Statistica Sinica **32** (2022), 1-22
doi:https://doi.org/10.5705/ss.202020.0461

# DOUBLE HAPPINESS: ENHANCING THE COUPLED GAINS OF L-LAG COUPLING VIA CONTROL VARIATES

Radu V. Craiu and Xiao-Li Meng

*University of Toronto and Harvard University*

*Abstract:* The recently proposed L-lag coupling for unbiased Markov chain Monte Carlo (MCMC) calls for a joint celebration by MCMC practitioners and theoreticians. For practitioners, it circumvents the thorny issue of deciding the burn-in period or when to terminate an MCMC sampling process, and opens the door for safe parallel implementation. For theoreticians, it provides a powerful tool to establish elegant and easily estimable bounds on the exact error of an MCMC approximation at any finite number of iterates. A serendipitous observation about the bias-correcting term leads us to introduce naturally available control variates into the L-lag coupling estimators. In turn, this extension enhances the gains of L-lag coupling, because it results in more efficient unbiased estimators, as well as a better bound on the total variation error of any MCMC iteration, albeit the gains diminish as L increases. Specifically, the new upper bound is theoretically guaranteed to never exceed the one given previously. We also argue that L-lag coupling represents a coupling for the future, breaking from the coupling-from-the-past type of perfect sampling, by reducing the generally unachievable requirement of being *perfect* to one of being *unbiased*, a worthwhile trade-off for ease of implementation in most practical situations. The theoretical analysis is supported by numerical experiments that show tighter bounds and a gain in efficiency when control variates are introduced.

*Key words and phrases:* Coupling from the Past, maximum coupling, median absolute deviation, parallel implementation, total variation distance, unbiased MCMC.

## 1. If Being Perfect is Impossible, Let's Try Being Unbiased

### 1.1. Perfect coupling – too much to hope for?

We thank Pierre Jacob and his team for a series of articles (e.g., Jacob, O'Leary and Atchadé (2020); Jacob, Lindsten and Schön (2020); Heng and Jacob (2019); Biswas, Jacob and Vanetti (2019)) that revitalized our experience (e.g., Murdoch and Meng (2001); Meng (2000); Craiu and Meng (2011); Stein and Meng (2013)) of working on coupling from the past (CFTP; Propp and Wilson (1996, 1998)) and, more generally, perfect sampling. The clever "cross-time coupling"

Corresponding author: Radu V. Craiu, Department of Statistics, University of Toronto, Ontario M5S 3G3, Cananda. E-mail: craiu@utstat.toronto.edu.

idea of Glynn and Rhee (2014), which can be considered a form of coupling for the future (CFTF), allows us to move away from the CFTP framework, which became popular around the turn of the century with its promise of providing perfect/exact Markov chain Monte Carlo (MCMC) samplers (e.g., see the annotated bibliography of Wilson (1998)). However, research progress on perfect or exact samplers has slowed significantly since then, because they are very challenging, if not impossible, to develop for many routine Bayesian computational problems (e.g., see Murdoch and Meng (2001)).

In its most basic form, a CFTP-type perfect sampler couples a Markov chain $\{X_t, t \geq 0\}$ with itself, but from different starting points, and runs two or more chains until they coalesce at a time $\tau$. This apparent convergence does not guarantee, in general, that $X_\tau$ is from the desired stationary distribution $\pi(x)$. By shifting the entire chain to "negative time"(i.e., the past), $\{X_t, t \leq 0\}$, Propp and Wilson (1996) have shown that if we follow this coalescent chain until it reaches the present time, that is, $t = 0$, then the resulting $X_0$ will be exactly from $\pi(x)$. Perhaps the most intuitive way to understand this scheme is to realize that running a chain from its infinite past ($t = -\infty$) to the present ($t = 0$) is mathematically equivalent to running the chain from the present ($t = 0$) to the infinite future ($t = +\infty$). The CFTP is a clever way of realizing this seemingly impossible task, relying on the fact that if the coalescence occurs regardless of how we have chosen the starting point, then the chain has "forgotten" its origin, and hence has settled in the perfect asymptotic distribution.

However, being perfect is never easy, especially in the mathematical sense. No error of any kind is allowed, and this requirement has manifested in two ways that greatly limit the practicality of perfect sampling. First, constructing a perfect sampler, especially for distributions with continuous and unbounded state spaces—which are ubiquitous in routine statistical applications—is a very challenging task in general, despite its great success for problems with some special structures, such as certain monotonic properties (see Berthelsen and Møller (2002); Corcoran and Tweedie (2002); Huber (2004, 2002); Ensor and Glynn (2000); Huber (2004); Murdoch and Takahara (2006)). Second, even if a perfect sampler is devised, it can be excruciatingly slow, because it refuses to deliver an output until it can guarantee its perfection, and one must devise problem-specific strategies to speed this up (e.g., Thönnes (Thönnes(1999); Dobrow and Fill (2003); Møller (1999); Dobrow and Fill (2003); Corcoran and Schneider (2005)).

### 1.2. Unbiased coupling – a new hope?

A relaxation of the exact sampling paradigm with important practical consequences has been proposed by Glynn and Rhee (2014) and Glynn (2016), who put forth strategies for the exact estimation of integrals using MCMC. The difference between exact sampling and exact estimation is a large conceptual leap that allows us to bypass most of the difficulties of perfect sampling, while maintaining some of its important benefits. Building on the work of Glynn and co-authors, the L-lag coupling of Biswas, Jacob and Vanetti (2019) and Jacob, O'Leary and Atchadé (2020) aims to deliver unbiased estimators of $\mathrm{E}[h(X_\pi)]$, for any (integrable) $h$, where $X_\pi$ denotes a random variable defined by $\pi(X)$. One may question if this is really a weaker requirement because the fact that $\mathrm{E}[h(X_\pi)] = \mathrm{E}[h(X_{\tilde{\pi}})]$ for all (integrable) $h$ immediately implies that $\pi(X) = \tilde{\pi}(X)$ (almost surely). This is where the innovation of L-lag coupling lies, because it does not couple a chain with itself from two or more starting points [e.g., two extreme states, as with monotone coupling; see Propp and Wilson (1996)]. Instead, it couples two chains that have the same transition probability and start from the same starting point or, more generally, the same initial distribution $\pi_0$, but are time-shifted by an integer lag, $L > 0$.

To illustrate, consider the case of $L = 1$, which was the focus of Jacob, O'Leary and Atchadé (2020). Two chains $\mathcal{X} = \{X_t, t \geq 0\}$ and $\mathcal{Y} = \{Y_t, t \geq 0\}$ are coupled in such a way that both of them have the same transition kernel (and, hence, the same target stationary distribution), and there exists with probability one a finite stopping time $\tau$, such that $X_t = Y_{t-1}$, for all $t \geq \tau$. This construction allows them to show that the following estimator based on both $\mathcal{X}$ and $\mathcal{Y}$,

$$H_k(\mathcal{X}, \mathcal{Y}) = h(X_k) + \sum_{j=k+1}^{\tau-1} [h(X_j) - h(Y_{j-1})], \tag{1.1}$$

is an unbiased estimator for $\mathrm{E}[h(X_\pi)]$, for any $k \geq 0$ (under mild conditions). Heuristically, this is because the sum in (1.1) is the same as $\sum_{j=k+1}^{\infty}[h(X_j) - h(Y_{j-1})]$, because any term with $j \geq \tau$ must be zero, by the coupling scheme. Furthermore, for the purpose of calculating expectations, we can replace $h(Y_{j-1})$ with $h(X_{j-1})$, for any $j$, because $X_{j-1}$ and $Y_{j-1}$ have identical distributions, by construction. However, $h(X_k) + \sum_{j=k+1}[h(X_j) - h(X_{j-1})]$ is nothing but $\lim_{t\to\infty} h(X_t)$, which has the same distribution as $h(X_\pi)$.

The cleverness of constructing an estimator based on *both* $\mathcal{X}$ and $\mathcal{Y}$ to ensure $\mathrm{E}[H_k(\mathcal{X}, \mathcal{Y})] = \mathrm{E}[h(X_\pi)]$, for any $h$, bypasses the requirement that $X_\tau$ itself must

be perfect. The series of illustrative and practical examples in Jacob, O'Leary and Atchadé (2020) and in Jacob, Lindsten and Schön (2020), Heng and Jacob (2019), and Biswas, Jacob and Vanetti (2019) provide good evidence of the practicality of this approach. The use of parallel computation for estimating $I = \mathrm{E}_\pi[h(X)]$ supports using $\mathrm{E}[\mathrm{E}_\pi[h(x)|\mathcal{U}_j]]$, where the inner expectation is the estimate obtained from the $j$th parallel process, $\mathcal{U}_j$, and the outer mean averages over all processes. However, if each inner mean is a biased estimator for $I$, then the accumulation of errors can be seriously misleading. This has been documented in the Monte Carlo literature extensively, for instance, in Glynn and Heidelberger (1991) and Nelson (2016). Hence, unbiased MCMC designs allow one to take full advantage of parallel computation strategies, without having to worry about the accumulation of bias as the number of parallel processes increases.

### 1.3. Using control variates – even higher hope?

The expression (1.1) also opens a path to explore further improvements, and that is the starting point of our exploration. In Craiu and Meng (2020), we noticed that (1.1) can be expressed equivalently as

$$H_k(\mathcal{X}, \mathcal{Y}) = h(X_{(\tau-1)\vee k}) + \sum_{j=k}^{\tau-2}[h(X_j) - h(Y_j)], \tag{1.2}$$

where $A \vee B = \max\{A, B\}$. Expression (1.1) renders the insight underlying Jacob, O'Leary and Atchadé (2020), which is that $H_k(\mathcal{X}, \mathcal{Y})$ achieves the desired unbiasedness by providing a *time-forward bias correction* to $h(X_k)$, whenever $\tau > k+1$; hence coupling for the future. (No correction is needed when $\tau \le k+1$.) The dual expression (1.2) indicates that $H_k(\mathcal{X}, \mathcal{Y})$ can also be viewed as a *time-backward bias correction* to $h(X_{\tau-1})$ for its imperfection, because $k < \tau - 1$.

Most intriguingly, each correcting term $\Delta_j \equiv h(X_j) - h(Y_j)$ in (1.2) has mean zero, by the construction of $\{\mathcal{X}, \mathcal{Y}\}$. However, the sum $\sum_{j=k}^{\tau-2}[h(X_j) - h(Y_j)]$ does not necessarily have mean zero, because $\tau$ is random and it depends critically on $\{\mathcal{X}, \mathcal{Y}\}$. Indeed, if this sum had mean zero, then $X_{(\tau-1)\vee k}$ would have been a perfect draw from $\pi(X)$, because then $\mathrm{E}[h(X_{(\tau-1)\vee k})] = \mathrm{E}[h(X_\pi)]$, for any (integrable) $h$, which would imply that $X_{(\tau-1)\vee k} \sim \pi$.

However, the fact that $\mathrm{E}(\Delta_j) = 0$ suggests that we can use any linear combination of $\Delta_j$ as a *control variate* for $H_k(\mathcal{X}, \mathcal{Y})$. Using control variates to reduce estimation errors is a well-known technique in the literature on improving MCMC samplers and estimators by using efficiency swindles, such as antithetic and control variates, Rao–Blackwellization, and so on, some of which we have explored

in the past (e.g., Van Dyk and Meng (2001); Craiu and Meng (2001, 2005); Craiu and Lemieux (2007); Yu and Meng (2011)). For example, for any finite constant $\eta > k+1$, the estimator

$$H_k^*(\mathcal{X},\mathcal{Y};\eta) = H_k(\mathcal{X},\mathcal{Y}) - \sum_{j=k}^{\eta-2}\Delta_j = h(X_{(\tau-1)\vee k}) + \sum_{j=k}^{\tau-2}\Delta_j - \sum_{j=k}^{\eta-2}\Delta_j \quad (1.3)$$

shares the mean of $H_k(\mathcal{X},\mathcal{Y})$, but can have a smaller variance, with a judicious choice of $\eta$. Intuitively, this reduction of variance is possible because of the potential partial cancellation (on average) of the $\Delta_j$ terms in the last two summations in (1.3).

Indeed, Section 2 investigates a more general class of control variates, and derives the optimal choice by establishing the minimal upper bound within the class on the total variation distance between the target $\pi$ and $\pi_k$, the distribution of $X_k$. This leads to both an improved theoretical bound over that of Biswas, Jacob and Vanetti (2019), as reported in Section 2, as well as a more efficient estimator than (1.1) owing to a parallel implementation. Section 3 describes the estimation methods and algorithms, and Section 4 provides examples and illustrations of both kinds of gains. Section 5 discusses some future work.

## 2. Theoretical Gains from Incorporating Control Variates

### 2.1. L-lag coupling: An elegant and powerful method

The scheme of L-lag coupling extends the coupling of $\{X_k, Y_{k-1}\}$ to the more general form of the coupling of $\{X_k, Y_{k-L}\}$, for some fixed $L \geq 1$, as detailed in Biswas, Jacob and Vanetti (2019). The significance of this extension can be best understood by expressing the L-lag coupling idea in its mathematically equivalent form of seeking $\tau_L$ such that $X_{k+L} = Y_k$, for all $k \geq \tau_L$, and letting $L \to \infty$ while keeping $k$ fixed. Heuristically, it is then clear that the larger $L$, the closer the distribution of $Y_{\tau_L}$ is to the target, because $X_{\tau_L+L}$ should converge to $X_\infty \sim \pi$ as $L \to \infty$, and $\mathcal{X}$ and $\mathcal{Y}$ share the same target $\pi$.

Indeed, by extending (1.1) to a general $L$, Biswas, Jacob and Vanetti (2019) show that (under mild regularity conditions) the total variation distance between $\pi_k$, the distribution of $X_k$, and $\pi$ is bounded by a very simple function of $\tau_L$ and $(k, L)$:

$$d_{\mathrm{TV}}(\pi_k,\pi) \leq \mathrm{E}[J_{k,L}], \quad \text{with} \quad J_{k,L} = \max\left\{0, \left\lceil \frac{\tau_L - L - k}{L} \right\rceil \right\}, \quad (2.1)$$

where $\lceil a \rceil$ denotes the smallest integer that is no less than $a$. We can clearly see the impact of increasing $L$ or $k$, because larger values of either of them make it more likely that $\tau_L - L - k < 0$, and hence $J_{k,L} = 0$. Perhaps a clear demonstration of this fact is when $\tau_L$ follows a geometric distribution with success probability $p$ and state space $\{L + i,\ i \geq 0\}$ (because $\tau_L \geq L$, by definition) or, equivalently, $\delta = \tau - (L-1) \sim \text{Geo}(p)$. Then, letting $q = 1 - p$, we have (see Biswas, Jacob and Vanetti (2019))

$$d_{\text{TV}}(\pi_k, \pi) \leq \text{E}[J_{k,L}] = \frac{q^{k+1}}{1 - q^L}. \tag{2.2}$$

We see that the bound is a decreasing function of both $k$ and $L$, though it decreases much faster with $k$, which controls the rate of convergence, than it does with $L$, which controls only the (constant) scaling factor. We also observe that the bound can be trivial, because it can be larger than one for small $k$ and/or $L$, whereas $d_{\text{TV}}$ cannot, suggesting there is room for improvement. Nevertheless, (2.1) is a remarkable bound because it encodes all the intricacies relevant for the convergence speed of $\mathcal{X}$, including the choice of $X_0$, into a univariate (truncated) coupling time $J_{k,L}$. In the case of (2.2), the bound also immediately establishes the geometric ergodicity of $\mathcal{X}$, and provides a rather practical way to assess the bound by estimating $p$ or, more generally, by assessing $J_{k,L}$ directly, say, from a parallel implementation (see Section 3).

It is perhaps even more remarkable to see that the left-hand side of (2.1) is a property of the marginal chain $\mathcal{X}$ (and, equivalently, of the $\mathcal{Y}$ chain), but its right-hand side depends on the construction of the joint chain $\{\mathcal{X}, \mathcal{Y}\}$. This suggests that we can seek improvement by better coupling. Furthermore, as we establish below, even without changing the coupling scheme, we can still obtain better bounds by using more efficient estimators than (1.1).

For a general $L$, the forward-correction expression in (1.1) becomes (Biswas, Jacob and Vanetti (2019))

$$H_{k,L}(\mathcal{X}, \mathcal{Y}) = h(X_k) + \sum_{j=1}^{J_{k,L}} \left[ h(X_{k+jL}) - h(Y_{k+(j-1)L}) \right], \tag{2.3}$$

and it is easy to verify that the backward-correction expression (1.2) takes the form

$$H_{k,L}(\mathcal{X}, \mathcal{Y}) = h(X_{k+LJ_{k,L}}) + \sum_{j=0}^{J_{k,L}-1} \left[ h(X_{k+jL}) - h(Y_{k+jL}) \right]. \tag{2.4}$$

**Remark 1.** The (random) subscript in $X_{k+JL}$ cannot be reduced to $(\tau - L) \vee k$ when $L > 1$, the most obvious extension of the index $(\tau - 1) \vee k$ in (1.2). This is

because $k + J_{k,L}L \geq (\tau - L) \vee k$, but the inequality can be strict when $\tau > k + L$. For example, if $\tau = L + k + M$, where $M$ is a positive integer less than $L$ (which does not exist when $L = 1$), $k + J_{k,L}L = k + L$, but $(\tau - L) \vee k = k + M$.

**Remark 2.** Whereas (2.3) and (2.4) are equivalent as equalities, they may lead to different inequalities depending on how we bound their respective right-hand sides. This is both a bonus and a trap, as we discuss below.

### 2.2. Deriving the optimal bound over choices of control variates

For notational simplicity, we drop the variables $k, L$ from the notation of $J_{k,L}$, and we let $\Delta_{k,j} = h(X_{k+jL}) - h(Y_{k+jL})$. Then, we know $\Delta_{k,j}$ has mean zero for any $\{k, j\}$ and $L$. This means that for any random sequence $\vec{\eta} \equiv \{\eta_j, j \geq 1\}$ such that: (A) it is independent of $\{\mathcal{X}, \mathcal{Y}\}$, and (B) $\sum_{j=1} \mathrm{E}_{\vec{\eta}}|\eta_j| < \infty$, we can use $C_\eta = \sum_{j\geq 1} \eta_j \Delta_{k,j}$ as a control variate for $H_{k,L} \equiv H_{k,L}(\mathcal{X}, \mathcal{Y})$, because $\mathrm{E}[C_\eta] = 0$. That is,

$$\tilde{H}^{(\vec{\eta})}_{k,L}(\mathcal{X}, \mathcal{Y}) = H_{k,L}(\mathcal{X}, \mathcal{Y}) - \sum_{j\geq 1} \eta_j \Delta_{k,j} \tag{2.5}$$

is also an unbiased estimator of $\mathrm{E}[h(X_\pi)]$. Next, we examine how to choose $\eta$.

To choose $\vec{\eta}$, instead of minimizing $\mathrm{Var}[\tilde{H}^{(\vec{\eta})}_{k,L}]$, which is not an easy task and will also likely produce an $h$-dependent solution, we first follow the argument used by Biswas, Jacob and Vanetti (2019) with a given $\eta$. We then minimize a class of bounds of $d_{\mathrm{TV}}(\pi_t, \pi)$ over the choice of $\vec{\eta}$ that satisfies (A) and (B). This leads to a sharper bound than (2.1), a special case corresponding to $\vec{\eta} = 0$, which, in general, is not an optimal choice, as shown below.

We proceed by using the same argument as in Biswas, Jacob and Vanetti (2019) for proving (2.1), but using (2.5) instead of (2.3). However, when applying (2.5), we must retain the expression of $H_{k,L}(\mathcal{X}, \mathcal{Y})$, as given by (2.3). (Interested readers are invited to try using (2.4).) Specifically, the unbiasedness of (2.5) implies that, for any $k \geq 1$,

$$\begin{aligned}
\mathrm{E}[h(X_\pi) - h(X_k)] &= \mathrm{E}\left\{ \sum_{j=1}^{J} \left[ h(X_{k+jL}) - h(Y_{k+(j-1)L}) \right] - \sum_{j\geq 1} \eta_j \Delta_{k,j} \right\} \\
&= \mathrm{E}\left\{ \sum_{j\geq 1} \left[ h(X_{k+jL}) - h(Y_{k+(j-1)L}) \right] 1_{\{j \leq J\}} - \sum_{j\geq 1} \eta_j \left[ h(X_{k+jL}) - h(Y_{k+jL}) \right] \right\}
\end{aligned} \tag{2.6}$$

$$= \mathrm{E}\left\{\sum_{j\geq 1} h(X_{k+jL})[1_{\{j\leq J\}} - \eta_j] + \sum_{j\geq 1} h(Y_{k+jL})[\eta_j - 1_{\{j+1\leq J\}}] - h(Y_k)1_{\{0<J\}}\right\}.$$

The interchanges of sum and expectation in the (infinite) sums hold under assumption (B) and the additional assumption that the $h$ function is bounded. To compute the total variation distance, let $h \in \mathcal{H} = \{h : \sup_x |h(x)| \leq 1/2\}$, as in Biswas, Jacob and Vanetti (2019). Consequently, (2.6) implies

$$\begin{aligned} d_{\mathrm{TV}}(\pi_k, \pi) &\leq \frac{1}{2}\left\{\sum_{j\geq 1} \mathrm{E}|1_{\{j\leq J\}} - \eta_j| + \sum_{j\geq 1} \mathrm{E}|\eta_j - 1_{\{j\leq J-1\}}| + \Pr(0 < J)\right\} \\ &= \sum_{j\geq 1} \mathrm{E}|1_{\{j\leq \tilde{J}\}} - \eta_j| + 0.5\Pr(J > 0), \end{aligned} \tag{2.7}$$

where $\tilde{J} = J - \xi$ and $\xi \sim \mathit{Bernoulli}(0.5)$ is independent of $J$. Note that the support for $\tilde{J}$ is $\{-1, 0, 1, \ldots\}$. Set

$$S_j = \Pr(\tilde{J} \geq j) = \Pr(J > j) + 0.5\Pr(J = j), \quad \text{for any } j \geq 0. \tag{2.8}$$

Recall that for any given random variable $V$, $\min_{U\perp V} E|V - U| = E|V - m_V|$, where $m_V$ is a median of $V$, and the notation $\min_{U\perp V}$ means to minimize over all $U$ that are independent of $V$. Hence, in order to minimize (2.7) over $\vec{\eta}$, we should set $\eta_j$ to be the median of the Bernoulli random variable $1_{\{j\leq \tilde{J}\}}$, that is, $\eta_j = 1_{\{S_j > 0.5\}}$.

Let $m_{\tilde{J}}$ be *the smallest integer* median of $\tilde{J}$. Then, for any $j > m_{\tilde{J}}$, $S_j = 1 - \Pr(\tilde{J} < j) \leq 1 - \Pr(\tilde{J} \leq m_{\tilde{J}}) \leq 1/2$ because $\Pr(\tilde{J} \leq m_{\tilde{J}}) \geq 1/2$, by the definition of $m_{\tilde{J}}$, implying $\eta_j = 0$. Therefore, we know the maximal number of nonzero $\eta_j$ cannot exceed $m_{\tilde{J}}$. However, other than the ideal case with $\Pr(J = 0) = 1$, $m_{\tilde{J}}$ can be zero, but not $-1$, because $\Pr(\tilde{J} = -1) = 0.5\Pr(J = 0) < 0.5$. This automatically implies that condition (B) is trivially satisfied. For this choice of $\vec{\eta}$, (2.7) yields our new bound for $d_{\mathrm{TV}}(\pi_k, \pi)$:

$$B_{k,L} = \sum_{j\geq 1} \min\{S_j, 1 - S_j\} + 0.5\Pr(J > 0) \tag{2.9}$$

$$= \sum_{j\geq 1} \min\left\{\Pr(J \geq j), \Pr(J \leq j)\right\}. \tag{2.10}$$

In deriving the last equality, we use the fact that $S_j + 0.5\Pr(J = j) = \Pr(J \geq j)$ and $1 - S_j + 0.5\Pr(J = j) = \Pr(J \leq j)$, and $\Pr(J > 0) = \sum_{j\geq 1} \Pr(J = j)$.

### 2.3. Understand and compare the bounds

It is immediate from expression (2.10) that our new bound cannot exceed the bound of Biswas, Jacob and Vanetti (2019), as given in (2.1), because (2.10) obviously cannot exceed $\sum_{j\geq 1}\Pr(J\geq j)$, which is $\mathrm{E}[J]$. The next result reveals alternative forms for the new bound, providing additional insights, including the optimality of the choice $\eta_j = 1_{\{j\leq m_{\tilde{J}}\}}$.

**Theorem 1.** *Under the same regularity conditions as in Biswas, Jacob and Vanetti* (2019)*, we have*

$$
\begin{aligned}
B_{k,L} &= \mathrm{E}\,|\tilde{J}_{k,L} - m_{\tilde{J}_{k,L}}| + \Pr(J_{k,L}>0) - 0.5 && (2.11)\\
&= \mathrm{E}\,|J_{k,L} - m_{J_{k,L}}| + \Pr(J_{k,L}>0) - S_{k,L} && (2.12)\\
&= 0.5\sum_{j\geq 1}[1 - |\Pr(\tau > k+(j+1)L) + \Pr(\tau > k+jL) - 1|] \\
&\quad + 0.5\Pr(\tau > k+L), && (2.13)
\end{aligned}
$$

*where* $S_{k,L} = \max\{\Pr(J_{k,L} > m_{J_{k,L}}), \Pr(J_{k,L} < m_{J_{k,L}})\} \leq 0.5$*, and* $m_{\tilde{J}_{k,L}}$ *and* $m_{J_{k,L}}$ *are the smallest integer medians of* $\tilde{J}_{k,L}$ *and* $J_{k,L}$*, respectively.*

*Proof.* To reduce the notation overload, we drop the $k, L$ for $J$, $\tilde{J}$, $m_J$, and $m_{\tilde{J}}$. We have already established that the optimal $\vec{\eta}$ must be of the form $\eta_j = 1_{\{j\leq m\}}$, for some $m \geq 0$. Note here that the use of $m = 0$ permits $\vec{\eta} = 0$ because $j \geq 1$. This is also consistent with setting $\eta_0 = 1$. We can minimize the right-hand side of (2.7) with respect to such a class, that is, with respect to the choice of $m$. However, it is easy to see that

$$
\begin{aligned}
\sum_{j\geq 1}\mathrm{E}|1_{\{j\leq\tilde{J}\}} - \eta_j| &= \sum_{j\geq 0}\mathrm{E}|1_{\{j\leq J\}} - 1_{\{j\leq m\}}| - \mathrm{E}[1 - 1_{\{0\leq\tilde{J}\}}] \\
&= \sum_{j\geq 0}\mathrm{E}\left[1_{\{\min\{\tilde{J},m\}<j\leq\max\{\tilde{J},m\}}\right] - \Pr(\tilde{J}=-1) \\
&= \mathrm{E}\left[\max\{\tilde{J},m\} - \min\{\tilde{J},m\}\right] - 0.5\Pr(J=0) \\
&= \mathrm{E}|\tilde{J}-m| - 0.5\Pr(J=0). && (2.14)
\end{aligned}
$$

It is clear from (2.14) that the optimal $m$ must be an integer median of $\tilde{J}$, and we choose the smallest one, $m_{\tilde{J}}$. With this choice of $\vec{\eta}$, substituting (2.14) into (2.7) yields the expression

$$
B_{k,L} = \mathrm{E}|\tilde{J} - m_{\tilde{J}}| - 0.5\Pr(J=0) + 0.5\Pr(J>0) \qquad (2.15)
$$

$$= \Pr(J > 0) + \mathrm{E}|\tilde{J} - m_{\tilde{J}}| - 0.5, \tag{2.16}$$

which proves (2.11).

In order to prove (2.12), we start from (2.10). Let

$$G(j) = \Pr(J \le j) - \Pr(J \ge j) = \Pr(J < j) - \Pr(J > j), \tag{2.17}$$

for $j \ge 0$. Then, it is easy to verify that $G(j)$ is a monotone increasing function, which means $G(j) - G(m_J)$ share the same sign with $j - m_J$, for all $j \ne m_J$. It follows that the sum in (2.10) can be decomposed into three parts, $A = \sum_{j=1}^{m_J-1} \Pr(J \le j)$, $B = 1_{\{m_J>0\}} \min\{\Pr(J \le m_J), \Pr(J \ge m_J)\}$, and $C = \sum_{j \ge m_J+1} \Pr(J \ge j)$. When $m_J = 0$, $C = \mathrm{E}[J]$, $B = 0$ because $1_{\{m_J>0\}} = 0$, and $A = 0$ by convention because $m_J - 1 < 1$. If $p_j = \Pr(J = j)$, then it is easy to see that whenever $m_J \ge 1$,

$$\begin{aligned} A &= \sum_{j=1}^{m_J-1} \sum_{h=1}^{j} p_h + (m_J - 1)p_0 = \sum_{h=1}^{m_J-1} \sum_{j=h}^{m_J-1} p_h + (m_J - 1)p_0 \\ &= \sum_{h=1}^{m_J-1} (m_J - h)p_h + (m_J - 1)p_0 = \sum_{h=0}^{m_J-1} (m_J - h)p_h - p_0; \end{aligned} \tag{2.18}$$

$$C = \sum_{j=m_J+1}^{\infty} \sum_{h=j}^{\infty} p_h = \sum_{h=m_J+1}^{\infty} \sum_{j=m_J+1}^{h} p_h = \sum_{h=m_J+1}^{\infty} (h - m_J)p_h. \tag{2.19}$$

Noting that $(m_J - h)p_h = 0$ when $h = m_J$, we see that when $m_J \ge 1$,

$$\begin{aligned} A + B + C &= \mathrm{E}|J - m_J| - p_0 + \min\{\Pr(J \ge m_J), \Pr(J \le m_J)\} \\ &= \mathrm{E}|J - m_J| + \Pr(J > 0) + \min\{\Pr(J \ge m_J), \Pr(J \le m_J)\} - 1 \\ &= \mathrm{E}|J - m_J| + \Pr(J > 0) - \max\{\Pr(J < m_J), \Pr(J > m_J)\}. \end{aligned} \tag{2.20}$$

When $m_J = 0$, $A = B = 0$, and $C = \sum_{h \ge 1} hp_h = \mathrm{E}[J]$, which is (2.12) because $S_{k,L} = \Pr(J > 0)$, cancelling exactly the $\Pr(J > 0)$ term. This completes the proof of (2.12).

The proof of (2.13) also follows from (2.10), using the identities $\max\{a, b\} = 0.5[a + b + |a - b|]$ and $\Pr(J \ge j) + \Pr(J \le j) = 1 + \Pr(J = j)$, for any $j$. This leads to

$$\sum_{j \ge 1} \min\{\Pr(J \ge j), \Pr(J \le j)\}$$

$$= 0.5 \sum_{j \geq 1} [1 + \Pr(J = j) - |\Pr(J \geq j) - 1 + \Pr(J > j)|]$$
$$= 0.5 \sum_{j \geq 1} [1 - |\Pr(J > j) + \Pr(J > j - 1) - 1|] + 0.5 \Pr(J > 0).$$

Expression (2.13) then follows because $\{J > j\} = \{\tau > k + (j+1)L\}$.

The above result tells us that, whenever $m_J = 0$, our bound is identical to the one given by Biswas, Jacob and Vanetti (2019). From (2.10), the two bounds are the same if and only if $G(1) \geq 0$, which is the same as $2p_0 \geq 1 - p_1$, where $p_k = \Pr(J = k)$. Clearly, this inequality is satisfied when $m_J = 0$, that is, when $p_0 \geq 1/2$. It also implies that $m_J \leq 1$, because for any $j < m_J$,

$$G(j) = \Pr(J < j) + \Pr(J \leq j) - 1 \leq 2 \Pr(J < m_J) - 1 < 0, \tag{2.21}$$

as $\Pr(J \leq m_J - 1) < 0.5$, because $m_J$ is the smallest integer median. Therefore, we have the following theorem.

**Theorem 2.** *Under the same regularity conditions as those of Theorem* 1*, a sufficient and necessary condition for the bound in Theorem* 1 *to equal* $\mathrm{E}[J]$ *is* $2p_0 \geq 1 - p_1$.

**Remark 3.** Theorem 2 implies that $m_J = 0$ is a sufficient condition and $m_J \leq 1$ is a necessary condition for the two bounds to be the same. However, the condition $m_J = 1$ itself is not sufficient.

**Remark 4.** An intriguing new insight provided by bound (2.12) is that not only the average coupling time matters, but the variation of the coupling time is important too. The $S_{k,L}$ term also suggests that even the symmetry matters, because $S_{k,L}$ achieves its maximum when the distribution is symmetrical locally around the median.

Let $\zeta = \tau - t$, which is the number of steps needed after time $t$ in order to couple (assuming the coupling has not already happened by time $t$). Then, the sufficient and necessary condition in Theorem 2 is the same as

$$\Pr(\zeta \leq L) \geq \Pr(\zeta > 2L), \tag{2.22}$$

suggesting that the new bound is more useful when the distribution of $\zeta$ places more mass on the right side of the coupling interval $(L, 2L]$ than it does on its left side, that is, when (2.22) is violated. The implication is that the improvement of the new bound, if any, will more likely come from those situations where either $t$ is small or $\tau$ is large (for fixed $L$), that is, when the mixing is poor.

## 3. Estimation and Practical Implementation

We assume that $Q > 1$ coupled processes $\{(X_t^{(q)}, Y_t^{(q)}) : 1 \le q \le Q\}$ are run in parallel and that, for all $1 \le q \le Q$, $\mathcal{X}^{(q)} = \{X_k^{(q)}\}_{k\ge 1}$ and $\mathcal{Y}^{(q)} = \{Y_k^{(q)}\}_{k\ge 1}$ have been successfully L-coupled. The latter implies that the chains $\mathcal{X}^{(q)}$ are run $L$ more steps than the $\mathcal{Y}^{(q)}$ chains, and there exists a stopping time $\{\tau^{(q)} : q = 1, \dots, Q\}$ such that $X_{t+L}^{(q)} = Y_t^{(q)}$, for all $t \ge \tau^{(q)}$.

### 3.1. Control variate estimators

We work with a modified version of (2.4) that incorporates control variates:

$$H_{k,L}^{*(q)}(\mathcal{X}^{(q)}, \mathcal{Y}^{(q)}) = h\left(X^{(q)}_{k+J_{k,L}^{(q)}L}\right) + \sum_{j=0}^{J_{k,L}^{(q)}-1}\left[h(X_{k+jL}^{(q)}) - h(Y_{k+jL}^{(q)})\right] - \sum_{j=0}^{m_{\tilde{J}_{k,L}^{(q)}}}\left[h(X_{k+jL}^{(q)}) - h(Y_{k+jL}^{(q)})\right], \tag{3.1}$$

where $m_{\tilde{J}}$ denotes the smallest integer median of $\tilde{J}$. Henceforth, in order to simplify the notation, we use $m_{k,L}^{(q)}$ and $\tilde{m}_{k,L}^{(q)}$ to denote $m_{J_{k,L}^{(q)}}$, and $m_{\tilde{J}_{k,L}^{(q)}}$, respectively. An unbiased estimator for $H_{k,L}^{*(q)}(\mathcal{X}^{(q)}, \mathcal{Y}^{(q)})$ is straightforward to produce, but additional care must be paid to maintain the independence between the estimator for $\tilde{m}_{k,L}^{(q)}$ (or $m_{k,L}^{(q)}$) and $(\mathcal{X}^{(q)}, \mathcal{Y}^{(q)})$. To satisfy the latter, we construct the unbiased estimators $m_{k,L}^{(q)}$ and $\tilde{m}_{k,L}^{(q)}$ from all coupled processes but the $q$th one, as described in Algorithm 1.

**Algorithm 1** Algorithm for computing $m_{k,L}^{(q)}$ and $\tilde{m}_{k,L}^{(q)}$ for a fixed $k$ and all $q \in \{1, 2, \dots, Q\}$.

1. Compute $J_{k,L}^{(q)}$;
2. Sample independently $\zeta^{(q)} \sim Bernoulli(0.5)$ and set $\tilde{J}_{k,L}^{(q)} = J_{k,L}^{(q)} - \zeta^{(q)}$;
3. Set $m_{k,L}^{(q)} = \lfloor med(\{J_{k,L}^{(h)} : 1 \le h \le Q,\ h \ne q\}\rfloor$ and $\tilde{m}_{k,L}^{(q)} = \lfloor med(\{\tilde{J}_{k,L}^{(h)} : 1 \le h \le Q,\ h \ne q\}\rfloor$, where $med(A)$ denotes the median of the values in set $A$ and $\lfloor\cdot\rfloor$ is the floor function.

When $L = 1$, in order to reduce the variance of the unbiased estimator $H_k$ in (1.1), Jacob, O'Leary and Atchadé (2020) recommend using the time-averaging

estimator

$$H_{k:r}(\mathcal{X}, \mathcal{Y}) = \frac{1}{r-k+1} \sum_{t=k}^{r} H_t(\mathcal{X}, \mathcal{Y}).$$

We follow the same strategy, and consider the time-averaging version of (2.4)

$$\begin{aligned} H^{(q)}_{k:r;L}(\mathcal{X}^{(q)}, \mathcal{Y}^{(q)}) = \frac{1}{r-k+1} \sum_{t=k}^{r} h(X^{(q)}_{t+J^{(q)}_{t,L}L}) \\ + \frac{1}{r-k+1} \sum_{t=k}^{r} \sum_{j=0}^{J^{(q)}_{t,L}-1} \left[ h(X^{(q)}_{t+jL}) - h(Y^{(q)}_{t+jL}) \right], \end{aligned} \tag{3.2}$$

and the average estimator that includes the control-variate swindle is then

$$\begin{aligned} & H^{*(q)}_{k:r;L}(\mathcal{X}^{(q)}, \mathcal{Y}^{(q)}) \\ & = H^{(q)}_{k:r;L}(\mathcal{X}^{(q)}, \mathcal{Y}^{(q)}) - \frac{1}{r-k+1} \sum_{t=k}^{r} \left\{ \sum_{j=0}^{\tilde{m}^{(q)}_{k,L}} \left[ h(X^{(q)}_{t+jL}) - h(Y^{(q)}_{t+jL}) \right] \right\}. \end{aligned} \tag{3.3}$$

Note that the original versions are obtained when $k = r$. Because each term in the control-variate term above, $h(X^{(q)}_{t+jL}) - h(Y^{(q)}_{t+jL})$, has mean zero, we expect that the gain from the control variate swindle diminishes when $r$ increases owing to the law of large numbers, leading to the overall control-variate term approaching zero. We see this phenomenon in Section 4.

### 3.2. Estimating the total variation bound

When estimating $B_{k,L}$, we can use (2.11), (2.12), or (2.13). In our numerical experiments we use (2.12), along the steps described in Algorithm 2.

In the next section, we investigate the performance of the control variate swindle and compare the new total variation bound with (2.1) provided by Biswas, Jacob and Vanetti (2019).

## 4. Examples and Illustrations

### 4.1. A theoretical comparison of the bounds: The geometric case

The distribution of the coupling time $\tau_L$ is, in general, unknown. However, there is one instance in which the distribution of $\tau_L$ is exactly geometric. Specifically, when coupling two independent Metropolis samplers, the maximal coupling procedure uses the same proposal for both transition kernels, and coupling occurs when both chains accept it. The tractability of derivations in the geometric case

**Algorithm 2** Algorithm for estimation of $B_{k,L}$, for any given $k$ and $L$.

1. Compute $J_{k,L}^{(q)}$ and $m_{k,L}^{(q)}$, for all $q = 1, \ldots, Q$;
2. Compute the empirical means

$$e_{k,L} = \frac{1}{Q}\sum_{q=1}^{Q}\left|J_{k,L}^{(q)} - m_{k,L}^{(q)}\right|, \quad p_{k,L} = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{1}_{\{J_{k,L}^{(q)}>0\}},$$
$$a_{k,L} = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{1}_{\{J_{k,L}^{(q)}>m_{k,L}^{(q)}\}}, \quad b_{k,L} = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{1}_{\{J_{k,L}^{(q)}<m_{k,L}^{(q)}\}};$$

3. Compute

$$\hat{B}_{k,L} = e_{k,L} + p_{k,L} - a_{k,L} \vee b_{k,L},$$

where $a \vee b$ denotes the maximum between $a$ and $b$.

allow us to better understand theoretically the relationship between the bound in Biswas, Jacob and Vanetti (2019) and ours.

We consider the case in which $\delta = \tau - (L-1) \sim Geo(p)$. Because $J = \max\{0, \lceil(\delta - k - 1)/L\rceil\}$, we see that

$$\begin{aligned}\Pr(J = 0) &= \Pr(\delta \leq k+1) = 1 - q^{k+1},\\ \Pr(J > j) &= \Pr(\delta > k+1+Lj) = q^{k+1+Lj}, \quad j = 0, 1, \ldots,\end{aligned} \tag{4.1}$$

where $q = 1 - p$. That is, the distribution of $J$ is a mixture of (i) the Dirac point measure $\delta_{\{0\}}$ with mixture proportion $1 - q^{k+1}$, and (ii) a geometric distribution with probability of success $1 - q^L$ with weight $q^{k+1}$. This implies immediately that the bound given in Biswas, Jacob and Vanetti (2019) has the expression (2.2).

For our new bound, in this case, it is easier to use expression (2.10) directly. Let $m$ be the largest integer such that $\Pr(J \geq m) \geq \Pr(J \leq m)$; that is, $m$ is the largest integer that ensures

$$q^{k+1+L(m-1)} + q^{k+1+Lm} \geq 1 \quad \Longleftrightarrow \quad m = \left\lfloor \frac{L-k-1}{L} - \frac{\log[1+q^L]}{L\log(q)} \right\rfloor. \tag{4.2}$$

Clearly, when $m \leq 0$, our $B_{k,L}(p)$ is the same as the old bound (2.2). When $m \geq 1$, we have

$$
\begin{aligned}
B_{k,L}(p) &= \sum_{j=1}^{m} \Pr(J \le j) + \sum_{j=m+1}^{\infty} \Pr(J \ge j) \\
\{\text{by (4.1)}\} &= \sum_{j=1}^{m} \left[1 - q^{k+1+Lj}\right] + \sum_{j=m+1}^{\infty} q^{k+1+L(j-1)} \\
&= m - \frac{q^{k+1+L}[1 - q^{mL}]}{1 - q^L} + \frac{q^{k+1+mL}}{1 - q^L} \\
&= m - \frac{q^{k+1+L}[1 - q^{mL} - q^{(m-1)L}]}{1 - q^L}. \qquad (4.3)
\end{aligned}
$$

We can also compute $B_{k,L}(p)$ directly from (2.13) as an infinite sum

$$B_{k,L}(p) = 0.5 \sum_{j \ge 1} [1 - |q^{k+1+L(j-1)} + q^{k+1+Lj} - 1|] + 0.5 q^{k+1}. \qquad (4.4)$$

Figure 1 compares the bounds (2.2) (dashed line) and (4.4) (solid line) for different values of $p$, $L$, and $t$. One can see that the new bound is sharper, and that only for larger values of $p$, corresponding to fast-mixing chains, are the two bounds indistinguishable. The horizontal line in Figure 1 marks the obvious bound, because $d_{\mathrm{TV}} \le 1$. Note too that for very small values of $p$, both bounds are vacuous, but the new bound has a larger range for being nonvacuous.

The simulations in the remaining two examples rely on the `unbiasedmcmc` package of Pierre Jacob, available from: `https://github.com/pierrejacob/unbiasedmcmc/tree/master/vignettes`. Additional programs for implementing the new ideas in this paper are available as supplemental material from the authors.

### 4.2. An empirical comparison of the bounds: Ising model

The Ising model example follows the setup in Biswas, Jacob and Vanetti (2019). We consider a $32 \times 32$ square lattice of pixels with values in $\{-1, 1\}$, and with periodic boundaries. A state of the system is then $x \in \{-1, 1\}^{32 \times 32}$, and the target probability is defined as $\pi_\beta(x) \propto \exp(\beta \sum_{i \sim j} x_i x_j)$, where $i \sim j$ means that $x_i$ and $x_j$ are pixel values in neighboring sites. This illustration uses the parallel tempering algorithm (PT, see Swendsen and Wang (1986)) coupled with a single site Gibbs (SSG) updating. It is known that larger values of $\beta$ increase the dependence between neighboring sites, and that this "stickiness" leads to slow mixing of the SSG. The target of interest corresponds to $\beta_0 = 0.46$ and we use 12 chains, each corresponding to a different $\pi_\beta(x)$, with $\beta$ values equally spaced between 0.3 and $\beta_0 = 0.46$. Figure 2 shows the total variation bounds, where that

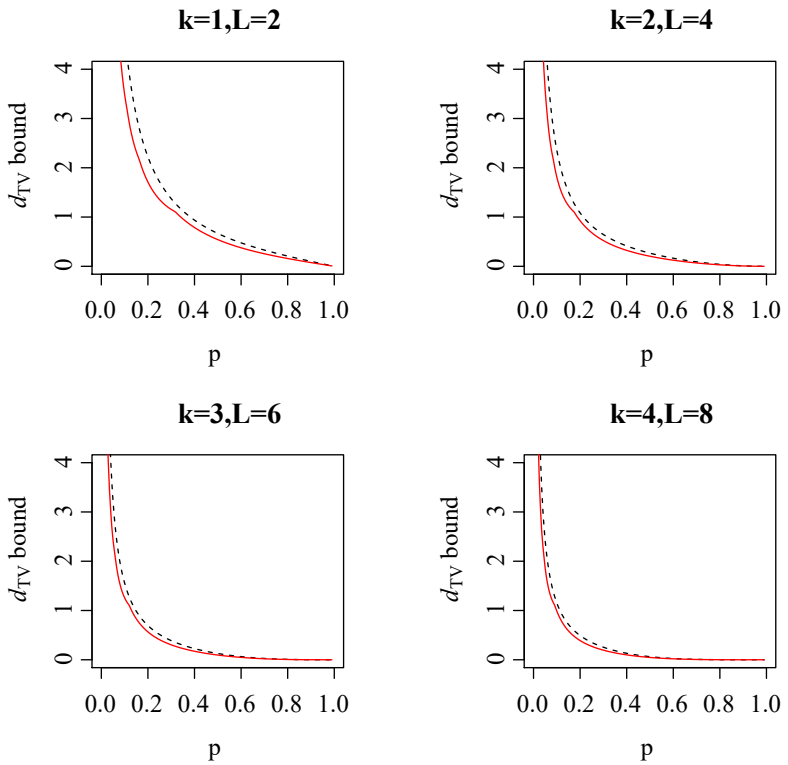


Figure 1. Comparison of the bound (2.1) provided by Biswas, Jacob and Vanetti (2019) (dashed line), and the new bound given in (2.16) (solid line). Note that for small values of $p$, both bounds are vacuous.

provided by (2.1) is shown as a dashed line and that from (2.16) is shown as a solid line. The bounds are derived for $1 \leq k < 25{,}000$ and $L \in \{1000, 2000, 3000, 4000\}$. For smaller values of $L$, the patterns are similar, but TV bounds are larger for smaller values of $k$. The new bound, (2.16), is computed from $Q = 50$ parallel runs, and is averaged over 20 independent replicates, while (2.1) is averaged over 1,000 independent replicates of a single coupled process.

Although the numerical results confirm that our new bound never exceeds the bound of Biswas, Jacob and Vanetti (2019), unfortunately, in this case, the improvement from our bound is visible only when it is not needed, that is, when both bounds exceed one. Whereas this is a disappointment for our effort to improve the bound with a real gain, it is good news for practitioners, because the bound in Biswas, Jacob and Vanetti (2019) is a bit simpler to use.

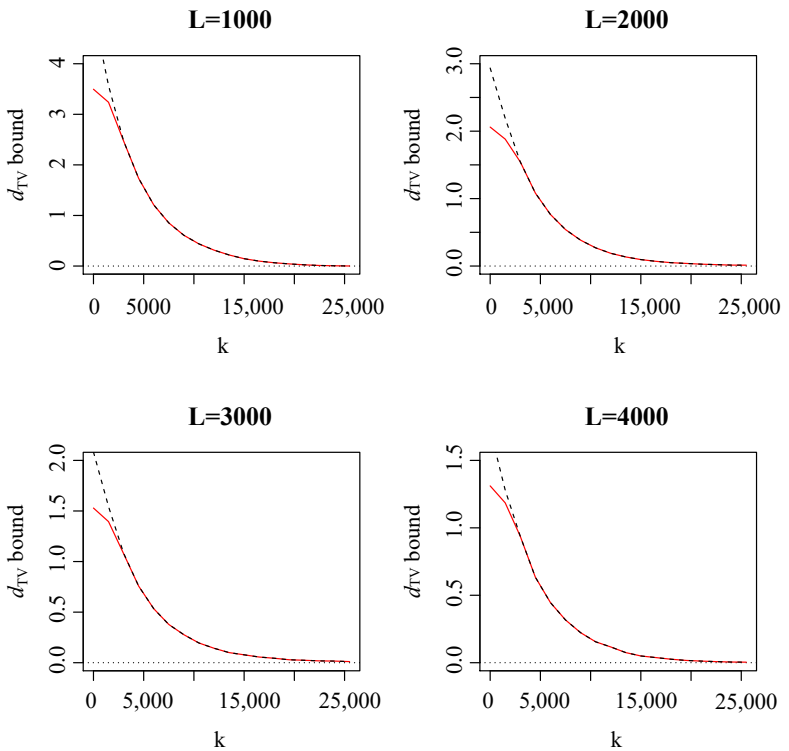


Figure 2. Ising Model: Comparison of TV bounds for the PT algorithm for SGS for $L \in \{1000, 2000, 3000, 4000\}$. The dashed line shows the bound (2.1) derived in Biswas, Jacob and Vanetti (2019), and the solid line shows the new bound given in (2.16).

### 4.3. Comparing bounds and estimators: A logistic regression example

To compare the bounds and the unbiased estimators, we follow Biswas, Jacob and Vanetti (2019) and consider a Bayesian logistic regression model for the German credit data of Lichman (2013). The data consist of $n = 1000$ binary responses, $\{Y_i : 1 \leq i \leq n\}$ and $d = 49$ covariates, $\{x_i \in \mathbf{R}^d; 1 \leq i \leq n\}$. The response $Y_i$ indicates whether the $i$th individual is fit to receive credit ($Y_i = 1$) or not ($Y_i = 0$). The logistic regression model frames the probabilistic dependence between the response and covariate as $\Pr(Y_i = 1|x_i) = [1 + \exp(-x_i^T \beta)]^{-1}$. The prior is set to $\beta \sim N(0, 10\mathbf{I}_d)$. Sampling from the posterior distribution is done using the Pólya-Gamma sampler of Polson, Scott and Windle (2013), using the R programs made available by Biswas, Jacob and Vanetti (2019) at `https://github.com/niloyb/LlagCouplings`. In Figure 3, we compare the bound of Biswas, Jacob and Vanetti (2019) (2.1) (dashed line) with our bound (2.16) (solid

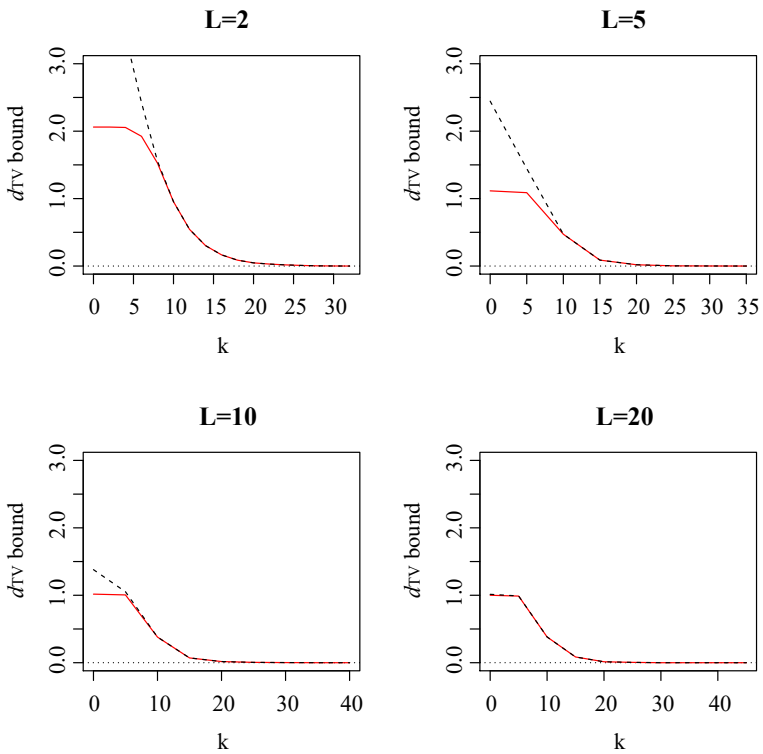


Figure 3. German credit Ddta: Comparison of TV bounds for the Pólya-Gamma sampler for $L \in \{2, 5, 10, 20\}$. The dashed line shows the bound (2.1) derived in Biswas, Jacob and Vanetti (2019), and the solid line shows the new bound given in (2.16). The bound from (2.16) is obtained from running 50 coupled chains in parallel and averaging over 40 independently replicated experiments. The bound from (2.1) is averaged over 2,000 independent replicates.

line). The bound in (2.1) is averaged over 2,000 independent replicates. The new bound is computed from running 50 coupled processes in parallel and averaged over 40 replicates, yielding the same number of runs. The difference between the two bounds is apparent for smaller values of $L$ when the new bound is sharper for small values of $k$, but the gain diminishes quickly as $L$ increases.

We are also interested in the gains in efficiency for the Monte Carlo estimators when implementing the control variate swindle. Using 500 independent replicates of a single coupled process with lag $L = 5$, we obtain Monte Carlo estimates of the posterior means for the regression coefficients. In Figure 4, we present the relative reduction in variance (RRV), computed as RRV$= \mathrm{Var}_{MCCV}(\hat{\beta})/\mathrm{Var}_{MC}(\hat{\beta})$, where $\hat{\beta}$ is the posterior mean of the regression coefficients, $\beta \in \mathbf{R}^{49}$, and $\mathrm{Var}_{MC}$ and

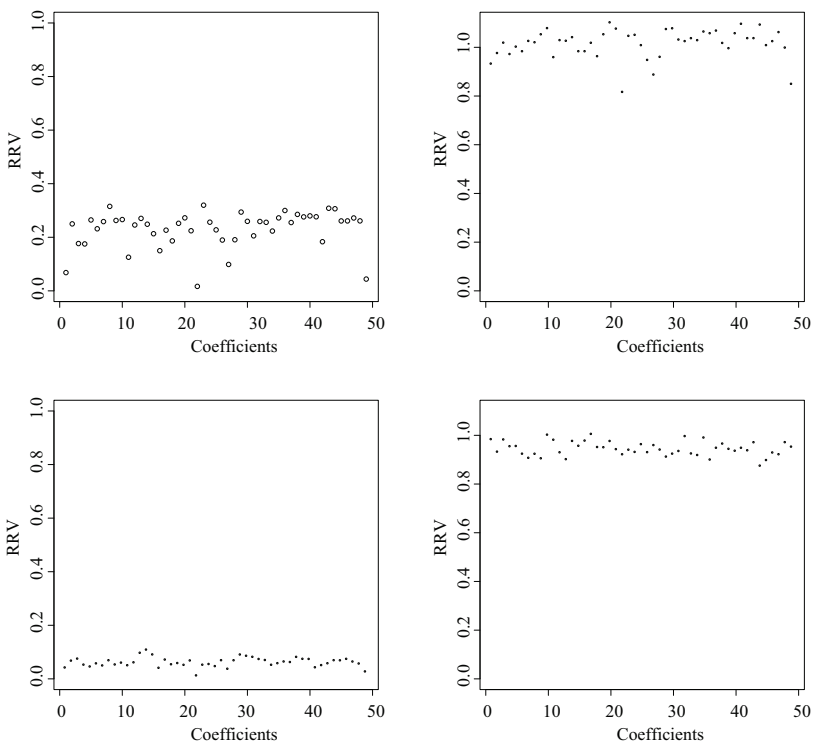


Figure 4. German credit data. Relative reduction in variance (RRV) for the 49 regression coefficients. *Top panels*: the lag is $L = 5$. *Bottom panels*: the lag is $L = 20$. *Left panels*: RRV is obtained from the single estimators without and with control variates, respectively, using $k = 5$ in (2.4) and (3.1). *Right panels*: RRV is obtained from the average estimators without and with control variates, respectively, using $k = 5$ and $r = 30$ in (3.2) and (3.3).

$\mathrm{Var}_{MCCV}$ denote the estimated Monte Carlo variances of $\hat{\beta}$ obtained without and with the control variates, respectively. The left panel shows the RRV when using the single-run estimators (2.4) and (3.1), while the right panel plots the RRV for the mean estimators (3.2) and (3.3). We see clearly that the gain is significant for $r = k = 5$ (left panels), but diminishes when $k = 5$ and $r = 30$ (right panels), as discussed in Section 3.

## 5. Can We Do Even Better?

The idea of L-lag coupling has opened multiple avenues for future research. The use of control variates is just one of them. Although the practical gain is small or possibly even negative when we take into account the increased com-

putation when computing the control variates, the theoretical gain is intriguing, because we obtain a theoretically superior bound without imposing any additional assumptions. This naturally raises the question of whether our bound is the best possible without further conditions. We do not know. We do not even know how to study such a question theoretically, because to the best of our knowledge, this is the first time a tighter *theoretical bound* has been obtained by a better *empirical estimator*. Whereas seeking other more efficient estimators seems to be a natural direction, we must keep in mind that they would likely incur additional computational costs.

One plausible direction is to go beyond linearly combining mean-zero control variates, although we had no success so far. However, even without seeking better bounds, our current bounds already offer the opportunity to investigate fresh perspectives for optimizing an MCMC kernel using adaptive ideas, and we intend to pursue these in future research.

## Acknowledgments

We are grateful to Pierre Jacob for many useful comments and his gracefully patient guidance through the package `unbiasedmcmc`, allowing us to perform the simulations in our study. We also thank Yves Atchadé, Tamas Papp, Christopher Sherlock, and Lei Sun for their helpful discussions and comments, and the NSERC of Canada (RVC) and NSF of USA (XLM) for their partial research support.

## References

Berthelsen, K. K. and Møller, J. (2002). A primer on perfect simulation for spatial point processes. *Bull. Braz. Math. Soc. (N.S.)* **33**, 351–367.

Biswas, N., Jacob, P. E. and Vanetti, P. (2019). Estimating convergence of Markov chains with L-lag couplings. In *Advances in Neural Information Processing Systems*, 7389–7399.

Corcoran, J. N. and Schneider, U. (2005). Pseudo-perfect and adaptive variants of the Metropolis-Hastings algorithm with an independent candidate density. *J. Stat. Comput. Simul.* **75**, 459–475.

Corcoran, J. N. and Tweedie, R. L. (2002). Perfect sampling from independent Metropolis-Hastings chains. *J. Statist. Plann. Inference* **104**, 297–314.

Craiu, R. V. and Lemieux, C. (2007). Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling. *Stat. Comput.* **17**, 109–120.

Craiu, R. V. and Meng, X.-L. (2001). Antithetic coupling for perfect sampling. In *Bayesian Methods, with Applications to Science, Policy and Official Statistics (Proceedings of the ISBA 2000 conference, Hersonnissos, Crete)* (Edited by E. I. George), 99-108. Office for Official Publications of the European Communities, Luxembourg.

Craiu, R. V. and Meng, X.-L. (2005). Multiprocess parallel antithetic coupling for backward and forward Markov chain Monte Carlo. *Ann. Statist.* **33**, 661–697.

Craiu, R. V. and Meng, X.-L. (2011). Perfection within reach: exact MCMC sampling. In *Handbook of Markov Chain Monte Carlo* (Edited by S. Brooks, A. Gelman, G. Jones and X.-L. Meng), 199–226. Chapman and Hall/CRC, New York.

Craiu, R. V. and Meng, X.-L. (2020). Discussion of "Unbiased Markov chain Monte Carlo with couplings" by Pierre E. Jacob, John O'Leary and Yves F. Atchadé. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 578–581.

Dobrow, R. P. and Fill, J. A. (2003). Speeding up the FMMR perfect sampling algorithm: A case study revisited. *Random Struct. Algorithms* **23**, 434–452.

Ensor, K. B. and Glynn, P. W. (2000). Simulating the maximum of a random walk. *J. Statist. Plann. and Inference* **85**, 127–135.

Glynn, P. W. (2016). Exact simulation vs exact estimation. In *2016 Winter Simulation Conference (WSC)*, 193–205. IEEE.

Glynn, P. W. and Heidelberger, P. (1991). Analysis of parallel replicated simulations under a completion time constraint. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **1**, 3–23.

Glynn, P. W. and Rhee, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.* **51**, 377–389.

Heng, J. and Jacob, P. E. (2019). Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika* **106**, 287–302.

Huber, M. L. (2002). A bounding chain for Swendsen-Wang. *Random Struct. Algorithms* **22**, 43–59.

Huber, M. L. (2004). Perfect sampling using bounding chains. *Ann. Appl. Probab.* **14**, 734–753.

Jacob, P. E., Lindsten, F. and Schön, T. B. (2020). Smoothing with couplings of conditional particle filters. *J. Amer. Statist. Assoc.* **115**, 721–729.

Jacob, P. E., O'Leary, J. and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo with couplings (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82** , 543–600.

Lichman, M. (2013). Uci machine learning repository, 2013.

Meng, X. L. (2000). Towards a more general Propp-Wilson algorithm: Multistage backward coupling. In *Monte Carlo Methods*, *Fields Institute Communications* (Edited by N. Madras), 85–93. American Mathematical Society, Providence.

Møller, J. (1999). Perfect simulation of conditionally specified models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61**, 251–264.

Murdoch, D. J. and Meng, X.-L. (2001). Towards perfect sampling for Bayesian mixture priors. In *Bayesian Methods, with Applications to Science, Policy and Official Statistics (Proceedings of the ISBA 2000 conference, Hersonnissos, Crete)* (Edited by E. I. George), 381-390. Office for Official Publications of the European Communities, Luxembourg.

Murdoch, D. J. and Takahara, G. (2006). Perfect sampling for queues and network models. *ACM Transactions on Modeling and Computer Simulation* **16**, 76–92.

Nelson, B. L. (2016). 'Some tactical problems in digital simulation' for the next 10 years. *Journal of Simulation* **10**, 2–11.

Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* **108**, 1339–1349.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct. Algorithms* **9**, 223–252.

Propp, J. G. and Wilson, D. B. (1998). How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *J. Algorithms* **27**, 170–217.

Stein, N. M. and Meng, X.-L. (2013). Practical perfect sampling using composite bounding chains: the Dirichlet-multinomial model. *Biometrika* **100**, 817–830.

Swendsen, R. H. and Wang, J.-S. (1986). Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **57**, 2607–2609.

Thönnes, E. (1999). Perfect simulation of some point processes for the impatient user. *Adv. in Appl. Probab.* **31**, 69–87.

Van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion). *J. Comput. Graph. Statist.* **10**, 1–50.

Wilson, D. B. (1998). Annotated bibliography of perfectly random sampling with Markov chains. In, *Microsurveys in Discrete Probability* (Edited by D. Aldous and J. Propp) **41**, 209–220. American Mathematical Society, Providence. Updated versions at `http://www.dbwilson.com/exact/`.

Yu, Y. and Meng, X.-L. (2011). To center or not to center: That is not the question—an Ancillarity–Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency (with discussion). *J. Comput. Graph. Statist.* **20**, 531–570.

Radu V. Craiu

Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3, Cananda.

E-mail: craiu@utstat.toronto.edu

Xiao-Li Meng

Department of Statistics, Harvard University, Cambridge, MA 02138-2901, USA.

E-mail: meng@stat.harvard.edu

(Received November 2020; accepted January 2021)