The Controlled Thermodynamic Integral for Bayesian Model Evidence Evaluation

Chris J. Oates, Theodore Papamarkou, and Mark Girolami

ABSTRACT
Approximation of the model evidence is well known to be challenging. One promising approach is based on thermodynamic integration, but a key concern is that the thermodynamic integral can suffer from high variability in many applications. This article considers the reduction of variance that can be achieved by exploiting control variates in this setting. Our methodology applies whenever the gradient of both the log-likelihood and the log-prior with respect to the parameters can be efficiently evaluated. Results obtained on regression models and popular benchmark datasets demonstrate a significant and sometimes dramatic reduction in estimator variance and provide insight into the wider applicability of control variates to evidence estimation. Supplementary materials for this article are available online.

Introduction
In hypothesis-driven research, we are presented with data \( y \) that is assumed to have arisen under one of two (or more) putative models \( m \), characterized by a probability density \( p(y|m) \). Given a priori model probabilities \( p(m) \), the data \( y \) induce a posteriori probabilities \( p(m|y) \) that are the basis for Bayesian model comparison. Since any prior probability distribution gets transformed to a posterior probability distribution through consideration of the data, the transformation itself represents the evidence provided by the data (Kass and Raftery 1995). For the simple case of two models, this transformation follows from Bayes’ rule as

\[
\frac{p(m_2|y)}{p(m_1|y)} = \frac{p(y|m_2)}{p(y|m_1)} \times \frac{p(m_2)}{p(m_1)}. \tag{1}
\]

Thus, the influence of the data on the posterior probability distribution is captured through that Bayes factor \( B_{21} \) in favor of Model 2 over Model 1. Rearranging, we can interpret the Bayes factor as the ratio of the posterior odds to the prior odds. A natural approach to computation of Bayes factors is to directly compute the evidence

\[
p(y|m_1) = \int p(y|\theta, m_1) p(\theta|m_1) d\theta, \tag{2}
\]

provided by data \( y \) in favor of model \( m_1 \), where \( \theta \) are parameters associated with model \( m_1 \). Yet for almost all models of interest, the evidence is unavailable in closed form and must be approximated. Numerous techniques have been proposed to approximate the model evidence (Equation (2)), a selection of which includes path sampling (Ogata 1989; Gelman and Meng 1998), thermodynamic integration (TI; Frenkel and Smit 2002), harmonic means (Gelfand and Dey 1994), Chib’s method (Chib and Jeliazkov 2001), nested sampling (Skilling 2006), particle filters (Del Moral, Doucet, and Jasra 2006), multicanonical algorithms (Marinari and Parisi 1992; Geyer and Thompson 1995), approximate Bayesian computation (Didelot et al. 2011), and variational approximations (Corduneanu and Bishop 2001). Alternatively, one can directly target the Bayes factor \( B_{21} \) that compares between two models. Here, too numerous methods have been proposed, including importance sampling (Gelman and Meng 1998; Chen, Shao, and Ibrahim 2000), ratio importance sampling (Torrie and Valleau 1977), bridge sampling (Gelman and Meng 1998; Chen, Shao, and Ibrahim 2000), sequential Monte Carlo (Zhou, Johansen, and Aston 2013), annealed importance sampling (Neal 2001), reversible-jump Markov chain Monte Carlo (MCMC; Green 1995), stochastic approximate Monte Carlo (Liang 2007), and also again approximate Bayesian computation (Toni et al. 2009). Recent reviews of these methodologies include V’yushemirsky and Girolami (2008), Marin and Robert (2011), and Friel and Wyse (2012).

Of the estimators of evidence that are based on Monte Carlo sampling, it remains the case that estimator variance can in general be extremely high. General approaches to reduction of Monte Carlo error that have been proposed in the literature include importance sampling, stratified sampling, post-stratification (Guha and MacEachern 2006), antithetic variables (Green and Han 1992), control variates and Rao–Blackwellization (Robert and Casella 2004), Riemann sums (Philippe and Robert 2001), and a plethora of MCMC schemes that aim to improve mixing (e.g., Girolami and Calderhead 2011). These methods could all be used to reduce the variance of estimators for model evidence that are based on computing Monte Carlo expectations. In this article, we extend the zero-variance (ZV) control variate technique, introduced in the physics literature by Assaraf and Caffarel (1999), to TI estima-
tors of model evidence. The methodology applies whenever the gradient of the log-likelihood (and the log-prior) can be evaluated and therefore can be used "for free" when differential geometric sampling schemes are employed in construction of the Markov chain (Papamarkou, Mira, and Girolami 2014). Theoretical results are provided that guide maximal variance reduction in practice. Results on popular benchmark datasets demonstrate a substantial reduction in variance compared to existing estimators and the method is shown to be exact in the special case of Bayesian linear regression.

The article proceeds as follows: Section 2 recalls key ideas from TI and ZV that we use in our methodology. In Section 3, we derive control variates for TI and provide theoretical results that guide maximal variance reduction in practice. Section 4 compares the proposed methodology to the state-of-the-art estimators of model evidence applied to popular benchmark datasets. Section 5 investigates scenarios where the proposed methodology is likely to fail. Finally, Section 6 provides more general evidence, drawing an important distinction between "equilibrium" and "nonequilibrium" estimators that determines whether or not control variates may be applicable.

2. Background

2.1 Thermodynamic Integration

Path sampling and the closely related technique of TI emerged from the physics community as a computational approach to compute normalizing constants (Gelman and Meng 1998). Recent empirical investigations, including Vyshemirsky and Girolami (2008) and Friell and Wyse (2012), have revealed that TI is among the most promising approach to estimation of model evidence. Below we provide relevant background on TI, referring the reader to Calderhead and Girolami (2009) for further details.

TI targets the model evidence directly; in what follows we therefore implicitly condition upon a model $m$ and aim to compute the evidence $p(y) = p(y|m)$ provided by data $y$ in favor of model $m$. Following the presentation of Friell and Pettitt (2008), the power posterior is defined as $p(\theta|y, t) = p(\theta|y)^t / Z_t(y)$, where the normalizing constant is given by $Z_t(y) = \int p(\theta|y)^t p(\theta) d\theta$. Here, $t$ is known as an inverse temperature parameter and by analogy the process of increasing $t$ is known as annealing. Note that $p(\theta|y, t = 0)$ is the density of the prior distribution, whereas $p(\theta|y, t = 1)$ is the density $p(\theta|y)$ of the posterior distribution. Varying $t \in (0, 1)$ produces a continuous path between these two distributions and in this article it is assumed that all intermediate distributions exist and are well defined. The normalizing constant $Z_t(y)$ is equal to one and $Z_1(y)$ is equal to $p(y)$, the model evidence that we aim to estimate.

The standard thermodynamic identity is

$$
\log(p(y)) = \int_0^1 \mathbb{E}_{\theta|y, t} \log(p(\theta|y)) dt,
$$

where the expectation in the integrand is with respect to the power posterior whose density is given above. The correctness of Equation (3) is established in, for example, Friel and Pettitt (2008). In TI, this one-dimensional integral is evaluated numerically using a quadrature approximation over a discrete temperature ladder, whereas in the related approach of path sampling this integral is evaluated using MCMC. Note that the use of quadrature methods introduces bias into the estimator of model (log-)evidence; it is therefore important to select an accurate quadrature approximation (Appendix A).

2.2 Control Variates and The ZV Technique

Control variates are often employed when we aim to estimate, with reduced variance, the expectation $\mathbb{E}_\pi[g(\theta)]$ of a function $g(\theta)$ of a random variable $\theta$ that is distributed according to a (possibly unnormalized) density $\pi(\theta)$. In this article, we focus on real-valued $\theta \in \Theta \subseteq \mathbb{R}^d$ and we aim to approximate

$$
\mathbb{E}_\pi[g(\theta)] \approx \frac{\int g(\theta) \cdot \pi(\theta) d\theta}{\int \pi(\theta) d\theta}.
$$

The generic control variate principle relies on constructing an auxiliary function $\tilde{g}(\theta) = g(\theta) + h(\theta)$ that satisfies $\mathbb{E}_\pi[h(\theta)] = 0$ and so $\mathbb{E}_\pi[\tilde{g}(\theta)] = \mathbb{E}_\pi[g(\theta)]$. Write $\nabla \pi[\tilde{g}(\theta)]$ for the variance of the function $g(\theta)$ of a random variable $\theta$ whose (unnormalized) density is $\pi(\theta)$. In many cases, it is possible to choose $h(\theta)$ such that $\nabla \pi[\tilde{g}(\theta)] < \nabla \pi[\pi(\theta)]$, leading to a reduction in Monte Carlo variance. Intuitively, greater variance reduction can occur when $h(\theta)$ is negatively correlated with $g(\theta)$ under $\pi(\theta)$, since much of the randomness "cancels out" in the auxiliary function $\tilde{g}(\theta)$. In classical literature, $h(\theta)$ is formed as a sum $\phi_1 h_1(\theta) + \cdots + \phi_m h_m(\theta)$, where the $h_i(\theta)$ have zero mean under $\pi(\theta)$ and are known as control variates, while $\phi_i$ are coefficients that must be specified. For estimation based on Markov chains, Andradóttir, Heyman, and Ott (1993) proposed control variates for discrete state spaces. Later Mira, Tencon, and Bressanini (2003) extended this approach to continuous state spaces, observing that the optimal choice of $h(\theta)$ is intimately associated with the solution of the Poisson equation $h(\theta) = \mathbb{E}_\pi[\tilde{g}(\theta)] - g(\theta)$ and proposing to solve this equation numerically. Further work on constructing control variates for Markov chains includes Hammer and Tjelmeland (2008) for Metropolis–Hastings chains and Dellaportas and Kontoyiannis (2012) for Gibbs samplers.

In this article, we consider the particularly tractable class of ZV control variates that are expressed as functions of the gradient $\nabla g(\theta)$ of the log-target density (i.e., the score function). Here $\nabla \theta = [\partial / \partial \theta_1, \ldots, \partial / \partial \theta_d]^T$. More specifically, Mira, Solgi, and Imparato (2013) proposed to use

$$
h(\theta) = \Delta_g[Q(\theta)] + \nabla \theta [Q(\theta)] \cdot u(\theta),
$$

where $\Delta_g = \nabla \theta \cdot \nabla \theta$ is the Laplacian operator, the trial function $Q(\theta)$ is a polynomial in $\theta$ and

$$
u(\theta) = \nabla g(\theta)
$$

is the score function. (In Mira, Solgi, and Imparato (2013) an unnecessary constant of proportionality was used, that we omit here.) In this article, we adopt the convention that both $\theta$ and $u(\theta)$ are $d \times 1$ vectors. The trial function $Q(\theta)$ should be viewed as approximating the function of interest $g(\theta)$ based on the value of the score $u(\theta)$. An optimal choice for the trial function $Q$ is
given by the solution of the Poisson equation

\[ \Delta_0[Q(\theta)] + \nabla_\theta[Q(\theta)] \cdot u(\theta) + g(\theta) = \mathbb{E}_\theta[g(\theta)]. \]  

For arbitrary functions \( g(\theta) \) and densities \( \pi(\theta) \), the solution of Equation (7) will not be available (since the right-hand side is unknown), but we may be able to obtain a good polynomial approximation to the solution; this is the basis of the ZV methodology.

The thermodynamic identity (Equation (3)) is based on expected values of log-likelihoods \( \log(\pi(\theta)) \). Since \( u(\theta) \) is closely related to \( \log(\pi(\theta)) \), ZV control variates appear as a natural strategy to achieve variance reduction in TI. As shown in Mira, Solgi, and Imparato (2013), ZV control variates arise closely related to \( \log(\pi(\theta)) \). \( \mathcal{ZV} \) has been applied essentially “for free” inside differential-geometric MCMC sampling schemes for which the score function is a prerequisite for sampling (Girolami and Calderhead 2011).

3. Methodology

In Section 3.1, we develop a control variate scheme for the estimation of model evidence, taking TI as our base estimator whose variance we propose to reduce. The main methodological challenge in this setting is the elicitation of both the optimal control variate coefficients \( \phi \) and the optimal temperature ladder that underlies TI. In Section 3.2, we derive optimal expressions for both these quantities and in Section 3.3 we describe how coefficients and temperature ladders are selected in practice.

3.1 The Controlled Thermodynamic Integral

Taking the target density \( \pi(\theta) \) to be the power posterior \( p(\theta|y, t) \), it follows from Equation (6) that

\[ u(\theta|y, t) = t \frac{\nabla_\theta p(y|\theta)}{p(y|\theta)} + \frac{\Delta_0[Q(\theta \phi(y, t))] \cdot u(\theta|y, t)}{\mathbb{E}_\theta[Q(\phi(y, t))]}. \]  

The ZV control variates (Equation (5)) are then

\[ h(\theta|y, t) = \Delta_0[Q(\theta \phi(y, t))] \cdot u(\theta|y, t), \]

where \( u(\theta|y, t) \) is as defined in Equation (8). Here, the notation \( Q(\theta \phi(y, t)) \) indicates that the polynomial \( Q \) has coefficients \( \phi \equiv \phi(y, t) \) that will in general depend on both the data \( y \) and inverse temperature \( t \). Integrating these control variates into TI, we obtain the “controlled thermodynamic integral” (CTI)

\[ \log(p(y)) = \int_0^1 \mathbb{E}_{\theta|y, t} [\log(p(y|\theta))] + h(\theta|y, t)]dt. \]

To use CTI to estimate the model (log-)evidence, we need to specify both (i) polynomial coefficients \( \phi(y, t) \) and (ii) an appropriate discretization \( 0 = t_0 < t_1 < \cdots < t_m = 1 \) (the temperature ladder) of the one-dimensional integral. Specification of both polynomial coefficients and temperature ladder should be targeted at minimizing the variance of CTI (see below).

3.2 Optimal Coefficients and Ladders

We derive the jointly optimal, variance-minimizing, polynomial coefficients, and temperature ladder. For the latter, note that there is a surjective mapping from partitions \( 0 = t_0 < t_1 < \cdots < t_m = 1 \) to probability distributions on \([0, 1]\) with density function \( p(t) \) that is given by \( \int_0^t p(s)ds = \frac{t}{m} \). For the development below, it is convenient to focus on optimizing the density \( p(t) \), mapping back to the temperature ladder during implementation (see Section 3.3). For clarity of the exposition, write \( \ell(\theta) = \log(p(y|\theta)) \) where we temporarily suppress dependence on both data \( y \) and model \( m \). The CTI identity can be rewritten as

\[ \log(p(y)) = \int_0^1 \mathbb{E}_{\theta|y, t} [\ell(\theta) + h(\theta|t)]dt \]

where the final expectation is taken with respect to the distribution with density \( p(\theta, t|y) = p(\theta|y, t)p(t) \). Under an approximation that Monte Carlo samples are obtained independently, so-called “perfect transitions,” the variance of the estimator of model (log-)evidence is given by

\[ \frac{1}{N} \int_0^1 \mathbb{E}_{\theta|y, t} [(\ell(\theta) + h(\theta|t))^2]dt - [\log(p(y))]^2 \],

where \( N \) is the number of Monte Carlo samples.

The optimal choice of polynomial coefficients \( \phi(t) \) and temperature ladder \( p(t) \) are defined as the pair that jointly minimize Equation (12). Specifically, we seek to minimize the Lagrangian

\[ \int_0^1 \mathbb{E}_{\theta|y, t} [(\ell(\theta) + h(\theta|t))^2]dt + \lambda \int_0^1 p(t) \]  

over \( (p, \phi) : [0, 1] \rightarrow \mathbb{R}^{r+1} \), where \( e \) is the dimension of \( \phi \) and depends on the degree of the polynomial \( Q(\theta|\phi) \) that is being employed. Here, \( \lambda \) is a Lagrange multiplier that will be used to ensure \( \int p(t)dt = 1 \). Below we consider degree 1 polynomials \( Q(\theta|\phi) = \theta^T \phi \) so that \( h(\theta|t) = \phi(t)^T u(\theta|t) \) but the derivation applies analogously to higher degree polynomials, as explained in Appendix C. The solution \( (p^*, \phi^*) \) of the Lagrangian optimization problem (Equation (13)) is

\[ \phi^*(t) = -\nabla_{\theta|y, t}^1[u(\theta)] \mathbb{E}_{\theta|y, t}[\ell(\theta)u(\theta)] \]

\[ p^*(t) \propto \sqrt{\mathbb{E}_{\theta|y, t}[\ell(\theta)^2] - \mathbb{E}_{\theta|y, t}[\ell(\theta)u(\theta)]^2} \sqrt{\mathbb{E}_{\theta|y, t}[u(\theta)]^2} \mathbb{E}_{\theta|y, t}[\ell(\theta)u(\theta)]. \]

where \( \nabla_{\theta|y, t}^1[u(\theta)] \) and \( \mathbb{E}_{\theta|y, t}[\ell(\theta)u(\theta)] \) denote, respectively, variance and cross-covariance matrices (since \( \mathbb{E}_{\theta|y, t}[u(\theta)] = 0 \)). Notice that the optimal temperature ladder for CTI is not the same as the optimal ladder for standard TI, which is given by \( p^*(t) \propto \sqrt{\mathbb{E}_{\theta|y, t}[\ell(\theta)^2]} \) (Calderhead and Girolami 2009).
It can be shown (Rubinstein and Marcus 1985) that this choice of polynomial coefficients $\phi = \phi^*$ is characterized as the minimizer of the variance ratio

$$R(t) := \frac{\nabla_{\theta_i}[\ell(\theta) + \phi(t)^T u(\theta)]}{\nabla_{\theta_i}[\ell(\theta)]}$$

(16)

and at this minimum

$$R(t) = 1 - \text{Corr}_{\theta_i} \left[ \ell(\theta), \phi^T u(\theta) \right]^2,$$

(17)

so that greater variance reduction is expected in the case where a linear combination of the elements of the vector $u(\theta)$ is highly correlated with the target function $\ell(\theta)$.

### 3.3 Implementation

For most models of interest both Equations (14) and (15) do not possess closed-form expressions and it becomes necessary to employ estimates or approximations to the optimal values. We begin by noting that Equation (14) actually defines the optimal, variance-minimizing, coefficients independently of the choice of temperature ladder $p(t)$; this is directly verified from the Euler–Lagrange equations applied to $\phi : [0, 1] \rightarrow \mathbb{R}^s$, where $p(t)$ is held fixed. This observation allows us to discuss these two aspects of the implementation separately.

#### 3.3.1 Polynomial Coefficients

Optimal coefficients for control variates are typically estimated based on the same sequence of MCMC samples that will subsequently be used to compute the controlled expectations (Robert and Casella 2004). Specifically, to estimate the optimal control variate coefficients $\phi^*(t)$ we exploit MCMC samples to estimate both the covariance $\nabla_{\theta_i}[u(\theta)]$ and the cross-covariance $\nabla_{\theta_i}[\ell(\theta) u(\theta)]$. These estimates are then plugged directly into Eq. (14) to obtain an estimate

$$\phi^*(t) \approx -\tilde{\nabla}_{\theta_i}[u(\theta)]^{-1} \tilde{\nabla}_{\theta_i}[\ell(\theta) u(\theta)]$$

(18)

for the optimal coefficients. Further discussion of “plug-in” estimators for control coefficients can be found in Dellaportas and Kontoyiannis (2012).

#### 3.3.2 Temperature Ladder

For estimating the optimal temperature ladder of Equation (15), one obvious numerical approach would be to first estimate $p^*(t)$ up to proportionality over a uniform grid $\{t_i\}$, using a preliminary MCMC run to estimate both $\nabla_{\theta_i}[\ell(\theta)^2]$ and the covariance and cross-covariance matrices $\nabla_{\theta_i}[u(\theta)]$ and $\nabla_{\theta_i}[\ell(\theta) u(\theta)]$. Then nonparametric density estimation could be applied to obtain an estimate for the optimal ladder $\{t_i\}$. However this two-step procedure is computationally burdensome. Neal (1996) showed that a geometric temperature ladder is optimal for annealing on the scale parameter of a Gaussian and Behrens, Friel, and Hurn (2012) extended this result to target distributions of the same form as $\ell(\theta)$, which includes Gaussians. In this article, we fix a quintic temperature ladder $t_i = (i/50)^5$ for use in all applications; this ladder is widely used in the TI literature and has demonstrated strong performance in empirical studies (e.g., Calderhead and Girolami 2009; Friel, Hurn, and Wyse 2014a). The question of how to select appropriate temperature ladders in practice is an ongoing area of research and the recent contributions of Miasojedow, Moulines, and Vihola (2012), Behrens, Friel, and Hurn (2012), Zhou, Johansen, and Aston (2013), and Friel, Hurn, and Wyse (2014a) are compatible with our methodology.

### 3.3.3 Quadrature

The second-order quadrature method of Friel, Hurn, and Wyse (2014a), described in Appendix A, requires us also to estimate the variance $\nabla_{\theta_i}[\ell(\theta)]$ at each step in the temperature ladder. In experiments below, we use ZV control variates to estimate this variance, using the identity

$$\nabla_{\theta_i}[\ell(\theta)] = \mathbb{E}_{\theta_i}[\ell(\theta) - \mathbb{E}_{\theta_i}[\ell(\theta)]]^2$$

(19)

and applying control variates in the estimation of each of these expectations.

### 4. Applications

We present several empirical studies that compare CTI to the state-of-the-art TI estimators of Friel, Hurn, and Wyse (2014a). In all applications below, we base estimation on the output of a population MCMC sampler ( Jasra, Stephens, and Holmes 2007) limited to $N$ iterations at each of the 51 rungs of the temperature ladder (a total of $51 \times N$ evaluations of the likelihood function). In brief, the within-temperature proposal was provided by the manifold Metropolis-adjusted Langevin algorithm (mMALA) of Girolami and Calderhead (2011), while the between-temperature proposal randomly chooses a pair of (inverse) temperatures $t_i$ and $t_j$, proposing to swap their state vectors with probability given by the Metropolis–Hastings ratio (Calderhead and Girolami 2009). To ensure fairness, the same samples were used as the basis for all estimators of model evidence, ensuring that all estimators require essentially the same amount of computation (since the score function is computed as a matter of course in mMALA). Moreover, to explore the statistical properties of the estimators themselves, we generated 100 independent realizations of the population MCMC and thus 100 realizations of each estimator. Full details are provided in the supplementary materials.

#### 4.1 Bayesian Linear Regression

##### 4.1.1 Known Precision

We begin with an analytically tractable problem in Bayesian linear regression. The (log-)likelihood function is given by

$$\log p(y|X, \beta, \sigma) = -\frac{n}{2} \log(2\pi \sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta),$$

(20)

where $y$ is $n \times 1$, $X$ is $n \times d$, and $\beta$ is $d \times 1$. In simulations below, we took $\sigma = 1$, $d = 3$, $\beta = [0, 1, 2]$. The design matrix $X$ was populated with $n = 100$ rows by drawing each entry independently from the standard normal distribution and then data $y$ were generated from $N(X\beta, \sigma^2 I_{n \times n})$; both $X$ and $y$ were then fixed for all experiments below. From the Bayesian perspective, we take a conjugate prior $\beta \sim N(0, \zeta^2 I_{d \times d})$ with $\zeta = 1$. 
In this section, we assume $\sigma$ is fixed and known, but we relax this assumption in the next section. Thus, the unknown model parameters here are $\theta = \beta \in \mathbb{R}^d$ and we aim to compute the evidence $p(y)$ by marginalizing over these parameters. This example is an ideal benchmark since it is possible to obtain many relevant quantities in closed form; see Appendix D.1 for full details.

Before applying CTI, we are required to check that the sufficient conditions for the unbiasedness of ZV estimators are satisfied (see Appendix B). This amounts to noticing that the tails of the power posterior $p(\beta|y, t)$ decay exponentially in $\beta$ (Appendix D.1). Using the plug-in estimates (Equation (18)), we obtain estimates for the optimal coefficients $\phi^*$, which are shown in supplementary Figure S6. For degree 2 polynomials, we see that the plug-in estimator is deterministic. Indeed, by direct calculation we see that $u(\beta|y, t)$ is an invertible affine transformation of the parameter vector

$$u(\beta|y, t) = \frac{1}{\sigma^2} X^T y - \Sigma(t)^{-1} \beta,$$

where $\Sigma(t) = (\frac{1}{\sigma^2} X^T X + \frac{1}{t^2} I)^{-1}$. This allows us to intuit that CTI based on degree 2 polynomials will produce an exact estimate of the (log-)evidence (up to quadrature error), as we explain below. Indeed, by another invertible affine transformation, we can map $u(\beta|y, t) \mapsto y - X \beta$, which, when multiplied by the polynomial $Q(\beta|\phi) = (y - X \beta)^T$ produces a quantity $(y - X \beta)^T (y - X \beta)$ that is perfectly correlated with the log-likelihood under the power posterior. It then follows from Equation (16) that CTI will possess zero variance. This argument is made rigorous in the supplementary materials.

In supplementary Figure S7, we plot 100 independent estimates of the integrand $\mathbb{E}_{\beta|y_t, [\ell(\beta)]}$ at each of the 51 temperatures in the ladder for polynomial trial functions of degree 0 (i.e., standard TI), 1, and 2. It is apparent that estimator variance is greatest at lower values of $t$; this motivates the heavily skewed temperature ladder used by ourselves and others, as we wish to target our computational effort on this high-variance region. We quantify the reduction in estimator variance at an (inverse) temperature $t$ using the variance ratio $R(t)$ as estimated from the MCMC samples. Figure 1 shows that degree 1 polynomials achieve (on average) variance reduction at all temperatures, with the greatest reduction occurring in the region where $t$ is small. This is encouraging as the region where $t$ is small is most important for variance reduction of TI, as discussed above. For degree 2 polynomials, we have $R(t) = 0$ for all $t$, which recapitulates the exactness of the CTI estimator in this example.

Finally, we explore the quality of the estimators of model evidence themselves. For this model (the log-)evidence is available in closed form (Appendix D.1) and this allows us to compute the mean square error (MSE) over all 100 independent realizations of each estimator. Results, shown in Table 1, demonstrate that CTI with degree 2 polynomials achieves a two-fold reduction in MSE compared to standard TI when both estimators are based on first-order quadrature. However, first-order quadrature is known to lead to significant estimator bias (Friel, Hurn, and Wyse 2014a) and when estimators are based instead on more accurate second-order quadrature, CTI is seen to be approximately 10,000× better that TI in terms of MSE; a dramatic difference. We also compared TI approaches against annealed importance sampling (AIS; Neal 2001), as described in the supplementary materials. In this case, CTI (degree 2) is over 10,000× more accurate compared to AIS (supplementary Figure S9(a)).

### 4.1.2 Unknown Precision (Radiata Pine)

We now relax the assumption of known precision $\tau = 1/\sigma^2$; we will see that in these circumstances CTI is no longer exact. Specifically, we consider data from Williams (1959) on $n = 42$ samples. We quantify the reduction in estimator variance at an (inverse) temperature $t$ using the variance ratio $R(t)$ as estimated from the MCMC samples. Figure 1 shows that degree 1 polynomials achieve (on average) variance reduction at all temperatures, with the greatest reduction occurring in the region where $t$ is small. This is encouraging as the region where $t$ is small is most important for variance reduction of TI, as discussed above. For degree 2 polynomials, we have $R(t) = 0$ for all $t$, which recapitulates the exactness of the CTI estimator in this example.

Finally, we explore the quality of the estimators of model evidence themselves. For this model (the log-)evidence is available in closed form (Appendix D.1) and this allows us to compute the mean square error (MSE) over all 100 independent realizations of each estimator. Results, shown in Table 1, demonstrate that CTI with degree 2 polynomials achieves a two-fold reduction in MSE compared to standard TI when both estimators are based on first-order quadrature. However, first-order quadrature is known to lead to significant estimator bias (Friel, Hurn, and Wyse 2014a) and when estimators are based instead on more accurate second-order quadrature, CTI is seen to be approximately 10,000× better that TI in terms of MSE; a dramatic difference. We also compared TI approaches against annealed importance sampling (AIS; Neal 2001), as described in the supplementary materials. In this case, CTI (degree 2) is over 10,000× more accurate compared to AIS (supplementary Figure S9(a)).

### Table 1. Bayesian linear regression with a) known precision and b) unknown precision.

| Precision | $N$ | deg(Q) | Quadr. | MSE | S.E. |
|-----------|----|--------|--------|-----|-----|
| a) Known  | 1e3| 0      | 1      | 2.3e-2 | 2.9e-3 |
|           |    | 2      | 2.1e-2 | 2.5e-3 |      |
|           |    | 1      | 1.8e-2 | 2.1e-3 |      |
|           |    | 2      | 2.0e-2 | 2.5e-3 |      |
|           | 5e3| 0      | 1      | 5.2e-3 | 7.2e-4 |
|           |    | 2      | 4.0e-3 | 5.9e-4 |      |
|           |    | 1      | 3.8e-3 | 6.0e-4 |      |
|           |    | 2      | 3.4e-3 | 5.3e-4 |      |
|           |    | 2      | 1.2e-3 | 0      |      |
|           |    | 2      | 2.0e-7 | 2.7e-8 |      |
| b) Unknown| 1e3| 0      | 1      | 7.9e-3 | 1.2e-3 |
|           |    | 2      | 7.7e-3 | 1.1e-3 |      |
|           |    | 1      | 7.6e-3 | 1.0e-3 |      |
|           |    | 2      | 7.6e-3 | 1.0e-3 |      |
|           | 5e3| 0      | 1      | 1.4e-5 | 2.0e-6 |
|           |    | 2      | 1.3e-5 | 1.6e-6 |      |
|           |    | 2      | 1.4e-3 | 1.8e-4 |      |
|           |    | 2      | 1.3e-3 | 1.8e-4 |      |
|           |    | 1      | 1.4e-3 | 2.0e-4 |      |
|           |    | 2      | 2.4e-6 | 3.0e-7 |      |
|           |    | 2      | 1.5e-6 | 2.0e-7 |      |

**NOTE:** The best performing estimators are highlighted in bold.
specimens of *radiata* pine. This dataset is well known in the multivariate statistics literature and was recently used by Friel and Wyse (2012) and Friel, Hurn, and Wyse (2014a) to benchmark estimators of model evidence. Data consist of the maximum compression strength parallel to the grain $y$, as a function of density $x$, and density adjusted for resin content $z$. Both density $x$ and adjusted-density $z$ have comparable standard deviations of 4.5445 and 4.6593, respectively. It is wished to determine whether the density or resin-adjusted density is a better predictor of compression strength parallel to the grain. Following Friel, Hurn, and Wyse (2014a), we consider Bayesian model comparison between a pair of competing models:

$$\text{Model 1: } y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \sim N(0, \tau^{-1}) \quad (22)$$

$$\text{Model 2: } y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i, \quad \eta_i \sim N(0, \lambda^{-1}) \quad (23)$$

Here $\bar{x}$ and $\bar{z}$ are the sample means of the $x_i$ and $z_i$, respectively. The priors for $(\alpha, \beta)$ and $(\gamma, \delta)$ are both Gaussian with common mean $B_0 = [3000, 185]^T$ and precisions $\tau Q_0, \lambda Q_0$ where $Q_0 = \text{diag}(0.06, 6)$. Both $\tau$ and $\lambda$ were assigned gamma priors with shape 6 and rate 4 $\times$ 3002. To compare between these models we consider estimates for the log-Bayes factor $\log(B_{21})$ that are obtained as the difference between independent estimates for the log-evidence of each model.

This application is interesting for two reasons: First, one can directly calculate the log-Bayes factor for this example as $\log(B_{21}) = 8.7086$, so that we have a gold standard performance benchmark. Second, when the precision $\tau$ (or $\lambda$) is unknown, ZV methods are no longer exact. We therefore have an opportunity to assess the performance of CTI in a nontrivial setting.

Formulas in Appendix D.2 demonstrate that the sufficient condition for unbiasedness of ZV methods is satisfied. Results in Figure 2 show that CTI (degree 1) achieves a modest reduction in variance across temperatures $t$, whereas CTI (degree 2) achieves a massive variance reduction. Computing the MSE relative to the true log-Bayes factor, we see that CTI (degree 2) is over 500 $\times$ more accurate compared to TI, though the variance of the estimator is not identically equal to zero in this case (Table 1).

As before, MSE is further reduced as a result of applying second-order quadrature. AIS performed slightly worse than the methods based on TI in this example (supplementary Figure S9(b)).

### 4.2 Bayesian Logistic Regression (Pima Indians)

Here, we examine data that contain instances of diabetes and a range of possible diabetes indicators for $n = 532$ women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona. This dataset is frequently used as a benchmark for supervised learning methods (e.g., Marin and Robert 2011). Friel, Hurn, and Wyse (2014a) considered five predictors of diabetes recorded for this group; number of pregnancies (NP); plasma glucose concentration (PGC); body mass index (BMI); diabetes pedigree function (DP), and age (AGE). Diabetes incidence $y_i$ in person $i$ is modeled by the binomial likelihood

$$p(y_i | \beta) = \prod_{j=1}^{n} p_i^y (1 - p_i)^{1-y_i}, \quad (24)$$

where the probability of incidence $p_i$ for person $i$ is related to the covariates $x_i = (1, x_{i1}, \ldots, x_{id})^T$ and the parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_d)$ by

$$\logit(p_i) = \log \left( \frac{p_i}{1 - p_i} \right) = x_i \cdot \beta. \quad (25)$$

Consider Bayesian comparison between the two candidate models

Model 1: $\logit(p) = \beta_0 + \beta_1 \cdot \text{NP} + \beta_2 \cdot \text{PGC}$

$$+ \beta_3 \cdot \text{BMI} + \beta_4 \cdot \text{DP} \quad (26)$$

Model 2: $\logit(p) = \beta_0 + \beta_1 \cdot \text{NP} + \beta_2 \cdot \text{PGC} + \beta_3 \cdot \text{BMI}$

$$+ \beta_4 \cdot \text{DP} + \beta_5 \cdot \text{AGE} \quad (27)$$

subject to the common prior belief $\beta \sim N(0, \tau^{-1} I)$. Following Friel and Wyse (2012), we set $\tau = 0.01$.

The unbiasedness criterion in Appendix B is seen to be satisfied and we have

$$u(\beta | y, t) = t X^T (y - p) - \tau \beta, \quad (28)$$

where the $i$th row of $X$ is $x_i \cdot$. In Figure 3, we see that degree 1 ZV methods achieve a greater variance reduction at smaller $t$, but moreover we see that degree 2 ZV methods continue to achieve a substantial variance reduction at all temperatures. In Table 2, we display the mean of each estimator, computed over all 100 independent runs of population MCMC, together with the
standard deviation of this collection of estimates. We see that this variance reduction transfers to estimates of the Bayes factor themselves, where the standard deviation of the CTI estimators is approximately 20 × lower compared to estimators based on TI. Although no exact expression is available for log(B21), Friel, Hurn, and Wyse (2014a) computed the log-evidence for both models using an extended run of 2000 temperatures and N = 20,000 iterations within standard TI. Their estimates were −257.2342 and −259.8519, respectively, for Models 1 and 2, corresponding to an estimate of the log-Bayes factor of log(B21) = −2.6177. This estimate, obtained at considerable computational expense, closely matches the estimates obtained by CTI (degree 2), which is based on 800 × fewer evaluations of the likelihood function. AIS performs comparably with standard TI in this example (supplementary Figure 9(c)).

Table 2. Estimates of the log-Bayes factor log(B21), based on 100 independent runs of population MCMC. (a) Bayesian logistic regression. The actual log-Bayes factor, as computed by Friel, Hurn, and Wyse (2014a), is log(B21) = −2.6177. (b) Nonlinear ODEs: Estimates of the log-Bayes factor log(B21), based on 10 independent runs of population MCMC. deg(Q) is the dimension of the ZV polynomial Q(θ), with 0 denoting standard TI. Quadr. is the order of numerical quadrature scheme. N is the number of MCMC iterations. Mean log-BF and SD are the mean and standard deviation of the estimated log-Bayes factors.

| Model               | N   | deg(Q) | Quadr. | Mean B.F. | S.D.   |
|---------------------|-----|--------|--------|-----------|--------|
| (a) Logistic regression | 1e3 |       |        |          |        |
|                     | 0   | 1      | −2.59  | 0.74      |        |
|                     | 1   | 1      | −2.44  | 0.70      |        |
|                     | 1   | 2      | −2.42  | 0.69      |        |
|                     | 2   | 1      | −2.62  | 0.050     |        |
|                     | 2   | 2      | −2.61  | 0.044     |        |
| (b) Nonlinear ODEs  | 5e3 |       |        |          |        |
|                     | 0   | 1      | −2.62  | 0.35      |        |
|                     | 1   | 1      | −2.58  | 0.35      |        |
|                     | 1   | 2      | −2.56  | 0.34      |        |
|                     | 2   | 1      | −2.64  | 0.016     |        |
|                     | 2   | 2      | −2.62  | 0.016     |        |
| (b) Nonlinear ODEs  | 1e3 |       |        |          |        |
|                     | 0   | 1      | −3.75  | 0.31      |        |
|                     | 1   | 1      | −3.69  | 0.31      |        |
|                     | 1   | 2      | −3.69  | 0.31      |        |
|                     | 2   | 1      | −3.57  | 0.27      |        |
|                     | 2   | 2      | −3.56  | 0.27      |        |

NOTE: The best performing estimators are highlighted in bold.

5. Limitations of CTI

We have demonstrated, using standard benchmark datasets, that CTI is well suited to Bayesian model comparison between regression models. Regression analyses continue to be widely applicable in disciplines such as econometrics, epidemiology, political science, psychology, and sociology, so that these findings have significant implications. Nevertheless in many disciplines such as engineering, geophysics, and systems biology, statistical models are significantly more complex, often based on a mechanistic understanding of the underlying process. Below we provide such an example and find that CTI offers little improvement over TI; this allows us to explore the limitations of our approach and, conversely, to understand in what circumstances it is likely to be successful.

5.1 A Negative Example (Goodwin Oscillator)

We consider nonlinear dynamical systems of the form

\[
\frac{dx}{ds} = f(x, s; \theta), \quad x(0) = x_0. \tag{29}
\]

We assume only a subset of the variables is observed under noise, so that \( x = [x_1, \ldots, x_d] \) and \( y \) is a \( d \times n \) matrix of observations of the variables \( x_d \). Model comparison for systems specified by nonlinear differential equations is known to be profoundly challenging, even when the number of models is small (Calderhead and Girolami 2011).

Write \( s_1 < s_2 < \cdots < s_n \) for the times at which observations are obtained, such that \( y(s_j) = y_j \). We consider a Gaussian observation process with likelihood

\[
p(y | \theta, x_0, \sigma) = \prod_{j=1}^{n} N(y(s_j) | x_d(s_j; \theta, x_0), \sigma^{-1} I), \tag{30}
\]

where \( x_d(s_j; \theta, x_0) \) denotes the solution of the system in Equation (29). For the Gaussian observation model, it can be shown that a sufficient condition for unbiasedness of ZV is that the prior density \( p(\theta) \) vanishes faster than \( r^{\dim \theta + k - 2} \) when \( r = \| \theta \|_1 \rightarrow \infty \). Here \( \dim \Theta = d \) and \( k = 1 \) is the degree of the polynomial that is being employed (see Appendix B).
Assuming the sufficient condition for \( ZV \) is satisfied, we have

\[
u_i(\theta) = \frac{t}{\sigma^2} \sum_{j=1}^{N} S_j \cdot \sum_{k=1}^{\text{dim } x_a} \left( y(s_j) - x_a(s_j; \theta, x_0) \right) + \nabla \log(p(\theta)),
\]

where \( S \) is a matrix of sensitivities with entries \( S_{j,k} = \frac{\partial y(s_j)}{\partial \theta_k} \). Note that in Equation (31), \( S_{j,k} \) ranges over indices \( 1 \leq k \leq \text{dim } x_a \) corresponding only to the observed variables. In general, the sensitivities \( S \) will be unavailable in closed form, but may be computed numerically by augmenting the system of ordinary differential equations (ODEs) in Equation (29) as described in Appendix E. Indeed, these sensitivities are already computed when differential-geometric sampling schemes are employed, so that the evaluation of Equation (31) incurs negligible computational cost.

We focus on a dynamical model of oscillatory enzymatic control due to Goodwin (1965), which was recently considered in the context of Bayesian model comparison by Calderhead and Girolami (2009). This kinetic model, specified by a system of 9 ODEs, describes how a negative feedback loop between protein expression and mRNA transcription can induce oscillatory dynamics as experimentally observed in circadian regulation (Locke, Millar, and Turner 2005). A full specification is provided in Appendix E. As shown in Calderhead and Girolami (2009), the Goodwin oscillator induces a highly multi-modal posterior distribution that renders estimation of the model evidence extremely challenging. We consider Bayesian comparison of two models; a simple model with one intermediate protein species (\( g = 3 \)) and a more complex model with two intermediate protein species (\( g = 4 \)).

Figure 4 demonstrates that in this extremely challenging example the benefits of control variate schemes that we have previously observed are heavily reduced. Since the variance ratio \( R(t) \) is related to the canonical correlation between control variates and the log-likelihood under the power posterior (Equation (17)), we hypothesize that the extreme multi-modality of the power posterior distribution is limiting the extent to which strong canonical correlation can be achieved. This is confirmed in Figure 5 where we plot values of the target function \( \ell(\theta) \) against the control variates \( h(\theta) \) that are obtained from MCMC sampling in the posterior. We observe much reduced correlation in the case of the Goodwin oscillator that is a consequence of the complex nature of the likelihood surface. Turning to the Bayes factor itself, in Table 2 we display the mean of each estimator of the log-Bayes factor, together with the standard deviation of this collection of estimates. We find that CTI (degree 1) provides negligible reduction in variance and CTI (degree 2) provides an insignificant 15% reduction in variance.

In this example, AIS consistently produced lower estimates for log-Bayes factors (supplementary Figure S9(d)). This likely reflects the low number \( N \) of Monte Carlo iterations that are characteristic of such computationally demanding applications.

6. Discussion

To the best of our knowledge, this is the first article to consider the use of control variates for the purpose of estimating the model evidence. Motivated by previous empirical studies, we focused on TI estimators for the model (log-)evidence. However, in general, control variate techniques could be leveraged whenever estimators of the evidence (or Bayes factors) take the form of a Monte Carlo expectation. General control variate schemes for MCMC rely on the fact that the expectation of the control variates along the MCMC sample path will be approximately zero. We thus draw a distinction between "equilibrium" Monte Carlo estimators for the model evidence, such as TI and path sampling, which require the underlying Markov chain to have converged, and "nonequilibrium" estimators such as AIS and sequential Monte Carlo that do not require convergence. The former class are amenable to existing control variate schemes whereas the latter are not. This motivates the "equilibration" of these nonequilibrium estimators.

Given its close connection with TI (Gelman and Meng 1998), we considered whether an equilibrated version of AIS, which
Figure 5. Comparing the likelihood surfaces and canonical correlations of different models. (Log-likelihood surfaces (top) for the (a) Radiata Pine and (b) Pima Indians examples can be well approximated by a Gaussian and induce strong canonical correlation (bottom) between the (degree 2) control variates \( h(\theta) \) and the log-likelihood \( l(\theta) \) in the posterior. On the other hand, the log-likelihood surface for (c) Goodwin Oscillator is highly multi-modal and there is much weaker canonical correlation between the control variates and the log-likelihood.)

jointly samples from all rungs of the temperature ladder at once, would benefit from application of ZV control variates. In contrast to CTI, the controlled AIS estimator (CAIS) demonstrated an increase in variance compared to standard AIS. Full details are provided in the supplementary materials, in addition to results on each of the applications considered in this article. To understand these counter-intuitive results, notice that control variates must be constructed simultaneously over all \( m \) rungs of the temperature ladder, so that for degree 1 polynomials we have to jointly estimate \( md \) coefficients, where \( d \) denotes the number of model parameters, and for degree 2 polynomials we have to jointly estimate \( m(d+3)/2 \) polynomial coefficients. To achieve this using the plug-in principle, we must estimate covariance matrices containing \( \mathcal{O}(m^2d^2) \) and \( \mathcal{O}(m^2d^4) \) entries, respectively. Our results are therefore consistent with the finding that poor estimation of the polynomial coefficients can actually increase estimator variance (Glasserman 2004). It remains unclear how to develop control variates for these nonequilibrium estimators.

We exploited the ZV control variate scheme due to Mira, Solgi, and Imparato (2013), which permits the automatic construction of control variates for any statistical model in which the gradient of the log-likelihood (and the log-prior) are available. More generally, we envisage that for models where these gradients are unavailable in closed form, the use of numerical approximations could provide a successful strategy (Calderhead and Sustik 2012). For a large class of intractable likelihood models, Friel, Mira, and Oates (2014b) showed how to obtain a stochastic approximation to ZV control variates that could be directly leveraged within our methodology. Results on benchmark datasets demonstrate that CTI outperforms standard TI, but that the difference in performance is reduced when the likelihood function is strongly multi-modal. A natural direction for further research is to explore whether alternative control variates are better suited to these challenging problems. The recent work of Oates, Girolami, and Chopin (2014) indicates that nonparametric alternatives to ZV control variates offer much potential.

CTI clearly inherits the theoretical and methodological challenges that are associated with control variates more generally. In particular ZV control variates are not parameterization-invariant and it is unclear how to select an optimal variance-minimizing parameterization. Pertinent to CTI in particular, the optimal coefficients \( \phi^*(t) \) will vary smoothly with (inverse) temperature \( t \) (supplementary Figure S6(b)), yet the conventional plug-in approach to estimation treats each rung \( t_i \) of the temperature ladder independently, leading to rough trajectories (supplementary Figure S6(a)). It would therefore be interesting to design an information sharing scheme that jointly estimates all coefficients.

The development of low-cost computational approaches to Bayesian model comparison is necessary for the widespread adoption of Bayesian methodology in hypothesis-driven research. When the number of models is not large and model dimensionality is not high, control variate strategies offer a promising route toward achieving this goal.
Appendix A: Quadrature for TI

Implementations of TI employ quadrature to approximate the one-dimensional integral in Equation (3). Friel and Pettitt (2008) originally employed a simple trapezoidal rule whereby the (inverse) temperature domain $t \in [0,1]$ was partitioned using $0 = t_0 < t_1 < \cdots < t_m = 1$ and the (log-)evidence was approximated by

$$
\log(p(y)) \approx \frac{1}{2} \sum_{i=0}^{m-1} \left( \frac{t_{i+1} - t_i}{2} \right) \left[ E_{t_{i+1/2}}[\ell(\theta)] + E_{t_{i-1/2}}[\ell(\theta)] \right].
$$

(A.1)

The use of quadrature introduces bias into the resulting estimator. To reduce this quadrature error and thus the estimator bias, Friel, Hurn, and Wyse (2014a) proposed the second-order correction term

$$
\frac{1}{12} \sum_{i=0}^{m-1} \left( t_{i+1} - t_i \right)^2 \left[ V_{t_{i+1/2}}[\ell(\theta)] - V_{t_{i-1/2}}[\ell(\theta)] \right]
$$

(A.2)

that is subtracted from Equation (A.1).

Appendix B: Asymptotic unbiasedness

Propositions 1 and 2 of Mira, Solgi, and Imiparto (2013) show that sufficient conditions for asymptotic unbiasedness of ZV control variates, that is, $E_x[h(\theta)] = 0$, is that, in the case where $\Theta$ is unbounded, the limit $\lim_{r \to \infty} I_{B_r} \pi \nabla Q \cdot \nabla \sigma$ as $r \to \infty$. This condition could be difficult to verify directly; below we contribute a sufficient condition for $\Theta = \mathbb{R}^d$ that is easily verified. Consider a $d$-dimensional hypercube $B_r = \{ \theta : |\theta| \leq r/2 \}$ with side length $r$ and surface area $2dr^{d-1}$ and let $k$ be the degree of the polynomial $Q(\theta)$. Then crude bounds give $I_{B_r} \pi \nabla Q \cdot \nabla \sigma \leq \sup_{B_{r/2}} \| \pi \nabla Q(\theta) \| | \pi \nabla \sigma(\theta) |$. Since $\sup_{B_{r/2}} \| \pi \nabla Q(\theta) \| = O(r^{d-1})$ it follows that a sufficient condition for unbiasedness of ZV is

$$
\sup_{|\theta| \geq r} \pi(\theta) \xrightarrow{d} \infty \quad \text{as} \quad r \to \infty.
$$

(B.1)

In practice, this requires that the tails of the (unnormalized) density $\pi(\theta)$ vanish sufficiently quickly, with faster convergence required when higher degree polynomials are to be used.

Appendix C: Second Degree Polynomials

Second degree polynomials can be expressed as $Q(\theta) = c^T \theta + \frac{1}{2} \theta^T B \theta$, where $c$ is a $d \times 1$ and $B$ is $d \times d$. This leads to ZV control variates of the form

$$
h(\theta) = \text{tr}(B) + (c^T B)^T u(\theta),
$$

(C.1)

where $c$ and $B$ denote the quadratic polynomial coefficients and $\text{tr}(B)$ is the trace of $B$. We assume that $B$ is symmetric, but this is not required in general. Following Mira, Solgi, and Imiparto (2013), it is possible to rearrange the terms on the right-hand side of Equation (C.1) into the form $\phi^T w(\theta)$ where the column vectors $\phi$, $w(\theta)$ have $d + (d+2)!$ elements each, and are defined as $\phi = [c^T, B^T]^T$, where $d$ is the diagonal of $B$ and $c$ is a column vector with $d - (d-1)/2$ elements, whose element in the $(2d - j)(j - 1)/2 + i - j$ position is the lower diagonal $(i, j)$ element of $B$, and $w := [u^T x, u^T y]^T$, where $x = \theta \circ u - \frac{1}{2}$, with $\circ$ denoting the Hadamard product and the unit vector, respectively, while $v$ is a column vector comprising $d(d-1)/2$ elements, whose element in the $(2d - j)(j - 1)/2 + i - j$ position equals $\theta_{k+i} + \theta_{i+k}$, $i \in \{1, 2, \ldots, d\}$, $i \in \{2, 3, \ldots, d\}$, $j < i$.

The same derivation used to obtain Equation (14) can be followed to deduce that optimal coefficients $\phi^*$ in the case of second-order polynomials are given by $\phi^* = -\nabla_q E_q[\ell(\theta)]$ and that at this point the variance reduction satisfies $R(t) = 1 - \text{Corr}(\theta, \ell(\theta), \phi^T w(\theta))^2$. Thus, the ZV strategy with degree 2 polynomials can be expected to reduce variance when a linear combination of the components of $w(\theta)$ is highly correlated with the log-likelihood $\ell(\theta)$.

Appendix D: Formulae for Bayesian Linear regression

D.1. Known Precision

The power posterior follows $p(y, t \sim N(\mu(t), \Sigma(t))$, where $\mu(t) = \frac{1}{t} \Sigma(t) X^T y$, $\Sigma(t) = \frac{1}{t} X^T X + 1/t I$, while the integrand $E_{\theta|y,t}[\log p(y|\beta, \sigma)]$ has the closed-form expression $-\frac{1}{2} \log(2\pi \sigma^2) - \frac{1}{2} \sigma y - X \mu(t)^T (y - X \mu(t)) - \frac{1}{2\sigma^2} tr(X^T X \Sigma(t))$ and the model evidence is

$$
p(y) = \frac{1}{(2\pi)^{n/2} |\Omega|^{1/2}} \exp \left[ -\frac{1}{2} y^T \Omega^{-1} y \right],
$$

(C.2)

where $\Omega = \sigma^2 I + \chi^2 X^T X$.

D.2. Unknown Precision

Using the transformation $\tau \sim \log(\tau)$, we can ensure that the posterior $p(\theta|y, t, m)$ is defined on $\mathbb{R}^d$ and has exponential tails so that, by Equation (B.1), the unbiasedness condition is satisfied. For Model 1, we have $u_1 = \theta \circ (\sum y_i - \beta \bar{x}_i) - \frac{1}{2} \sigma \theta (\alpha - 3000)$, $u_2 = \theta \circ (\sum (y_i - \alpha - \beta \bar{x}_i)) - e^{-\frac{t}{2}} (\sum (y_i - \alpha - \beta \bar{x}_i)^2 + 1 + \delta \theta^2 [b_0 + \frac{1}{2} \sigma (\alpha - 3000)^2 + \frac{1}{2} (\beta - 185)^2]$, where the components are ordered with respect to $\theta = \theta \circ (\alpha, \beta, \eta)$.

Write $X$ for the design matrix with $i$th row $[1, x_i]$. The model evidence, which is the object we wish to estimate, is given by

$$
p(y) = \frac{1}{(2\pi)^{n/2} |\Omega|^{1/2}} \left[ |\Omega_1| \Gamma(\alpha_1) \right. \left. \times \left\{ b_0 + \frac{1}{2} |y - B_1 \theta + B_2 \eta| \right\} \right]^{-\alpha_1},
$$

(C.3)

where $\alpha_0 = \alpha_0 + \frac{1}{2}$, $Q_0 = Q_0 + X^T X$ and $B_1 = Q_1^{-1} (X^T y + Q_0 B_0)$. Derivations for Model 2 are analogous.

Appendix E: Formulas for Goodwin Oscillator

The Goodwin oscillator with $g$ species is given by

$$
\begin{align*}
\frac{dx_1}{dt} &= a_1 - \frac{a_2}{1 + x_1^g} - a_3 x_1 \\
\frac{dx_2}{dt} &= k_1 x_1 - a_2 x_2 \\
&\vdots \\
\frac{dx_g}{dt} &= k_{g-1} x_{g-1} - a_g x_g.
\end{align*}
$$

(E.4)

Here, $x_1$ represents the concentration of mRNA for a target gene and $x_2$ represents its corresponding protein product. Additional variables $x_3, \ldots, x_g$ represent intermediate protein species that facilitate a cascade of enzymatic activation that ultimately leads to a negative feedback via $x_3$, on the rate at which mRNA is transcribed. The solution $x(t; \theta, x_0)$ of this dynamical system depends upon synthesis rate constants $a_1, k_1, \ldots, k_{g-1}, a_g$ and degradation rate constants $a_2, a_3$. The Goodwin oscillator permits oscillatory solutions only when $g > 8$. Following Calderhead and Girolami (2009), we therefore set $g = 10$ as a fixed parameter. A-g variable Goodwin model as described above therefore has $g + 2$ uncertain parameters $(a_1, a_2, k_1, \ldots, k_{g-1}, a)$. The Goodwin oscillator does not permit a closed-form solution, meaning that each evaluation of the likelihood function requires the numerical integration of the system in Equation (E.4). Due to the substantive computational challenges associated with model comparison in this setting, we considered only 10 independent runs of population MCMC, each using only $N = 1000$ iterations.
We consider a realistic setting where only mRNA and protein product are observed, corresponding to $x_k = [x_1, x_2]$. We assume $x_0 = [0, \ldots, 0]$ and $\sigma = 0.1$ are both known and take sampling times to be $s = 41, \ldots, 80$. Parameters were assigned independent $\Gamma(2, 1)$ prior distributions. We generated data using $a_1 = 1, a_2 = 3, k_1 = 2, k_2, \ldots, k_{c-1} = 1, a = 0.5$, which produce oscillatory dynamics that do not depend heavily upon initial conditions (supplementary Figure S8).

In practice, we work with the log-transformed parameters $\theta$. In particular, this allows us to verify that ZV methods are valid, since the tails of $p(\theta)$ vanish exponentially quickly. Sensitivities $S_{j,k}$, defined in the main text, satisfy

$$S_{j,k}^i = \frac{\partial f_k}{\partial x_j} + \sum_{l=1}^{J} \frac{\partial f_k}{\partial x_l} w_{l,k,j} \tag{E.5}$$

where $\lambda_{i,j} = 0$ at $z = 0$. Equation (E.5) provides a route to compute the sensitivities numerically, when they cannot be obtained analytically, by augmenting the state vector of the dynamical system to include the $S_{j,k}$.

**Supplementary Materials**

The supplementary materials contain additional theoretical details and empirical results.

**Acknowledgments**

The authors are grateful to Christian Robert for discussions on the use of control variates and to Nial Friel for comments on an earlier draft of this work.

**Funding**

This work was supported by UK EPSRC EP/D002060/1, EP/J016934/1, EU Grant 259348 (Analyzing and Striking the Sensitivities of Embryonal Tumors), and a Royal Society Wolfson Research Merit Award.

**References**

Andradyottir, S., Heyman, D. P., and Ott, T. J. (1993), “Variance Reduction Through Smoothing and Control Variates for Markov Chain Simulations,” ACM Transactions on Modeling and Computer Simulation, 3, 167–189. [635]

Assaraf, R., and Caffarel, M. (1999), “Zero-Variance Principle for Monte Carlo Algorithms,” Physical Review Letters, 83, 4682–4685. [634]

Behrens, G., Friell, N., and Hurn, M. (2012), “Tuning Tempered Transitions,” Statistics and Computing, 22, 65–78. [637]

Calderhead, B., and Girolami, M. (2009), “Estimating Bayes Factors via Thermodynamic Integration and Population MCMC,” Computational Statistics & Data Analysis, 53, 4028–4045. [635,636,641]

Calderhead, B., and Girolami, M. (2011) (2011), “Statistical Analysis of Nonlinear Dynamical Systems Using Differential Geometric Sampling Methods,” Interface Focus, 1, 821–835. [640]

Calderhead, B., and Sustik, M. (2012), “Sparse Approximate Manifolds for Differential Geometric MCMC,” Advances in Neural Information Processing Systems, 25, 2888–2896. [642]

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000), Monte Carlo Methods in Bayesian Computation, New York: Springer. [634]

Chib, S., and Jeliazkov, I. (2001), “Marginal Likelihood From the Metropolis-Hastings Output,” Journal of the American Statistical Association, 96, 270–281. [634]

Corduneanu, A., and Bishop, C. M. (2001), “Variational Bayesian Selection for Mixture Distributions,” Proceedings of the 8th International Conference on Artificial Intelligence and Statistics, 27–34. [634]

Dellaportas, P., and Kontoyiannis, I. (2012), “Control Variates for Estimation Based on Reversible Markov Chain Monte Carlo Samplers,” Journal of the Royal Statistical Society, Series B, 74, 133–161. [635,637]

Didelot, X., Everitt, R. G., Johansen, A. M., and Lawson, D. J. (2011), “Likelihood-Free Estimation of Model Evidence,” Bayesian Analysis, 6, 49–76. [634]

Frenkel, D., and Smit, B. (2002), Understanding Molecular Simulation: From Algorithms to Applications (2nd ed.), San Diego, CA: Academic Press. [634]

Friel, N., Hurn, M. A., and Wyse, J. (2014a), “Improving Power Posterior Estimation of Statistical Evidence,” Statistics and Computing, 24, 709–723. [637,638,639,643]

Friel, N., Mira, A., and Oates, C. J. (2014b), “Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods,” Statistics and Computing, 24, 709–723. [642]

Friel, N., and Pettitt, A. N. (2008), “Marginal Likelihood Estimation via Power Posteriors,” Journal of the Royal Statistical Society, Series B, 70, 589–607. [635,643]

Friel, N., and Wyse, J. (2012), “Estimating the Statistical Evidence – A Review,” Statistical Science, 66, 288–308. [634,635,639]

Gelfand, A. E., and Dey, D. K. (1994), “Bayesian Model Choice: Asymptotics and Exact Calculations,” Journal of the Royal Statistical Society, Series B, 56, 501–514. [634]

Gelman, A., and Meng, X.-L. (1998), “Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling,” Statistical Science, 13, 163–185. [634,635,641]

Geyer, C. J., and Thompson, E. A. (1995), “Annealing Markov Chain Monte Carlo With Applications to Ancestral Inference,” Journal of the American Statistical Association, 90, 909–920. [634]

Girolami, M., and Calderhead, B. (2011), “Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods,” Journal of the Royal Statistical Society, Series B, 73, 1–37. [634,636,637]

Glasserman, P. (2004), Monte Carlo Methods in Financial Engineering, New York: Springer-Verlag. [642]

Goodwin, B. (1965), “Oscillatory Behavior in Enzymatic Control Processes,” Advances in Enzyme Regulation, 3, 425–438. [641]

Green, P., and Han, X. (1992), “Metropolis Methods, Gaussian Proposals, and Antithetic Variables,” Lecture Notes in Statistics, Stochastic Methods and Algorithms in Image Analysis, 74, 142–164. [634]

Green, P. J. (1995), “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination,” Biometrika, 82, 711–732. [634]

Guha, S., and MacEachern, S. N. (2006), “Generalized Poststratification and Importance Sampling for Subsampled Markov Chain Monte Carlo Estimation,” Journal of the American Statistical Association, 101, 1175–1184. [634]

Hammer, H., and Tjelmeland, H. (2008), “Control Variates for the Metropolis-Hastings Algorithm,” Scandinavian Journal of Statistics, 35, 400–414. [635]

Jasra, A., Stephens, D., and Holmes, C. (2007), “On Population-Based Simulation for Static Inference,” Statistics and Computing, 17, 263–279. [637]

Kass, R. E., and Raftery, A. E. (1995), “Bayes Factors,” Journal of the American Statistical Association, 90, 773–795. [634]

Liang, F., Liu, C., and Carroll, R. J. (2007), “Stochastic Approximation in Monte Carlo Computation,” Journal of the American Statistical Association, 102, 305–320. [634]

Locke, J., Millar, A., and Turner, M. (2005), “Modelling Genetic Networks With Noisy And Varied Experimental Data: The Circadian Clock in Arabidopsis Thaliana,” Journal of Theoretical Biology, 234, 383–393. [641]

Marin, J. M., and Robert, C. P. (2011), “Importance Sampling Methods for Bayesian Discrimination Between Embedded Models,” in Frontiers of Statistical Decision Making and Bayesian Analysis, eds. M.-H. Chen, P. Müller, D. Sun, K. Ye, and D. Dey, pp. 513–527, New York: Springer. [634,639]

Marniani, E., and Parisi, G. (1992), “Simulated Tempering: A New Monte Carlo Scheme,” Europhysics Letters, 19, Article 451. [634]

Miasojedow, B., Moulines, E., and Vihola, M. (2012), “Adaptive Parallel Tempering Algorithm,” Journal of Computational and Graphical Statistics, 22, 649–664. [637]

Mira, A., Solgi, R., and Imparato, D. (2013), “Zero Variance Markov Chain Monte Carlo for Bayesian Estimators,” Statistics and Computing, 23, 653–662. [635,636,642,643]
Mira, A., Tenconi, P., and Bressanini, D. (2003), “Variance Reduction for MCMC,” Technical Report 2003/29, Università degli Studi dell’ Insubria, Italy. [635]

Moral, P., Doucet, A., and Jasra, A. (2006), “Sequential Monte Carlo Samplers,” Journal of the Royal Statistical Society, Series B, 68, 411–436. [634]

Neal, R. M. (1996), “Sampling From Multimodal Distributions Using Tempered Transitions,” Statistics and Computing, 6, 353–366. [637]
—— (2001), “Annealed Importance Sampling,” Statistics and Computing, 11, 125–139. [634,638]

Oates, C. J., Girolami, M., and Chopin, N. (2014), “Control Functionals for Monte Carlo Integration,” CRiSM Working Paper Series 14-22, University of Warwick. [642]

Ogata, Y. (1989), “A Monte Carlo Method for High Dimensional Integration,” Numerical Mathematics, 55, 137–157. [634]

Papamarkou, T., Mira, A., and Girolami, M. (2014), “Zero Variance Differential Geometric Markov Chain Monte Carlo Algorithms,” Bayesian Analysis, 9, 97–128. [635,636]

Philippe, A., and Robert, C. (2001), “Riemann Sums for MCMC Estimation and Convergence Monitoring,” Statistics and Computing, 11, 103–115. [634]

Robert, C., and Casella, G. (2004), Monte Carlo Statistical Methods (2nd ed.), New York: Springer-Verlag. [634,637]

Rubinstein, R. Y., and Marcus, R. (1985), “Efficiency of Multivariate Control Variates in Monte Carlo Simulation,” Operations Research, 33, 661–677. [637]

Skilling, J. (2006), “Nested Sampling for General Bayesian Computation,” Bayesian Analysis, 1, 833–859. [634]

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009), “Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems,” Journal of the Royal Society Interface, 6, 187–202. [634]

Torrie, G. M., and Valleau, J. P. (1977), “Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling,” Journal of Computational Physics, 23, 187–199. [634]

Vyshemirsky, V., and Girolami, M. A. (2008), “Bayesian Ranking of Biochemical System Models,” Bioinformatics, 24, 833–839. [634,635]

Williams, E. (1959), Regression Analysis, New York: Wiley. [638]

Zhou, Y., Johansen, A. M., and Aston, J. A. D. (2013), “Towards Automatic Model Comparison an Adaptive Sequential Monte Carlo Approach,” CRiSM Technical Report No. 13-04, University of Warwick. [634,637]