JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–17: Theory and Methods
https://doi.org/10.1080/01621459.2020.1825447

Warp Bridge Sampling: The Next Generation

Lazhi Wang[a], David E. Jones[b], and Xiao-Li Meng[c]

[a]Two Sigma Investments, LP, New York, NY; [b]Department of Statistics, Texas A&M University, College Station, TX; [c]Department of Statistics, Harvard University, Cambridge, MA

ABSTRACT

Bridge sampling is an effective Monte Carlo (MC) method for estimating the ratio of normalizing constants of two probability densities, a routine computational problem in statistics, physics, chemistry, and other fields. The MC error of the bridge sampling estimator is determined by the amount of overlap between the two densities. In the case of unimodal densities, Warp-I, II, and III transformations are effective for increasing the initial overlap, but they are less so for multimodal densities. This article introduces Warp-U transformations that aim to transform multimodal densities into unimodal ones (hence "U") without altering their normalizing constants. The construction of a Warp-U transformation starts with a normal (or other convenient) mixture distribution ϕ_{mix} that has reasonable overlap with the target density p, whose normalizing constant is unknown. The stochastic transformation that maps ϕ_{mix} back to its generating distribution $\mathcal{N}(0,1)$ is then applied to p yielding its Warp-U version, which we denote $\tilde{p}$. Typically, $\tilde{p}$ is unimodal and has substantially increased overlap with ϕ. Furthermore, we prove that the overlap between $\tilde{p}$ and $\mathcal{N}(0,1)$ is guaranteed to be no less than the overlap between p and ϕ_{mix}, in terms of any f-divergence. We propose a computationally efficient method to find an appropriate ϕ_{mix}, and a simple but effective approach to remove the bias which results from estimating the normalizing constants and fitting ϕ_{mix} with the same data. We illustrate our findings using 10 and 50 dimensional highly irregular multimodal densities, and demonstrate how Warp-U sampling can be used to improve the final estimation step of the Generalized Wang–Landau algorithm, a powerful sampling and estimation approach. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2019
Accepted September 2020

KEYWORDS
Bridge sampling; f-Divergence; Monte Carlo integration; Normal mixture; Normalizing constants; Optimal transport; Stochastic transformation; Wang–Landau algorithm

1. Motivation

Markov chain Monte Carlo (MCMC) methods, such as the Metropolis–Hastings algorithm, enable us to simulate from an unnormalized density without knowing its normalizing constant. However, in many scientific and statistical studies the very quantities of interest are normalizing constants or ratios of them (see, e.g., Voter 1985; Kass and Raftery 1995; Meng and Wong 1996; DiCiccio et al. 1997; Gelman and Meng 1998; Shao and Ibrahim 2000; Tan 2013). A well-known example from physics and chemistry is the computation of partition functions, which describe the statistical properties of a system in thermodynamic equilibrium. A partition function is the integral of an unnormalized density $q(\omega; T, \nu) = \exp\{-H(\omega,\nu)/(kT)\}$, where T is temperature, k is Boltzmann's constant, ν is a vector of system characteristics, and $H(\omega,\nu)$ is the energy function. Because of the high dimensionality of H, Monte Carlo (MC) methods are often the only feasible tool for estimating a partition function, that is, the normalizing constant of q (see Bennett 1976; Voter and Doll 1985; Ceperley 1995).

Two key objects in statistics which can be expressed as normalizing constants are observed-data likelihoods and Bayes factors. Focusing on the latter, let Y be our data, and let M_0 and M_1 be two plausible models parameterized by Θ_0 and Θ_1, respectively. The Bayes factor is then the ratio of the model likelihoods, $P(Y|M_0)$ and $P(Y|M_1)$, where

$$P(Y|M_i) = \int P(Y|\Theta_i, M_i)P(\Theta_i|M_i)\mathbf{u}(\mathrm{d}\Theta_i)$$

is the normalizing constant of the unnormalized density, $P(\Theta_i, Y|M_i) \propto P(Y|\Theta_i, M_i)P(\Theta_i|M_i)$, for $i = 1, 2$. In most Bayesian analyses, MC draws of Θ_i from $P(\Theta_i|Y, M_i)$ are made for the purpose of statistical inference, often using MCMC methods. Hence, to estimate $P(Y|M_i)$, it is desirable to use methods that require only these available draws (plus perhaps some draws from another convenient distribution).

One such method is the bridge sampling approach introduced by Bennett (1976) and generalized and popularized by Meng and Wong (1996). In this article, we propose a method to improve the efficiency of bridge sampling in the multimodal context. Specifically, we introduce a class of stochastic transformations, Warp-U transformations, that can warp two multimodal densities into densities having substantial overlap but without changing their respective normalizing constants. For bridge sampling, an increase in distributional overlap implies superior statistical efficiency. The key idea of Warp-U transformations is to approximate the unnormalized density of interest q by a mixture distribution (e.g., a normal mixture), and then to

CONTACT David E. Jones djones@stat.tamu.edu Department of Statistics, Texas A&M University, College Station, TX 77843.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

construct a coupling which allows us to (stochastically) map q to a unimodal density in the same way that the approximating mixture can be mapped back to a single generative density (e.g., a single normal density). Our work builds on the warp transformations (centering, scaling, and symmetrizing) for unimodal densities that were proposed by Meng and Schilling (2002). Our method also has an intrinsic connection with the theory of optimal transport (e.g., Villani 2003), albeit here we typically seek only a reasonable transport (from one density to another) which can achieve a beneficial compromise between statistical efficiency and computational efficiency.

The utility of Warp-U transformations is especially promising because bridge sampling is similar to many other mixture sampling approaches, which may also benefit from the strategy. Indeed, a number of methods in the literature turn out to be special cases of bridge sampling or adaptations of it, as demonstrated by Mira and Nicholls (2004). For instance, the marginal likelihood approaches of Chib (1995) and Chib and Jeliazkov (2001) based on Metropolis–Hastings output correspond to bridge sampling with specific choices of the bridge density. Similarly, the defensive sampling method of Hesterberg (1995) for estimating normalizing constants can be directly interpreted as bridge sampling. The "balance weight heuristic" introduced by Veach and Guibas (1995) is a generalization of bridge sampling where the unnormalized densities to be integrated are not necessarily included among the sampling densities. This more general bridge sampling is also covered by the likelihood approach proposed by Kong et al. (2003), which reformulates MC integration as an estimation problem with the dominating measure as the estimand. Their likelihood framework provides a unified way of deriving and characterizing various methods for boosting the statistical efficiency of MC estimation strategies, such as the control variates approach of Owen and Zhou (2000); see Tan (2004), Meng (2005), and Kong et al. (2006) for details and illustrations. There is a possibility to cast Warp-U bridge sampling methods under the same likelihood framework and thereby make further connections with other methods, but that is a topic for future exploration; see Section 6.2.

Our proposed Warp-U bridge sampling method can also be used to improve the accuracy of powerful adaptive importance sampling based algorithms which combine sampling and estimation. In Section 5, we illustrate this application of our method in the special case of the GWL algorithm (Liang 2005), an extension of the discrete algorithm proposed by Wang and Landau (2001). There are numerous other adaptive importance sampling methods that have been proposed and these could potentially benefit from a similar combination with Warp-U bridge sampling, for example, dynamic weighting (e.g., Wong and Liang 1997; Liu, Liang, and Wong 2001) and layered adaptive importance sampling (Martino et al. 2017).

Our article is organized as follows. Section 2 briefly overviews bridge sampling and the warp transformations of Meng and Schilling (2002), highlighting their power for increasing distribution overlap. Section 3 defines and illustrates the Warp-U transformation we propose and then establishes the theoretical result that Warp-U transformations never reduce distribution overlap. Section 4 outlines a computationally efficient strategy for finding a specific Warp-U transformation and studies the properties of the corresponding estimator. Section 5 demonstrates that the estimation performance of the aforementioned GWL algorithm can be improved by combining it with Warp-U bridge sampling. Section 6 discusses limitations and future work. The appendices in the online supplementary materials provide a proof of Theorem 1, computational details, and guidance on selecting the tuning parameters of our method.

2. The Basics of Warp Bridge Sampling

Bridge sampling (Bennett 1976; Meng and Wong 1996) estimates the ratio of the normalizing constants of two unnormalized densities by leveraging the overlap between the two densities. Any method that can increase this overlap has the potential to reduce the MC error. The warp bridge sampling of Meng and Schilling (2002) explored this idea by transforming the original MC draws so that the densities of the transformed draws have substantially more overlap.

Let q_i be an unnormalized density with normalizing constant c_i, for $i = 1, 2$. Furthermore, let $\mathbf{u}$ be the underlying measure common to both densities, typically the Lebesgue or counting measure. We use p_i to denote the normalized density, that is, $p_i(\omega) = c_i^{-1} q_i(\omega)$, for $\omega \in \Omega_i$, where Ω_i is the support of q_i. Our goal is to estimate the ratio $r = c_1/c_2$ or $\lambda = \log(r)$, using the draws, $\{w_{i,1}, w_{i,2}, \ldots, w_{i,n_i}\}$, from p_i, for $i = 1, 2$. In some instances, we wish only to estimate one normalizing constant c_1, in which case we will select $q_2 = p_2$ to be a convenient density with $c_2 = 1$ (discussed in Section 3). Below, we begin by assuming that draws from p_1 and p_2 are given, but in Section 5 we combine our estimation strategy with the GWL sampling algorithm.

2.1. Bridge Sampling

Here, we review the key aspects of bridge sampling that will be used in this article. For a complete treatment, the reader is referred to Meng and Wong (1996) and the practical introduction by Gronau et al. (2017). An R package developed by Gronau, Singmann, and Wagenmakers (2017) is available at *https://cran.r-project.org/web/packages/bridgesampling/*.

Bridge sampling relies on the fact that for any function, α, defined on $\Omega_1 \cap \Omega_2$ and satisfying $0 < \left| \int_{\Omega_1 \cap \Omega_2} \alpha(\omega) p_1(\omega) p_2(\omega) \mathbf{u}(\mathrm{d}\omega) \right| < \infty$, the following identity holds:

$$r = \frac{c_1}{c_2} = \frac{E_2[q_1(\omega)\alpha(\omega)]}{E_1[q_2(\omega)\alpha(\omega)]}, \tag{1}$$

where E_i represents expectation with respect to p_i. Here, α serves as a "bridge" connecting p_1 and p_2. The bridge sampling estimator of r is the sample counterpart of (1), that is,

$$\hat{r}_\alpha = \frac{n_2^{-1} \sum_{j=1}^{n_2} q_1(w_{2,j})\alpha(w_{2,j})}{n_1^{-1} \sum_{j=1}^{n_1} q_2(w_{1,j})\alpha(w_{1,j})}. \tag{2}$$

For example, both importance sampling and geometric bridge sampling are special cases of bridge sampling, with $\alpha_{\text{imp}} \propto 1/q_2$ and $\alpha_{\text{geo}} \propto 1/\sqrt{q_1 q_2}$, respectively.

Different choices of α lead to estimators with different statistical efficiency, which we quantify by the asymptotic variance of $\hat{\lambda}_\alpha = \log(\hat{r}_\alpha)$, or equivalently, the asymptotic relative variance of $\hat{r}_\alpha$, $E(\hat{r}_\alpha - r)^2/r^2$. Under the assumption that all the MC draws used in (2) are identically and independently distributed (iid), Meng and Wong (1996) derived the first-order asymptotic variance of $\hat{\lambda}_\alpha$, from which they found that the optimal bridge has the form

$$\alpha_{\text{opt}}(\omega) \propto \frac{1}{s_1 q_1(\omega) + r s_2 q_2(\omega)}, \quad \text{where } s_i = \frac{n_i}{n},\ i = 1, 2. \tag{3}$$

Before we proceed, we emphasize that the bridge sampling method itself does not require the assumption of iid sampling; otherwise the method would be too limited to deserve a general R package. The iid assumption was invoked by Meng and Wong (1996) to make the theoretical calculation both feasible and insightful, in the sense that the resulting optimal bridge (3) takes an appealing mixture form which provides practical guidance. Indeed, regardless of whether the iid assumption holds, (3) provides a very effective bridge. In contrast, without the iid assumption the optimal bridge has a very involved expression (Romero 2003), and offers little practical guidance. Therefore, for the rest of the article we invoke the iid assumption only for theoretical claims (e.g., when we refer to the "optimal" approach) or for simulation simplicity.

Because α_{opt} depends on the unknown quantity r, Meng and Wong (1996) proposed an iterative sequence that rapidly converges to $\hat{r}_{\text{opt}}$, that is,

$$\hat{r}_{\text{opt}}^{(t+1)} = \frac{\frac{1}{n_2}\sum_{j=1}^{n_2}\left[\frac{l_{2,j}}{s_1 l_{2,j} + s_2 \hat{r}_{\text{opt}}^{(t)}}\right]}{\frac{1}{n_1}\sum_{j=1}^{n_1}\left[\frac{1}{s_1 l_{1,j} + s_2 \hat{r}_{\text{opt}}^{(t)}}\right]}, \tag{4}$$

where $l_{i,j} = q_1(w_{i,j})/q_2(w_{i,j})$, for $i = 1, 2$, and $j = 1, 2, \ldots, n_i$. Meng and Wong (1996) showed that, under the iid assumption, the asymptotic variance of $\hat{\lambda}_{\text{opt}} = \log(\hat{r}_{\text{opt}})$ is

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left[\left(1 - H_{\text{A}}(p_1, p_2)\right)^{-1} - 1\right] + o\left(\frac{1}{n_1 + n_2}\right), \tag{5}$$

which is the same as the asymptotic variance of the unobtainable optimal estimator $\hat{\lambda}_{\alpha_{\text{opt}}} = \log(\hat{r}_{\alpha_{\text{opt}}})$. Here, $H_{\text{A}}(p_1, p_2)$ is the sample-size adjusted harmonic divergence between p_1 and p_2:

$$H_{\text{A}}(p_1, p_2) = 1 - \int_{\Omega_1 \cap \Omega_2} \left[w_1 p_1^{-1}(\omega) + w_2 p_2^{-1}(\omega)\right]^{-1} \mathbf{u}(\mathrm{d}\omega), \tag{6}$$

with $w_i = s_i^{-1}/(s_1^{-1} + s_2^{-1}), i = 1, 2$. Using a likelihood that treats the baseline measure $\mathbf{u}$ as the (infinite dimensional) parameter, Kong et al. (2003) showed that $\hat{r}_{\text{opt}}$ is the maximum likelihood estimator for r (again, under the iid assumption), thereby further confirming its optimality.

2.2. Warp Bridge Sampling

For $i = 1, 2$, consider a transformation $\mathcal{F}_i$ of $w_{i,j}$ such that (a) the unnormalized density, $\tilde{q}_i$, of the transformed draws, $\tilde{w}_{i,j} = \mathcal{F}_i(w_{i,j})$, has the same normalizing constant as q_i, and (b) $H_{\text{A}}(\tilde{p}_1, \tilde{p}_2) < H_{\text{A}}(p_1, p_2)$. Then by (5), the optimal bridge sampling estimator based on the transformed draws $\{(\tilde{w}_{i,1}, \ldots, \tilde{w}_{i,n_i}); i = 1, 2\}$ will have smaller asymptotic variance than that based on the original draws $\{(w_{i,1}, \ldots, w_{i,n_i}); i = 1, 2\}$, assuming the draws are independent. This observation motivated the Warp transformations proposed by Meng and Schilling (2002), whose contribution also demonstrated empirically the benefit of Warp transformations under general MCMC settings (i.e., without requiring iid draws).

The simple idea of Warp-I transformations is to increase overlap among densities (e.g., in terms of H_A in (6)) by shifting them so that they share a common location. Specifically, let μ_i be a location parameter (e.g., mean or mode) of p_i, for $i = 1, 2$, and suppose that the dominating measure (e.g., the Lebesgue measure) is invariant to translation. Let $\tilde{w}_{i,j}^{(\text{I})} = w_{i,j} - \mu_i$ and denote the corresponding unnormalized density by $\tilde{q}_i^{(\text{I})}(w) = q_i(w + \mu_i)$; clearly this density has the same normalizing constant c_i as the original target $q_i(w)$, for $i = 1, 2$. The Warp-I bridge sampling estimator is then obtained by replacing $w_{i,j}$ and q_i in (2) with $\tilde{w}_{i,j}^{(\text{I})}$ and $\tilde{q}_i^{(\text{I})}$, respectively.

The next obvious transformation is to match both the location and the spread. Let μ_i be a location parameter and $\mathcal{S}_i$ be a scaling parameter, for $i = 1, 2$. The Warp-II transformation then sets $\tilde{w}_{i,j}^{(\text{II})} = \mathcal{S}_i^{-1}(w_{i,j} - \mu_i)$ and $\tilde{q}_i^{(\text{II})}(\omega) = |\mathcal{S}_i| q_i(\mathcal{S}_i\omega + \mu_i)$. The dash-dot curve in the left panel of Figure 1 illustrates that $\tilde{p}_1^{(\text{II})}$ overlaps more with p_2 than p_1 does. It also overlaps more than the Warp-I transformed density $\tilde{p}_1^{(\text{I})}$ does (not shown).

Warp-III transformations increase overlap further by making the densities in question symmetric via a *stochastic transformation*. Specifically, a Warp-III transformation sets $\tilde{w}_{i,j}^{(\text{III})} = \xi_j \mathcal{S}_i^{-1}(w_{i,j} - \mu_i)$, where ξ_j takes on the value 1 or -1 with equal probability (independently of $w_{i,j}$). The unnormalized density of $\tilde{w}^{(\text{III})}$ is $\tilde{q}_i^{(\text{III})}(\omega) = |\mathcal{S}_i|\left[q_i(\mu_i - \mathcal{S}_i\omega) + q_i(\mu_i + \mathcal{S}_i\omega)\right]/2$, an example of which is shown in the right panel of Figure 1 (dash-dot curve). Below we show that stochastic transformations are also very powerful in dealing multimodality, a challenging issue in MC based estimation and indeed in statistical inference more generally.

3. Warp-U Bridge Sampling

Consider a unimodal density, ϕ, such as a $\mathcal{N}(0, I_d)$ or t-distribution. The key idea of our approach is to construct a stochastic transformation of the original MC draws such that the density for the transformed draws is much closer to ϕ. To simplify the exposition, we consider the problem of estimating a single normalizing constant and fix the other density used in the bridge sampling estimator (2) to be ϕ. The problem of estimating a ratio of two normalizing constants can then be handled in the following two ways. First, we can use two bridge sampling estimators, one in the numerator and one in the denominator, based on the Warp-U transformed draws $\{\tilde{w}_{i,1}, \ldots, \tilde{w}_{i,n_i}\} \sim \tilde{p}_i$

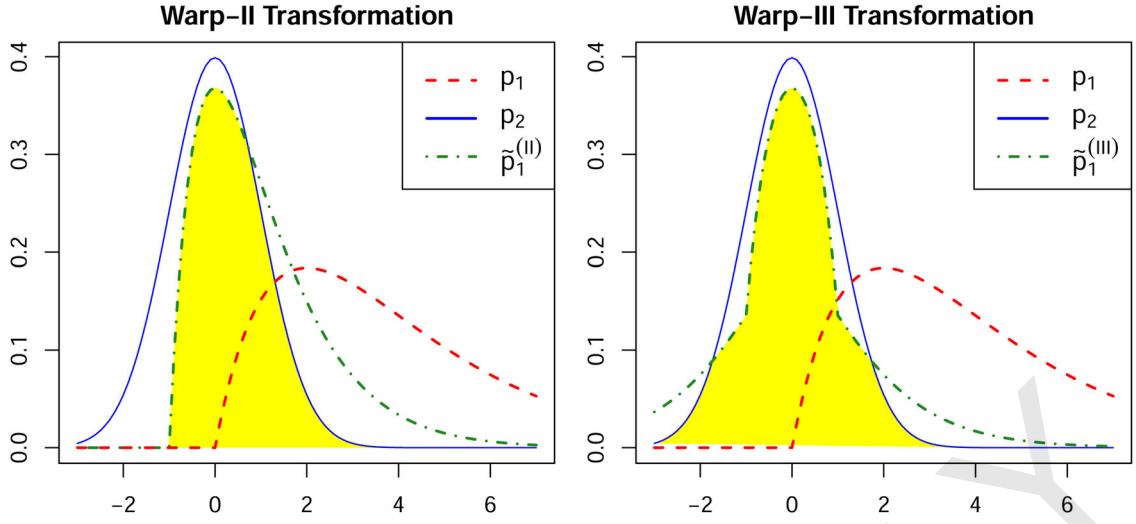

Figure 1. Graphical illustration of Warp-II (left) and Warp-III transformations (right). The dashed and the solid lines are the curves of p_1 and p_2. The dash-dot lines are $p_1^{(\mathrm{II})}$ (left) and $p_1^{(\mathrm{III})}$ (right), obtained by Warp-II and Warp-III transformation, respectively. The shaded areas highlight the much increased overlap between the warp-transformed densities and the reference density $p_2 = N(0, 1)$.

and draws from the convenient unimodal *auxiliary* distribution $\{z_{i,1}, \ldots, z_{i,m_i}\} \sim \phi$, for $i = 1, 2$, respectively. The two estimators can share the same auxiliary distribution ϕ, or even the same set of auxiliary draws. We emphasize again here that we do not require any of these draws to be iid, though typically those from the auxiliary distribution are iid by design. Second, we could disregard ϕ after the warp transformation and then use one bridge sampling estimator of the ratio r based only on the full set of the transformed draws $\{\tilde{w}_{i,1}, \ldots, \tilde{w}_{i,n_i},\ i = 1, 2\}$. This second strategy is effective because if $\tilde{p}_1$ and $\tilde{p}_2$ both overlap significantly with ϕ then they are likely to also have substantial overlap with each other.

Since we focus on a single unnormalized density q, we drop the double indices and let $\{w_1, \ldots, w_n\}$ be n draws from $p = c^{-1}q$, where p is assumed to be a continuous density on $\mathbb{R}^d$. Similarly, we use $\{z_1, \ldots, z_m\}$ to denote m iid draws from ϕ. For concreteness, we set $\phi = \mathcal{N}(0, I_d)$, but other choices of ϕ can work equally well or even better. For instance, if p is heavy-tailed then ϕ_{mix} may require many Gaussian components to well approximate p, in which case using t-distribution components will likely be more parsimonious and computationally efficient. In general, a good choice for the components will be fast to evaluate (compared with p) and have tails similar to p. More discussion on the choice of the mixture components is given in Section 6.2. Importantly, a precise match to p is not required, only a reasonable approximation.

3.1. Constructing Warp-U Transformations

When q is multimodal, we could approximate it by a Gaussian mixture ϕmix and then perform standard bridge sampling using q and ϕmix. Warp-U bridge sampling aims to improve on this approach but begins in the same way. Let

$$\phi_{\mathrm{mix}}(x; \boldsymbol{\zeta}) = \sum_{k=1}^{K} \phi^{(k)}(x) = \sum_{k=1}^{K} \pi_k \left|\mathcal{S}_k\right|^{-1} \phi\left(\mathcal{S}_k^{-1}(x - \mu_k)\right), \tag{7}$$

where $\phi^{(k)}$ represents the kth component in ϕ_{mix}, including its weight π_k, for $k = 1, \ldots, K$, and $\boldsymbol{\zeta}$ collects the transformation parameters $\{\pi_k, \mu_k, \mathcal{S}_k,\ k = 1, \ldots, K\}$. Alspach and Sorenson (1972) showed that any piecewise continuous density can be approximated arbitrarily well by a Gaussian mixture of the form (7) as $K \to \infty$ (specifically, they demonstrated uniform convergence). In practice, for a reasonable choice of K, it is usually possible to find a ϕ_{mix} that has substantial overlap with p. Section 4 will discuss how to estimate ϕ_{mix}. Here, we assume that ϕ_{mix} is known.

The Warp-U transformation uses a coupling between augmented random variables drawn from ϕmix and p, and we now specify this relationship. Suppose $X \sim \phi_{\mathrm{mix}}$, depicted in Figure 2(a) as the solid line, then we can write $X = \mathcal{S}_\Theta Z + \mu_\Theta$, where $Z \sim \phi$ and is independent of Θ, a discrete random variable distributed such that $P(\Theta = k) = \pi_k$ for $k = 1, \ldots, K$. Figure 2(b) shows the joint distribution of Θ and X, with their marginal distributions on the two faded vertical plates. The random index Θ induces a random transformation

$$\mathcal{F}_\Theta(x) = \mathcal{S}_\Theta^{-1}(x - \mu_\Theta). \tag{8}$$

It follows trivially that if we draw (x, θ) from the joint distribution of (X, Θ), then $\tilde{x} = \mathcal{F}_\theta(x) \sim \phi$.

Next, let W be a random variable from p and Ψ be a random index. We create a coupling between (W, Ψ) and (X, Θ) by requiring that $\Psi|W$ and $\Theta|X$ have the same distribution, that is,

$$\begin{aligned} \varpi(k|\omega) &\triangleq P(\Psi = k|W = \omega) \equiv P(\Theta = k|X = \omega) \\ &= \phi^{(k)}(\omega)/\phi_{\mathrm{mix}}(\omega), \quad k = 1, \ldots, K. \end{aligned} \tag{9}$$

We can then decompose p into K components, that is, $p(\omega) = \sum_{k=1}^{K} p^{(k)}(\omega)$, where

$$p^{(k)}(\omega) = p(\omega, \Psi = k) = p(\omega)\frac{\phi^{(k)}(\omega)}{\phi_{\mathrm{mix}}(\omega)}. \tag{10}$$

Figure 2(c) shows the joint distribution of (W, Ψ) (thick dashed curves) and their marginal distributions (thin dash curves in the two vertical plates). The Warp-U transformation is then

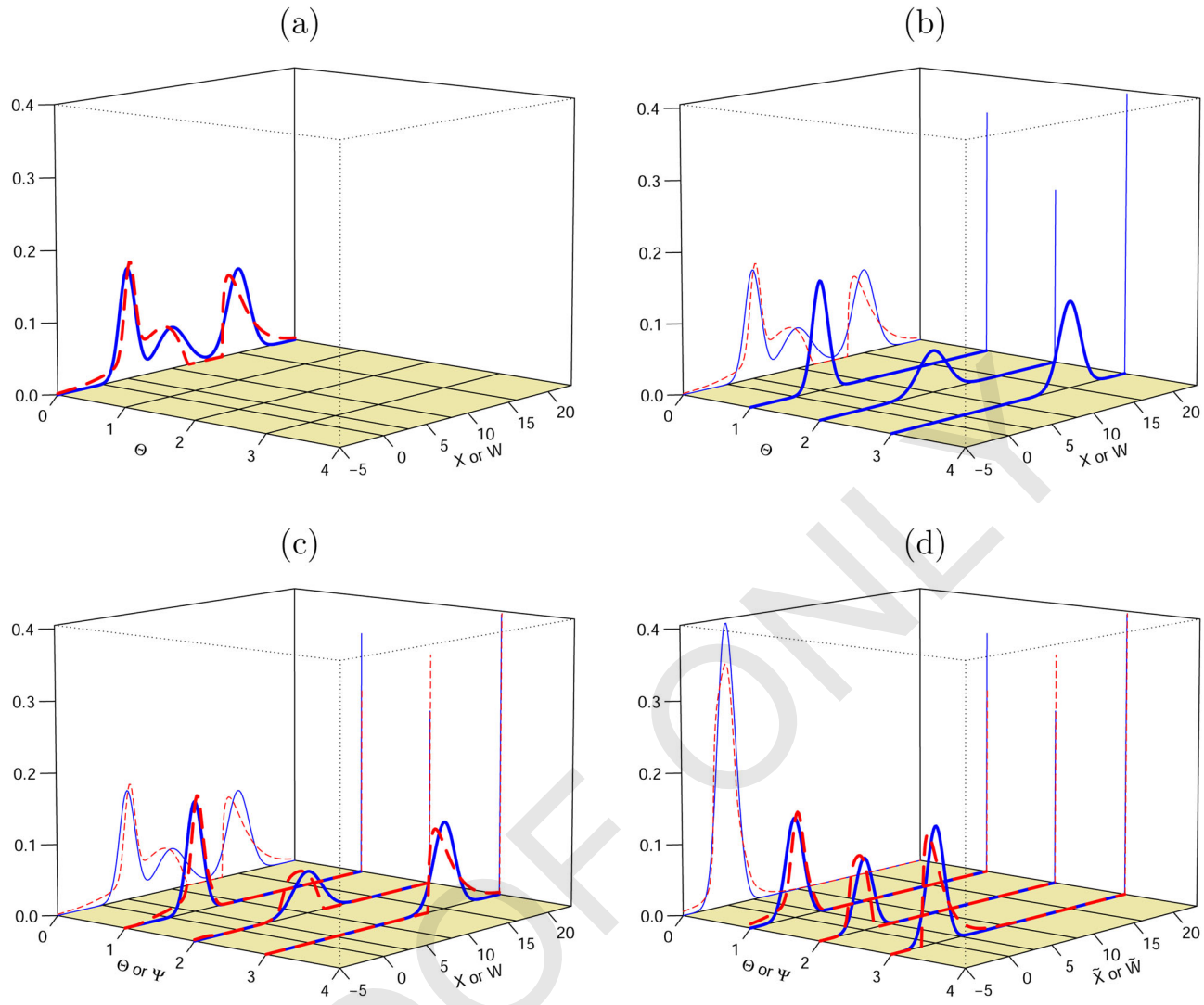

Figure 2. Illustration of Warp-U transformation. (a) ϕ_{mix} (solid line) and p (dashed line); (b) the joint and marginal distributions of X and Θ (solid line); (c) the joint and marginal distributions of W and Ψ (dashed line); (d) the joint and marginal distributions of Θ and $\widetilde{X}$ (solid line) and those of Ψ and $\widetilde{W}$ (dashed line), where $\widetilde{X}$ and $\widetilde{W}$ are obtained via Warp-U transformation.

constructed by again using the map in (8) but with (W, Ψ) in place of (X, Θ),

$$\widetilde{W} = \mathcal{F}_\Psi(W) = \mathcal{S}_\Psi^{-1}(W - \mu_\Psi) \sim \widetilde{p}. \tag{11}$$

Intuitively, because the transformation $\mathcal{F}$ maps the multimodal ϕmix back to the original unimodal (generating) density ϕ, when it is applied to the multimodal p, it can achieve similar a "unimodalizing" effect because ϕmix was chosen to approximate p.

In practice, to apply a Warp-U transformation to w_j, we calculate $\varpi(\cdot|w_j)$ according to (9), draw ψ_j from $\varpi(\cdot|w_j)$, and finally apply the deterministic transformation $\mathcal{F}_{\psi_j}$ to w_j. Graphically, each $p^{(k)}$ in Figure 2(c) is recentered and rescaled, like its counterpart, $\phi^{(k)}$. The dashed lines in Figure 2(d) are the joint distribution of Ψ and the Warp-U transformed variable, $\widetilde{W}$. In the faded left vertical panel of Figure 2(d), we see that the distribution of $\widetilde{W}$ overlaps substantially with ϕ.

When $K = 1$, the Warp-U transformation is the same as the Warp-II transformation provided that we choose ϕ_{mix} to be a location-scale family. For $K > 1$, Theorem 1 in Section 3.2 ensures that there will be additional overlap between $\widetilde{p}$ and ϕ compared to the overlap between p and ϕ_{mix}, except for in trivial cases, for example, when $p = \phi_{\text{mix}}$.

3.2. Theoretical Guarantee for General Warp-U Transformations

Figure 3 summarizes the key variables and distributions underlying a general Warp-U transformation, which does not assume that ϕ is the normal density. We do still require that ϕ shares the same support Ω as our target p. Another generalization included in Figure 3 is that the "index variable" Θ (and hence also Ψ) is permitted to take on any distribution π with support Π and dominating measure $\mathbf{v}$, and in particular Θ (and Ψ) is no longer required to be discrete.

For all $\theta \in \Pi$, the map $\mathcal{F}_\theta$ in Figure 3 is required to be one-to-one, onto, and almost surely differentiable, and to satisfy $\Omega = \mathcal{F}_\theta(\Omega)$. We denote its inverse map by $\mathcal{H}_\theta$. Since we specify $X \sim \mathcal{H}_\theta(Z)$, where $Z \sim \phi$, the conditional distribution $X\big|\Theta = \theta$ is

$$\phi_{X|\Theta}(x|\theta) = \phi(\mathcal{F}_\theta(x))\left|\mathcal{F}_\theta'(x)\right|, \quad x \in \Omega \tag{12}$$

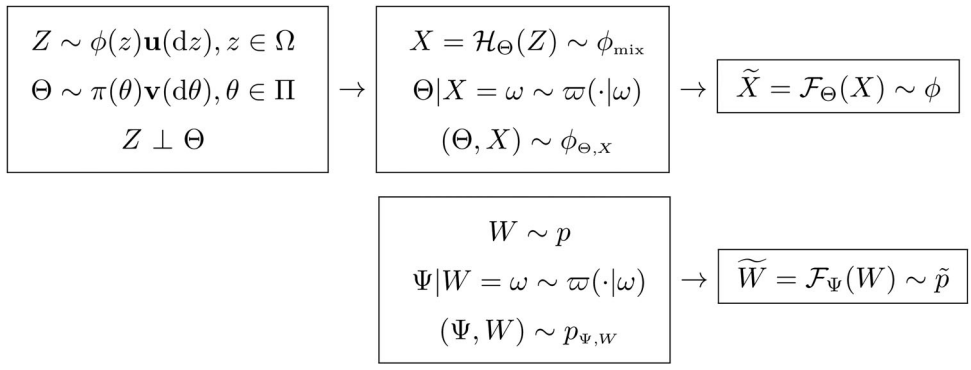

Figure 3. Relationships among the random variables and their distributions for Warp-U transformation. Here for almost surely (with respect to $\mathbf{v}$) all values of $\theta \in \Pi$, $\mathcal{F}_\theta$ and its inverse $\mathcal{H}_\theta$ are one-to-one, onto, and almost surely (with respect to $\mathbf{u}$) differentiable maps from $\Omega \to \Omega$.

and the (marginal) density of X is

$$\begin{aligned}\phi_{\mathrm{mix}}(x) &= \int_\Pi \phi_{X|\Theta}(x|\theta)\pi(\theta)\mathbf{u}(\mathrm{d}\theta) \\ &= \int_\Pi \phi(\mathcal{F}_\theta(x))\left|\mathcal{F}'_\theta(x)\right|\pi(\theta)\mathbf{u}(\mathrm{d}\theta). \qquad (13)\end{aligned}$$

Let $\varpi(\cdot|x)$ be the conditional distribution $\Theta|X = x$,

$$\varpi(\theta|x) = \frac{\phi_{X|\Theta}(x|\theta)\pi(\theta)}{\phi_{\mathrm{mix}}(x)}, \quad \theta \in \Pi, \qquad (14)$$

and, as before, let the variable Ψ be defined through $P(\Psi = \theta|W = \omega) = \varpi(\theta|\omega)$. The joint distributions of (Ψ, W) and (Θ, X) therefore share the same conditional specification:

$$\begin{aligned}&p_{\Psi,W}(\theta,\omega) = \varpi(\theta|\omega)p(\omega) \quad \text{and} \\ &\phi_{\Theta,X}(\theta,\omega) = \varpi(\theta|\omega)\phi_{\mathrm{mix}}(\omega), \quad (\omega,\theta) \in \Omega \times \Pi. \qquad (15)\end{aligned}$$

Considering this shared structure, here and in what follows we sometimes use the dummy variables (ω, θ) to refer to realizations of both (W, Ψ) and (X, Θ), and prevent confusion through our notation for the density functions in question.

The key consequence of the coupling (15) is that the overlap between ϕ and the density of the Warp-U transformed W: $\widetilde{W} = \mathcal{F}_\Psi(W) \sim \tilde{p}$, is greater than that between ϕ_{mix} and p. To prove this mathematically, we need a measure or multiple measures of overlap. The notion of f-divergence, or more precisely its complement (since small divergence corresponds to large overlap), serves well for our purposes. For any (nontrivial) *convex* function f on $[0, \infty)$ such that $f(1) = 0$, the corresponding f-divergence between two probability densities p_1 and p_2, when p_1 is absolutely continuous with respect to p_2, is defined as

$$\mathcal{D}_f(p_1||p_2) = \int_\Omega p_2(\omega) f\left(\frac{p_1(\omega)}{p_2(\omega)}\right)\mathbf{u}(\mathrm{d}\omega). \qquad (16)$$

Theorem 1 states that Warp-U transformations can never increase any f-divergence, and typically reduces them unless the transformation or f is trivially chosen; the proof is given in Appendix A in the supplementary materials.

Theorem 1. Let the Warp-U transformation $\mathcal{F}_\Psi$ be defined as in Figure 3, with the conditions given in the caption. The following results then hold.

(I) For any f-divergence $\mathcal{D}_f$, we have $\mathcal{D}_f(\tilde{p}||\phi) \leqslant \mathcal{D}_f(p||\phi_{\mathrm{mix}})$.

(II) If f is strictly convex, then the equality in (I) holds if and only if $\ell(\theta;\tilde{\omega}) \equiv \frac{p(\mathcal{H}_\theta(\tilde{\omega}))}{\phi_{\mathrm{mix}}(\mathcal{H}_\theta(\tilde{\omega}))}$ is free of θ (almost surely with respect to $\mathbf{v} \times \mathbf{u}$).

The Hellinger distance, the weighted harmonic divergence in (6), and the L_1 distance are all f-divergences, with $f_{He}(t) = 0.5(1-\sqrt{t})^2, f_{Ha}(t) = w_1(1-t)/(w_1+w_2t)$, and $f_{L_1}(t) = |1-t|$, respectively. The weighted harmonic divergence in (6) is an especially important case because it determines the asymptotic variance of bridge sampling estimators; see (5). Consequently, Theorem 1 says that the bridge sampling estimator based on $\tilde{p}$ and ϕ has smaller asymptotic variance than that based on p and ϕ_{mix}, thus supporting the use of Warp-U transformations (Section 3.3 gives the explicit form of these two estimators). Interestingly, inequality (I) does not necessarily hold for L_p distance when $\mathrm{p} \neq 1$ (and hence L_p distance is not an f-divergence when $\mathrm{p} \neq 1$). As a simple counter-example, let $K = 1$ in (7) and therefore $\phi_{\mathrm{mix}}(\omega) = |\mathcal{S}|^{-1}\phi\left(\mathcal{S}^{-1}(\omega-\mu)\right)$. Then $\tilde{p}(\omega) = |\mathcal{S}|p(\mathcal{S}\omega+\mu)$, and

$$\begin{aligned}L_\mathrm{p}(\tilde{p},\phi) &= \left(\int \left||\mathcal{S}|p(\mathcal{S}\tilde{\omega}+\mu) - \phi(\tilde{\omega})\right|^\mathrm{p}\mathbf{u}(\mathrm{d}\tilde{\omega})\right)^{\mathrm{p}^{-1}} \\ &= |\mathcal{S}|^{1-\mathrm{p}^{-1}}L_\mathrm{p}(p,\phi_{\mathrm{mix}}),\end{aligned}$$

so $L_\mathrm{p}(\tilde{p},\phi) > L_\mathrm{p}(p,\phi_{\mathrm{mix}})$ whenever $|\mathcal{S}|^{1-\mathrm{p}^{-1}} > 1$ (and $L_\mathrm{p}(p,\phi_{\mathrm{mix}}) > 0$).

Part (II) of Theorem 1 means that a Warp-U transformation will always result in real gain, as measured by any strictly convex f-divergence, unless one of two situations occur: (A) ϕ_{mix} is a perfect fit to p, in which case obviously $\ell(\theta,\tilde{\omega}) = 1$; or (B) $p \neq \phi_{\mathrm{mix}}$, but the Warp-U transformation $\mathcal{F}_\Theta$ is unfortunately (or unwisely) chosen such that it renders the "likelihood ratio" $\ell(\theta;\tilde{\omega})$ flat as a function of θ. Situation (B) includes the trivial cases where $\mathcal{F}_\theta$ does not depend on θ, or θ does not vary because π is a singleton, as well as some more complex scenarios.

An illustration of Theorem 1 is given in Figure 4 and Table 1 for the case of a tri-modal target distribution p and an approximating density ϕ_{mix} with $K = 2$. The green region in Figure 4(f) shows that the final overlap between $\tilde{p}$ and ϕ after Warp-U transformation is greater than the sum of the overlaps between $p^{(k)}$ and $\phi^{(k)}$, for $k = 1, 2$. We call the additional overlap *cross-overlap*, and it is this phenomenon which is key to Warp-U transformations. Table 1 lists several f-divergences for the pairs

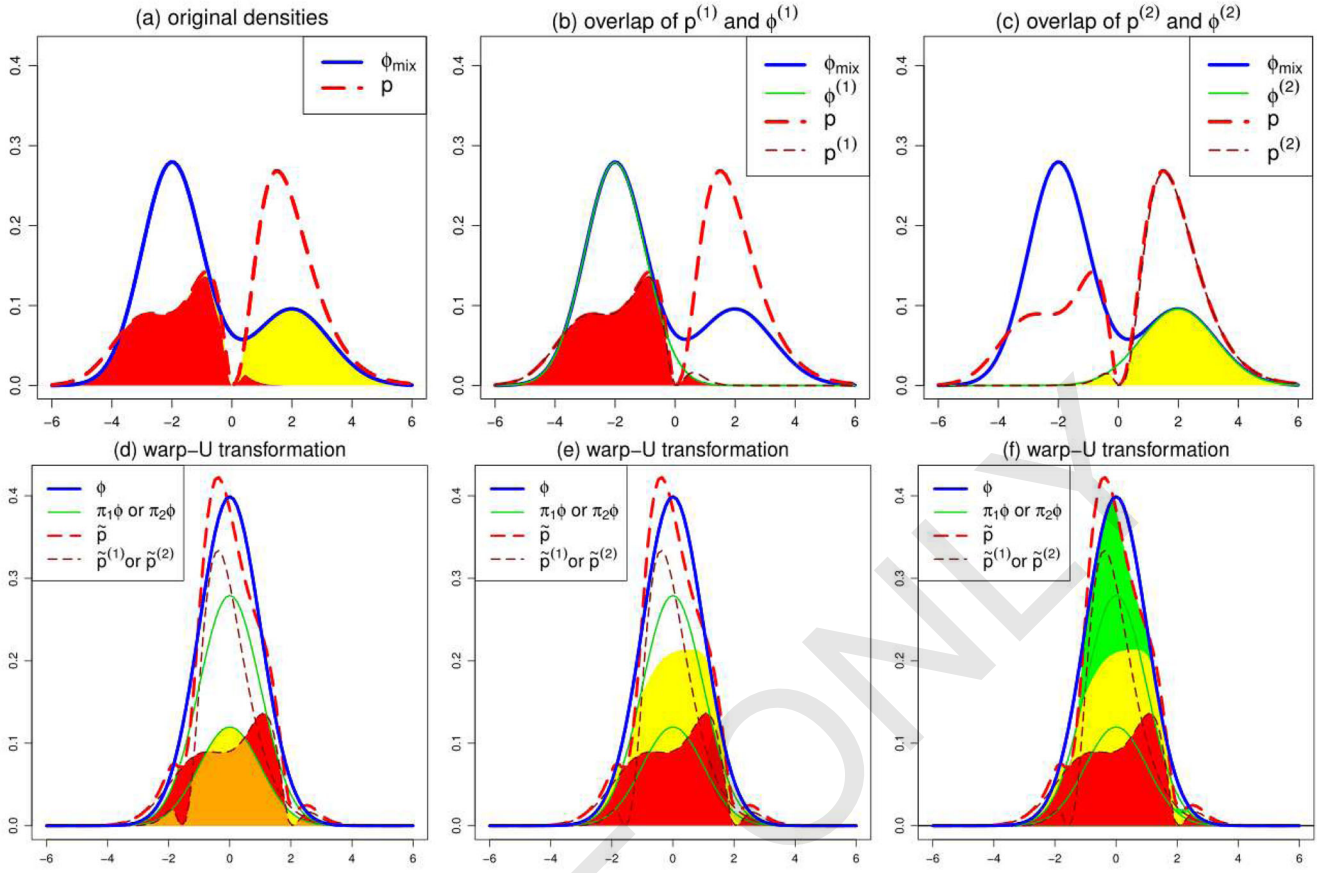

Figure 4. Illustration of the increase in the area of the overlapping region after Warp-U transformation. (a) p (dashed line) and ϕ_{mix} (solid line); (b) the 1st component of p, denoted as $p^{(1)}$ (thin dashed line), the 1st component of ϕ_{mix}, denoted as $\phi^{(1)}$ (thin solid line), and their overlap (shaded in red); (c) $p^{(2)}$, $\phi^{(2)}$, and their overlap (shaded in yellow); (d) the corresponding curves and shaded areas after Warp-U transformation; (e) the yellow region is added on top of the red region; (f) the green area shows the additional cross-overlap between the 1st and 2nd components induced by the Warp-U transformation.

Table 1. The overlapping area, and the distances between p and ϕ_{mix} and between $\tilde{p}$ and ϕ.

Densities	Overlap area	L_1 distance	Hellinger distance	Harmonic divergence
(p, ϕ_{mix})	0.66	0.68	0.28	0.145
$(\tilde{p}, \phi)$	0.92	0.16	0.08	0.013

(p, ϕ_{mix}) and $(\tilde{p}, \phi)$, and it confirms that the f-divergences are much lower in the latter case. Indeed, due to cross-overlap, the overlapping area for the pair of densities $(\tilde{p}, \phi)$ is nearly 40% larger than that for (p, ϕ_{mix}).

3.3. Warp-U Bridge Sampling

After the parameters $\boldsymbol{\zeta}$ for ϕ_{mix} have been chosen, the Warp-U transformation is determined. The unnormalized density of the transformed draws $\{\tilde{w}_1, \ldots, \tilde{w}_n\}$ can then be expressed as

$$\begin{aligned}\tilde{q}(\tilde{w}; \boldsymbol{\zeta}) &= \sum_{k=1}^{K} cp^{(k)}(\omega = \mathcal{S}_k \tilde{w} + \mu_k) \\ &= \phi(\tilde{w}) \sum_{k=1}^{K} \frac{q(\mathcal{S}_k \tilde{w} + \mu_k)}{\phi_{\text{mix}}(\mathcal{S}_k \tilde{w} + \mu_k)} \pi_k. \end{aligned} \tag{17}$$

Clearly, the normalizing constants of $\tilde{q}$ and q are both c, and hence we can estimate c with the bridge sampling estimator based on $\{\tilde{w}_1, \ldots, \tilde{w}_n\} \sim \tilde{p}$ and $\{z_1, \ldots, z_m\} \sim \phi$, that is,

$$\hat{c}_\alpha^{(\text{U})} \equiv \hat{r}_\alpha^{(\text{U})} = \frac{m^{-1} \sum_{j=1}^{m} \tilde{q}(z_j; \boldsymbol{\zeta}) \alpha(z_j; \tilde{p}, \phi)}{n^{-1} \sum_{j=1}^{n} \phi(\tilde{w}_j) \alpha(\tilde{w}_j; \tilde{p}, \phi)}. \tag{18}$$

As mentioned in Section 2.1, the optimal choice of $\alpha(\cdot; \tilde{p}, \phi)$ is proportional to $(s_1 \tilde{p} + s_2 \phi)^{-1}$. Since ϕ_{mix} also has some overlap with p, the normalizing constant can alternatively be estimated with the bridge sampling estimator based on $\{w_1, \ldots, w_n\} \sim p$ and $\{x_1, \ldots, x_m\} \sim \phi_{\text{mix}}$, that is,

$$\hat{c}_\alpha^{(\text{mix})} \equiv \hat{r}_\alpha^{(\text{mix})} = \frac{m^{-1} \sum_{j=1}^{m} q(x_j) \alpha(x_j; p, \phi_{\text{mix}})}{n^{-1} \sum_{j=1}^{n} \phi_{\text{mix}}(w_j; \boldsymbol{\zeta}) \alpha(w_j; p, \phi_{\text{mix}})}. \tag{19}$$

Theorem 1 implies $\mathcal{D}(\tilde{p}, \phi) \leqslant \mathcal{D}(p, \phi_{\text{mix}})$ when $\mathcal{D}$ is the weighted harmonic divergence in (6), so the asymptotic variance of $\hat{\lambda}_\alpha^{(\text{U})} = \log\left(\hat{c}_\alpha^{(\text{U})}\right)$ is smaller than that of $\hat{\lambda}_\alpha^{(\text{mix})} = \log\left(\hat{c}_\alpha^{(\text{mix})}\right)$ under the optimal choice of α, when the draws are independent. Even when we choose some other α (e.g., the geometric mean $\sqrt{p_1 p_2}$) or the draws are not independent, we can still expect that the increased overlap obtained by the Warp-U transformation helps (18) to outperform (19), at least when both use the same n and m.

In challenging situations, our choice of ϕ_{mix} may be a poor match with p. This will clearly impact the quality of the corresponding Warp-U transformation, but $\mathcal{D}(\tilde{p}, \phi) \leqslant \mathcal{D}(p, \phi_{\text{mix}})$

will nevertheless still hold. Perhaps the most important case where the Warp-U transformation is not beneficial is when p has many highly isolated modes. The cross-overlap discussed in Section 3.2 will then be small, and so will the reduction achieved by $\mathcal{D}(\tilde{p}, \phi)$. Consequently, in this scenario, the extra computation required by the Warp-U transformation is not worthwhile.

4. Warp-U Computation Details and Numerical Examples

The key step in applying Warp-U bridge sampling is to identify a mixture density ϕ_{mix} that adequately overlaps with p, under reasonable constraints on computation. In relatively low dimensional ($\leqslant$10) problems, we can obtain ϕ_{mix} based on the expression for q, for example, using iterated Laplace approximations (see Bornkamp 2011; Gelman et al. 2013). However, these methods are too costly and unstable in high dimensions. Below we outline a simple method which uses the draws $\{w_1, \ldots, w_n\}$, can capture a good proportion of the mass of p, and has computational cost that is linear in dimensionality. We then adopt another practical strategy to remove an over-fitting bias due to this simple method.

4.1. Fitting ϕ_{mix}: Diagonal Covariance Matrices

Suppose that p is D dimensional and that our draws from p reasonably represent the regions of nonnegligible density. We seek a normal mixture ϕmix in the form of (7) to approximate p, where $\mathcal{S}_k$ is a positive definite *diagonal* matrix, $\mathcal{S}_k = \text{Diag}\{\sigma_{k,1}, \sigma_{k,2}, \ldots, \sigma_{k,D}\}$, for $k = 1, \ldots, K$, and hence $\boldsymbol{\zeta} = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \mathcal{S}_1, \ldots, \mathcal{S}_K)$. Unlike usual statistical inference problems where ignoring correlations can have very serious consequences, for Warp-U transformations using diagonal covariance matrices is often an acceptable compromise between computational efficiency and MC efficiency. Indeed, as discussed in Section 3, it is not necessary for ϕ_{mix} to be a great fit to p in order for us to benefit significantly from Warp-U transformations. In the next section, we provide further empirical evidence to illustrate this point.

Since a mixture of normal components without suitable restrictions has unbounded likelihood (Kiefer and Wolfowitz 1956; Day 1969), we estimate $\boldsymbol{\zeta}$ by the penalized MLE proposed by Chen, Tan, and Zhang (2008). In particular, we make use of the EM procedure proposed by Chen and Tan (2009), but with a "robustified" penalty function

$$\mathbf{p_n}(\boldsymbol{\zeta}) = -\frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{d=1}^{D} \left\{ \frac{\widehat{\text{IQ}}_d^2}{\sigma_{k,d}^2} - \log(\sigma_{k,d}^2) \right\},$$

where $\widehat{\text{IQ}}_d$ is the inter-quantile range of the draws from p in the dth dimension. Because EM tends to become trapped at local modes, we apply it M times, randomly generating a new initial point $\boldsymbol{\zeta}^{(0)}$ for each repetition as follows. The initial values for the π_k's and $\mathcal{S}_k$'s are $\pi_k^{(k)} = K^{-1}$ and $\sigma_{k,d}^2 = 1.5\widehat{\text{IQ}}_d^2$ for all k and all M replications. For the mean parameters μ_k, for the first $M/2$ replications, we randomly sample K of available draws from p (without replacement) to be the initial values. For the second $M/2$ replications, along the dimension with the largest estimated variance, we first identify a region where 95% of the draws from p reside and divide it into K subregions so that each subregion contains approximately the same number of draws. We then sample one draw from each of the K subregions to set the initial mean parameters. Our EM stopping criterion is $|1 - (l_n^{(t)}/l_n^{(t-1)})| < 10^{-6}$, where $l_n^{(t)}$ is the value of the (un-penalized) log-likelihood at iteration t. In our simulations, the EM usually stopped within 100 iterations. After obtaining M estimates of $\boldsymbol{\zeta}$, we choose the one with the largest likelihood value to be the parameter, $\widetilde{\boldsymbol{\zeta}}$, for Warp-U bridge sampling. Simulations show that M as small as 2 to 10 is sufficient to obtain a local maxima that serves well for the purpose of ensuring adequate overlap between p and ϕ_{mix}.

4.2. Overcoming Adaptive Bias and Setting Tuning Parameters

Let $\widetilde{\boldsymbol{\zeta}}_{\mathcal{D}}$ be the estimate of $\boldsymbol{\zeta}$ obtained by applying EM to all the draws from p, $\mathcal{D} = \{w_1, \ldots, w_n\}$, and let $\hat{\lambda}_{\mathcal{D}}^{(\text{U})} = \log\left(\hat{c}_{\mathcal{D}}^{(\text{U})}\right)$ be the corresponding Warp-U bridge sampling estimator. Because $\widetilde{\boldsymbol{\zeta}}_{\mathcal{D}}$ is a function of the draws from p, the distribution of the corresponding Warp-U transformed draws, $\{\tilde{w}_1, \ldots, \tilde{w}_n\}$, is no longer proportional to $\tilde{q}(\cdot; \boldsymbol{\zeta})$ in (17) when we substitute $\boldsymbol{\zeta} = \widetilde{\boldsymbol{\zeta}}_{\mathcal{D}}$. In other words, $\hat{\lambda}_{\mathcal{D}}^{(\text{U})}$ has an adaptive bias induced by the dependence of $\widetilde{\boldsymbol{\zeta}}_{\mathcal{D}}$ on $\mathcal{D}$, demonstrated in Figure 7 (see Section 4.3).

Since the additional bias of $\hat{\lambda}_{\mathcal{D}}^{(\text{U})}$ is due to the dependence of $\widetilde{\boldsymbol{\zeta}}_{\mathcal{D}}$ on the draws from p, an obvious remedy is to use two disjoint subsets of the draws from p for estimating $\boldsymbol{\zeta}$ and for bridge sampling. We can then switch the roles of these subsets to gain more statistical efficiency. Figure 5 depicts the subsampling strategy we use to obtain two separate bridge sampling estimators, $\hat{\lambda}_{\text{H}_i}^{(\text{U})}, i = 1, 2$. Each $\hat{\lambda}_{\text{H}_i}^{(\text{U})}$ is obtained by using $L \leqslant n/2$ of the draws from p to estimate $\boldsymbol{\zeta}$ and the other 50% of the draws for the Warp-U bridge sampling specified by the estimated $\boldsymbol{\zeta}$. Our final estimator $\hat{\lambda}_{\text{H}}^{(\text{U})}$ is the average of $\hat{\lambda}_{\text{H}_1}^{(\text{U})}$ and $\hat{\lambda}_{\text{H}_2}^{(\text{U})}$. Clearly, both $\hat{\lambda}_{\text{H}_1}^{(\text{U})}$ and $\hat{\lambda}_{\text{H}_2}^{(\text{U})}$ are individually valid. The only concern is that the two estimators may be highly correlated causing their average to have high variance. However, under the setting of

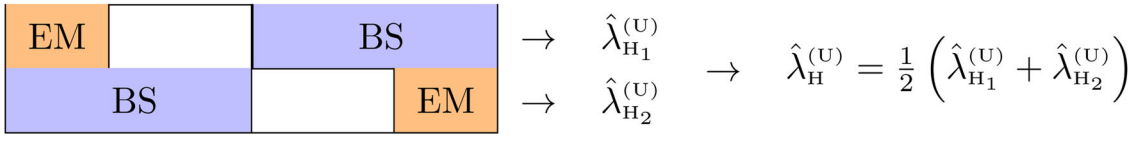

Figure 5. A strategy for removing the adaptive bias without (unduly) increasing the variance of the Warp-U bridge sampling estimator. Each $\hat{\lambda}_{\text{H}_i}^{(\text{U})}, i = 1, 2$ uses up to 50% of the draws from p for estimating $\boldsymbol{\zeta}$ and the other 50% for Warp-U bridge sampling. We then average the two estimators.

iid draws from p and ϕ, our empirical investigations detailed in Appendix B in the supplementary materials suggest that the correlation between $\hat{\lambda}^{(U)}_{H_1}$ and $\hat{\lambda}^{(U)}_{H_2}$ is often very small, for example, <0.06. Thus, when the iid assumption approximately holds, the variance of $\hat{\lambda}^{(U)}_{H}$ is nearly half that of $\hat{\lambda}^{(U)}_{H_i}$, for $i = 1, 2$. Appendix B in the supplementary materials gives further details and suggests a strategy for approximating the variance of the final estimator $\hat{\lambda}^{(U)}_{H}$.

As depicted in Figure 5, we may choose $L < n/2$ to reduce the EM computation, which is a reasonable strategy given that ϕ_{mix} does not need to be a very precise approximation to p. The number of components K can be chosen using standard model selection criteria such as the Bayesian information criterion (BIC) applied to the L data points used for EM. In particular, such model selection criteria protect against over-fitting by use of a penalty term and can be expected to provide a mixture that is a good approximation to p, at relatively low computational cost. In terms of statistical efficiency, a good approximation to p is what is required because the asymptotic variance of both standard bridge sampling and Warp-U bridge sampling decreases as the approximation ϕ_{mix} improves, see (5) and Theorem 1. On the other hand, given that greater K implies greater computational cost per iteration of Warp-U bridge sampling (and standard bridge sampling to a lesser extent), we may anticipate that once computation is accounted for the best choices of K will be those around where the curve of BIC against K begins to level out, indicating diminishing returns, as opposed to the value of K at exactly the minimum BIC value. Figure 6 shows that this is indeed the case for the 10-dimensional example to be discussed in Section 4.3: the BIC curve (left panel) starts to flatten out around $K = 20$–40, and the precision per CPU second (right panel) shows that this is also the optimal range of K for Warp-U bridge sampling in terms of computational efficiency.

Appendix C in the supplementary materials provides further practical guidance for setting K, L, and m (the number of draws from ϕ), which we now summarize. First, we suggest setting $K \leq n/100$ because our simulations suggest that ϕ_{mix} tends to overfit for $K > n/100$ which can even cause the RMSE of $\hat{\lambda}^{(U)}_{H}$ to increase. Another reason to avoid large K is that computational cost increases quadratically with K. Next, we found that a reasonable choice of L is $\min(50K, n/2)$ because the reductions in the RMSE of $\hat{\lambda}^{(U)}_{H}$ are relatively small when we increase L past $50K$. Lastly, in terms of precision per CPU second (*PpS*, defined as the reciprocal of RMSE×CPU seconds), when K is already large, increasing m is a more efficient strategy for reducing the variance of $\hat{\lambda}^{(U)}_{H}$ than increasing K further. However, when K is small it is often more efficient to increase K rather than m, because if there are fewer components in ϕ_{mix} than there are major modes of p, or if the modes of p are asymmetric or heavy tailed, then large reductions in RMSE can usually be obtained by increasing K.

Setting $L = n/2$, assuming diagonal covariance matrices (see Section 4.1), and treating K as fixed, the computational complexity of our complete method is $O((n+m)Kg(d))$, where $g(d) \geq O(d)$ is the cost of a single evaluation of q. Here, we have assumed that the number of EM iterations and initializations are fixed. If K needs to be chosen then the EM part of our algorithm, which costs $O(nKd)$, has to be repeated for each plausible value of K. However, in practice, unless the only suitable values of K are very large and completely unknown, then it is the number of evaluations of q (cost $g(d)$) that is most crucial, and this is therefore the focus of our computational efficiency comparisons in Section 5.

4.3. Examples in 10 and 50 Dimensions

To illustrate the effectiveness of using diagonal covariance matrices and the above bias reduction strategy, we first consider a 10 dimensional example where p is set to be a mixture of 25 multivariate skew-t distributions, whose density is given in the R package "sn" by Azzalini (2011); also see Azzalini (2013). We specify the degrees of freedom of the 25 skew-t distributions to take various values between 1 and 4, the skew parameters to take values between -100 and 200, and the scale matrices to be non-sparse. To evaluate our methods properly, we simulate

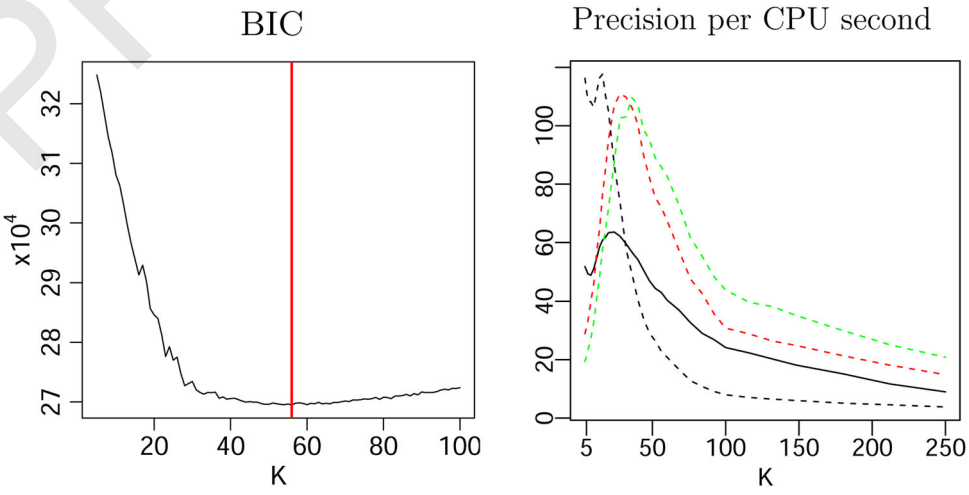

Figure 6. BIC as a function of K (left) and the precision per CPU second as a function of K (right) of the optimal bridge sampling estimators $\hat{\lambda}^{(U)}_{H}$ (solid lines, $m = n$) and $\hat{\lambda}^{(\text{mix})}_{H}$ (dashed lines) with $m = n$ (black), $16n$ (red), and $32n$ (green). Recall that "mix" refers to the ordinary bridge sampling estimator using p and ϕ_{mix}, that is, the logarithm of (19).

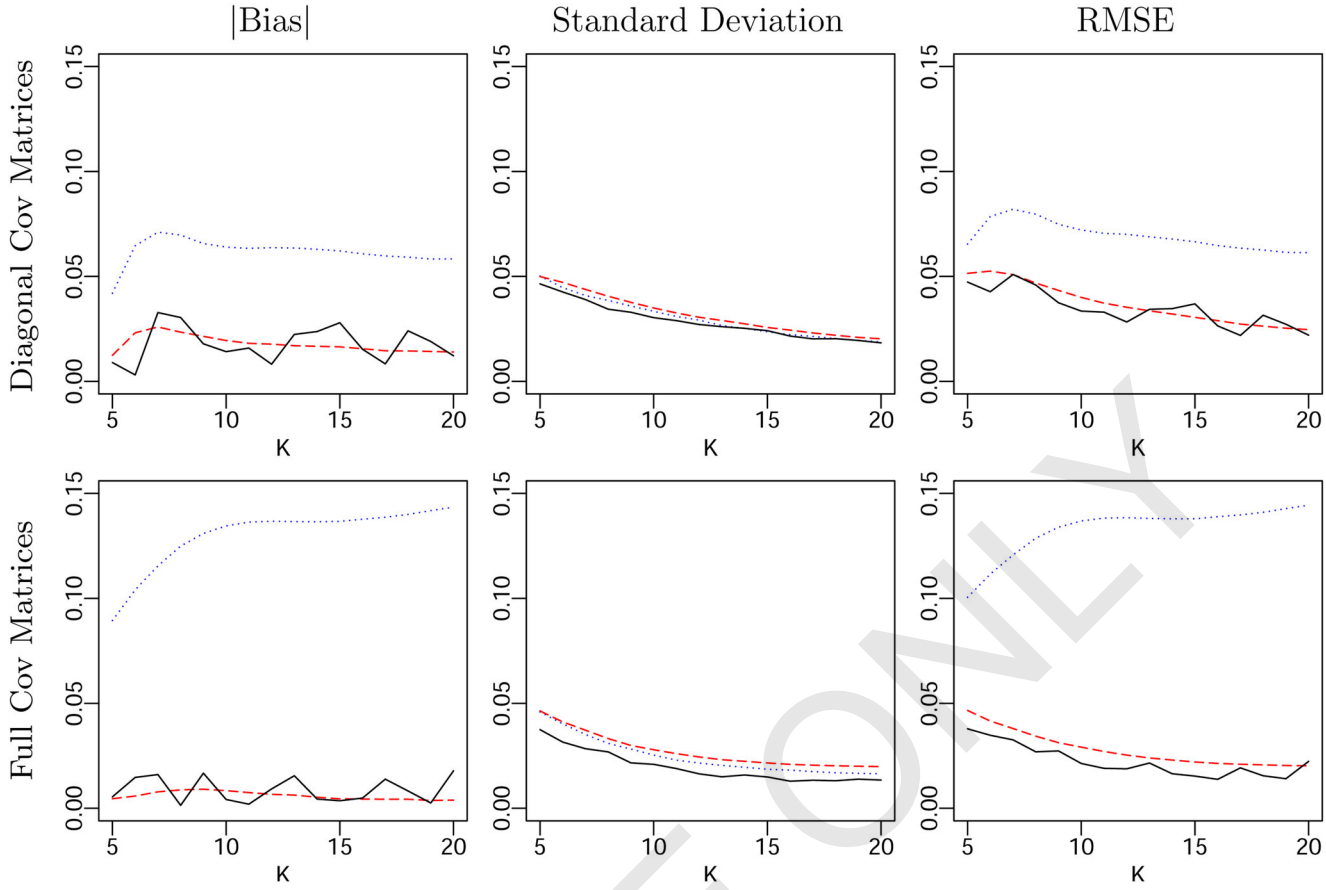

Figure 7. The columns show the |bias|, standard deviation, and RMSE of (i) $\hat{\lambda}^{(U)}_{\mathcal{D},\mathcal{Z}} = \log(\hat{c}^{(U)}_{\mathcal{D},\mathcal{Z}})$ (dotted lines), the Warp-U estimator specified by $\widetilde{\boldsymbol{\zeta}}_{\mathcal{D}}$, which is estimated from $\mathcal{D} = \{w_1, \ldots, w_n\}$, (ii) $\hat{\lambda}^{(U)}_{\mathcal{I},\mathcal{Z}} = \log(\hat{c}^{(U)}_{\mathcal{I},\mathcal{Z}})$ (solid lines), the Warp-U specified by $\widetilde{\boldsymbol{\zeta}}_{\mathcal{I}}$, which is independent of $\mathcal{D}$, and (iii) (dashed lines) the average of two Warp-U bridge sampling estimators with half of the draws from p for estimating $\boldsymbol{\zeta}$ and the other half for bridge sampling. The subscript "$\mathcal{Z}$" indicates "Diag" (top row) or "Full" (bottom row) covariance matrices in the Gaussian mixture model.

10^4 replicate datasets, each of which contains 2500 independent draws from p.

We consider three Warp-U bridge sampling estimators, with diagonal covariance matrices for ϕ_{mix}. They are $\hat{\lambda}^{(U)}_{\mathcal{D},\mathcal{D}\text{iag}}$, $\hat{\lambda}^{(U)}_{\text{H},\mathcal{D}\text{iag}}$, and $\hat{\lambda}^{(U)}_{\text{I},\mathcal{D}\text{iag}}$, where the first subscript specifies whether $\hat{\lambda}$ is computed by estimating $\boldsymbol{\zeta}$ using all the draws from p ($\mathcal{D}$), by setting $L = n/2$ and using the scheme in Section 4.2 (H), or by estimating $\boldsymbol{\zeta}$ from an independent set of draws ($\mathcal{I}$). For all three estimators, we use the optimal choice of α and set $m = 2500$, that is, the number of independent draws from the auxiliary $\phi = \mathcal{N}(0, I_{10})$. The estimator $\hat{\lambda}^{(U)}_{\mathcal{I},\text{Diag}}$ serves as a benchmark for comparison because it is free of adaptive bias.

The lines in the top row of Figure 7 show the bias (left panel), the standard deviation (center panel), and the RMSE (right panel) of the three estimators: $\hat{\lambda}^{(U)}_{\mathcal{D},\mathcal{D}\text{iag}}$ (dotted lines), $\hat{\lambda}^{(U)}_{\text{H},\mathcal{D}\text{iag}}$ (dashed lines), and $\hat{\lambda}^{(U)}_{\text{I},\mathcal{D}\text{iag}}$ (solid lines). Results are plotted for all values of K between 5 and 20 inclusive. Larger values of K generally represent a better approximation to p but more computation (and potentially less gain from using Warp-U bridge sampling as opposed to standard bridge sampling between p and ϕ_{mix}). The top left panel of Figure 7 shows the excessive bias of $\hat{\lambda}^{(U)}_{\mathcal{D},\mathcal{D}\text{iag}}$ compared with $\hat{\lambda}^{(U)}_{\mathcal{I},\mathcal{D}\text{iag}}$. In contrast, the bias of our bias adjusted estimator, $\hat{\lambda}^{(U)}_{\text{H},\mathcal{D}\text{iag}}$, is as low as that of the benchmark $\hat{\lambda}^{(U)}_{\mathcal{I},\mathcal{D}\text{iag}}$. In the top center panel of Figure 7, we see that the variances of all three estimators are very similar, and decrease as K increases. The decrease is because on average larger K corresponds to more overlap between p and the calibrated ϕ_{mix}, and thus more overlap between $\tilde{p}$ and ϕ. The top right panel of Figure 7 shows the RMSE of the estimators which is similar for $\hat{\lambda}^{(U)}_{\text{H},\mathcal{D}\text{iag}}$ and $\hat{\lambda}^{(U)}_{\mathcal{I},\mathcal{D}\text{iag}}$, but much larger for $\hat{\lambda}^{(U)}_{\mathcal{D},\mathcal{D}\text{iag}}$ because of its large bias.

The bottom row of Figure 7 shows similar results to those discussed above, but in the case where the covariance matrices of the components of ϕ_{mix} are not constrained to be diagonal. In this setting, we denote the three estimators by $\hat{\lambda}^{(U)}_{\mathcal{D},\text{Full}}$, $\hat{\lambda}^{(U)}_{\text{H,Full}}$, and $\hat{\lambda}^{(U)}_{\text{I,Full}}$. The results broadly match those in the top row of Figure 7, except that the bias (bottom left panel) and RMSE (bottom right panel) of $\hat{\lambda}^{(U)}_{\mathcal{D},\text{Full}}$ are even larger than those of $\hat{\lambda}^{(U)}_{\mathcal{D},\mathcal{D}\text{iag}}$. This is because with full covariance matrices, we have significantly more parameters to be estimated, and hence more substantial over-fitting bias. However, Figure 7 clearly shows that our method removes the adaptive bias regardless of its magnitude, and differences between using full and diagonal covariance matrices when fitting ϕ_{mix} are minor (compare the dashed lines in the top and bottom panels). Since fitting ϕ_{mix}

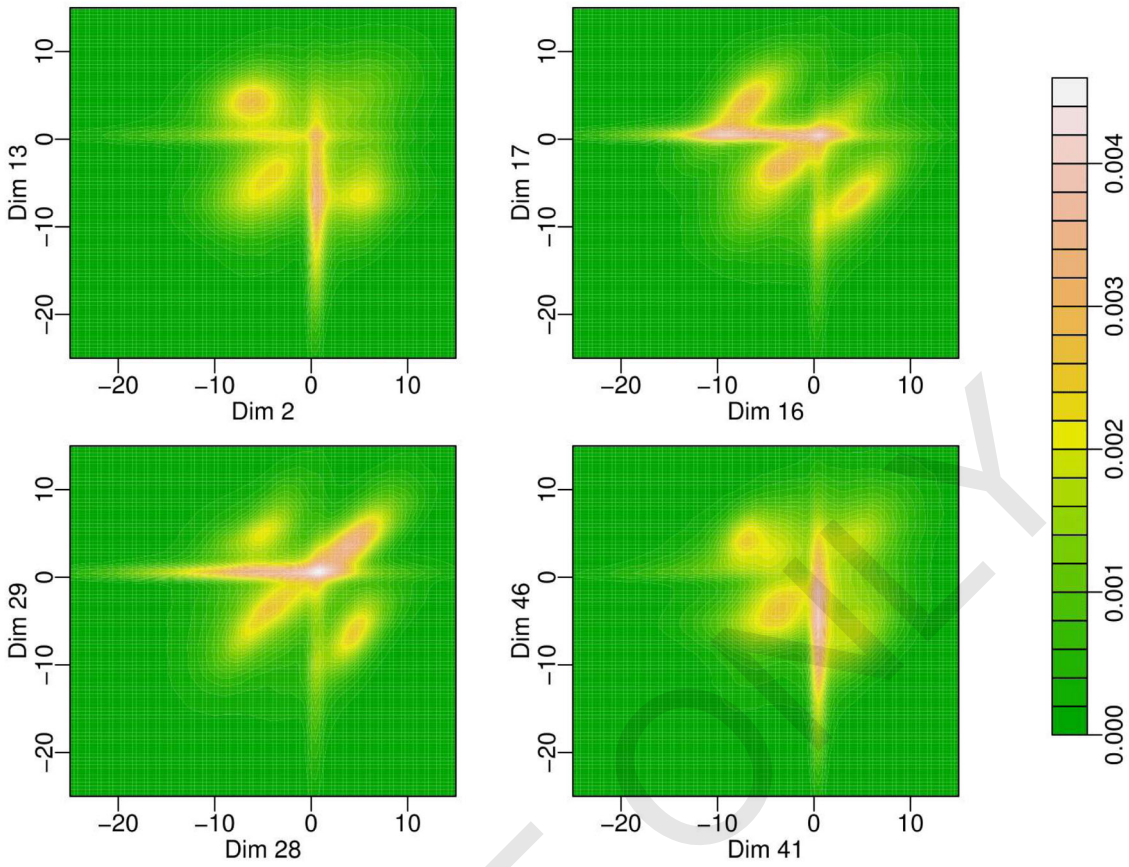

Figure 8. Contours of the density p projected onto different pairs of dimensions.

with diagonal covariance matrices is computationally much less expensive, $\hat{\lambda}^{(\mathrm{U})}_{\mathrm{H},\mathcal{D}\mathrm{iag}}$ achieves better RMSE per CPU second than $\hat{\lambda}^{(\mathrm{U})}_{\mathrm{H,Full}}$; see Appendix D in the supplementary materials for further demonstration. Hence, from hereon we always use diagonal covariance matrices and the estimation strategy in Section 4.2 and denote the final estimator by $\hat{\lambda}^{(\mathcal{X})}_{\alpha} = \frac{1}{2}\left(\hat{\lambda}^{(\mathcal{X})}_{\alpha,1} + \hat{\lambda}^{(\mathcal{X})}_{\alpha,2}\right)$, where $\mathcal{X} = U$ or "mix." Recall that "mix" refers to the ordinary bridge sampling estimator using p and ϕ_{mix}, that is, the logarithm of (19).

Next we consider a 50 dimensional example. For this example, p is a mixture of 30 distributions, including normal distributions, t-distributions (including Cauchy distributions), and multivariate distributions with gamma and/or exponential marginal distributions and normal copulas. The four two-dimensional projection contour plots of p in Figure 8 show the density has very long tails and is quite skewed in some directions. Evaluating p is about 700 times more costly than evaluating ϕ (the auxiliary density). The simulation results are based on 10^4 replications, and in each replication, $n = 10^4$ samples were drawn from p.

Figure 9 shows the total computational cost, the RMSE, and the *PpS* of $\hat{\lambda}^{(\mathcal{X})}_{\mathrm{opt}}$. As in the 10 dimensional example, the RMSE decreases as K increases up to $n/100$, and when $K > n/100$, the mixture model overfits the data (i.e., the draws from p), resulting in a slight increase in the RMSE of $\hat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$. On average, log(RMSE) of $\hat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ is about 60% of that of $\hat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$, but the computational cost of $\hat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ is 4.7 times that of $T^{(\mathrm{mix})}_{\mathrm{opt}}$, so in terms of the *PpS*, $\hat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$ is superior to $\hat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$. In addition, for large K, when we increase m from n (black lines) to $16n$ (red) and $32n$ (green), the total computational cost of $\hat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$ increases by only a small fraction, but the gain in statistical efficiency is substantial. Thus, in this illustration $\hat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$ is preferred to $\hat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$. However, such preferences vary with machines and implementations, because we have not optimized the function evaluation routines or other aspects of the code. Furthermore, the relatively high computational cost of $\hat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ seen in Figure 9 is partly due to obtaining the Warp-U transformed draws $\{\tilde{\omega}_1, \ldots, \tilde{\omega}_n\}$, which does not require any target evaluations. In other scenarios, evaluations of q may dominate the computational cost more, and then computational efficiency would depend mostly on the number of target evaluations. In particular, in such cases, increasing m from n to say $16n$ would represent an 8.5 (i.e., $(16 + 1)/2$) multiplicative increase in computation, as opposed to the relatively modest increase seen in the left panel of Figure 9, and therefore the computational cost of $\hat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$ would be more similar to that of $\hat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ for a given log(RMSE). We consider measuring computational efficiency by the number of target evaluations in Section 5, and find $\hat{\lambda}^{(\mathrm{mix})}_{\mathrm{opt}}$ and $\hat{\lambda}^{(\mathrm{U})}_{\mathrm{opt}}$ to be closely comparable. Opportunities for reducing the computational cost associated with Warp-U bridge sampling will be discussed in Section 6.

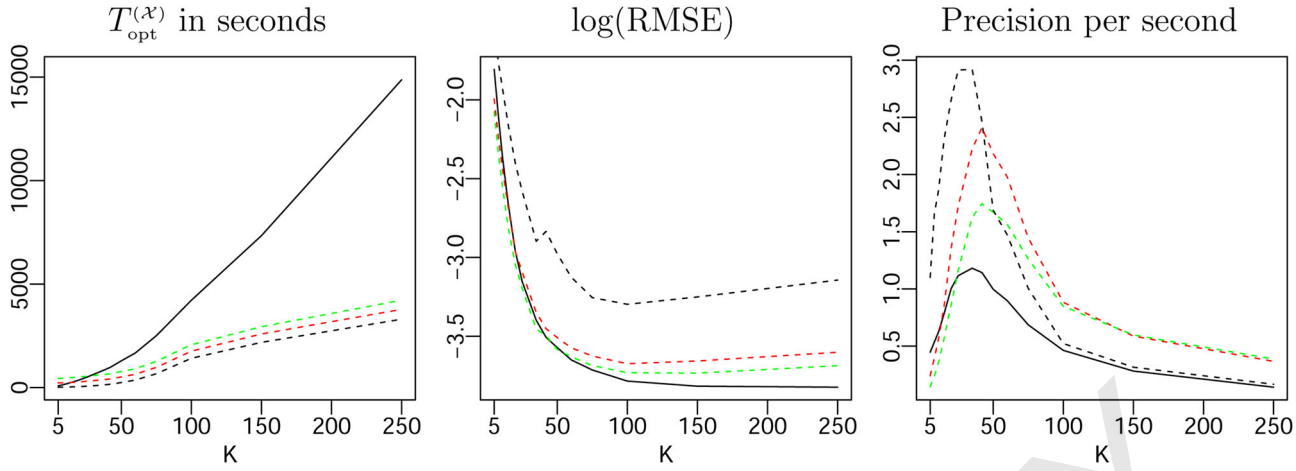

Figure 9. The total computational cost (left), the log(RMSE) (middle), and the *PpS* (right) of $\hat{\lambda}_{\text{opt}}^{(\text{U})}$ (solid lines, $m = n$) and $\hat{\lambda}_{\text{opt}}^{(\text{mix})}$ (dashed lines) with $m = n$ (black), $16n$ (red), and $32n$ (green).

5. Warp-U Extension of GWL Method

As we emphasized earlier, bridge sampling is applicable to general MCMC settings. We therefore do not need the draws to be independent, as long as they are from the target p, or at least in the long run. There are cases, however, where the available draws are from a distribution that is known to be different from the target p. Indeed, some of the most promising approaches for estimating normalizing constants combine the tasks of sampling and estimation into one coherent algorithm, but often do not directly sample from p. An important case of such a combined approach is the GWL algorithm proposed by Liang (2005). GWL is particularly useful in the current context because of its ability to efficiently sample from highly multimodal distributions. We therefore take it as a benchmark, and illustrate how it can be combined with Warp-U bridging sampling to obtain an improved estimator for normalizing constants. More generally, our strategy of incorporating a Warp-U bridge sampling step can be tried on other algorithms that combine sampling and estimation of normalizing constants.

5.1. GWL Algorithm

We begin by briefly describing GWL, which shares some similarities with other energy based methods such as the equi-energy sampler (Kou, Zhou, and Wong 2006). Suppose that we want to compute the integral $\int_S q(\omega)\mathbf{u}(\mathrm{d}\omega)$, denoted by a set function $g(S)$, for some unnormalized density q and bounded region S. Typically S is the region over which q has nonnegligible density, that is, $q(S) \approx 1$. We divide S into r subregions $S_1, \ldots, S_r$, which are defined by target energy bins; that is, within each S_i, the energy level, defined by $-\log q$, is roughly the same. Let the current estimate of the integral $\int_{S_i} q(\omega)\mathbf{u}(\mathrm{d}\omega)$ be denoted by $\hat{g}(S_i)$, and set the initial estimate to be $\hat{g}(S_i) = 1$, for $i = 1, \ldots, r$. GWL takes the inputs n_0, δ_0, and T (e.g., $n_0 = 1000$, $\delta_0 = e - 1 \approx 1.718$, $T = 25$) and proceeds as detailed below. (Here, we have ignored an additional tuning parameter in Liang (2005) that is not needed for computing normalizing constants.)

GWL algorithm.
For stage $t = 1, \ldots, T$:

1. Set $n_t = n_{t-1}(1.1)^{t-1}$, $\delta_t = \sqrt{1 + \delta_{t-1}} - 1$, and $\hat{g}^{(t,1)}(S_i) = \hat{g}^{(t-1,n_{t-1})}(S_i)$, for $i = 1, \ldots, r$.
2. For $k = 1, \ldots, n_t$ do the following:

   (i) Use a Metropolis–Hastings step, with proposal density h, to draw a sample ω from the current target density

   $$\psi^{(t,k)}(\omega) \propto \sum_{i=1}^{r} \frac{q(\omega)}{\hat{g}^{(t,k)}(S_i)} I(\omega \in S_i). \tag{20}$$

   The (t,k) superscripts indicate that the current target and the estimate of $g(S_i)$, for $i = 1, \ldots, r$, are updated in each iteration within each stage.

   (ii) Update $\hat{g}^{(t,k)}(S_{I_\omega})$ to $(1 + \delta_t)\hat{g}^{(t,k)}(S_{I_\omega})$, where I_ω is the index such that $\omega \in S_{I_\omega}$, that is, a regional mass estimate $\hat{g}^{(t,k)}(S_i)$ is increased only if S_i contains the draw ω.

It should be clear from the description above that, in the limit, GWL samples the subregions $S_1, \ldots, S_r$ with equal probability, and within each S_i, it samples according to q. Therefore, its stationary distribution is not the targeted q, but what can intuitively be described as a "redistributed" q that equalizes the masses of the energy bins:

$$\psi(\omega) = \sum_{i=1}^{r} \frac{q(\omega)}{g(S_i)} I(\omega \in S_i). \tag{21}$$

Liang (2005) verified this convergence assuming that the intermediate densities of (20) can be sampled from exactly. Practically, we can sample from them approximately, say by repeating the Metropolis–Hastings step many times between each update of the estimate of $g(S_i)$, for $i = 1, \ldots, r$. However, Liang (2005) used only one Metropolis–Hastings update in his illustrations, and we follow this practice (but with an ideal proposal, see Appendix E in the supplementary materials). His proof assumed n_t grows sufficiently fast with t, but the multiplicative factor 1.1 he suggested may not always be adequately large, though it is computationally problematic to increase it much further.

Since GWL samples the subregions uniformly, the final estimators $\hat{g}(S_i) = \hat{g}^{(T,n_T)}(S_i)$ estimate the integrals $g(S_i)$ only up to a common constant, denoted A, which depends on δ_0 and the number of iterations made at each stage of the algorithm. To estimate the log normalizing constant of q, namely $\lambda = \log(c)$, we must remove A, which can be done by running GWL with a modified version of q as we now explain. Choose $S_2, \ldots, S_r$ to be such that $\left(\bigcup_{i=2}^r S_i\right)^c$ has negligible mass under q, and choose $S_1 \subset \left(\bigcup_{i=2}^r S_i\right)^c$ such that its volume $|S_1|$ is finite and known (Liang 2005). Next, run GWL with q replaced by

$$q_{\text{mod}}(\omega) = \begin{cases} q(\omega) & \text{for } \omega \in \bigcup_{i=2}^r S_i, \\ \frac{1}{|S_1|} & \text{for } \omega \in S_1, \\ 0 & \text{otherwise.} \end{cases}$$

The basic idea is that on S_1 we can treat q as uniform since its actual distribution contributes little to the normalizing constant of q. Lastly, for each $S \in \{S_2, \ldots, S_r\}$, Liang (2005) estimated the integral $\int_S q(\omega)d\omega$ by $\hat{g}(S)/\hat{g}(S_1)$. Assuming convergence of GWL, a consistent estimate of λ is thus given by $\hat{\lambda}_{\text{GWL}} = \log(\sum_{i=2}^r \hat{g}(S_i)/\hat{g}(S_1))$.

The key strengths of GWL are its adaptive nature and that it exhibits good mixing properties even for multimodal targets, the latter property being a benefit of asymptotic uniform sampling across energy bins. In practice, a limitation of the algorithm is that the MSE of $\hat{\lambda}_{\text{GWL}}$ is bounded below for fixed n_0 and the specified geometric growth in n_t, as can be seen in the top left panel of Figure 10 (discussed in Section 5.3). (Liang, Liu, and Carroll (2007) attempted to mitigate this phenomenon, but the convergence properties of the updated algorithm again may not be ideal, and further developments are still being made; see for example Jacob and Ryder (2014).) Here, we simply view GWL as a related method to compare against and combine with. For these purposes the lower bound on the convergence of $\hat{\lambda}_{\text{GWL}}$ does not play a large role because the number of target evaluations we allow is approximately equal to or lower than the number required by GWL to achieve its minimum MSE. For further details of GWL, the reader is referred to Liang (2005), Liang, Liu, and Carroll (2007), Bornn et al. (2013), and Jacob and Ryder (2014).

5.2. Combining GWL With Warp-U

Consider a situation where GWL has been run for $T^* < T$ stages. We suspect that it is near convergence, and want to determine if we can reduce the MSE by using the computation to complete only the remaining $T - T^*$ stages (the GWL-only approach) or to make the use of Warp-U bridge sampling to obtain the estimator $\hat{\lambda}_{\text{opt}}^{(\text{U})}$ (the GWL+Warp approach). Draws from our target p are required to implement the latter approach, but can be obtained from the GWL run without any additional target evaluations. To see this, first let G_i denote the set of samples collected from subregion S_i during the T^* stages of the GWL run, for $i = 1, \ldots, r$. With this notation, we propose the following addition to GWL.

Warp-U addition.

1. For $l = 1, \ldots, n$, repeat the following two steps:
   (i) Sample a subregion index $k \in \{2, \ldots, r\}$ using the probabilities $b_i \propto \hat{g}(S_i)1_{\{G_i \neq \emptyset\}}$, for $i = 2, \ldots, r$.
   (ii) With uniform sampling, select a sample $\omega_l \in G_k$, that is, select one of the samples in subregion S_k collected by GWL.
2. Apply Warp-U bridge sampling with $\{\omega_1, \ldots, \omega_n\}$ and m draws from ϕ to obtain the estimate $\hat{\lambda}_{\text{opt}}^{(\text{U})}$.

The first step obtains draws from p restricted to $\bigcup_{i=2}^r S_i$, and the second step applies Warp-U bridge sampling using these draws. The indicator $1_{\{G_i \neq \emptyset\}}$ in b_i indicates that we sample only regions from which there are samples during the GWL run. Since $q(\omega_l)$, for $l = 1, \ldots, n$, has already been evaluated during the GWL run, the only new target evaluations required for the above Warp-U addition are related to Step 2 not Step 1: in particular, $n(K-1) + mK$ evaluations are needed to compute the $K-1$ terms of $\tilde{q}$ in (17) (i.e., terms of the form $q(\mathcal{S}_k\tilde{w} + \mu_k)$) for which $\mathcal{S}_k\tilde{w}_l + \mu_k \neq w_l$, for $l = 1, \ldots, n$, and the full K terms of $\tilde{q}$ for the m draws from ϕ. In Step 2, we could alternatively apply standard bridge sampling to p and ϕ_{mix} to obtain $\hat{\lambda}_{\text{opt}}^{\text{mix}}$, as described in Section 3.3. We refer to this alternative approach as GWL+BS.

5.3. Illustration of the GWL+Warp-U Algorithm

We consider the 25 skewed-t mixture example from Section 4.2. We run the GWL-only algorithm for $T = 25$ stages, and again set $\delta_0 = e - 1$. We try $n_0 = 10^3, 10^4, 10^5$ and find that larger n_0 requires more stages for the estimator $\hat{\lambda}_{\text{GWL}}$ to converge but leads to lower RMSE; see the top left panel of Figure 10. We choose $n_0 = 10^4$ as a compromise between RMSE and computational cost. The value of n_0 is not the target of our comparisons, nor is the choice of the proposal density h or the partition $S_1, \ldots, S_r$. We therefore use our knowledge of the true target p to configure these components to favor the GWL-only algorithm; see Appendix E in the supplementary materials for details.

The solid lines in the top left panel of Figure 10 indicate the RMSE when the GWL-only algorithm (with $n = 10^4$) is run for 5, 9, and 11 stages, and we see the RMSE is stabilized after stage 9. For reference, the same three RMSE values are indicated by a horizontal line in the top right, bottom left, and bottom right panels of Figure 10, respectively. Running GWL for more than 9 stages does not improve the RMSE, but we can improve it using the Warp-U addition detailed in Section 5.2, even if we do not increase the overall computational cost. In the top right panel of Figure 10, the hollow symbols show the RMSE for the GWL+Warp-U estimator $\hat{\lambda}_{\text{opt}}^{(\text{U})}$, in the case where the GWL component was run for only 10 stages, thereby saving the computation needed for an 11th stage to be used for Warp-U sampling. Of course, we may decide we do not need to use all of the saved computation. The x-axis of Figure 10 gives the number of target evaluations we use for the Warp-U step, in units of $\log_{10}$ of the proportion of target evaluations needed for an 11th stage of GWL, that is, at 0 on the x-axis the computational cost of the Warp-U sampling matches that of an 11th stage. The different hollow shapes correspond to different values of K (number of mixture components). For comparison, the horizontal line indicates the RMSE under the GWL-only

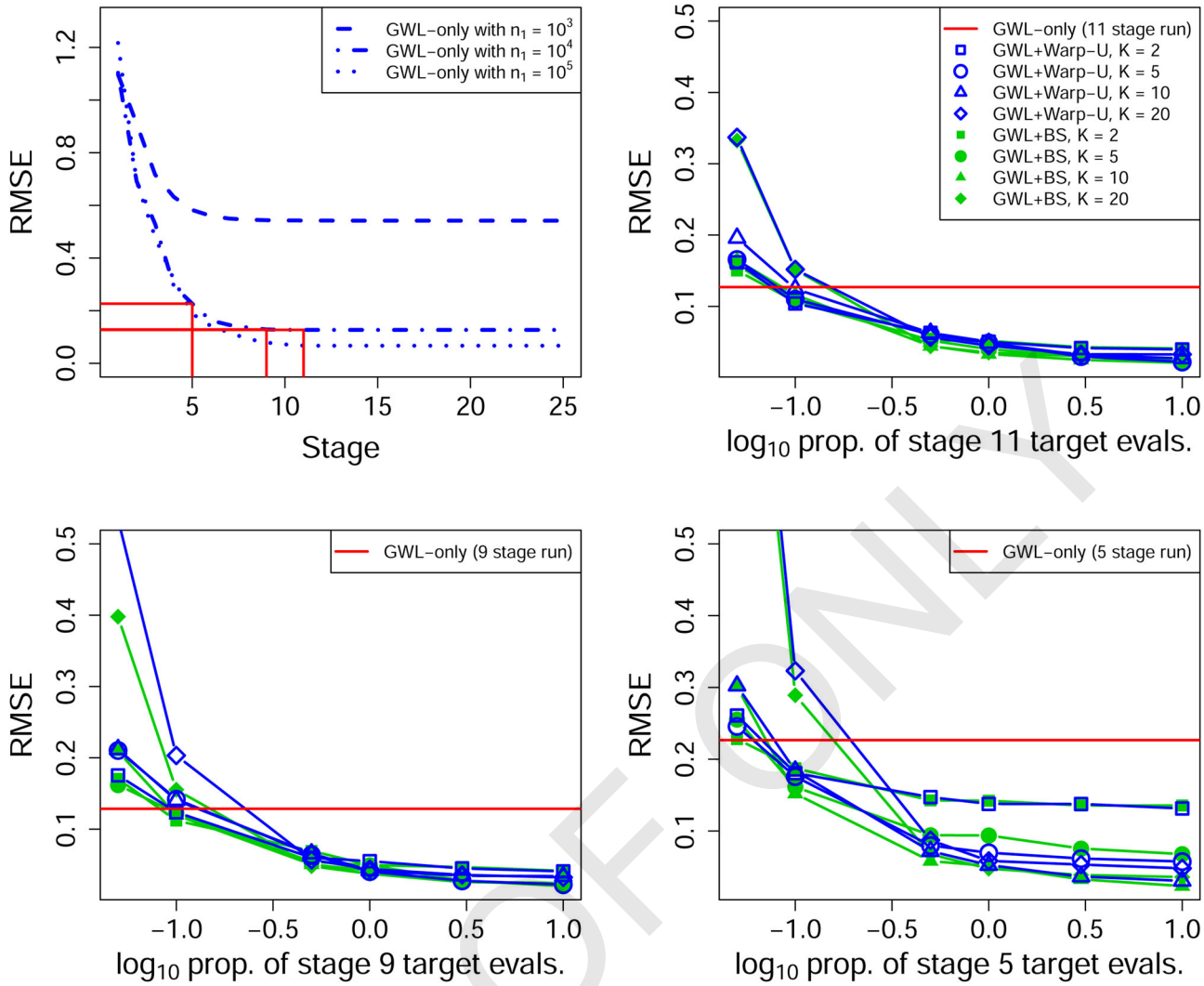

Figure 10. The top left panel shows the RMSE of $\hat{\lambda}_{\text{GWL}}$ at each of 25 stages, for $n_0 = 10^3, 10^4, 10^5$. The solid lines indicate the results for 11-stage, 9-stage, and 5-stage runs of the GWL-only algorithm with $n_0 = 10^4$. The top right panel shows the RMSE of the GWL+Warp-U estimator (hollow symbols) and the GWL+BS estimator (solid symbols), where the initial GWL run is for 10 stages. The number of target evaluations used by the Warp-U or ordinary bridge sampling step is given on the x-axis as a proportion of the target evaluations that would be needed for an eleventh stage of GWL (on a $\log_{10}$ scale). The different shapes correspond to different settings of K. The bottom left and right panels show similar results where the initial GWL run is for 8 and 4 stages, respectively.

algorithm after 11 stages (for which the x-axis is irrelevant). At 0 on the x-axis the GWL-only and GWL+Warp-U methods use the same number of target evaluations, but the GWL+Warp-U method yields substantially lower RMSE for all four values of K (2, 5, 10, and 20). Indeed, the RMSE obtained is even lower than that achieved by the GWL-only algorithm with $n_0 = 10^5$ after 11 (or 25) stages, which uses a factor of 10 more target evaluations than used by the GWL+Warp-U runs (with $n_0 = 10^4$). This example suggests that applying Warp-U when GWL is at or close to convergence offers a way to substantially lower RMSE with only the number of target evaluations required to run one more stage of GWL (or even fewer). The improvements offered by GWL+Warp-U are thus almost free because we can stop GWL one stage early, and in any case the number of target evaluations required by the Warp-U part is relatively low.

For further comparison, the solid shapes in the top right panel of Figure 10 show results for the GWL+BS method, that is, where the two bridge sampling densities are p and ϕ_{mix}. Note that, at any given point on the x-axis the number of target evaluations is the same for the GWL+BS and GWL+Warp-U algorithms, which is achieved by setting $m = n(2K - 1)$ in the GWL+BS algorithm and $m = n$ in the GWL+Warp-U algorithm. We see that GWL+BS again achieves substantially lower RMSE than is obtained by simply running an 11th stage of the GWL-only algorithm. The RMSE under the GWL+BS method is also seen to be marginally lower than that under the GWL+Warp-U method, but now that we have controlled the number of target evaluations, we see the performance is very similar.

Naturally, Warp-U or standard bridge sampling can be applied after any number of GWL stages; we do not have to wait until we are sure of convergence. To investigate this, we ran the GWL+Warp-U and GWL+BS algorithms again but where the initial GWL run was shorter. The bottom left and bottom right panel of Figure 10 show results in the case where the initial GWL run was 8 and 4 stages, respectively. These results are qualitatively similar to before, except that GWL+Warp-U occasionally performs slightly better than GWL+BS (e.g., when

$K = 5$ in the bottom right panel). The main constraint on when Warp-U or standard bridge sampling can be used is that GWL needs to have been run for long enough to have explored regions where p has high density. Otherwise, weighted resampling of the GWL samples will not yield approximate draws from p. This is not a major issue in practice because, as can be seen in the top left panel of Figure 10, GWL substantially reduces RMSE in its initial stages, and it is likely to be the latter stages where applying Warp-U or standard bridge sampling is most appealing.

6. Strategies for Making Further Improvements

We have seen that stochastic Warp-U transformations can improve overlap between multimodal densities by transforming them into approximately unimodal ones, and Theorem 1 implies that in terms of statistical efficiency, Warp-U transformations are always beneficial. However, this says little about computational efficiency, which is an important direction for further research, as is identifying a good approximating mixture distribution ϕ_{mix}. Below we briefly discuss both.

6.1. Reducing Computational Cost

Perhaps one of the most promising approaches for reducing computation is to construct a *complementary* sampling method that is itself efficient but also computes many of the quantities needed in the final estimation step. In general, it is reasonable to expect that posterior sampling will require substantially more computational resources than the final Warp-U estimation step (which should nevertheless be as efficient as possible). Indeed, in Section 5.3, we saw that sampling using the GWL algorithm required substantially more computation than the final estimation step.

Another approach is to reduce the cost of evaluations of the transformed density $\tilde{q}$. From (17), it can be seen that the density $\tilde{q}$ is at least K times more expensive to evaluate than q, the latter being used in ordinary bridge sampling. It is therefore of interest to find ways to reduce the computational cost of evaluating $\tilde{q}$, as compared to q. Direct density approximation is unlikely to be fruitful in high dimensions. A plausible strategy is to randomly select only some of the K terms in (17) to evaluate, as is done in the mixture sampling method of Elvira et al. (2019); they randomly partitioned the set of sampling densities and then used these partitions in computing importance weights. Specifically, Elvira et al. (2019) set the importance weight for a sample $\omega \sim p_s$ to be the reciprocal of the average of the sampling densities in the same partition as p_s evaluated at ω. We could apply a similar approach by viewing the terms $c\tilde{p}^{(k)}(\tilde{\omega}) = cp^{(k)}(S_k\tilde{\omega} + \mu_k))$ in (17) as the weighted and unnormalized sampling densities. However, we do not know from which weighted component $c\tilde{p}^{(k)}$ each sample $\tilde{\omega} \sim \tilde{p}$ originated. Furthermore, the $c\tilde{p}^{(k)}$ terms incorporate unknown weights, so the unweighted mixture components of $\tilde{q}$ are not available and therefore we cannot combine them using their empirical weights, as is required by (generalized) bridge sampling, for example, the weights s_i, for $i = 1, 2$, in (4). Regarding the unknown sampling component $c\tilde{p}^{(k)}$, it may be sufficient to stochastically impute the index of the "true" component by drawing from the conditional distribution $\varpi(\Psi|\omega)$ in (14). A possible solution to the second difficulty is to use the theoretical component weights incorporated in the $c\tilde{p}^{(k)}$ in the final bridge sampling estimator as opposed to the observed weights. These possibilities need to be explored.

6.2. Base and Mixture Distribution Selection

Theorem 1 gives us the freedom to choose the base distribution ϕ (now viewed as a generic density). As mentioned earlier, for a heavy-tailed target p, a t-distribution may be more efficient as a base density than the standard normal, rendering fewer mixture components and thereby reducing computational costs. In other contexts, the support of p may be bounded and then a base density with bounded support would be more appropriate. There are also cases where it would be beneficial for ϕ_{mix} to be a mixture of several different base densities, though this would require some modifications to the development here. With a different choice of the base function, a different method for fitting ϕ_{mix} would be needed, for example, for the t-distribution, we could use the approach of Peel and McLachlan (2000) to fit ϕ_{mix}.

Second, although our approach for fitting the Warp-U parameters $\boldsymbol{\zeta}$ is promising in practice, it is almost certainly not optimal. Kong et al. (2003) showed that a standard bridge sampling estimator is in fact a maximum likelihood estimator (MLE), and it may be possible to use the same likelihood framework to find an optimal estimator for $\boldsymbol{\zeta}$. More specifically, let ϕ_i be the pdf of $\mathcal{N}(\mu_i, \Sigma_i)$, for $i = 1, \ldots, K$, and $\phi_{\text{mix}} = \sum_{i=1}^{K} \pi_i \phi_i$. Then the maximum likelihood estimator of c (with $p = q/c$ as before) identified by Kong et al. (2003) is

$$\hat{c} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \frac{q(w_{i,j})}{n_1 \hat{c}^{-1} q(w_{i,j}) + n_2 \phi_{\text{mix}}(\omega_{i,j})}, \tag{22}$$

where $\{\omega_{1,1}, \ldots, \omega_{1,n_1}\}$ are draws from p, and $\{\omega_{2,1}, \ldots, \omega_{2,n_2}\}$ are draws from ϕ_{mix}. The estimator (22) is the same as $\hat{c}_{\text{opt}}^{(\text{mix})} = \exp(\hat{\lambda}_{\text{opt}}^{(\text{mix})})$. If Warp-U transformations can be correctly incorporated into this likelihood framework then we can use the MLE of $(c, \boldsymbol{\zeta})$ to improve upon our current approach. Jones (2015) provided initial insights into this likelihood formulation, but also identified challenges for making this likelihood approach fruitful for warp bridge sampling.

6.3. It's Time to Build a Bridge

The vast majority of the MC literature is about improving MC *sampling* efficiency, that is, how to design MC sampling algorithms most effectively. In contrast, bridge sampling, with or without warp transformations, is about improving MC *inference* efficiency, that is, how to gain more precision with a given set of MC draws. Adding warp bridge sampling to GWL provides an example of bridging the sampling and analysis approaches, a strategy we believe has much more to offer than the current literature recognizes. We therefore invite interested readers to join us in laying further foundations for this much needed bridge.

Supplementary Materials

Q4

Acknowledgments

The authors thank an anonymous associate editor and two anonymous referees for constructive comments and suggestions that improved the article. The authors also gratefully acknowledge helpful conversations with members of the Department of Statistics at Harvard University and at Texas A&M University, constructive comments from the audience of the 2016 MCQMC conference at Stanford University.

Disclaimer

The views expressed herein are solely the views of the author(s) and are not necessarily the views of Two Sigma Investments, LP or any of its affiliates. They are not intended to provide, and should not be relied upon for, investment advice. Please see the full disclaimer on page 10 of the online supplementary materials.

Funding

The authors thank partial financial support from NSF and JTF.

References

Alspach, D. L., and Sorenson, H. W. (1972), "Nonlinear Bayesian Estimation Using Gaussian Sum Approximations," *IEEE Transactions on Automatic Control*, 17, 439–448. [4]

Azzalini, A. (2011), "R Package sn: The Skew-Normal and Skew-*t* Distributions" (version 0.4-17), available at *http://azzalini.stat.unipd.it/SN*. [9]

——— (2013), *The Skew-Normal and Related Families*, New York: Cambridge University Press. [9]

Bennett, C. H. (1976), "Efficient Estimation of Free Energy Differences From Monte Carlo Data," *Journal of Computational Physics*, 22, 245–268. [1,2]

Bornkamp, B. (2011), "Approximating Probability Densities by Iterated Laplace Approximations," *Journal of Computational and Graphical Statistics*, 20, 656–669. [8]

Bornn, L., Jacob, P. E., Del Moral, P., and Doucet, A. (2013), "An Adaptive Interacting Wang–Landau Algorithm for Automatic Density Exploration," *Journal of Computational and Graphical Statistics*, 22, 749–773. [13]

Ceperley, D. M. (1995), "Path Integrals in the Theory of Condensed Helium," *Reviews of Modern Physics*, 67, 279–355. [1]

Chen, J., and Tan, X. (2009), "Inference for Multivariate Normal Mixtures," *Journal of Multivariate Analysis*, 100, 1367–1383. [8]

Chen, J., Tan, X., and Zhang, R. (2008), "Inference for Normal Mixtures in Mean and Variance," *Statistica Sinica*, 18, 443–465. [8]

Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321. [2]

Chib, S., and Jeliazkov, I. (2001), "Marginal Likelihood From the Metropolis–Hastings Output," *Journal of the American Statistical Association*, 96, 270–281. [2]

Day, N. E. (1969), "Estimating the Components of a Mixture of Normal Distributions," *Biometrika*, 56, 463–474. [8]

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, 92, 903–915. [1]

Elvira, V., Martino, L., Luengo, D., and Bugallo, M. F. (2019), "Generalized Multiple Importance Sampling," *Statistical Science*, 34, 129–155. [15]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, Boca Raton, FL: CRC Press. [8]

Gelman, A., and Meng, X.-L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185. [1]

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017), "A Tutorial on Bridge Sampling," *Journal of Mathematical Psychology*, 81, 80–97. [2]

Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (2017), "Bridgesampling: An R Package for Estimating Normalizing Constants," arXiv no. 1710.08162. [2]

Hesterberg, T. (1995), "Weighted Average Importance Sampling and Defensive Mixture Distributions," *Technometrics*, 37, 185–194. [2]

Jacob, P. E., and Ryder, R. J. (2014), "The Wang–Landau Algorithm Reaches the Flat Histogram Criterion in Finite Time," *The Annals of Applied Probability*, 24, 34–53. [13]

Jones, D. E. (2015), "Likelihood Methods for Monte Carlo Estimation," Ph.D. qualifying paper, Harvard University, Department of Statistics, pp. 1–30. [15]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [1]

Kiefer, J., and Wolfowitz, J. (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *The Annals of Mathematical Statistics*, 27, 887–906. [8]

Kong, A., McCullagh, P., Meng, X.-L., and Nicolae, D. (2006), "Further Explorations of Likelihood Theory for Monte Carlo Integration," in *Advances in Statistical Modeling and Inference: Essays in Honor of Kjell A. Doksum*, ed. V. Nair, Singapore: World Scientific Press, pp. 563–592. [2]

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., and Tan, Z. (2003), "A Theory of Statistical Models for Monte Carlo Integration" (with discussions), *Journal of the Royal Statistical Society*, Series B, 65, 585–604. [2,3,15]

Kou, S., Zhou, Q., and Wong, W. H. (2006), "Equi-Energy Sampler With Applications in Statistical Inference and Statistical Mechanics," *The Annals of Statistics*, 34, 1581–1619. [12]

Liang, F. (2005), "A Generalized Wang–Landau Algorithm for Monte Carlo Computation," *Journal of the American Statistical Association*, 100, 1311–1327. [2,12,13]

Liang, F., Liu, C., and Carroll, R. J. (2007), "Stochastic Approximation in Monte Carlo Computation," *Journal of the American Statistical Association*, 102, 305–320. [13]

Liu, J. S., Liang, F., and Wong, W. H. (2001), "A Theory for Dynamic Weighting in Monte Carlo Computation," *Journal of the American Statistical Association*, 96, 561–573. [2]

Martino, L., Elvira, V., Luengo, D., and Corander, J. (2017), "Layered Adaptive Importance Sampling," *Statistics and Computing*, 27, 599–623. [2]

Meng, X.-L. (2005), "Comment: Computation, Survey and Inference," *Statistical Science*, 20, 21–28. [2]

Meng, X.-L., and Schilling, S. (2002), "Warp Bridge Sampling," *Journal of Computational and Graphical Statistics*, 11, 552–586. [2,3]

Meng, X.-L., and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860. [1,2,3]

Mira, A., and Nicholls, G. (2004), "Bridge Estimation of the Probability Density at a Point," *Statistica Sinica*, 14, 603–612. [2]

Owen, A., and Zhou, Y. (2000), "Safe and Effective Importance Sampling," *Journal of the American Statistical Association*, 95, 135–143. [2]

Peel, D., and McLachlan, G. J. (2000), "Robust Mixture Modelling Using the *t*-Distribution," *Statistics and Computing*, 10, 339–348. [15]

Romero, M. (2003), "On Two Topics With No Bridge: Bridge Sampling With Dependent Draws and Bias of the Multiple Imputation Variance Estimator," Ph.D. thesis, University of Chicago, Department of Statistics. [3]

Shao, Q.-M., and Ibrahim, J. G. (2000), *Monte Carlo Methods in Bayesian Computation*, Springer Series in Statistics, New York: Springer. [1]

Tan, Z. (2004), "On a Likelihood Approach for Monte Carlo Integration," *Journal of the American Statistical Association*, 99, 1027–1036. [2]

——— (2013), "Calibrated Path Sampling and Stepwise Bridge Sampling," *Journal of Statistical Planning and Inference*, 143, 675–690. [1]

Veach, E., and Guibas, L. J. (1995), "Optimally Combining Sampling Techniques for Monte Carlo Rendering," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, ACM, New York, NY, pp. 419–428. [2]

Villani, C. (2003), *Topics in Optimal Transportation* (Vol. 58), Providence, RI: American Mathematical Society. [2]

Voter, A. F. (1985), "A Monte Carlo Method for Determining Free-Energy Differences and Transition State Theory Rate Constants," *The Journal of Chemical Physics*, 82, 1890–1899. [1]

Voter, A. F., and Doll, J. D. (1985), "Dynamical Corrections to Transition State Theory for Multistate Systems: Surface Self-Diffusion in the Rare-Event Regime," *The Journal of Chemical Physics*, 82, 80–92. [1]

Wang, F., and Landau, D. (2001), "Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States," *Physical Review Letters*, 86, 2050–2053. [2]

Wong, W. H., and Liang, F. (1997), "Dynamic Weighting in Monte Carlo and Optimization," *Proceedings of the National Academy of Sciences of the United States of America*, 94, 14220–14224. [2]