhong, M. (2019). "Neural Control Variates for Variance Reduction." In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. 868, 883

Zhuo, J., Liu, C., Shi, J., Zhu, J., Chen, N., and Zhang, B. (2018). "Message passing Stein variational gradient descent." In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, 6018–6027. PMLR. 873

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320. MR2137327. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x. 882

Acknowledgments

The authors thank anonymous referees and the associate editor for helpful comments. The authors also wish to thank Nial Friel for the suggestion to reduce the variance of the SMC evidence estimator using ZV-CV and for comments on an earlier draft. LFS and CD are associated with the ARC Centre of Excellence for Mathematical & Statistical Frontiers (ACEMS). LFS would like to thank Matthew Sutton for useful discussions about penalized regression methods. Computational resources used in this work were provided by the HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia and by the High End Computing facility at Lancaster University.