# The Coupled Rejection Sampler

Adrien Corenflos[1] and Simo Särkkä[1]

[1]Department of Electrical Engineering and Automation, Aalto University.

March 11, 2022

### Abstract

We propose a coupled rejection-sampling method for sampling from couplings of arbitrary distributions. The method relies on accepting or rejecting coupled samples coming from dominating marginals. Contrary to existing acceptance-rejection coupling methods, the variance of the execution time of the proposed method is limited and stays finite as the two target marginals approach each other in the sense of the total variation norm. In the important special case of coupling multivariate Gaussians with different means and covariances, we derive positive lower bounds for the resulting coupling probability of our algorithm, and we then show how the coupling method can be optimized in closed form. Finally, we show how we can modify the coupled rejection-sampling method to propose from coupled ensemble of proposals, so as to asymptotically recover a maximal coupling. We then apply the method to the problem of coupling rare events samplers, derive a parallel coupled resampling algorithm to use in particle filtering, and show how the coupled rejection-sampler can be used to speed up unbiased MCMC methods based on couplings.

**Keywords:** coupling methods; parallel resampling; ensemble methods; unbiased Markov chain Monte Carlo

## 1 Introduction

Given two probability distributions $p$ and $q$, a coupling (see, e.g., Thorisson, 2000; Lindvall, 2002) of $p$ and $q$ is defined as a pair of random variables $(X, Y)$, defined on a joint probability space, that are marginally distributed according to $p$ and $q$, respectively. Typical examples of such couplings, exhibiting different uses and properties, comprise (i) the trivial independent coupling, where $X \sim p$ and $Y \sim q$ with $X$ independent of $Y$; (ii) optimal transport couplings, for which the joint distribution $\Gamma_{OT}$ of $(X, Y)$ minimizes a given objective over all possible couplings $\Gamma_{OT} = \operatorname{argmin}_\Gamma \int c(x, y)\Gamma(\mathrm{d}x, \mathrm{d}y)$, (Villani, 2009); (iii) common random number couplings, where the same uniform/random seed is used to sample the marginals, e.g., if we know the quantile functions $F^-$ and $G^-$ of $p$ and $q$, we can sample $X = F^-(U)$, $Y = G^-(U)$ where $U$ is uniformly distributed (Whitt, 1976; Gal et al., 1984); (iv) similarly to common random numbers, antithetic variables, which are constructed in the same way as the common random numbers couplings but by using $X = F^-(U)$, $Y = G^-(1 - U)$ instead (Hammersley and Morton, 1956).

In this article, we consider another class of couplings, which preserve mass over the diagonal of the joint distribution, that is, couplings $(X, Y)$ such that $\mathbb{P}(X = Y) > 0$. Such couplings are ubiquitous in proving a number of inequalities used, for example, to study the convergence of Monte Carlo algorithms (see, e.g., Reutter and Johnson, 1995; Lindvall, 2002). More recently, they have also have been used as a computational tool rather than a theoretical one. In particular, the seminal paper of Glynn and Rhee (2014) showed how to compute unbiased estimates of expectations using coupled Markov chains, allowing to then compute these to an arbitrary precision using distributed hardware (Jacob et al., 2020b). Following this work, several extensions have been developed, both in

the classical Markov chain Monte Carlo (MCMC) methods (Jacob et al., 2020b; Heng and Jacob, 2019; Xu et al., 2021; Wang et al., 2021), as well as in pseudo-marginal and particle Markov chain Monte Carlo (PMCMC) methods (Jacob et al., 2020a; Middleton et al., 2019, 2020).

If $X_t$ and $Y_t$ are two Markov chains, the coupling (or meeting) time (e.g. Thorisson, 2000; Lindvall, 2002) is defined as $\inf\{t \geq 0 \mid X_t = Y_t\}$. In the context of MCMC methods, a common requirement of debiasing coupling methods is that the constructed coupled Markov chains have a coupling time with finite expectation $\mathbb{E}[\tau] < \infty$ (e.g. Jacob et al., 2020b) (more precisely, it needs to have polynomial tails (Middleton et al., 2019)). To construct such Markov chains, one typically relies on implementing a maximum coupling of the proposal distributions $q(\cdot \mid X_t)$, $q(\cdot \mid Y_t)$ or of the Metropolis-Hastings kernel directly (Wang et al., 2021).

When considering coupling the proposals, a maximal coupling is a pair of random variables $(X^*, Y^*)$ which are marginally distributed according to $q(\cdot \mid X_t)$ and $q(\cdot \mid Y_t)$, respectively, such that $\mathbb{P}(X^* = Y^*)$ is maximized over the (infinite) set of all possible couplings. When $q(\cdot \mid X_t)$ and $q(\cdot \mid Y_t)$ are Gaussian distributions with the same covariance, this can be done, for example, using the reflection-maximal coupling introduced in Bou-Rabee et al. (2020). However, in the case when the covariances are different, one needs to resort to using a rejection sampling algorithm (see, e.g. Thorisson, 2000; Johnson, 1998). An important drawback of the latter algorithm is that the variance of its (random) run time goes to infinity as the total variation distance of the marginal distributions goes to zero (Gerber and Lee, 2021). This property limits its application in contractive schemes, such as unbiased MCMC methods, where the aim is to get the marginals to converge to one another. However, as pointed out in Gerber and Lee (2021), if one is willing to compromise on the maximal property, that is, sample $X^*$ and $Y^*$ with a suboptimal diagonal mass $\mathbb{P}(X^* = Y^*)$, then it is possible to control the variance of the run time. This can be used as a trade-off between computational cost and optimality (Jacob, 2021, Ch. 2).

The first contribution of this article is to present an algorithm for sampling from couplings of arbitrary distributions by using an acceptance-rejection scheme on proposal samples coming from a dominating coupling. The algorithm has the advantage that the variance of the execution time stays finite as the coupling marginals approach each other. The second contribution is to specialize this to Gaussians with different means and covariances, derive a principled convex optimisation scheme to find a good dominating proposal, and show that the coupling probability is lower-bounded by a positive (parametric) constant. The lower bounds we derive for the acceptance probability also lead to upper bounds for the total variation distance between Gaussian distributions that are always less than 1. The third contribution is to modify the acceptance-rejection scheme by considering a coupled ensemble proposal such that, as the size of the proposal ensemble increases, the resulting algorithm asymptotically produces samples from a maximal coupling. We finally apply the methods introduced to derive a novel parallel coupled resampling algorithm for particle filtering, and show the attractiveness of our algorithms on several sampling problems.

The structure of the article is the following. In Section 2 we describe the coupling algorithm for generic distributions, prove its correctness, and analyse its theoretical properties. In Section 3, by extending Deligiannidis et al. (2020), we subsequently derive a coupled ensemble rejection sampler that extends the coupled rejection-sampler algorithm designed in Section 2. We derive asymptotic bounds for the coupling success probability of the resulting coupling that show how this new method asymptotically recovers a maximal coupling. In Section 4 we then discuss the special case of designing efficient coupling of Gaussian distributions with different means and covariances. We derive tight positive lower bounds for the coupling success probability and show how to design and compute efficient dominating Gaussian couplings. Finally, in Section 5, we (i) show how the method can be used to sample from couplings of Gaussian tails, (ii) implement a coupled parallel resampling algorithm, (iii) compare our method with the algorithm discussed in Gerber and Lee (2021) on a Gibbs sampling problem, and finally (iv) compare our algorithm to Wang et al. (2021) on a Bayesian logistic regression problem using a manifold Metropolis-adjusted Langevin algorithm sampler.

## Assumptions and Notations

In this article we suppose that all the random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and take value in a measurable space $\chi$ being either $\mathbb{Z}$ or $\mathbb{R}^d$ for some $d \geq 1$. It is, however, worth noting that the results and algorithms of this article directly extend to any measurable space for which the event $\{X = Y\}$ is measurable for all random variables $X, Y$.

When considering that $X$ is sampled from a probability distribution $p$ we write $X \sim p$. For the sake of notational simplicity, we will assume throughout this article that all distributions, apart from coupling distributions, have a density with respect to a common measure (e.g., Lebesgue measure for densities defined on $\mathbb{R}^d$ and the counting measure on the set of integers $\mathbb{Z}$), and, where no confusion is possible, we will identify the probability measure with its density. We, however, emphasize that all of our results can easily be rewritten in terms of Radon–Nikodym derivatives. Moreover, for the sake of conciseness, we use a slight abuse of notation by denoting $\int (\cdot) \, \mathrm{d}x = \int (\cdot) \, \mathrm{d}\mu$ for an integral with respect to an implicit unnamed reference measure $\mu$.

The uniform distribution on $[0,1]$ is denoted by $\mathcal{U}(0,1)$. The probability distribution of a general coupling of $(X, Y)$ is denoted by $\Gamma$ and the probability distribution of a given dominating coupling is denoted by $\hat{\Gamma}$. When there is no ambiguity we interchangeably refer to the pair $(X, Y)$ or its distribution as the coupling. Finally, for any random variable $X$ and measurable event $E$, we write $X \mid E$ for a random variable with law $\mathbb{P}((X \mid E) \in A) = \mathbb{P}(X \in A \mid E)$.

When dealing with sequences $u_N$ and $v_N$, we say that they are equivalent and write $u_N \sim v_N$ if we have $\frac{u_N}{v_N} \to 1$ as $N$ goes to infinity. Furthermore, we write $\limsup_{N \to \infty} u_N$ for the quantity $\lim_{N \to \infty} \sup_{K \geq N} u_K$ when it exists, and similarly for $\liminf$.

We denote the Gaussian distribution with mean $\mu$ and covariance $\Sigma$ by $\mathcal{N}(\mu, \Sigma)$ and the corresponding density by $\mathcal{N}(x; \mu, \Sigma)$. For a given covariance matrix $\Sigma$ we write $\Sigma^{1/2}$ for its lower Cholesky factor and $\Sigma^{-1/2}$ for the inverse thereof.

## 2 The coupled rejection-sampling method

Let $p$ and $q$ be two probability densities defined on $\chi$. We define a diagonal coupling of $p$ and $q$ in the following way.

**Definition 1** (Diagonal coupling). *A pair of random variables $(X, Y)$ defined on the same probability space and taking values in $\chi$ is a diagonal coupling of the probability densities $p$ and $q$ if and only if it is a coupling, that is, if*

$$\mathbb{P}(X \in A) = \int_A p(x) \, \mathrm{d}x, \quad \mathbb{P}(Y \in A) = \int_A q(x) \, \mathrm{d}x,$$

*for all measurable sets $A \subset \chi$, and if $\mathbb{P}(X = Y) > 0$.*

With this definition, a maximal coupling (Thorisson, 2000; Lindvall, 2002) is a diagonal coupling with maximal mass on the diagonal event $\{X = Y\}$. The maximal coupling also has a connection with the total variation distance between the distributions $\mathbb{P}(X)$ and $\mathbb{P}(Y)$. If we define total variation distance as $\|p - q\|_{TV} = \sup_A P(X \in A) - P(Y \in A) = 1 - \int \min(p(x), q(x)) \, dx$ then for a maximal coupling $(X, Y)$, we have

$$\mathbb{P}(X = Y) = 1 - \|p - q\|_{TV} = \int \min(p(x), q(x)) \, dx. \tag{1}$$

In order to construct the acceptance-rejection scheme, we need the following definition.

**Definition 2** (Dominating pair). *Let $(\hat{p}, \hat{q})$ be a pair of probability densities. We say that it dominates $(p, q)$ (or that $(p, q) \preceq (\hat{p}, \hat{q})$) if and only if there exists $1 \leq M(p, \hat{p}) < \infty$ and $1 \leq M(q, \hat{q}) < \infty$ such that for any $x \in \mathbb{R}^d$ we have*

$$p(x) \leq M(p, \hat{p}) \, \hat{p}(x), \quad q(x) \leq M(q, \hat{q}) \, \hat{q}(x). \tag{2}$$

It is worth noting that even if $(\hat{p}, \hat{q})$ dominates $(p, q)$, a coupling $\hat{\Gamma}$ of $(\hat{p}, \hat{q})$ does not necessarily dominate a coupling $\Gamma$ of $(p, q)$. Indeed, there might not exist $M(\Gamma, \hat{\Gamma})$ such that $\frac{d\Gamma}{d\hat{\Gamma}} \leq M(\Gamma, \hat{\Gamma})$ uniformly. However, in a slight abuse of language and notation, we say that a coupling $\hat{\Gamma}$ dominates the coupling $\Gamma$ (i.e., $\hat{\Gamma} \succeq \Gamma$) when its marginals do.

Now, suppose that we have a diagonal coupling $\hat{\Gamma}$ of $\hat{p}$ and $\hat{q}$ dominating the independent coupling $(p \otimes q)(x, y) = p(x) \, q(y)$. It turns out that we can produce samples from a diagonal coupling $\Gamma$ with marginals $p$ and $q$ by using an acceptance-rejection scheme with the coupling $\hat{\Gamma}$ as the proposal. Algorithm 1 shows how do this.

---

**Algorithm 1:** Rejection-coupling of $(p, q)$

---

**1 Function** RejectionCoupling($\hat{\Gamma}$, $p$, $q$):

      // Supposing $\hat{\Gamma} \succeq p \otimes q$ is a coupling of $\hat{p}$ and $\hat{q}$.

**2**    Set $A_X = 0$ and $A_Y = 0$ // Acceptance flags

**3**    **while** $A_X = 0$ *and* $A_Y = 0$ **do**

**4**        Sample $X_1, Y_1 \sim \hat{\Gamma}$, $U \sim \mathcal{U}(0, 1)$

**5**        **if** $U < \frac{p(X_1)}{M(p, \hat{p})\hat{p}(X_1)}$ **then** set $A_X = 1$

**6**        **if** $U < \frac{q(Y_1)}{M(q, \hat{q})\hat{q}(Y_1)}$ **then** set $A_Y = 1$

**7**    Sample $X_2, Y_2$ from $p \otimes q$

      **Return:** $X = A_X \, X_1 + (1 - A_X) \, X_2$, $Y = A_Y \, Y_1 + (1 - A_Y) \, Y_2$

---

Let us denote by $\tau$ the number of times $\hat{\Gamma}$ is sampled from. We call this quantity the number of steps of Algorithm 1 as it is roughly proportional to the run time of the algorithm, noting that in reality, sampling from $p \otimes q$ may be expensive too. The following proposition shows that Algorithm 1 samples from a coupling of $p$ and $q$, and the expectation and variance of its number of steps are bounded.

**Proposition 1.** *Suppose that $\hat{\Gamma} \succeq p \otimes q$. Then the random variables $X$ and $Y$ generated by Algorithm 1 are marginally distributed according to $p$ and $q$, respectively, and the mean and variance of the random number of steps $\tau$ of Algorithm 1 verify*

$$
\begin{aligned}
\mathbb{E}[\tau] &\leq \min\left(M(p, \hat{p}), M(q, \hat{q})\right), \\
\mathbb{V}[\tau] &\leq \min\left(M(p, \hat{p}), M(q, \hat{q})\right)^2 - \min\left(M(p, \hat{p}), M(q, \hat{q})\right).
\end{aligned}
\tag{3}
$$

*Finally, when the coupling for $X_2, Y_2$ is such that $\mathbb{P}(X_2 = Y_2) = 0$, we also have $\mathbb{P}(X = Y) \leq \mathbb{P}(X_1 = Y_1)$.*

*Proof.* The proof, mostly on classical rejection sampling theory, can be found in Appendix A. $\qquad \square$

**Remark 1.** *The upper bounds on the mean and variance of number of steps $\tau$ reduce to $1$ and $0$ when the proposals recover the targets.*

**Proposition 2.** *Suppose $\hat{\Gamma} \succeq p \otimes q$ is a diagonal coupling of $\hat{p}$ and $\hat{q}$. Furthermore, suppose that for all $x \in \mathbb{R}^d$, we have $\hat{p}(x) > 0 \Rightarrow p(x) > 0$ and similarly for $\hat{q}$ and $q$. In this case, Algorithm 1 defines a diagonal coupling of $p$ and $q$.*

*Proof.* Proposition 1 ensures that Algorithm 1 defines a coupling of $p$ and $q$. Now, using the notations of Algorithm 1,

$$
\begin{aligned}
\mathbb{P}(X = Y) &\geq \mathbb{P}(X_1 = Y_1 \ \& \ A_X = A_Y = 1) \\
&= \mathbb{P}(A_X = A_Y = 1 \mid X_1 = Y_1)\mathbb{P}(X_1 = Y_1) \\
&= \mathbb{P}(X_1 = Y_1) \int c(x) \min\left(\frac{p(x)}{M(p, \hat{p})\hat{p}(x)}, \frac{q(x)}{M(q, \hat{q})\hat{q}(x)}\right) dx,
\end{aligned}
\tag{4}
$$

4

where $c(x)$ is the density of $X_1 \mid X_1 = Y_1$.

Because $\mathbb{P}(X_1 = Y_1) > 0$, we must have a measurable set $A$ with $c(A) > 0$ and, for any $x \in A$, $\hat{p}(x) > 0$ and $\hat{q}(x) > 0$. This ensures that $\int c(x) \min\left(\frac{p(x)}{M(p,\hat{p})\hat{p}(x)}, \frac{q(x)}{M(q,\hat{q})\hat{q}(x)}\right) \mathrm{d}x > 0$, and finally that $\mathbb{P}(X = Y) > 0$. $\qquad\square$

Proposition 2 provides weak conditions for Algorithm 1 to return a diagonal coupling of $p$ and $q$, however, in practice, the coupling success probability will depend greatly on $\mathbb{P}(A_X = A_Y = 1 \mid X_1 = Y_1)$. This means $M(p,\hat{p})$ and $M(q,\hat{q})$ need to be as close as possible to 1 for the rejection step to not degrade the coupling success probability too much. This, unfortunately, may come at the cost of making the dominating coupling less successful, and it is not clear what the best trade-off is in this respect. In the next section we discuss how we can completely bypass this trade-off at the cost of additional (parallelisable) computations, while in Section 4 we will see how we can construct a tight dominating maximal coupling for arbitrary Gaussian distributions at low computational cost.

## 3 The ensemble coupled rejection-sampling method

As mentioned in Section 2, the success probability of the coupling defined by Algorithm 1 is highly dependent on the acceptance probability of the marginal rejection subroutine. In fact, if the latter goes to 0 (which is typically the case when the dimension of the sampling space increases), the former will too. Thankfully, in Deligiannidis et al. (2020), the authors introduce an ensemble rejection sampler which increases the acceptance probability of a rejection-sampling algorithm by considering not one, but several proposals. Formally, if $p_{\mathrm{RS}}$ is the probability of accepting a sample in the classical rejection-sampling algorithm, then their ensemble rejection-sampler with $N$ proposals has probability $p_{\mathrm{ERS}} \geq \frac{N p_{\mathrm{RS}}}{1+(N-1)p_{\mathrm{RS}}}$ to result in an accepted sample (Deligiannidis et al., 2020, Proposition 1).

In order to do so, the authors form the self-normalized importance sample $\sum_{i=1}^{N} W_i \delta_{X_i}$, where $X_i \sim \hat{p}, i = 1, \ldots, N$ are i.i.d., $w_i = \frac{p(X_i)}{\hat{p}(X_i)}$, $W_i = \frac{w_i}{\sum_{j=1}^{N} w_i}$, for $i = 1, \ldots, N$. When this is done, they sample $I$ from a categorical distribution $\mathrm{Cat}((W_i)_{i=1}^{N})$, and then accept the proposal $X_I$ with probability $\frac{\sum_{i=1}^{N} w_i}{M(p,\hat{p})+\sum_{i \neq I} w_i}$.

In order to improve the acceptance rate of Algorithm 1, we could decide to replace the rejection-sampling step by the ensemble rejection-sampling one and Propositions 1 and 2 would essentially remain unchanged. However, in practice, sampling $I$ and $J$ independently from the categorical distributions corresponding to the $p$ and $q$ component, respectively, would not result in a large probability for the event $I = J$, even if the corresponding weights are close and using common random numbers.

Thankfully, it is easy to sample from a maximal coupling of multinomial distributions (see, e.g., Thorisson, 2000, Section 4.2), and we can therefore write a coupled version of the ensemble rejection sampler. This is summarized in Algorithm 2.

In the remainder of this section, all the quantities $X, Y, \mathbb{P}(X = Y)$, and $\mathbb{P}(I = J)$ implicitly depend on the number of proposals $N$ in the sampled ensemble. However, we do not notationally emphasize this dependency for the sake of readibility.

Similarly to Proposition 1, if we now define $\tau$ to be the number of times we sample $N$ i.i.d. variables from $\hat{\Gamma}$, we have the following result.

**Proposition 3.** *Suppose that $\hat{\Gamma} \succeq p \otimes q$. Then, for any $N > 0$, the random variables $X$ and $Y$ generated by Algorithm 2 are distributed according to $p$ and $q$, respectively. Furthermore, the mean and variance of the random number of steps $\tau$ of Algorithm 2 verify*

$$
\mathbb{E}[\tau] \leq \frac{N + \min\left(M(p,\hat{p}), M(q,\hat{q})\right) - 1}{N},
$$
$$
\mathbb{V}[\tau] \leq \left(\frac{N + \min\left(M(p,\hat{p}), M(q,\hat{q})\right) - 1}{N}\right)^2 - \frac{N + \min\left(M(p,\hat{p}), M(q,\hat{q})\right) - 1}{N}.
$$

(5)

---

**Algorithm 2:** Ensemble rejection coupling of $(p, q)$

---

**1 Function** EnsembleRejectionCoupling($\hat{\Gamma}$, $p$, $q$, $N$):

 // Operations involving the index $i$ need to be done $N$ times

**2**   Set $A_X = 0$ and $A_Y = 0$ // Acceptance flags

**3**   **while** $A_X = 0$ *and* $A_Y = 0$ **do**

**4**     Sample $\hat{X}_i, \hat{Y}_i \sim \hat{\Gamma}$, $U \sim \mathcal{U}(0, 1)$

**5**     Set $w_i^X = \frac{p(\hat{X}_i)}{\hat{p}(\hat{X}_i)}$, $w_i^Y = \frac{q(\hat{Y}_i)}{\hat{p}(\hat{Y}_i)}$, $\hat{Z}_X = \frac{1}{N}\sum_{n=1}^{N} w_n^X$, and $\hat{Z}_Y = \frac{1}{N}\sum_{n=1}^{N} w_n^Y$

**6**     Set $W_i^X = \frac{w_i^X}{N\hat{Z}_X}$ and $W_i^Y = \frac{w_i^Y}{N\hat{Z}_Y}$

**7**     Sample $(I, J)$ from a maximal coupling of $\mathrm{Cat}((W_i^X)_{i=1}^N)$ and $\mathrm{Cat}((W_i^Y)_{i=1}^N)$

**8**     Set $\bar{Z}_X = \hat{Z}_X + \frac{1}{N}(M(p, \hat{p}) - w_I^X)$ and $\bar{Z}_Y = \hat{Z}_Y + \frac{1}{N}(M(q, \hat{q}) - w_J^Y)$

**9**     Set $X_1 = \hat{X}_I$ and $Y_1 = \hat{Y}_J$

**10**    **if** $U < \frac{\hat{Z}_X}{\bar{Z}_X}$ **then** set $A_X = 1$

**11**    **if** $U < \frac{\hat{Z}_Y}{\bar{Z}_Y}$ **then** set $A_Y = 1$

**12**   Sample $X_2, Y_2$ from $p \otimes q$

   **Return:** $X = A_X X_1 + (1 - A_X) X_2$, $Y = A_Y Y_1 + (1 - A_Y) Y_2$

---

*Proof.* As outlined in the proof of Proposition 1 in Deligiannidis et al. (2020), the ensemble rejection sampler is a rejection sampler on an extended space. As a consequence, because marginally both $X$ and $Y$ in Algorithm 2 are sampled from an ensemble rejection sampler, the analysis in the proof of Proposition 1 still holds. The marginals are therefore preserved, and the bounds (5) follow from replacing $p_{\mathrm{RS}}$ by $p_{\mathrm{ERS}}$. $\square$

Moreover Algorithm 2 preserves the diagonal property as well, and, in fact, contrarily to Algorithm 1, the coupling success probability of Algorithm 2 can be quantified, and we can show that it asymptotically recovers the maximal coupling probability. This result is given in the following proposition and subsequent corollary.

**Proposition 4.** *Let $(\hat{X}, \hat{Y}) \sim \hat{\Gamma}$ be distributed according to the dominating coupling. Suppose that the hypotheses of Proposition 2 are verified, then there exists a deterministic sequence $l_N$, such that for all $N$ we have $0 < l_N \leq \mathbb{P}(X = Y)$ and*

$$l_N \sim \mathbb{P}(\hat{X} = \hat{Y}) \cdot \frac{\int \min(p(x), q(x))\, \mathrm{d}x}{1 + u \cdot \frac{\sqrt{2N \log\log N}}{N}} \cdot \frac{N}{N - 1 + \frac{1}{p_X}} \cdot \frac{N}{N - 1 + \frac{1}{p_Y}}, \tag{6}$$

*where $u = \max\left(\mathbb{V}\left[\frac{p(\hat{X})}{\hat{p}(\hat{X})}\right]^{1/2}, \mathbb{V}\left[\frac{q(\hat{Y})}{\hat{q}(\hat{Y})}\right]^{1/2}\right)$, and*

$$p_X = \mathbb{P}\left(\frac{p(\hat{X})}{M(p, \hat{p})\hat{p}(\hat{X})} \mid \hat{X} = \hat{Y}\right) > 0, \quad p_Y = \mathbb{P}\left(\frac{q(\hat{Y})}{M(q, \hat{q})\hat{q}(\hat{X})} \mid \hat{X} = \hat{Y}\right) > 0 \tag{7}$$

*are the conditional acceptance probabilities of the individual rejection samplers. Additionally, when the distributions $p$ and $q$ do not have atoms, there exists a sequence $u_N$, such that for all $N$, $\mathbb{P}(X = Y) \leq u_N$ and*

$$u_N \sim \mathbb{P}(\hat{X} = \hat{Y}) \cdot \frac{\int \max(p(x), q(x))\, \mathrm{d}x}{1 - u \cdot \frac{\sqrt{2N \log\log N}}{N}}. \tag{8}$$

*Finally, when the coupling $\hat{\Gamma}$ is maximal, we have $p_X \geq \frac{\mathbb{E}[p(\hat{Y})]}{\mathbb{P}(\hat{X} = \hat{Y}) M(p, \hat{p})}$, and similarly, $p_Y \geq \frac{\mathbb{E}[q(\hat{X})]}{\mathbb{P}(\hat{X} = \hat{Y}) M(q, \hat{q})}$, with the identity $p_X = \frac{1}{M(p, \hat{p})}$ and $p_Y = \frac{1}{M(q, \hat{q})}$ when we also have $\hat{p} = \hat{q}$.*

*Proof.* The proof of Proposition 4, relying on the law of iterated logarithm, can be found in Appendix B.
□

Proposition 4 implies the following corollary.

**Corollary.** *Suppose that the hypotheses of Proposition 2 are verified and that the distributions p and q do not have atoms. Using the same notations, we have*

$$\mathbb{P}(X = Y) \xrightarrow[N\to\infty]{} \mathbb{P}(\hat{X} = \hat{Y}) \int \min(p(x), q(x)) \, \mathrm{d}x \,. \tag{9}$$

**Remark 2.** *At first sight, the result of Proposition 4 might seem disappointing. Indeed, we are proving that by throwing an infinite amount of computing power at the coupling algorithm, we can lower bound the coupling success probability by a quantity lower than the maximal one. However, we need to remember that the choice of $\hat{\Gamma}$ is in fact arbitrary, and that we can often choose it in such a way that it increases $\mathbb{P}(\hat{X} = \hat{Y})$. Furthermore, we can also use a common proposal $\gamma$ for p and q, in which case using the dominating coupling $(X, Y)$, where we have $X \sim \gamma$ and $Y = X$ results in $\mathbb{P}(\hat{X} = \hat{Y}) = 1$, and therefore the asymptotics of Proposition 3 recover the maximal coupling probability.*

Proposition 4 also provides a clear rule of thumb on how to choose $N$ to achieve a desired coupling probability. For simplicity, suppose that we are using the dominating coupling $\hat{\Gamma}$ corresponding to the duplicated distribution $\gamma$. In this case, in order to obtain a coupling success of $\alpha \cdot \int \min(p(x), q(x)) \, \mathrm{d}x$, $0 < \alpha < 1$, we can simply solve $N$ for

$$\alpha = \frac{1}{1 + u \cdot \frac{\sqrt{2N \log \log N}}{N}} \Leftrightarrow u \cdot \frac{\sqrt{2N \log \log N}}{N} = \frac{1}{\alpha} - 1, \tag{10}$$

which approximately corresponds to $N \approx 2 \left( \frac{u}{\frac{1}{\alpha} - 1} \right)^2$.

A natural critique that may be raised when using ensemble methods is its computational cost. In fact, in Deligiannidis et al. (2020), the authors argue that using an ensemble proposal to do rejection sampling is less efficient than running a classical rejection sampler for longer. However, we discuss in Appendix J that this may not be the case when the algorithm is run on parallel hardware, at least for small enough acceptance probabilities. The theoretical analysis done in Appendix J was also confirmed empirically in Section 5.2 (see Figure 3).

## 4 The multidimensional Gaussian case

In Sections 2 and 3, we provided a general framework for sampling from diagonal couplings of arbitrary probability distributions by using a dominating coupling $\hat{\Gamma}$ as a proposal. We now focus on the important special case of coupling $d$-dimensional Gaussian distributions with different means and covariances $\mathcal{N}(\mu_p, \Sigma_p)$ and $\mathcal{N}(\mu_q, \Sigma_q)$.

The problem of coupling Gaussian distributions has appeared in (Eberle et al., 2019; Bou-Rabee et al., 2020) and efficient solutions are known for the case when $\Sigma_p = \Sigma_q$. When it is not the case, one can use Thorisson's algorithm (Thorisson, 2000), or its finite variance version (Gerber and Lee, 2021) in order to sample from a (sub-)maximal coupling of these. This last solution, however, presents several inconveniences: first, as already discussed, it has a run time variance that increases as the marginals get closer. While this may be acceptable on distributed systems (such as the one considered in Jacob et al. (2020b)), the context of parallel hardware, which we mainly consider here, requires all parallel executions to finish to proceed. Second, when used with a MCMC algorithm, Thorisson's algorithm does not result in strong contractive properties (Wang et al., 2021), thereby increasing the total run time of the method, as well as the variance of the estimate. The goal of this section is therefore to provide a fast and efficient way to generalize the reflection approach of Bou-Rabee et al. (2020); Eberle

et al. (2019) to Gaussians with different covariance so as to reduce the variance of the run time, and preserve the efficient geometric properties that they induce.

Our aim is now to use Algorithm 1 to sample from a coupling of $\mathcal{N}(\mu_p, \Sigma_p)$ and $\mathcal{N}(\mu_q, \Sigma_q)$, with different means and covariances, by using a maximal coupling between Gaussians sharing the same covariance matrix (Bou-Rabee et al., 2020) as the proposal $\hat{\Gamma}$. For the sake of completeness, we reproduce the method of Bou-Rabee et al. (2020), called "reflection-maximal" coupling in Appendix D. A natural candidate for a diagonal dominating coupling of $\mathcal{N}(\mu_p, \Sigma_p)$ and $\mathcal{N}(\mu_q, \Sigma_q)$ now consists in a reflection-maximal coupling of $\mathcal{N}(\mu_p, \hat{\Sigma})$ and $\mathcal{N}(\mu_q, \hat{\Sigma})$. For this coupling to be dominating, the covariance matrix $\hat{\Sigma}$ needs to verify $\hat{\Sigma}^{-1} \preceq \Sigma_p^{-1}$ and that $\hat{\Sigma}^{-1} \preceq \Sigma_q^{-1}$ in the sense of Loewner ordering. This ensures that we have $\mathcal{N}(x; \mu_p, \Sigma_p) \leq \frac{\det(2\pi \hat{\Sigma})^{1/2}}{\det(2\pi \Sigma_p)^{1/2}} \mathcal{N}(x; \mu_p, \hat{\Sigma})$ and $\mathcal{N}(x; \mu_q, \Sigma_q) \leq \frac{\det(2\pi \hat{\Sigma})^{1/2}}{\det(2\pi \Sigma_q)^{1/2}} \mathcal{N}(x; \mu_q, \hat{\Sigma})$.

The resulting algorithm now preserves diagonality of the coupling, and we have explicit upper bounds for the mean and variance of the number of steps $\tau$ .

**Proposition 5.** *Let $\hat{\Sigma}$ be a covariance matrix with $\hat{\Sigma}^{-1} \preceq \Sigma_p^{-1}$ and $\hat{\Sigma}^{-1} \preceq \Sigma_q^{-1}$, and let $\hat{\Gamma}$ be a reflection-maximal coupling of $\mathcal{N}(\mu_p, \hat{\Sigma})$ and $\mathcal{N}(\mu_q, \hat{\Sigma})$. In this case, we have $M(p, \hat{p}) = \frac{\det(2\pi \hat{\Sigma})^{1/2}}{\det(2\pi \Sigma_p)^{1/2}}$ and $M(q, \hat{q}) = \frac{\det(2\pi \hat{\Sigma})^{1/2}}{\det(2\pi \Sigma_q)^{1/2}}$ and Algorithm 1 for $\hat{\Gamma}$, $\mathcal{N}(\mu_p, \Sigma_p)$, and $\mathcal{N}(\mu_q, \Sigma_q)$ returns a diagonal coupling of $\mathcal{N}(\mu_p, \Sigma_p)$.*

*Proof.* This is a direct consequence of Proposition 1 applied to Gaussian distributions. $\square$

**Remark 3.** *When $\Sigma_p = \Sigma_q$, one can take $\hat{\Sigma} = \Sigma_p = \Sigma_q$, so that the coupling proposal is always accepted and the resulting algorithm recovers the reflection-maximal coupling.*

A covariance matrix $\hat{\Sigma}$ with the properties $\hat{\Sigma}^{-1} \preceq \Sigma_p^{-1}$ and $\hat{\Sigma}^{-1} \preceq \Sigma_q^{-1}$ always exists, as one can take, for example, $\hat{\Sigma}_{\max} = \max([\sigma(\Sigma_p), \sigma(\Sigma_q)]) I_d$, where $\sigma(M)$ is the spectrum of the matrix $M$. Given a chosen matrix $\hat{\Sigma}$, we could, in theory, compute the probability of a successful coupling for Algorithm 1, but the expression is computationally intractable. Therefore, we resort to finding lower and upper bounds for the probability $\mathbb{P}(X = Y)$.

**Proposition 6.** *Under the hypotheses of Propositions 5, we have*

$$0 < \frac{\det(2\pi H)^{1/2}}{\det\left(2\pi\hat{\Sigma}\right)^{1/2}} \left[\exp(-\beta/2)\mathcal{F}(\alpha, H) + \exp(-\gamma/2)\left[1 - \mathcal{F}(\delta, H)\right]\right] \tag{11}$$

$$\leq \mathbb{P}(X = Y) \leq 2F\left(-\frac{1}{2}\left\|\hat{\Sigma}^{-1/2}(\mu_p - \mu_q)\right\|_2\right), \tag{12}$$

*where*

$$H = \left(\Sigma_p^{-1} + \Sigma_q^{-1} - \hat{\Sigma}^{-1}\right)^{-1},$$

$$\alpha = H\left(\Sigma_p^{-1}\mu_p + \left(\Sigma_q^{-1} - \hat{\Sigma}^{-1}\right)\mu_q\right), \quad \beta = \mu_p^\top \Sigma_p^{-1}\mu_p + \mu_q^\top\left(\Sigma_q^{-1} - \hat{\Sigma}^{-1}\right)\mu_q - \alpha^\top H^{-1}\alpha, \tag{13}$$

$$\delta = H\left(\Sigma_q^{-1}\mu_q + \left(\Sigma_p^{-1} - \hat{\Sigma}^{-1}\right)\mu_p\right), \quad \gamma = \mu_q^\top \Sigma_q^{-1}\mu_q + \mu_p^\top\left(\Sigma_p^{-1} - \hat{\Sigma}^{-1}\right)\mu_p - \delta^\top H^{-1}\delta,$$

*along with*

$$\mathcal{F}(u, V) := F\left(\frac{1}{2}\frac{\mu_p^\top \hat{\Sigma}^{-1}\mu_p - \mu_q^\top \hat{\Sigma}^{-1}\mu_q - 2u^\top \hat{\Sigma}^{-1}(\mu_p - \mu_q)}{\left\|\left[V^{1/2}\right]^\top \hat{\Sigma}^{-1}(\mu_p - \mu_q)\right\|_2}\right), \tag{14}$$

*where $F$ is the cumulative distribution function of a standard normal random variable.*

*Proof.* The right hand side directly follows from the fact that the rejection coupling is at best as successful as the dominating diagonal coupling used, which in this case is maximal with success probability given by $2F\left(-\frac{1}{2}\left\|\hat{\Sigma}^{-1/2}(\mu_p - \mu_q)\right\|_2\right)$. The proof of the left hand side can be found in Appendix C. $\quad\square$

**Remark 4.** *The bound in Proposition 6 is tight in the sense that the lower and upper bounds coincide when $\Sigma_p = \Sigma_q = \hat{\Sigma}$. In that case, $H = P$, $\alpha = \mu_p$, $\beta = 0$, $\delta = \mu_q$, and $\gamma = 0$, resulting in*

$$\mathcal{F}(\alpha, H) = F\left(-\frac{1}{2}\left\|\hat{\Sigma}^{-1/2}(\mu_p - \mu_q)\right\|_2\right), \quad \mathcal{F}(\beta, H) = 1 - F\left(-\frac{1}{2}\left\|\hat{\Sigma}^{-1/2}(\mu_p - \mu_q)\right\|_2\right), \qquad (15)$$

*which give* (12) *when substituted into* (11).

As a direct corollary, Proposition 6 provides a new class of upper bounds for the total variation distance between two Gaussians via (1). These upper bounds are always less than 1. It is however not clear how they compare to already known bounds (see, e.g., Devroye et al. (2018) for a recent arXiv report on the subject), in particular in the high dimensional regime, and comparing these is left for future work.

An important aspect of the algorithm is the suitable selection of $\hat{\Sigma}$ so that the coupling probability is as high as possible. Although $\hat{\Sigma}_{\max} = \max([\sigma(\Sigma_p), \sigma(\Sigma_q)]) I_d$ is one possible choice for $\hat{\Sigma}$, it might not be a good choice in practice as it may result in a low acceptance probability. Another possible option would be to optimize the lower bound in (11) with respect to $\hat{\Sigma}$, which unfortunately leads to a computationally heavy (and likely non-convex) optimisation problem. However, a suitable criterion for maximising the chance of coupling is to simply maximize $\mathbb{P}(A_X = 1)\,\mathbb{P}(A_Y = 1)$. As functions of $\hat{\Sigma}^{-1}$ the logarithms of marginal acceptance probabilities are both of form $\frac{1}{2}\log\det\left(\hat{\Sigma}^{-1}\right) + \text{constant}$, and therefore a suitable optimisation problem for $\hat{\Sigma}^{-1}$ is

$$\begin{aligned}\max & \log\det\left(\hat{\Sigma}^{-1}\right) \\ & \hat{\Sigma}^{-1} \preceq \Sigma_p^{-1}, \quad \hat{\Sigma}^{-1} \preceq \Sigma_q^{-1}, \quad \hat{\Sigma}^{-1} \succeq 0.\end{aligned} \qquad (16)$$

**Proposition 7.** *A solution to* (16) *is given by*

$$\hat{\Sigma}_{opt} = C\,V\,U\,V^\top C^\top \qquad (17)$$

*where $C = \Sigma_q^{1/2}$, $V\,D\,V^\top = C^\top \Sigma_p^{-1} C$ is such that $V$ is orthonormal and $D$ is diagonal, and $U$ is diagonal with $U_{ii} = 1/\min(1, D_{ii})$ for all $i$.*

*Proof.* The proof of Proposition 7 can be found in Appendix E $\qquad\square$

The most expensive operation involved in computing the solution covariance matrix given by Proposition 7 is a spectral decomposition, which typically has the same complexity as computing a Cholesky decomposition. As a consequence, computing $\hat{\Sigma}_{opt}$ only incurs a constant additional cost compared to using $\hat{\Sigma}_{max}$. The benefit of optimising for $\hat{\Sigma}$ as compared to simply using $\hat{\Sigma}_{max}$ is illustrated in Appendix F.

## 5  Applications

To illustrate the scope and attractiveness of our proposed method, we apply it on a number of examples covering extreme events sampling, particle filtering, Gibbs, and manifold MCMC methods. The code used to generate the examples in this section can be found at the following **address**. For consistency, all the examples in this Section were run using a Nvidia GeForce GTX 1080 Ti graphics processing unit with 3584 cores and 11 GB of memory.

## 5.1 Coupling tails of Gaussian distributions

The first illustrative example consists in coupled sampling from the tails of Gaussian distributions. This sampling is known to be challenging even without the coupling (Robert, 1995). In particular, let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ be Gaussian random variables and let $p_\mu(x)$ and $q_\eta(y)$ be the densities corresponding to random variables $X_\mu = X \mid X > \mu$ and $Y_\eta = Y \mid Y > \eta$, respectively. When $\mu$ and $\eta$ are small enough, we can sample from $p$ and $q$ by computing the corresponding inverse CDF formulas. However, when $\mu$ and $\eta$ are large ( $\gtrsim 5$) this method becomes unreliable and rejection sampling methods are needed (Robert, 1995). Sampling from a maximal coupling of $p_\mu$ and $q_\eta$ in this case would in fact be even harder than sampling from the marginals independently. Fortunately, we can use our methodology to construct a coupled version of the Gaussian tail rejection sampling algorithm in Robert (1995), by constructing a maximal coupling of translated exponentials as the dominating coupling.

Robert (1995) uses translated exponential distributions $\text{Exp}(x; m, \lambda) = \lambda e^{-\lambda(x-m)} \mathbb{1}_{x \geq m}$ as the proposal distributions for rejection sampling from the Gaussian tails. Let us now suppose that $\eta > \mu$. As per Robert (1995), the optimal exponential proposals for $p(x)$ and $q(y)$ are given by $\hat{p}(x) = \text{Exp}(x; \mu, \alpha(\mu))$ and $\hat{q}(x) = \text{Exp}(x; \eta, \alpha(\eta))$, where $\alpha(z) = \frac{z + \sqrt{z^2+4}}{2}$ for all $z$ and hence we have

$$\frac{p(x)}{M(p, \hat{p})\hat{p}(x)} = e^{-\frac{1}{2}(x - \alpha(\mu))^2}, \quad \frac{q(x)}{M(q, \hat{q})\hat{q}(x)} = e^{-\frac{1}{2}(x - \alpha(\eta))^2}. \tag{18}$$

Sampling from a maximal coupling of $\hat{p}$ and $\hat{q}$ can then be done explicitly by using the following algorithm (see, e.g, (Thorisson, 2000; Lindvall, 2002)):

1. Draw $u \sim \mathcal{U}(0,1)$, then

   (a) If $u \leq \int \min(\hat{p}(x), \hat{q}(x))\, dx$: sample $x$ from $c(x) \propto \min(\hat{p}(x), \hat{q}(x))$ and put $y = x$.

   (b) Else sample $x$ from $\tilde{p}(x) \propto \hat{p}(x) - \min(\hat{p}(x), \hat{q}(x))$ and $y$ from $\tilde{q}(y) \propto \hat{q}(y) - \min(\hat{p}(y), \hat{q}(y))$.

All the probabilities and probability densities above can be evaluated in closed form and the samplings can be done with inverse CDF method. The only difficult parts are the distributions $\tilde{p}(x)$ and $\tilde{q}(y)$ whose inverse CDFs are not available in closed form. However, sampling from these distributions can be implemented with a fast line search method which numerically solves the values of the inverse CDF, see Appendix G for more details on the implementation.

The above method along with the ratios defined in (18) allow us to use Algorithm 1 or 2 to sample from a diagonal coupling of Gaussian tails without ever having to compute the CDF or the inverse CDF, as per the motivation in Robert (1995). In order to assess the performance of the resulting coupled rejection-sampler, we use Algorithm 1 to sample $100\,000$ from a coupling of $p_6$ and $q_\eta$ for different values of $\eta \geq 6$ using the dominating coupling given by the translated exponentials described in (18).

In Figure 1a we report both the maximal coupling probability of $p_6$ and $q_\eta$, as computed numerically by $\int \min(p_6, q_\eta)$, as well as the empirical coupling probability recovered by our coupled rejection-sampler. In Figure 1b we also report the resulting run time for different values of the truncation constants. As shown by Figure 1a, for this specific example, the coupling probability of the samples obtained by Algorithm 1 is essentially indistinguishable from its theoretical maximum. Furthermore, Figure 1b shows that the resulting run time of the method is well-behaved, even as $\eta \to 6$, which in the case of using Thorisson's algorithm would have resulted in a run time with infinite variance (see Section 1).

It is worth noting that even in the case when we are not worried about the variance of the run time when $\eta \to 6$, using Thorisson's algorithm would have required sampling from the marginals $p_6$ and $q_\eta$, which would have required running a rejection sampling algorithm anyway and would have required computing the ratio of the truncated normal PDFs.

10
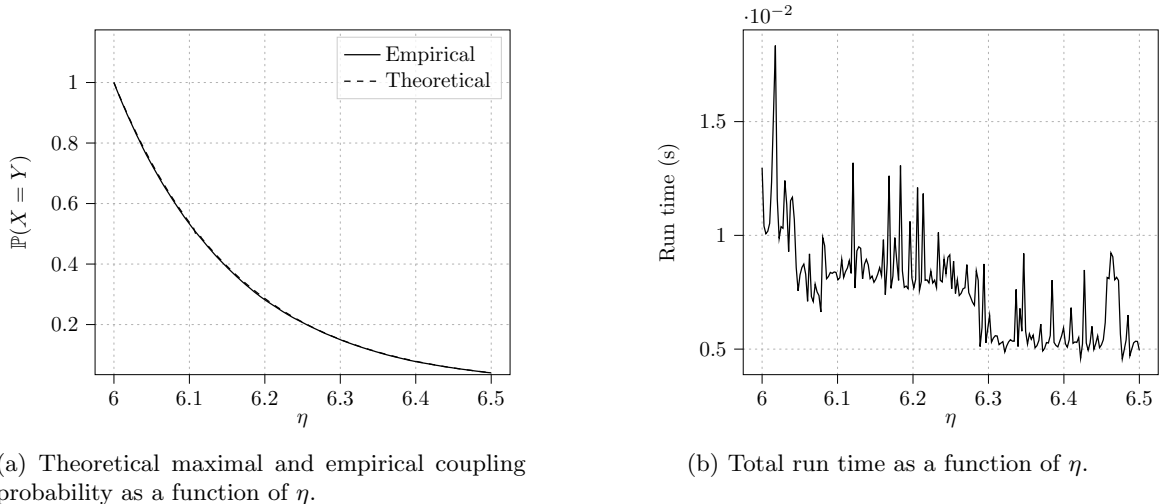
(a) Theoretical maximal and empirical coupling probability as a function of $\eta$.



(b) Total run time as a function of $\eta$.

Figure 1: Coupling statistics 1a and run time 1b (computed over $100\,000$ samples) of using Algorithm 1 with shifted exponential proposals to sample a coupling of Gaussian tails $p_6$ and $q_\eta$. The average coupling probability observed $\mathbb{P}(X = Y)$ is almost indistinguishable from its theoretical maximal for all values of $\eta$. On the other hand, the total run time for sampling from Algorithm 1 stays limited as $\eta \to 6$.

## 5.2 Coupled parallel resampling

Resampling is a key ingredient in sequential Monte Carlo (SMC) methods (for a recent review, see, e.g., Chopin and Papaspiliopoulos, 2020) as it allows to avoid path degeneracy by forgetting unsuccessful trajectories. Formally, given an empirical distribution with unnormalized weights $(w_i^Z)_{i=1}^M$ and locations $(Z_i)_{i=1}^M$, resampling methods typically form a random empirical distribution with uniform weights $\frac{1}{M}$ and locations $(Z_{B_i^Z})_{i=1}^N$, where $\mathbb{P}(B_i^Z = j) \propto w_j^Z$ for all $i, j = 1, \ldots, M$.

While most operations in SMC are trivially parallelisable, standard techniques for resampling, such as multinomial, systematic, or stratified resampling Carpenter et al. (1999) require communication and prevent full parallelisation of the method. In Murray et al. (2016), however, the authors show how the resampling operation can be performed by a tweaked parallel rejection sampling algorithm provided that one knows an upper bound to the unnormalized weights $(w_i^Z)_{i=1}^M$.

On the other hand, Chopin and Singh (2015) discuss how we can construct a maximal coupling of multinomial distributions by first computing the overlap term $\int \min(p, q)$ and then sampling from a mixture representation of the maximal coupling (this coupling, also used in Jacob et al. (2020a), is however much older and can found at least in Lindvall (2002); Thorisson (2000)). While sampling from the maximal coupling – once it has been formed – can be then done in parallel using Murray et al. (2016), computing the mass overlap term $\int \min(p, q)$ breaks the parallelisation obtained via Murray et al. (2016) and therefore the algorithm needs to be modified to retain parallelisation.

Our Algorithms 1 and 2 can readily be adapted to provide a parallel coupled version of the rejection sampler in Murray et al. (2016)[1] and we summarize the modification in Appendix H. Moreover, using the ensemble version also helps with reducing both the expectation and variance of the individual rejection sampler run time. Because in the parallel context variable length tasks need to be synchronized (Murray et al., 2016), using an ensemble rejection method should therefore improve the effective performance of the resulting algorithm. This, however, also comes at the cost of additional computation on $M \ll N$ auxiliary weights, thereby resulting in an increased memory consumption as well as a total parallel computational complexity of $O(\log(M))$ instead of $O(1)$.

---

[1]The only subtlety residing in the fact that the initial proposal in Murray et al. (2016) is deterministic.

In order to illustrate our method on this problem, we consider a similar setting as in Murray et al. (2016). Let $M = 2^{14}$, and $x_i$ and $z_i$ be identically distributed samples from $\mathcal{N}(0,1)$, for $i = 1, \ldots, M$. For $y = 0, 1, 2, 3$, we then take the weight function to be $\omega_y(a) = \mathcal{N}(y - a; 0, 1)$, forming two unnormalized samples $w_{i,y}^X = \omega_y(x_i)$ and $w_{i,y}^Z = \omega_y(z_i)$ which are both upper bounded by $1/\sqrt{2\pi}$. The experiment consists of applying the coupled rejection resampling algorithm 100 times for the two fixed sets of normally distributed particles and the different values of $y$ and $N$.
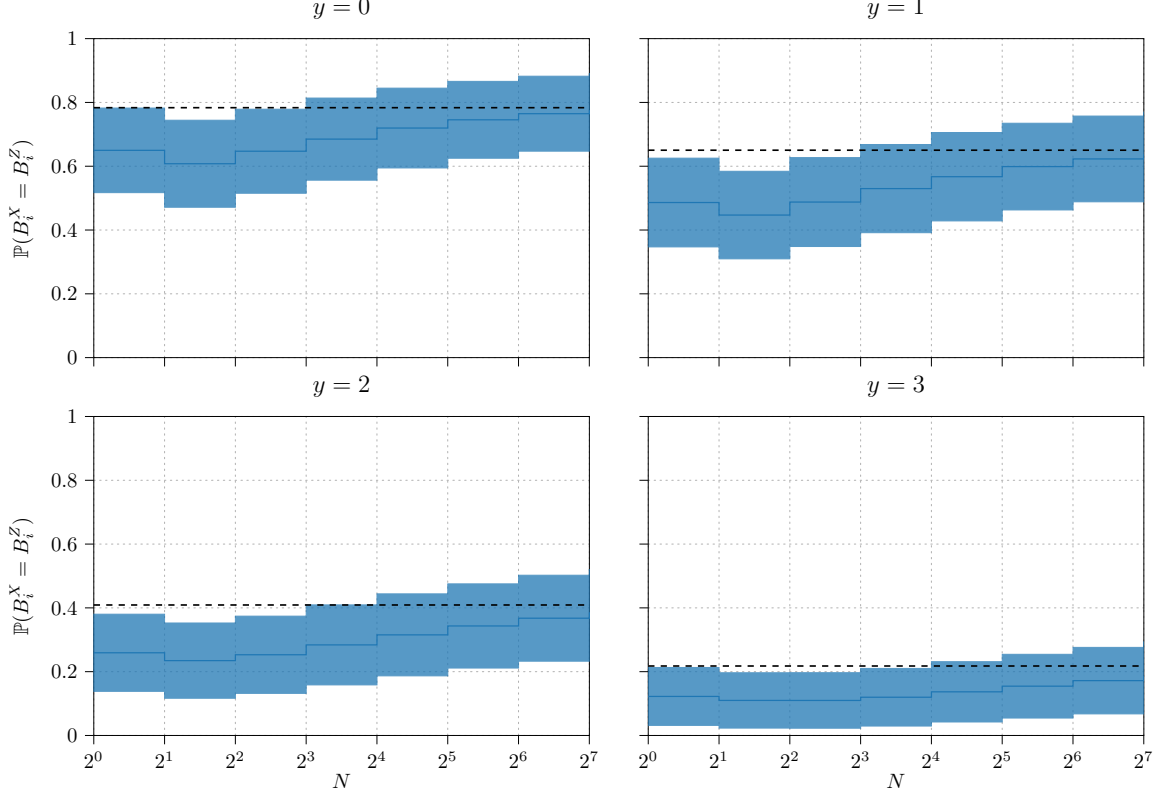


Figure 2: Coupling probability $\mathbb{P}(B_i^X = B_i^Z)$ for the parallel rejection resampling in Section 5.2 as a function of the number of proposals $N$ in Algorithm 2. The blue full line and shaded area correspond to the average coupling probability of the resampling experiments, as well as the 95% confidence interval of these. The dashed line corresponds to the theoretical maximal coupling probability.

In Figure 2 and Figure 3 we respectively report the empirical coupling success $\mathbb{P}(B_i^X = B_i^Z)$ and the run time of running Appendix 4 as $y$ changes (and therefore the variance of the weights increases) for different ensemble proposal sizes $N$. In Figure 2, the empirical coupling probability is compared with the theoretical maximal coupling probability as given by $\sum_i \min(W_{i,y}^X, W_{i,y}^Z)$, where $W_{i,y}^X = \frac{w_{i,y}^X}{\sum_{j=1}^M w_{j,y}^X}$ and $W_{i,y}^Z \frac{w_{i,y}^Z}{\sum_{j=1}^M w_{j,y}^Z}$ are the self-normalized versions of the weights $w_{i,y}^X$ and $w_{i,y}^Z$. The experiment was run 100 times.

As can be seen in Figure 2, the empirical coupling probability seems to converge to the theoretical one for the large number of proposals $N$ regime. This behavior is in agreement with the asymptotics of Proposition 4 as the dominating coupling is here given by the trivial coupling $\hat{X} = \hat{Y} \sim \mathcal{U}(\{1, \ldots, N\})$, so that Remark 2 applies. As shown by the average run time reported in Figure 3 increasing the ensemble size in Algorithm 2, at the same time, lowers the total run time of the resampling routine for higher weights variance, corresponding to $y = 2$ and $y = 3$. For a discussion on the computational
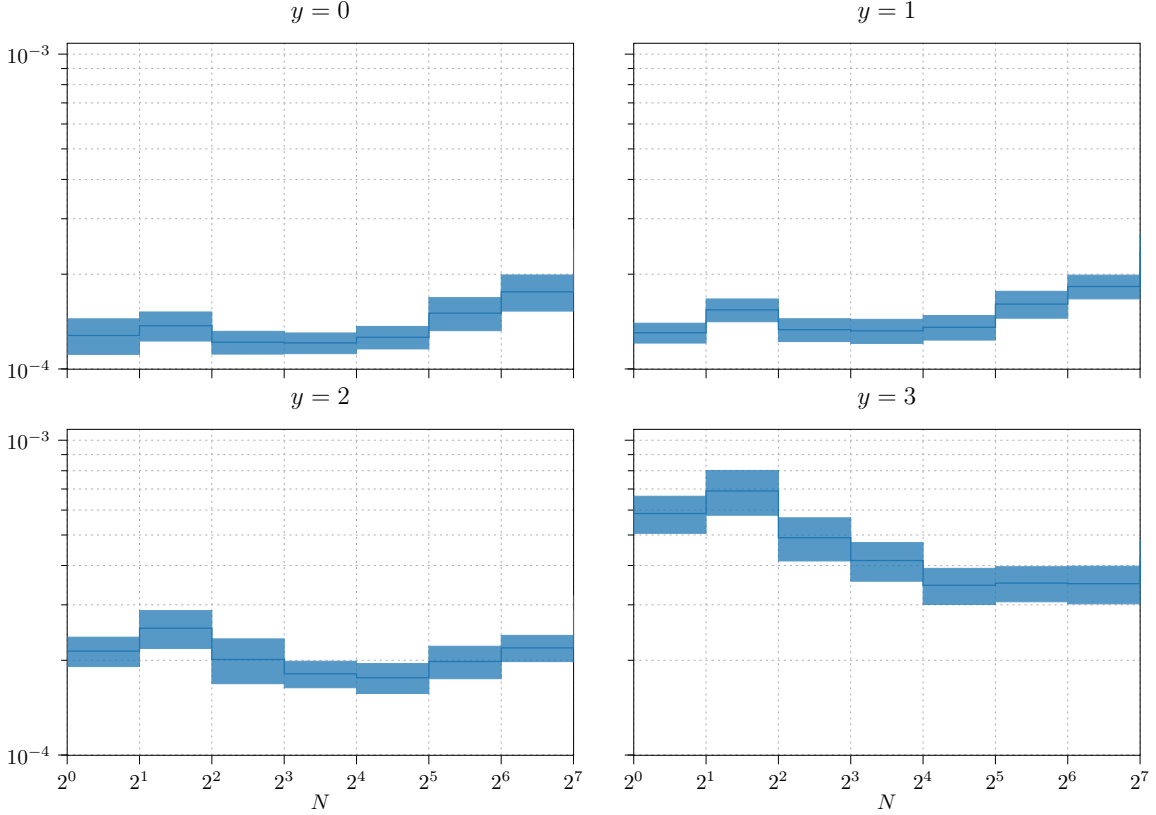
Figure 3: Run time (in seconds) for the parallel rejection resampling in Section 5.2 as a function of the number of proposals $N$ in Algorithm 2. The blue full line and shaded area correspond to the average run time of the resampling experiments, as well as its 95% confidence interval.

benefits of using ensemble proposals in the GPU context we refer the reader to Appendix J.

The run time results reported in Figure 3 suggest that using an ensemble rejection method in the parallel context may lead to broader applicability of parallel rejection resampling than the recommendation in Section 3.4 of Murray et al. (2016), in particular in the high weight variance regime. More precisely, when the resampling routine is used as part of a particle filtering loop, it is highly likely that the number of threads necessary to compute the weights is fairly large. This means that, once these have been computed, the number of threads available at resampling time is typically higher than the total number of particles. Allowing to improve the run time by leveraging the idle compute capacity therefore seems valuable in most cases, and particularly so in the coupling context, due to its additional benefit of increasing the coupling probability.

## 5.3 Coupled Gibbs sampling

Let us now consider the following density $p(x, y) \propto \exp\left(-\frac{1}{2}\left(x^\top x \cdot y^\top y + x^\top x + y^\top y\right)\right)$, $x, y \in \mathbb{R}^d$. We can express the conditional densities $p_{x|y}$ and $p_{y|x}$ in the following way:

$$p_{x|y}(x \mid y) = \mathcal{N}(x; 0, P_y), \quad P_y = \frac{1}{1 + y^\top y} I_d, \tag{19}$$

$$p_{y|x}(y \mid x) = \mathcal{N}(y; 0, P_x), \quad P_x = \frac{1}{1 + x^\top x} I_d. \tag{20}$$

13

This decomposition ensures that we can sample from $p(x, y)$ using Gibbs sampling (see, e.g., Robert and Casella, 2004, Ch. 4) by sampling from $p_{x|y}$ and $p_{y|x}$ alternatively. As described in Jacob et al. (2020b), to couple two resulting Gibbs chains, we only need to couple their conditional transition densities. That is, if we have two states $(x, y)$ and $(x', y')$, we need the ability to sample from a coupling of $p_{x|y}$ and $p_{x|y'}$ and from a coupling of $p_{y|x}$ and $p_{y|x'}$. Because the densities have different covariance matrices for different states, we cannot use the reflection-maximal coupling of Bou-Rabee et al. (2020) described Appendix D, but, instead, we need to resort to either Thorisson's algorithm Thorisson (2000) in its modified form Gerber and Lee (2021) (so as to have finite run time variance) or to our proposed method. For the sake of completeness, we reproduce the modified version of Thorisson's algorithm in Appendix I.
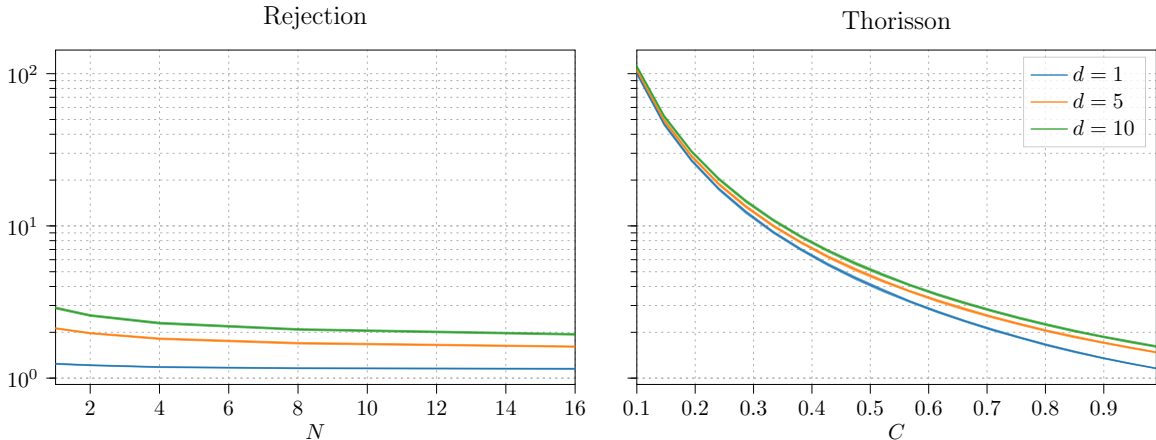


Figure 4: Average coupling time (computed over 50 experiments) of the coupled Gibbs chains for our rejection method and Thorisson's modified algorithm. The "Thorisson" x-axis represents the suboptimality parameter $C$ in the method (see Appendix I).

In Figure 4, we report the average meeting time Jacob et al. (2020b) of $10\,000$ coupled chains and, in Figure 5, the average run time of an experiment, as well of the 95% confidence of these (we repeated the experiment 50 times to measure the uncertainty in these quantities). Our results were generated by using the dominating covariance matrix $Q_{max}$, which in this case actually corresponds to the optimal matrix $Q_{opt}$.

We can see that our proposed method outperforms the modified version of Thorisson's algorithm in run time (Figure 5) for all sizes of ensembles and almost all values of $C$, on the other hand, Thorisson's algorithm results in slightly faster coupling times for $C \approx 1$, at the cost of a longer running algorithm than our method (due to the increase in variance in the run time). We believe, similarly to the analysis in Wang et al. (2021) that the mostly faster contraction of the Gibbs chains when using our method is explained by the fact that our coupling preserves the reflection properties the reflection-maximal coupling.

## 5.4   Coupled manifold Markov chain Monte Carlo

Adaptation for the proposal distribution (for a review see, e.g., Andrieu and Thoms, 2008) typically improves the convergence of MCMC algorithm by adapting the sampling mechanism to the target at hand, either globally, or locally. Coupling two adaptive or pre-conditioned algorithms however requires to be able to couple MCMC methods with different proposal distributions. In the literature, only two methods have been proposed to do this in the general case: coupling proposals using Thorisson's algorithm, or sampling from a maximal coupling of Metropolis–Hastings kernels (Wang et al., 2021).

We now consider the simplified manifold Metropolis-adjusted Langevin algorithm (mMALA) ap-
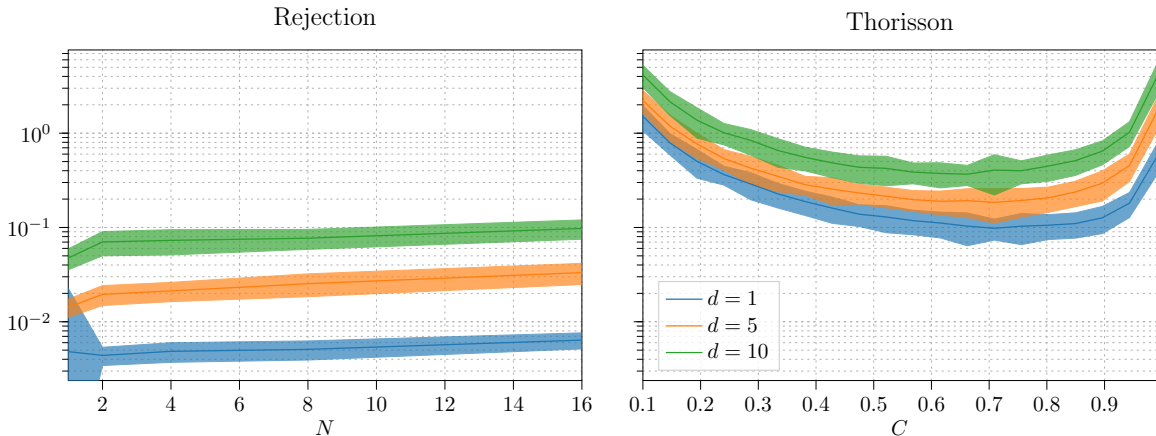
Figure 5: Average time in seconds (computed over 50 experiments) taken for the coupled Gibbs chains to meet for our rejection method and Thorisson's modified algorithm. The "Thorisson" x-axis represents the suboptimality parameter $C$ in the method (see Appendix I).

plied to the logistic regression example as described in Girolami and Calderhead (2011). Specifically, we consider the "heart disease" binary regression dataset (Detrano et al., 1989), and the same setup as Girolami and Calderhead (2011) apart from the starting point of the Markov chains, which we take to be randomly distributed according to $\mathcal{N}(0, 1/4^2)$. Because Riemannian manifold MCMC methods rely on sampling Gaussian distributions, the parameters of which depend on the Fisher information matrix computed at the current state of the chain, applying the reflection-maximal coupling method directly is impossible. This means that we can only hope to couple the two chains by using (i) Thorisson's (modified) algorithm to couple proposals, (ii) the "full-kernel coupling" with independent or reflected marginals in Wang et al. (2021, Algorithms 3 and 4), or by coupling the proposals using the coupled rejection sampler introduced in this article [2]. In accordance to Wang et al. (2021) and the rest of our experiments, using Thorisson's algorithm underperformed the other alternatives, so we do not report the results below.

We run two mMALA chains with coupled proposals, using either the "full-kernel coupling with reflection residuals" Metropolis-Hastings method of Wang et al. (2021, Algorithm 4), or the ensemble coupled rejection sampler with reflection-maximal proposals using the optimal covariance matrix described in Section 4.

|                   | $N$ | mean | std  | 5% | 50% | 95% |
|-------------------|-----|------|------|----|-----|-----|
| This paper        | 1   | 11.0 | 15.4 | 1  | 5   | 41  |
|                   | 4   | 7.7  | 10.8 | 1  | 4   | 28  |
|                   | 16  | 6.3  | 8.4  | 1  | 4   | 21  |
|                   | 64  | 5.8  | 7.7  | 1  | 3   | 19  |
| Wang et al. (2021)|     | 7.5  | 8.5  | 1  | 5   | 24  |

Table 1: Coupling time statistics of the coupled ensemble rejection sampler on the simplified manifold MALA algorithm of (Girolami and Calderhead, 2011) for different values of $N$ and Algorithm 4 in Wang et al. (2021). The percentages represent empirical quantiles of the distribution.

---

[2]It is also worth noting that the empirical evaluation of the different methods proposed in Wang et al. (2021) suffers from an implementation bug, and that, contrarily to what was reported in the article, Algorithm 4 results in faster coupling times than Algorithm 5 in high dimensional problems. This further justifies comparing with Algorithm 4 only.

|              | $N$ | mean    | std     | 5%      | 50%     | 95%     |
|--------------|-----|---------|---------|---------|---------|---------|
| This paper   | 1   | 9.3e-03 | 1.3e-02 | 1.0e-03 | 4.4e-03 | 3.5e-02 |
|              | 4   | 7.3e-03 | 1.0e-02 | 1.1e-03 | 3.8e-03 | 2.6e-02 |
|              | 16  | 6.1e-03 | 8.0e-03 | 1.1e-03 | 3.7e-03 | 2.0e-02 |
|              | 64  | 5.8e-03 | 7.4e-03 | 1.1e-03 | 3.5e-03 | 1.9e-02 |
| Wang et al. (2021) |  | 1.5e-02 | 1.9e-02 | 6.8e-04 | 8.7e-03 | 5.2e-02 |

Table 2: Run time statistics of the coupled ensemble rejection sampler on the simplified manifold MALA algorithm of (Girolami and Calderhead, 2011) for different values of $N$ and Algorithm 4 in Wang et al. (2021). The percentages represent empirical quantiles of the distribution.

In Table 1 and Table 2, we report the resulting meeting times and run times, respectively, over 100 000 experiments for the ensemble rejection sampler proposal coupling with different $N$'s as well as the "full-kernel coupling with reflection residuals" (Wang et al., 2021, Algorithm 4). On average, the coupled rejection sampler method runs faster and results in smaller coupling times Algorithm 4 in Wang et al. (2021) method as soon as $N > 4$. As we increase $N$, it also starts exhibiting smaller variance for both quantities, highlighting its attractiveness for implementing coupled chains in parallel environments. This improvement is the same phenomenon as the one discussed in Wang et al. (2021): using high quality proposals (when possible) may sometimes beat the full-kernel coupling, even when this one has reflected residuals.

# 6    Conclusion and discussion

In this article, we have introduced a novel coupled-rejection algorithm and an ensemble version of it for sampling from a target coupling using an acceptance-rejection scheme on samples from a dominating coupling. We have derived upper and lower bounds for their coupling success probabilities as well as the expectation and variance of their number of steps. In particular, our method provably preserves positive coupling probabilities of dominating coupling proposals and, in the ensemble case, asymptotically recovers the maximal coupling.

An important special case of our method consists in the multivariate Gaussian case where the algorithm allows to sample from coupled Gaussian distributions with different means and covariances with a low (finite) run time variance. This also allowed us to derive a family of upper bounds to the total variation distance between Gaussians. We believe that these bounds can be further improved by a more careful study of the quantities involved, possibly leading to even more useful bounds.

The proposed method is, however, not limited to the Gaussian case, and, in fact, if an extension of the reflection-maximal coupling of Bou-Rabee et al. (2020) to Gaussians with different means and covariances was derived, it could be used as a proposal in our algorithm. This would in turn simply expand and improve the applications of our proposed method for sampling diagonal couplings of sub-exponential distributions with high success probability.

We have also demonstrated the empirical usefulness of our method in coupling rare events, parallel resampling in particle filtering, Gibbs samplers, as well as Riemannian manifold MCMC methods. Possible future research include applications such as fully parallel unbiased particle smoothing (Jacob et al., 2020a), and possible extensions such as coupled empirical rejection sampling or coupling of state-space models for which the ensemble rejection sampling method was originally introduced (Deligiannidis et al., 2020). An interesting conclusion of our experimental evaluation, analysed in Appendix J, is that using ensemble proposals not only increases the coupling probability, as predicted by the theory, but also improves upon the total run time of the method when using parallel hardware such as a GPU. This is the consequence of two compounding factor: first, a computational improvement coming from reducing the variance in the run time of the rejection sampler (Section 5.2, Appendix J), second, a

decrease in the expected meeting time, reducing the total number of operations run before two chains meet (Sections 5.3 and 5.4).

Another direction of work is concerned with the special case of Gaussian distributions. In particular, finding computationally advantageous alternatives to $Q_{opt}$, or even computing a covariance matrix maximising the actual coupling probability instead of the surrogate objective introduced in Section 4 would result in more successful and computationally feasible coupling. Another question consists in computing tighter bounds to the coupling probability of our algorithm, automatically resulting in better upper bounds for the TV norm between Gaussians.

It is worth noting that while the focus of this article has been with implementing couplings distributions for use in standard MCMC algorithms, such as Metropolis–Hastings, or Gibbs sampling, Wang et al. (2021) introduced direct maximal couplings of Metropolis–Hastings type kernels. These have however been shown to be suboptimal in higher dimensions, probably due to the lack of contractive behavior of the resulting chains. This behavior was confirmed in Section 5.4, where the "full-kernel coupling with reflection residuals" of Wang et al. (2021) under-performed the simple Metropolis–Hastings using our coupled rejection sampling method, thereby hinting to the same conclusion that using good proposals is more important than optimality for coupling MCMC chains. To improve upon these, it could be possible to combine both approaches so as to increase the performance of maximal (or otherwise) couplings of MH kernels, for instance by coupling multiple-try MCMC methods (Liu et al., 2000).

Finally, our method can be used to propagate arbitrary couplings from a dominating pair of marginals to the target ones. For example, suboptimal Wasserstein-like couplings (for a review see, e.g., Villani, 2009) could be sampled efficiently from a dominating optimal transport coupling of Gaussian distributions. Similarly, couplings other than the maximal coupling can be used for sampling the indices in Algorithm 2. For example using optimal transport coupling of the indices based on a well-chosen metric may lead to faster converging algorithms at the cost of submaximality. Studying the theoretical and empirical properties of these different extensions is left for future works.

# Individual contributions

The original idea, implementation, and redaction of this article are all due to Adrien Corenflos. Simo Särkkä contributed Proposition 7, and helped with reviewing and finalising the manuscript. Adrien Corenflos and Simo Särkkä jointly developed the example of Section 5.1.

# Acknowledgements

# References

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive mcmc. *Statistics and computing*, 18(4):343–373.

Bou-Rabee, N., Eberle, A., and Zimmer, R. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 30(3):1209–1250.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, 146(1):2–7.

Chandrupatla, T. R. (1997). A new hybrid quadratic/bisection algorithm for finding the zero of a nonlinear function without using derivatives. *Advances in Engineering Software*, 28(3):145–149.

Chopin, N. and Papaspiliopoulos, O. (2020). *An Introduction to Sequential Monte Carlo*. Springer.

Chopin, N. and Singh, S. S. (2015). On particle Gibbs sampling. *Bernoulli*, 21(3):1855–1883.

Deligiannidis, G., Doucet, A., and Rubenthaler, S. (2020). Ensemble rejection sampling. *arXiv preprint arXiv:2001.09188*.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., and Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310.

Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*.

Eberle, A., Guillin, A., and Zimmer, R. (2019). Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010.

Gal, S., Rubinstein, R., and Ziv, A. (1984). On the optimality and efficiency of common random numbers. *Mathematics and computers in simulation*, 26(6):502–512.

Gerber, M. and Lee, A. (2021). Discussion on the paper "Unbiased Markov chain Monte Carlo methods with couplings" by Jacob, O'Leary and Atchadè. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):584–585.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Glynn, P. W. and Rhee, C.-h. (2014). Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389.

Hammersley, J. and Morton, K. (1956). A new Monte Carlo technique: antithetic variates. *Mathematical proceedings of the Cambridge philosophical society*, 52(3):449–475.

Heng, J. and Jacob, P. E. (2019). Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika*, 106(2):287–302.

Jacob, P. E. (2021). Couplings and Monte Carlo. Lecture notes of course STAT 248 at Harvard University.

Jacob, P. E., Lindsten, F., and Schön, T. B. (2020a). Smoothing with couplings of conditional particle filters. *Journal of the American Statistical Association*, 115(530):721–729.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020b). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600.

Johnson, V. E. (1998). A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93(441):238–248.

Lindvall, T. (2002). *Lectures on the Coupling Method*. Dover.

Liu, J. S., Liang, F., and Wong, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2019). Unbiased smoothing using particle independent Metropolis-Hastings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2378–2387. PMLR.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2020). Unbiased Markov chain Monte Carlo for intractable target distributions. *Electronic Journal of Statistics*, 14(2):2842–2891.

Murray, L. M., Lee, A., and Jacob, P. E. (2016). Parallel resampling in the particle filter. *Journal of Computational and Graphical Statistics*, 25(3):789–805.

Pagès, G. (2018). *Numerical Probability: An Introduction with Applications to Finance*. Springer International Publishing, Cham.

Reutter, A. and Johnson, V. E. (1995). General strategies for assessing convergence of MCMC algorithms using coupled sample paths. Technical report, Duke University.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and computing*, 5(2):121–125.

Robert, C. P. and Casella, G. (2004). *The Two-Stage Gibbs Sampler*, pages 337–370. Springer New York, New York, NY.

Shiryaev, A. N. (1996). *Probability*. Springer, 2nd edition.

Thorisson, H. (2000). *Coupling, Stationarity, and Regeneration*. Springer-Verlag.

Vandenberghe, L., Boyd, S., and Wu, S.-P. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM journal on matrix analysis and applications*, 19(2):499–533.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wang, G., O'Leary, J., and Jacob, P. (2021). Maximal couplings of the Metropolis-Hastings algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 1225–1233. PMLR.

Whitt, W. (1976). Bivariate distributions with given marginals. *The Annals of statistics*, 4(6):1280–1289.

Xu, K., Fjelde, T. E., Sutton, C., and Ge, H. (2021). Couplings for multinomial Hamiltonian Monte Carlo. In *International Conference on Artificial Intelligence and Statistics*, pages 3646–3654. PMLR.

# A Proof of Proposition 1

Let us first prove that the algorithm samples from the correct marginal distributions. In the case when $A_X = 0$, we simply sample from $p$. Otherwise, when $A_X = 1$, sampling of $X$ again corresponds to sampling from $p$, but now with rejection sampling with proposal $\hat{p}$ and ratio bound $M(p, \hat{p})$. Therefore the marginal of $X$ is $p$. With a similar reasoning we can deduce that the marginal of $Y$ is $q$.

Now, following classical rejection sampling theory (see, e.g., Pagès, 2018, Ch. 1), we know that Algorithm 1 termination time $\tau$ follows a geometric law with acceptance probability $\frac{1}{\mathbb{E}[\tau]}$. In order for the algorithm to finish, we need either $A_X = 1$ or $A_Y = 1$. This implies that $1 \leq \tau \leq \min(\tau_X, \tau_Y)$ where $\tau_X$ and $\tau_Y$ are the time until acceptance of a sample for $X$ and $Y$, respectively. Similarly as for $\tau$, we know that $\tau_X$ and $\tau_Y$ are marginally distributed according to Geometric laws with expectation $\mathbb{E}[\tau_X] = M(p, \hat{p})$ and $\mathbb{E}[\tau_Y] = M(q, \hat{q})$. Now, due to the fact that $\tau$ follows a geometric disribution, we know that $\mathbb{V}[\tau] = \mathbb{E}[\tau]^2 - \mathbb{E}[\tau] = \mathbb{E}[\tau](\mathbb{E}[\tau] - 1)$, which yields the result. Finally, when $p$ and $q$ do not have atoms, then $X = Y$ only happens when $X_1 = Y_1$ and $A_X = 1, A_Y = 1$. This implies that we have $\mathbb{P}(X = Y) \leq \mathbb{P}(X_1 = Y_1)$.

# B Proof of Proposition 4

Similarly as for the proof of Proposition 2, using the notations of Algorithm 2, we have

$$
\begin{aligned}
\mathbb{P}(X = Y) &\geq \mathbb{P}(\hat{X}_I = \hat{Y}_J \,\&\, A_X = A_Y = 1 \,\&\, I = J) \\
&= \mathbb{P}(A_X = A_Y = 1 \mid I = J, \hat{X}_I = \hat{Y}_J) \, \mathbb{P}(\hat{X}_I = \hat{Y}_J \mid I = J) \, \mathbb{P}(I = J)
\end{aligned}
\tag{21}
$$

as well as, when the distributions $p$ and $q$ do not have atoms,

$$
\begin{aligned}
\mathbb{P}(X = Y) &\leq \mathbb{P}(\hat{X}_I = \hat{Y}_J \,\&\, I = J) \\
&= \mathbb{P}(\hat{X}_I = \hat{Y}_J \mid I = J) \, \mathbb{P}(I = J).
\end{aligned}
\tag{22}
$$

Following Deligiannidis et al. (2020), we can lower bound $\mathbb{P}(A_X = A_Y = 1 \mid I = J, \hat{X}_I = \hat{Y}_J)$ by

$$
\begin{aligned}
&\mathbb{P}(A_X = A_Y = 1 \mid I = J, \hat{X}_I = \hat{Y}_J) \\
&\geq \mathbb{P}(A_X = 1 \mid I = J, \hat{X}_I = \hat{Y}_J) \cdot \mathbb{P}(A_Y = 1 \mid I = J, \hat{X}_I = \hat{Y}_J) \\
&\geq \frac{N}{N - 1 + \frac{1}{p_X}} \cdot \frac{N}{N - 1 + \frac{1}{p_Y}},
\end{aligned}
\tag{23}
$$

where $p_X = \mathbb{P}\left(\frac{p(\hat{X})}{M(p, \hat{p})\hat{p}(\hat{X})} \mid \hat{X} = \hat{Y}\right) > 0$ and $p_Y = \mathbb{P}\left(\frac{q(\hat{Y})}{M(q, \hat{q})\hat{q}(\hat{X})} \mid \hat{X} = \hat{Y}\right) > 0$, for $(\hat{X}, \hat{Y}) \sim \hat{\Gamma}$. Furthermore, using the maximality property of $\hat{\Gamma}$, the density of $\hat{X}$ conditionally on $\hat{X} = \hat{Y}$ is $\min(\hat{p}, \hat{q})$ (Wang et al., 2021, Lemma 2), so that $p_X \geq \frac{1}{\mathbb{P}(\hat{X} = \hat{Y}) \, M(p, \hat{p})} \mathbb{E}\left[p(\hat{Y})\right]$, and similarly, $p_Y \geq \frac{1}{\mathbb{P}(\hat{X} = \hat{Y}) \, M(q, \hat{q})} \mathbb{E}\left[q(\hat{X})\right]$, with the identity $p_X = \frac{1}{M(p, \hat{p})}$ and $p_Y = \frac{1}{\mathbb{P}(M(q, \hat{q}))}$ when we also have $\hat{p} = \hat{q}$.

Now we have

$$\mathbb{P}(I=J)$$

$$= \mathbb{E}\left[\sum_{n=1}^{N}\min(W_n^X, W_n^Y)\right]$$

$$= \mathbb{E}\left[\sum_{n=1}^{N}\min\left(\frac{\frac{p(\hat{X}_n)}{\hat{p}(\hat{X}_n)}}{\sum_{m=1}^{N}\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)}}, \frac{\frac{q(\hat{Y}_n)}{\hat{q}(\hat{Y}_n)}}{\sum_{m=1}^{N}\frac{q(\hat{Y}_m)}{\hat{q}(\hat{Y}_m)}}\right)\right] \tag{24}$$

$$= N\mathbb{E}\left[\min\left(\frac{\frac{p(\hat{X}_1)}{\hat{p}(\hat{X}_1)}}{\sum_{m=1}^{N}\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)}}, \frac{\frac{q(\hat{Y}_1)}{\hat{q}(\hat{Y}_1)}}{\sum_{m=1}^{N}\frac{q(\hat{Y}_m)}{\hat{q}(\hat{Y}_m)}}\right)\right]$$

$$= \frac{N}{\sqrt{2N\log\log N}}\mathbb{E}\left[\min\left(\frac{\frac{p(\hat{X}_1)}{\hat{p}(\hat{X}_1)}}{\frac{1}{\sqrt{2N\log\log N}}\sum_{m=1}^{N}\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)}}, \frac{\frac{q(\hat{Y}_1)}{\hat{q}(\hat{Y}_1)}}{\frac{1}{\sqrt{2N\log\log N}}\sum_{m=1}^{N}\frac{q(\hat{Y}_m)}{\hat{q}(\hat{Y}_m)}}\right)\right] > 0,$$

the positivity following from the same argument as for Proposition 2.

We can notice that

$$\frac{1}{\sqrt{2N\log\log N}}\sum_{m=1}^{N}\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)} = \frac{1}{\sqrt{2N\log\log N}}\sum_{m=1}^{N}\left[\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)} - 1\right] + \frac{N}{\sqrt{2N\log\log N}} \tag{25}$$

so that

$$\frac{1}{\sqrt{2N\log\log N}}\sum_{m=1}^{N}\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)} \le \frac{N}{\sqrt{2N\log\log N}} + \sup_{K\ge N}\frac{1}{\sqrt{2K\log\log K}}\sum_{m=1}^{K}\left[\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)} - 1\right], \tag{26}$$

and

$$\frac{1}{\sqrt{2N\log\log N}}\sum_{m=1}^{N}\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)} \ge \frac{N}{\sqrt{2N\log\log N}} + \inf_{K\ge N}\frac{1}{\sqrt{2K\log\log K}}\sum_{m=1}^{K}\left[\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)} - 1\right]. \tag{27}$$

Furthermore, the law of the iterated logarithm (see, e.g., Shiryaev, 1996) ensures that

$$\limsup_{N\to\infty}\frac{1}{\sqrt{2N\log\log N}}\sum_{m=1}^{N}\left[\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)} - 1\right] = \mathbb{V}\left[\frac{p(\hat{X})}{\hat{p}(\hat{X})}\right]^{1/2} \tag{28}$$

and

$$\liminf_{N\to\infty}\frac{1}{\sqrt{2N\log\log N}}\sum_{m=1}^{N}\left[\frac{p(\hat{X}_m)}{\hat{p}(\hat{X}_m)} - 1\right] = -\mathbb{V}\left[\frac{p(\hat{X})}{\hat{p}(\hat{X})}\right]^{1/2}, \tag{29}$$

a similar result being true for the $\hat{Y}$ part. By rearranging the inequalities together, this ensures that there exist two sequences $l_N$ and $u_N$ such that for any $N$ we have

$$l_N \le \mathbb{P}(I=J) \le u_N, \tag{30}$$

with

$$l_N \sim \frac{\int\min(p(x),q(x))\,\mathrm{d}x}{1 + u\frac{\sqrt{2N\log\log N}}{N}}, \quad u_N \sim \frac{\int\min(p(x),q(x))\,\mathrm{d}x}{1 - u\frac{\sqrt{2N\log\log N}}{N}}, \tag{31}$$

where $u = \max\left(\mathbb{V}\left[\frac{p(\hat{X})}{\hat{p}(\hat{X})}\right]^{1/2}, \mathbb{V}\left[\frac{q(\hat{Y})}{\hat{q}(\hat{Y})}\right]^{1/2}\right)$.

21

# C    Proof of Proposition 6

Let $\hat{X}, \hat{Y}$ be the output of the reflection-maximal coupling of Algorithm 3. Because it is a maximal coupling (Wang et al., 2021, Lemma 2), the reflection-maximal coupling is successful with probability

$$\mathbb{P}(\hat{X} = \hat{Y}) = 1 - \left\| \mathcal{N}(\mu_p, \hat{\Sigma}) - \mathcal{N}(\mu_q, \hat{\Sigma}) \right\|_{TV} = 2F\left( -\frac{1}{2} \left\| Q^{-1/2}(\mu_p - \mu_q) \right\|_2 \right), \tag{32}$$

where $F$ is the cumulative distribution function of a standard Gaussian and we have

$$\begin{aligned}
\mathbb{P}(\hat{X} \in A \mid \hat{X} = \hat{Y}) &= \mathbb{P}(\hat{Y} \in A \mid \hat{X} = \hat{Y}) \\
&= \frac{1}{\mathbb{P}(\hat{X} = \hat{Y})} \int \mathbb{1}(x \in A) \min\left( \mathcal{N}(x; \mu_p, \hat{\Sigma}), \mathcal{N}(x; \mu_q, \hat{\Sigma}) \right) \mathrm{d}x\,.
\end{aligned} \tag{33}$$

For the coupling of Algorithm 1 to be successful, we need the dominating coupling to be successful and for both $\hat{X}$ or $\hat{Y}$ to be accepted, that is,

$$\mathbb{P}(X = Y) \geq \mathbb{P}(\hat{X} = \hat{Y})\mathbb{P}(A_X = A_Y = 1 \mid \hat{X} = \hat{Y}). \tag{34}$$

Now, using generic distributions notations,

$$\begin{aligned}
&\mathbb{P}(A_X = A_Y = 1 \mid \hat{X}) \\
&= \frac{1}{\mathbb{P}(\hat{X} = \hat{Y})} \int \min\left( \frac{p(x)}{M(p, \hat{p})\hat{p}(x)}, \frac{q(x)}{M(q, \hat{q})\hat{q}(x)} \right) \min(\hat{p}(x), \hat{q}(x)) \, \mathrm{d}x \\
&\leq \frac{1}{\mathbb{P}(\hat{X} = \hat{Y})} \int \frac{p(x)}{M(p, \hat{p})\hat{p}(x)} \cdot \frac{q(x)}{M(q, \hat{q})\hat{q}(x)} \min(\hat{p}(x), \hat{q}(x)) \, \mathrm{d}x\,,
\end{aligned} \tag{35}$$

because both terms in the minimum are at most 1.

**Remark 5.** *The last line would have corresponded to using a different uniform sample $U$ to accept $\hat{X}$ and $\hat{Y}$ in Algorithm 1.*

We can now compute the following integral explicitly:

$$\begin{aligned}
&\frac{p(x)}{M(p, \hat{p})\hat{p}(x)} \cdot \frac{q(x)}{M(q, \hat{q})\hat{q}(x)} \\
&= \exp\left( -\frac{1}{2} \left[ (x - \mu_p)^\top \left( \Sigma_p^{-1} - \hat{\Sigma}^{-1} \right) (x - \mu_p) + (x - \mu_q)^\top \left( \Sigma_q^{-1} - \hat{\Sigma}^{-1} \right) (x - \mu_q) \right] \right).
\end{aligned} \tag{36}$$

Now, we have

$$\begin{aligned}
&\mathcal{N}(x; \mu_p, \hat{\Sigma}) < \mathcal{N}(x; \mu_q, \hat{\Sigma}) \\
&\Leftrightarrow x^\top \hat{\Sigma}^{-1}(\mu_p - \mu_q) < \frac{1}{2} \left[ \mu_p^\top \hat{\Sigma}^{-1} \mu_p - \mu_q^\top \hat{\Sigma}^{-1} \mu_q \right].
\end{aligned} \tag{37}$$

Moreover, when we have $\mathcal{N}(x; \mu_p, \hat{\Sigma}) < \mathcal{N}(x; \mu_q, \hat{\Sigma})$, the integrand in (35) becomes

$$\begin{aligned}
&\frac{p(x)}{M(p, \hat{p})\hat{p}(x)} \cdot \frac{q(x)}{M(q, \hat{q})\hat{q}(x)} \mathcal{N}(x; \mu_p, \hat{\Sigma}) \\
&= \frac{1}{\det\left( 2\pi\hat{\Sigma} \right)^{1/2}} \exp\left( -\frac{1}{2} \left[ (x - \mu_p)^\top \Sigma_p^{-1} (x - \mu_p) + (x - \mu_q)^\top \left( \Sigma_q^{-1} - \hat{\Sigma}^{-1} \right) (x - \mu_q) \right] \right).
\end{aligned} \tag{38}$$

22

Now,

$$(x - \mu_p)^\top \Sigma_p^{-1}(x - \mu_p) + (x - \mu_q)^\top \left(\Sigma_q^{-1} - \hat{\Sigma}^{-1}\right)(x - \mu_q)$$
$$= (x - \alpha)^\top H^{-1}(x - \alpha) + \beta, \tag{39}$$

where

$$H = \left(\Sigma_p^{-1} + \Sigma_q^{-1} - \hat{\Sigma}^{-1}\right)^{-1}$$
$$\alpha = H\left(\Sigma_p^{-1}\mu_p + \left(\Sigma_q^{-1} - \hat{\Sigma}^{-1}\right)\mu_q\right) \tag{40}$$
$$\beta = \mu_p^\top \Sigma_p^{-1}\mu_p + \mu_q^\top \left(\Sigma_q^{-1} - \hat{\Sigma}^{-1}\right)\mu_q - \alpha^\top H^{-1}\alpha.$$

This means that when $\mathcal{N}(x; \mu_p, \hat{\Sigma}) < \mathcal{N}(x; \mu_q, \hat{\Sigma})$, the integrand in (35) becomes

$$\exp(-\beta/2)\frac{\det(2\pi H)^{1/2}}{\det\left(2\pi\hat{\Sigma}\right)^{1/2}}\mathcal{N}(x; \alpha, H). \tag{41}$$

Under $\mathcal{N}(\alpha, H)$, the event

$$x^\top \hat{\Sigma}^{-1}(\mu_p - \mu_q) < \frac{1}{2}\left[\mu_p^\top \hat{\Sigma}^{-1}\mu_p - \mu_q^\top \hat{\Sigma}^{-1}\mu_q\right]$$
$$\Leftrightarrow (x - \alpha)^\top H^{-\top/2}H^{\top/2}\hat{\Sigma}^{-1}(\mu_p - \mu_q) < \frac{1}{2}\left[\mu_p^\top \hat{\Sigma}^{-1}\mu_p - \mu_q^\top \hat{\Sigma}^{-1}\mu_q - 2\alpha^\top \hat{\Sigma}^{-1}(\mu_p - \mu_q)\right] \tag{42}$$

has probability

$$\mathcal{F}(\alpha, H) := F\left(\frac{1}{2}\frac{\mu_p^\top \hat{\Sigma}^{-1}\mu_p - \mu_q^\top \hat{\Sigma}^{-1}\mu - 2\alpha^\top \hat{\Sigma}^{-1}(\mu_p - \mu_q)}{\left\|H^{\top/2}\hat{\Sigma}^{-1}(\mu_p - \mu_q)\right\|_2}\right), \tag{43}$$

where $F$ is the cumulative distribution of a standard normal distribution. Symmetrically, when $\mathcal{N}(x; \mu_p, \hat{\Sigma}) > \mathcal{N}(x; \mu_q, \hat{\Sigma})$, the integrand in (35) becomes

$$\exp(-\gamma/2)\frac{\det(2\pi H)^{1/2}}{\det\left(2\pi\hat{\Sigma}\right)^{1/2}}\mathcal{N}(x; \delta, H), \tag{44}$$

where

$$\delta = H\left(\Sigma_q^{-1}\mu_q + \left(\Sigma_p^{-1} - \hat{\Sigma}^{-1}\right)\mu_p\right),$$
$$\gamma = \mu_q^\top \Sigma_q^{-1}\mu_q + \mu_p^\top \left(\Sigma_p^{-1} - \hat{\Sigma}^{-1}\right)\mu_p - \delta^\top H^{-1}\delta. \tag{45}$$

and the event $\mathcal{N}(x; \mu_p, \hat{\Sigma}) > \mathcal{N}(x; \mu_q, \hat{\Sigma})$ has probability $1 - \mathcal{F}(\delta, H)$ under $\mathcal{N}(\delta, H)$. Therefore, we finally obtain

$$\mathbb{P}(X = Y) \geq \frac{\det(2\pi H)^{1/2}}{\det\left(2\pi\hat{\Sigma}\right)^{1/2}}\left[\exp(-\beta/2)\mathcal{F}(\alpha, H) + \exp(-\gamma/2)\left[1 - \mathcal{F}(\delta, H)\right]\right]. \tag{46}$$

# D    Reflection maximal coupling

---

**Algorithm 3:** Reflection-maximal coupling

---

**1** **Function** `ReflectionCoupling`($a$, $b$, $\Sigma$)**:**

**2**  $\quad z = \Sigma^{-1/2}(a - b)$

**3**  $\quad e = z/\|z\|$

**4**  $\quad$ Sample $\dot{X} \sim \mathcal{N}(0, I)$ and $U \sim \mathcal{U}(0, 1)$

**5**  $\quad$ **if** $\mathcal{N}(\dot{X}; 0, I)\, U < \mathcal{N}(\dot{X} + z; 0, I)$

**6**  $\quad$ **then**

**7**  $\quad\quad \mid$ Set $\dot{Y} = \dot{X} + z$

**8**  $\quad$ **else**

**9**  $\quad\quad \mid$ Set $\dot{Y} = \dot{X} - 2\left\langle e, \dot{X}\right\rangle e$

**10**  $\quad$ Set $X = a + \Sigma^{1/2}\,\dot{X}$ and $Y = b + \Sigma^{1/2}\,\dot{Y}$

 $\quad$ **Return:** $X$, $Y$

---

# E    Proof of Proposition 7

Consider the problem in (16). Following Vandenberghe et al. (1998), for any invertible matrix $C$, maximising $\log \det \left( \hat{\Sigma}^{-1} \right)$ is equivalent to maximising $\log \det \left( C^\top \hat{\Sigma}^{-1} C \right)$ and $\hat{\Sigma}^{-1} \preceq \Sigma_q^{-1} \Leftrightarrow C^\top \hat{\Sigma}^{-1} C \preceq C^\top \Sigma_q^{-1} C$, and similarly for the other constraints.

We can now consider $C = \Sigma_q^{1/2}$ so that solving (16) is equivalent to solving

$$
\begin{aligned}
&\max \log \det(X) \\
&X \preceq I, \quad X \preceq Y, \quad X \succeq 0,
\end{aligned}
\tag{47}
$$

for $X = C^\top \hat{\Sigma}^{-1} C$ with $Y = C^\top \Sigma_p^{-1} C$. Now let $Y = V D V^\top$ be a diagonalisation of $Y$, with orthogonal matrix $V$. We can now use $V$ to transform (47) via $U = V^\top X V$ into the equivalent problem

$$
\begin{aligned}
&\max \log \det(U) \\
&U \preceq I, \quad U \preceq D, \quad U \succeq 0.
\end{aligned}
\tag{48}
$$

The solution to this problem is given by a diagonal matrix $\bar{U}$ with diagonal entries being $\bar{U}_{ii} = \min(1, D_{ii})$ for all $i$. This allows us to reconstruct a solution $\hat{\Sigma}_{opt}^{-1} = C^{-\top} V \bar{U} V^\top C^{-1}$, that is, $\hat{\Sigma}_{opt} = C V \bar{U}^{-1} V^\top C^\top$ to (16).
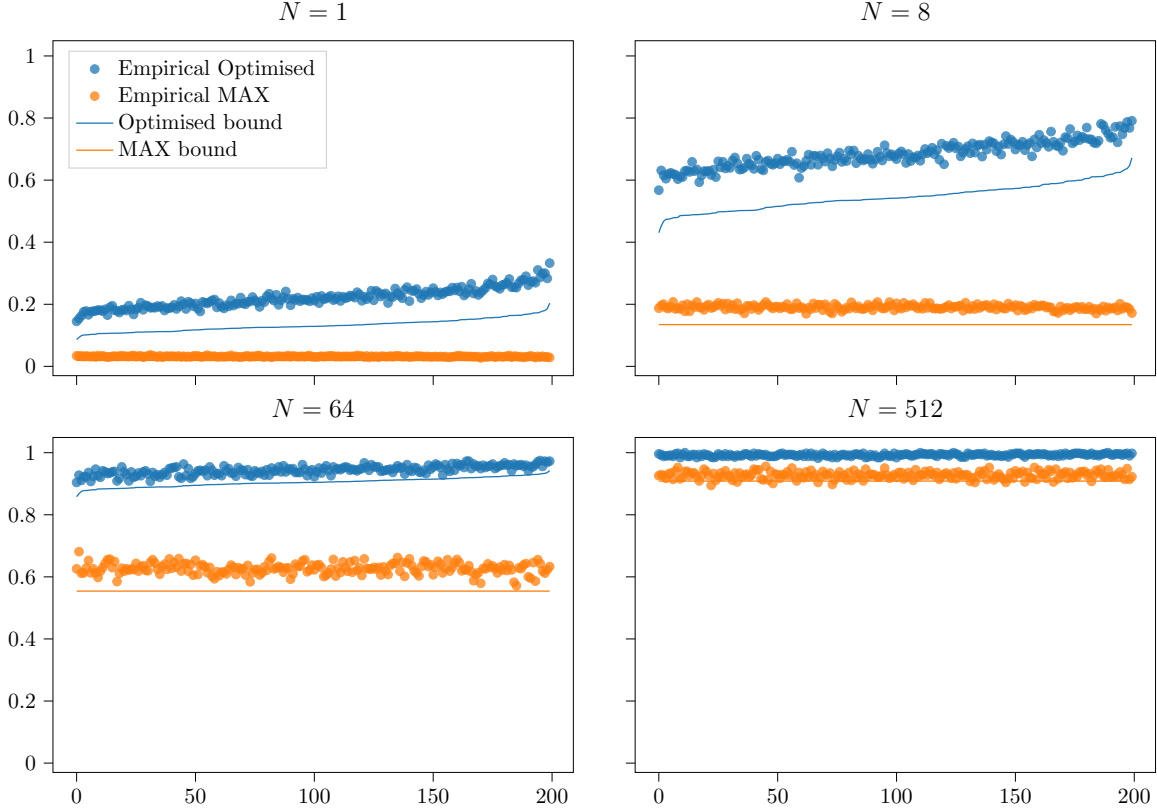
# F    Illustration of Gaussian couplings



Figure 6: Average number of steps $1/\tau$ $N$ taken by Algorithm 2 to sample from the coupling using the reflection-maximal dominating coupling with $\hat{\Sigma}_{opt}$ and $\hat{\Sigma}_{max}$ and an ensemble proposal of size $N$. For better visualisation, the points have been ordered according to the value of the bound derived in Proposition 1 using $\hat{\Sigma}_{opt}$ as the dominating covariance matrix.

To generate Figures 6 and 7, we take the dimension $d$ for the underlying space $\mathbb{R}^d$ to be 10, $\mu_p = \mu_q = 0$, $\Sigma_p = \text{diag}(1, 2, \cdots, d)$ and we draw 200 random $\Sigma_q$'s by sampling a random orthogonal matrix $U$ (given by the QR decomposition of a $d \times d$ matrix of i.i.d. standard Gaussian samples) and setting $\Sigma_q = U\Sigma_p U^\top$. We then run Algorithm 4 with a reflection-maximal dominating coupling 500 times for each pair of covariance matrices generated to compute the coupling and the acceptance probabilities.

Figure 6 shows that the number of trials required to sample from Algorithms 1 and 2 is positively impacted by using $\hat{\Sigma}_{opt}$ instead of $\hat{\Sigma}_{max}$ for all values of $N$. On the other hand, Figure 7 shows that optimising for $\hat{\Sigma}$ does not yield any benefit for $N = 1$, but starts doing so for any $N > 1$. This is because, for $N = 1$, the coupling success probability is driven mostly by that of the dominating coupling, while for $N > 1$, the term $\max\left(\mathbb{V}\left[\frac{p(\hat{X})}{\hat{p}(\hat{X})}\right]^{1/2}, \mathbb{V}\left[\frac{q(\hat{Y})}{\hat{q}(\hat{Y})}\right]^{1/2}\right)$, appearing in Proposition 4, drives the improvement and is directly impacted by $\det\left(\hat{\Sigma}\right)$.
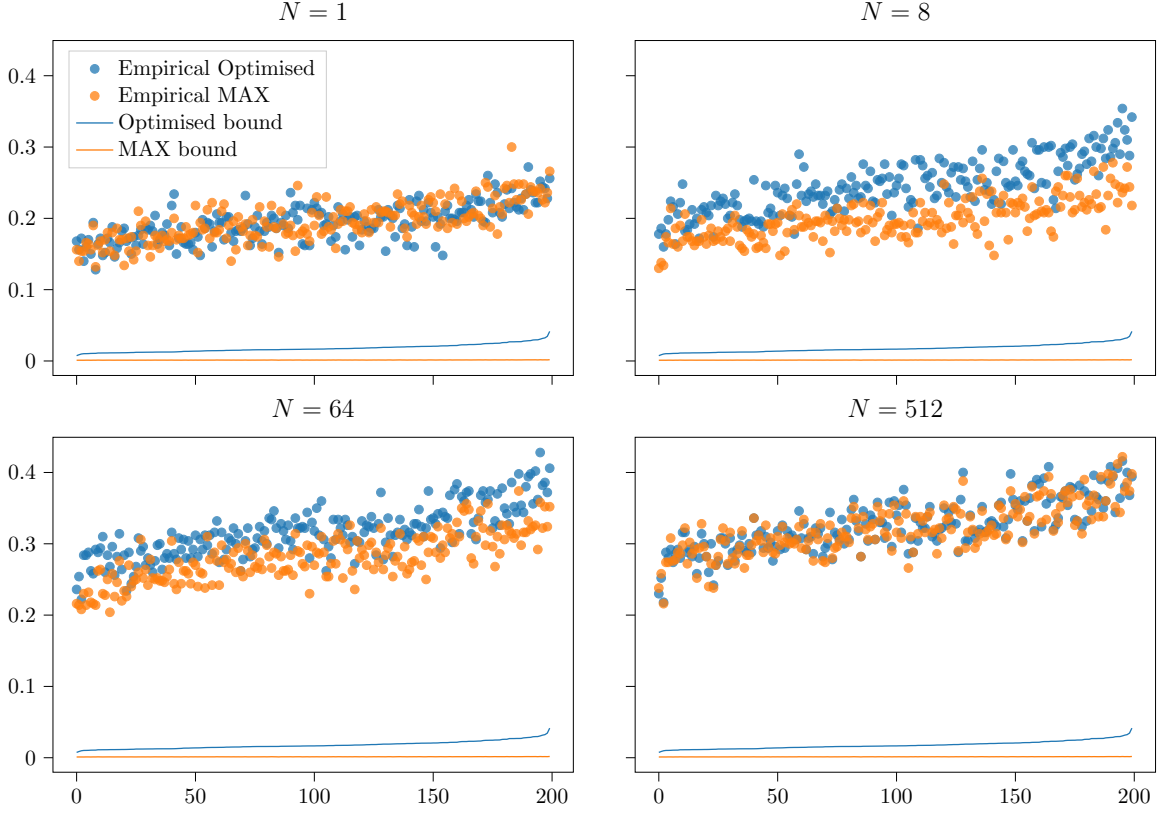
Figure 7: Empirical coupling success probability of Algorithm 2 using the reflection-maximal dominating coupling with $\hat{\Sigma}_{opt}$ and $\hat{\Sigma}_{max}$ and an ensemble proposal of size $N$. For better visualisation, the points have been ordered according to the value of the bound derived in Proposition 6 using $\hat{\Sigma}_{opt}$ as the dominating covariance matrix.

# G    Maximal coupling of translated exponential distributions

In the experiment of Section 5.1 we use maximal coupling of translated exponential distributions as the proposal for the rejection coupling algorithm. The aim here is now to present details on how to sample from this maximal coupling. The method uses the following proposals:

$$
\begin{aligned}
\hat{p}(x) &= \mathrm{Exp}(x; \mu, \alpha(\mu)) = \alpha(\mu)e^{-\alpha(\mu)(x-\mu)}, && x \geq \mu, \\
\hat{q}(x) &= \mathrm{Exp}(x; \eta, \alpha(\eta)) = \alpha(\eta)e^{-\alpha(\eta)(x-\eta)}, && x \geq \eta,
\end{aligned}
\tag{49}
$$

where $\alpha(z) = \frac{z+\sqrt{z^2+4}}{2}$. We assume that $\eta > \mu$ which also implies that $\alpha(\eta) > \alpha(\mu)$.

Then we have

$$
c(x) = \frac{1}{Z^c} \min(\hat{p}(x), \hat{q}(x)) = \begin{cases} \frac{1}{Z^c}\alpha(\mu)e^{-\alpha(\mu)(x-\mu)}, & x \geq \eta \text{ and } x \leq \gamma, \\ \frac{1}{Z^c}\alpha(\eta)e^{-\alpha(\eta)(x-\eta)}, & x \geq \gamma, \\ 0, & \text{otherwise,} \end{cases}
\tag{50}
$$

where

$$
\gamma = \frac{\log \alpha(\eta) - \log \alpha(\mu) + \alpha(\eta)\eta - \alpha(\mu)\mu}{\alpha(\eta) - \alpha(\mu)}.
\tag{51}
$$

We now get that

$$Z^c = \int \min(p(x), q(x))\, dx = \exp(-\alpha(\mu)(\eta - \mu)) - \exp(-\alpha(\mu)(\gamma - \mu)) + \exp(-\alpha(\eta)(\gamma - \eta)),$$

$$Z_1^c = \int_\eta^\gamma \alpha(\mu) e^{-\alpha(\mu)(x-\mu)}\, dx = \exp(-\alpha(\mu)(\eta - \mu)) - \exp(-\alpha(\mu)(\gamma - \mu)), \tag{52}$$

$$Z_2^c = \int_\gamma^\infty \alpha(\eta) e^{-\alpha(\eta)(x-\eta)}\, dx = \exp(-\alpha(\eta)(\gamma - \eta)),$$

where $Z^c$ is also the maximal coupling probability needed in the method. Both of the non-zero branches of (50) now have closed form inverse cumulative density functions (CDF):

$$C_1^{-1}(u) = \mu - \frac{1}{\alpha(\mu)} \log\left[ \exp(-\alpha(\mu)(\eta - \mu)) - u \left[\exp(-\alpha(\mu)(\eta - \mu)) - \exp(-\alpha(\mu)(\gamma - \mu))\right]\right],$$

$$C_2^{-1}(u) = \gamma - \frac{1}{\alpha(\eta)} \log(1 - u), \tag{53}$$

and therefore we can sample from $c(x)$ by first drawing the branch with probabilities $Z_1^c/Z^c$ and $Z_2^c/Z^c$, and then by sampling with inverse CDF method by using the CDFs given above.

Let us then look at the density

$$\tilde{p}(x) = \frac{1}{Z_p} \left[p(x) - \min(p(x), q(x))\right]$$

$$= \begin{cases} \frac{1}{Z^p} \left[\alpha \exp(-\alpha(\mu)(x - \mu)) - \alpha(\eta) \exp(-\alpha(\eta)(x - \eta))\right], & x \geq \gamma, \\ \frac{1}{Z^p}\alpha(\mu) \exp(-\alpha(\mu)(x - \mu)), & \mu \leq x \leq \eta, \\ 0, & \text{otherwise}, \end{cases} \tag{54}$$

where the overall normalisation constant and the normalisation constants of the non-zero branches are given as

$$Z^p = \int \left[p(x) - \min(p(x), q(x))\right]\, dx$$

$$= 1 - \exp(-\alpha(\mu)(\eta - \mu)) + \exp(-\alpha(\mu)(\gamma - \mu)) - \exp(-\alpha(\eta)(\gamma - \eta)),$$

$$Z_1^p = \int_\gamma^\infty \left[\alpha(\mu) \exp(-\alpha(\mu)(x - \mu)) - \alpha(\eta) \exp(-\alpha(\eta)(x - \eta))\right]\, dx \tag{55}$$

$$= \exp(-\alpha(\mu)(\gamma - \mu)) - \exp(-\alpha(\eta)(\gamma - \eta)),$$

$$Z_2^p = \int_\mu^\eta \alpha(\mu) \exp(-\alpha(\mu)(x - \mu))\, dx$$

$$= 1 - \exp(-\alpha(\mu)(\eta - \mu))$$

We can now again first draw the branch with probabilities $Z_1^p/Z^p$ and $Z_2^p/Z^p$, and then use inverse CDF method in each of the branches. The second branch has a closed form inverse CDF

$$\tilde{P}_2^{-1}(u) = \mu - \frac{1}{\alpha(\mu)} \log\left[1 - u \left[1 - \exp(-\alpha(\mu)(\eta - \mu))\right]\right]. \tag{56}$$

However, the first branch does not have a closed form inverse CDF. Instead we need to solve the value of the inverse CDF $\tilde{P}_1^{-1}(u)$ at each $u$ by numerically solving the equation

$$\frac{1}{Z_1^p} \exp(-\alpha(\mu)(x - \mu)) - \frac{1}{Z_1^p} \exp(-\alpha(\eta)(x - \eta)) = u, \qquad x \geq \gamma. \tag{57}$$

For that purpose, we use Chandrupatla's method (Chandrupatla, 1997).

Finally, we need to sample from the distribution

$$\tilde{q}(x) = \frac{1}{Z^q} \left[ q(x) - \min(p(x), q(x)) \right]$$

$$= \begin{cases} \frac{1}{Z^q} \left[ \alpha(\eta) \exp(-\alpha(\eta)(x - \eta)) - \alpha(\mu) \exp(-\alpha(\mu)(x - \mu)) \right], & x \geq \eta \text{ and } x \leq \gamma, \\ 0, & \text{otherwise,} \end{cases} \qquad (58)$$

where the normalisation constant is

$$Z^q = 1 - \exp(-\alpha(\mu)(\eta - \mu)) + \exp(-\alpha(\mu)(\gamma - \mu)) - \exp(-\alpha(\eta)(\gamma - \eta)). \qquad (59)$$

We can again sample from $\tilde{q}(x)$ by inverse CDF method although the inverse CDF is not available in closed form. However, we can solve values of the inverse CDF $\tilde{Q}^{-1}(u)$ by numerically solving the equation

$$1 - \exp(-\alpha(\eta)(x - \eta)) - \exp(-\alpha(\mu)(\eta - \mu)) + \exp(-\alpha(\mu)(x - \mu)) = Z^q u, \qquad (60)$$

for $\eta \leq x \leq \gamma$. For this purpose we again used the Chandrupatla's method.

# H   Coupled rejection resampling algorithm

In this section we reproduce the coupled (parallel) rejection resampling algorithm which is a modified version of the algorithm (Code 3) in Murray et al. (2016). For the sake of concision, we only reproduce the non-ensemble version, the ensemble one being a direct modification following Algorithm 2.

---

**Algorithm 4:** Coupled (parallel) rejection resampling of $w^X \in [0, \infty)^M$ and $w^Y \in [0, \infty)^M$

---

1 **Function** `CoupledRejectionResampling`($w^X$, $w^Y$, $\bar{w}^X$, $\bar{w}^Y$)**:**

     `// Supposing that for all` $m \in \{1, \ldots, M\}$, $w_m^X \leq \bar{w}^X < \infty$ `and` $w_m^Y \leq \bar{w}^Y < \infty$`.`

     `// The loop below can be done in parallel`

2    **for** $m = 1, \ldots, M$ **do**

3        Set $B_m^X = m$ and $B_m^Y = m$

4        Set $A_m^X = 0$ and $A_m^Y = 0$ `// Acceptance flags`

5        Sample $U \sim \mathcal{U}(0, 1)$

6        **if** $U < w_{B_m^X}^X / \bar{w}^X$ **then**

7           Set $A_m^X = 1$

8        **if** $U < w_{B_m^Y}^Y / \bar{w}^Y$ **then**

9           Set $A_m^Y = 1$

10       **while** $A_m^X = 0$ *and* $A_m^Y = 0$ **do**

11          Sample $U \sim \mathcal{U}(0, 1)$

12          Sample $B_m^X = B_m^Y$ from $\mathcal{U}(\{1, \ldots, N\})$

13          **if** $U < w_{B_m^X}^X / \bar{w}^X$ **then**

14             Set $A_m^X = 1$

15          **if** $U < w_{B_m^Y}^Y / \bar{w}^Y$ **then**

16             Set $A_m^Y = 1$

17       **if** $A_m^X = 0$ **then**

18          Continue the rejection loop for $w^X$ only

19       **if** $A_m^Y = 0$ **then**

20          Continue the rejection loop for $w^Y$ only

     **Return:** $B^X$, $B^Y$, $A^X$, $A^Y$

---

**Remark 6.** *It is worth noting that the Metropolis-Hastings based parallel resampling method in Murray et al. (2016, Code 2) can also be coupled by coupling the individual chains appearing in the loop. This observation, although important, does not follow from the methods introduced in the current article, but is a direct consequence of the work of Jacob et al. (2020b). However, it is not directly clear if the resulting algorithm would target – or even recover – the maximal coupling between categorical distributions. Studying the behavior of the resulting resampling algorithm falls outside of the scope of this article.*

# I   Thorisson algorithm

In this section we reproduce the modified version of Thorisson's algorithm (Thorisson, 2000; Gerber and Lee, 2021). This version of the algorithm introduces a parameter $0 < C < 1$ corresponding to the acceptable suboptimality of the coupling success. Formally, if one wants to sample $(X, Y)$ from a coupling of $(p, q)$, the resulting algorithm will have $\mathbb{P}(X = Y) = C \int \min(p, q)$. Introducing this variable enables control of the run time variance of the algorithm. $C = 1$ corresponds to Thorisson's algorithm with infinite run time variance (see Gerber and Lee, 2021, for details). The algorithm is given in Algorithm 5.

---

**Algorithm 5:** Modified Thorisson algorithm

---

1  **Function** `ThorissonCoupling`($p$, $q$, $C$):
2      Sample $X \sim p$
3      Sample $U \sim \mathcal{U}(0, 1)$
4      **if** $U < \min(\frac{q(X)}{p(X)}, C)$ **then**
5          | Set $Y = X$
6      **else**
7          Set $A = 0$
8          **while** $A \neq 1$ **do**
9              Sample $U \sim \mathcal{U}(0, 1)$
10              Sample $Z \sim q$
11              **if** $U > \min\left(1, C\frac{p(Z)}{q(Z)}\right)$ **then**
12                  | Set $A = 1$
13              Set $Y = Z$
       **Return:** $X, Y$

---

# J  Run time efficiency of ensemble rejection sampling

We now provide a theoretical argument as to why using an ensemble proposal in the rejection sampler results in better overall run time when running on a parallel hardware such as a GPU. For the sake of simplicity, we make the following assumptions:

1. $p_{\mathrm{RS}}$ is the acceptance probability of the standard rejection sampler with one proposal.

2. The cost of sampling a proposal and computing the importance weight is given by a constant $K > 0$, and sampling $N$ proposals in parallel comes at the constant same cost $K$.

3. The acceptance probability of the ensemble rejection sampler with $N$ particles is exactly (and not $\geq$) $p_{\mathrm{ERS}} = \frac{N p_{\mathrm{RS}}}{(N-1)p_{\mathrm{RS}}+1}$.

4. The run time of one step in the ERS algorithm is $K + \log_2(N)$, that is, the cost of sampling and computing weights $N$ times fully in parallel, and the cost of sampling one candidate using a prefix-sum type method (Murray et al., 2016, Supplementary material). We note that this recovers the case of single proposal RS when $N = 1$.

5. The number of cores available is infinite.

Let us understand the total run time of getting $M$ samples in parallel. Because parallelisation on a GPU is blocking [3], the expected run time in the case of using simple RS consists in the expectation of the maximum of $M$ i.i.d. Geometric distributions with parameter $p_{\mathrm{RS}}$ times the cost of one step, and similarly for the ERS version with $p_{\mathrm{ERS}}$ instead. To fix notation, let us call $C(M,N)$ this quantity, and keep in mind that $N = 1$ recovers the RS algorithm case. With these notations, the expected total run time of an algorithm sampling $M$ times in parallel with $N$ proposals per sample is therefore proportional to $[K + \log_2(N)]C(M,N)$.

In Figures 8a, 8b, and 8c we show the expected effective run time $N$, $M$, and $p_{\mathrm{RS}}$ for $K = 1/2$, $K = 1$, and $K = 2$, respectively. We computed the (intractable) order statistics mean $C(M,N)$ using Monte Carlo.
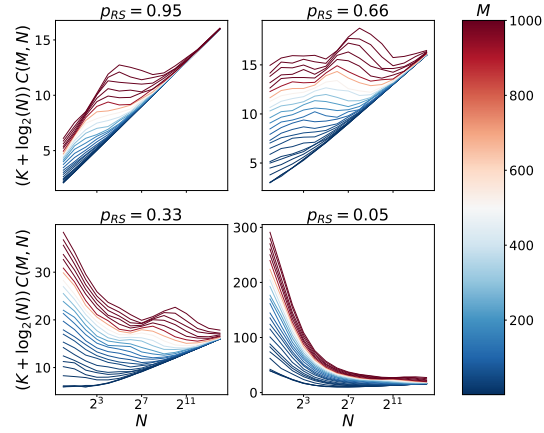
As outlined by these, when either the cost of selecting a candidate in the ensemble is higher than the cost of sampling a single proposal ($K < 1$), or when the acceptance probability $p_{\mathrm{RS}}$ is large, using an ensemble proposal is detrimental to the total run time of the algorithm. However, as soon as the cost of sampling one proposal is large $K > 1$, and/or the acceptance probability $p_{\mathrm{RS}}$ is small, using an ensemble proves beneficial, with larger gains appearing when trying to sample from larger sample sizes $M$ from the target distribution. While analysing the actual parallel time complexity of the algorithm in the more realistic scenario of a finite number of cores is beyond the scope of the current paper, this analysis provides some intuition so as to why ensemble proposals may improve sampling speed on parallel hardware. This behavior is empirically confirmed in Section 5.2.

---

[3]For an informal discussion on the matter, in the context of rejection sampling, we refer the reader to `https://cas.ee.ic.ac.uk/people/dt10/research/rngs-gpu-rejection.html`, last accessed on 22/02/2022.
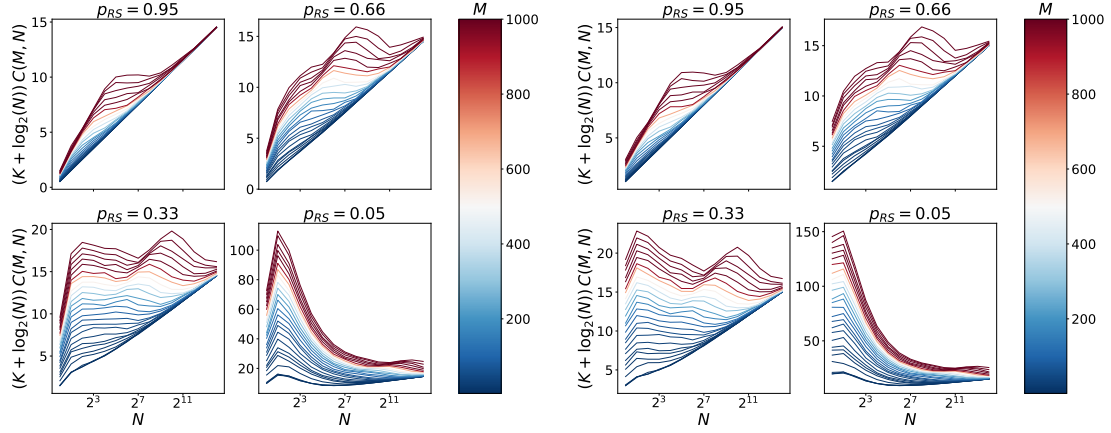
(a) Cost analysis of the ensemble rejection sampler for $K = 1/2$.



(b) Cost analysis of the ensemble rejection sampler for $K = 1$.



(c) Cost analysis of the ensemble rejection sampler for $K = 2$.