Journal of Machine Learning Research 24 (2023) 1-59 Submitted 11/22; Revised 5/23; Published 6/23

Improving multiple-try Metropolis with local balancing

Philippe Gagnon PHILIPPE.GAGNON.3@UMONTREAL.CA
Department of Mathematics and Statistics
Université de Montréal
Montreal, Canada

Florian Maire FLORIAN.MAIRE@UMONTREAL.CA
Department of Mathematics and Statistics
Université de Montréal
Montreal, Canada

Giacomo Zanella GIACOMO.ZANELLA@UNIBOCCONI.IT
Department of Decision Sciences and BIDSA
Bocconi University
Milan, Italy

Editor: Anthony Lee

Abstract

Multiple-try Metropolis (MTM) is a popular Markov chain Monte Carlo method with the appealing feature of being amenable to parallel computing. At each iteration, it samples several candidates for the next state of the Markov chain and randomly selects one of them based on a weight function. The canonical weight function is proportional to the target density. We show both theoretically and empirically that this weight function induces pathological behaviours in high dimensions, especially during the convergence phase. We propose to instead use weight functions akin to the *locally-balanced* proposal distributions of Zanella (2020), thus yielding MTM algorithms that do not exhibit those pathological behaviours. To theoretically analyse these algorithms, we study the high-dimensional performance of *ideal* schemes that can be thought of as MTM algorithms which sample an infinite number of candidates at each iteration, as well as the discrepancy between such schemes and the MTM algorithms which sample a finite number of candidates. Our analysis unveils a strong distinction between the convergence and stationary phases: in the former, local balancing is crucial and effective to achieve fast convergence, while in the latter, the canonical and novel weight functions yield similar performance. Numerical experiments include an application in precision medicine involving a computationally-expensive forward model, which makes the use of parallel computing within MTM iterations beneficial.

Keywords: Bayesian statistics, Markov chain Monte Carlo, parallel computing, random-walk Metropolis, scaling limit, weak convergence.

1. Introduction

1.1 Multiple-try Metropolis

In this paper, we study a specific Markov chain Monte Carlo (MCMC) method introduced by Liu et al. (2000) called *Multiple-try Metropolis* (MTM). It can be seen as a generalization of the Metropolis–Hastings (MH, Metropolis et al. (1953) and Hastings (1970)) algorithm:

at each iteration, several candidates for the next state of the Markov chain *(instead of one)* are sampled, hence the name of the algorithm. MTM can be used to sample from an *intractable* distribution π for Monte Carlo integration purposes, where *intractable* here refers to the impossibility to compute integrals exactly with respect to that distribution. In a sampling context, such a distribution is called the *target* distribution. This distribution often is a posterior distribution of model parameters resulting from a Bayesian model. In the following, we assume for simplicity that π admits a strictly positive probability density function (PDF) with respect to Lebesgue measure, implying that the model parameters in a Bayesian context are continuous random variables; to simplify the notation, we will also use π to denote the PDF. When the latter is a posterior density, it is proportional to the product of the likelihood function and a prior density. Its normalizing constant is not available, but it is assumed here that the target density can be evaluated pointwise (up to that normalizing constant).

In its simplest and most popular form, an iteration of MTM is as follows[1]:

1. N values $\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{R}^d$ are sampled independently from a proposal distribution with density $q_\sigma(\mathbf{x}, \cdot\,)$, where $\mathbf{x} \in \mathbb{R}^d$ is the current state of the chain and $\sigma > 0$ is a fixed scale parameter;

2. one of the $\mathbf{y}_i$'s is randomly selected to be the proposal for the next state of the chain, say $\mathbf{y}_j$, with probability proportional to a weight $w(\mathbf{x}, \mathbf{y}_j)$, where w is a strictly positive weight function;

3. $N-1$ values $\mathbf{z}_1, \ldots, \mathbf{z}_{N-1}$ are sampled independently from $q_\sigma(\mathbf{y}_j, \cdot\,)$;

4. the proposal is accepted, meaning that the next state is set to be $\mathbf{y}_j$, with probability

$$\alpha(\mathbf{x}, \mathbf{y}_j) := 1 \wedge \frac{\pi(\mathbf{y}_j)\, q_\sigma(\mathbf{y}_j, \mathbf{x})\, w(\mathbf{y}_j, \mathbf{x}) \Big/ \Big(\sum_{i=1}^{N-1} w(\mathbf{y}_j, \mathbf{z}_i) + w(\mathbf{y}_j, \mathbf{x})\Big)}{\pi(\mathbf{x})\, q_\sigma(\mathbf{x}, \mathbf{y}_j)\, w(\mathbf{x}, \mathbf{y}_j) \Big/ \Big(\sum_{i=1}^{N} w(\mathbf{x}, \mathbf{y}_i)\Big)}, \qquad (1)$$

   where $a \wedge b := \min(a, b)$ and the dependence of $\alpha(\mathbf{x}, \mathbf{y}_j)$ on $\mathbf{y}_i, i \neq j$, and $\mathbf{z}_1, \ldots, \mathbf{z}_{N-1}$ is made implicit to simplify the notation; when the proposal is rejected the chain remains at $\mathbf{x}$.

Step 3 and the specific form of the acceptance probability (1) are crucial ingredients to make the resulting Markov chains reversible with respect to π and thus to ensure that π is an invariant distribution and the algorithm is valid (see Liu et al. (2000) for full details). Typically, in Step 1, the sampling is performed through a random-walk scheme, and in particular, using a normal with a mean $\mathbf{x}$ and a diagonal covariance matrix with diagonal elements given by σ^2, denoted by $q_\sigma(\mathbf{x}, \cdot\,) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$, with $\mathbb{I}_d$ being the identity matrix of size d (we also use $q_\sigma(\mathbf{x}, \cdot\,)$ to denote the distribution to simplify). Note that the terms $q_\sigma(\mathbf{y}_j, \mathbf{x})$ and $q_\sigma(\mathbf{x}, \mathbf{y}_j)$ in (1) cancel each other when the density is symmetric as with the normal. In Steps 2 and 4, the weight function $w(\mathbf{x}, \mathbf{y})$ is typically a function of the ratio $\pi(\mathbf{y})/\pi(\mathbf{x})$, the most popular choice by far being $w(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x}) \propto \pi(\mathbf{y})$.

1. We follow the formulation of Bédard et al. (2012).

Randomly choosing among the candidates $\mathbf{y}_1, \ldots, \mathbf{y}_N$ based on a function of their target-density evaluations $\pi(\mathbf{y}_1), \ldots, \pi(\mathbf{y}_N)$ makes MTM an *informed* scheme, in the sense of Zanella (2020), meaning that it leverages target-distribution information in the proposal mechanism. Such schemes can improve in terms of asymptotic variance of ergodic averages and mixing properties over their uninformed counterparts.

1.2 Potential of MTM and parallel computing in MCMC

MTM is appealing in situations where it is computationally expensive to evaluate the target density (typically because of the likelihood function), and it is not possible to obtain an explicit form of its gradient. In this situation, gradient-based MCMC methods may be either inapplicable or computationally intensive, e.g., if repeated numerical derivative approximations are required. Practitioners facing such a situation may naturally consider using a MH sampler with a random-walk proposal mechanism q_σ, or its MTM generalization.[2] The weights in Steps 2 and 4 in MTM can be computed in parallel, which results in a significant computation-time reduction compared to serial computation when the iteration cost is largely dominated by that of evaluating the target density. In this situation, the iteration cost of MTM is roughly twice that of a MH sampler using the proposal distribution q_σ.[3]

We refer to the type of parallelization employed in MTM as *in-step* parallelization, given that parallel computing is used within each algorithm iteration. It is to be contrasted with a basic use of parallel computing where one runs N MH algorithms in parallel each using the proposal distribution q_σ. The advantage of the latter is that it is an embarrassingly parallel workload with essentially no communication cost (see, e.g., Rosenthal (2000) and Jacob et al. (2020)). It however does not reduce the burn-in required by each chain to reach stationarity, and thus its computational speed-up compared to running one MH algorithm is fundamentally limited if the chains have a large mixing time. As a result, in-step parallelization approaches that reduce burn-in can lead to significant gain in efficiency in the convergence phase; see, e.g., discussions in Neal (2003), Tjelmeland (2004), Frenkel (2004), Calderhead (2014) and Holbrook (2023).

In order to reduce the burn-in significantly enough to gain in efficiency with MTM, one has to choose carefully the weight function w. As mentioned, the most popular weight function is $w(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x}) \propto \pi(\mathbf{y})$. The intuition behind choosing this weight function is to select $\mathbf{y}_j$ among the candidates $\mathbf{y}_1, \ldots, \mathbf{y}_N$ proportionally to its "probability" under the target. Although intuitively sensible, this choice lacks theoretical justification and, in fact, has been observed to yield an MTM algorithm with pathological behaviours. For example, Martino and Louzada (2017) highlights that MTM with $w(\mathbf{x}, \mathbf{y}) \propto \pi(\mathbf{y})$ may have issues of convergence if initialized in the tails of the target density because of acceptance probabilities that are near zero in this area, resulting in Markov chains that get stuck. Also,

2. One could also use MTM with an informed gradient-based proposal distribution, but this is less common in practice.
3. When the iteration cost is largely dominated by that of evaluating the target density, we know that the computational cost of the other operations (like sampling from $q_\sigma(\mathbf{x}, \cdot)$) is negligible compared with that of evaluating the target density. This implies that the iteration cost of the MH sampler roughly corresponds to that of evaluating the target density once for evaluating the acceptance probability ($\pi(\mathbf{x})$ can have been recorded from the previous iteration), and the iteration cost of MTM roughly corresponds to that of evaluating the target density twice (because the weights in Steps 2 and 4 in MTM can be computed in parallel).

performance often decreases as N increases, which is counter-intuitive as the algorithm has a larger group of candidates to select the proposal from at each iteration. In this paper, we show that the cause of those issues is precisely the choice of weight function $w(\mathbf{x}, \mathbf{y}) \propto \pi(\mathbf{y})$ which makes the resulting MTM *globally balanced*, a qualifier coined by Zanella (2020) and that will be justified. Note that alternative choices are discussed in, e.g., Liu et al. (2000) and Pandolfi et al. (2010), but these are similar in spirit to the choice $w(\mathbf{x}, \mathbf{y}) \propto \pi(\mathbf{y})$ and are exposed to the same pathologies.

The global objective of this work is to identify effective weight functions based on theoretical arguments and to study the resulting MTM algorithms. The effective weight functions that we identify, such as $w(\mathbf{x}, \mathbf{y}) \propto \sqrt{\pi(\mathbf{y})}$, yield MTM algorithms that have a connection with the *locally-balanced* samplers proposed by Zanella (2020) in the context of discrete state-spaces. Therein, a large class of informed samplers are studied, with proposal distributions resulting from a combination of a random-walk proposal scheme and what is referred to as a *balancing function*. A specific choice of balancing function makes the sampler *locally balanced*. As already noted in Zanella (2020), the weight function in MTM and the balancing function play an analogous role. Given the domination of the locally-balanced sampler over the *globally-balanced* one shown in Zanella (2020), it is thus natural to consider the use of locally-balanced weight functions in MTM, especially given that the resulting samplers have the same computational cost as their globally-balanced counterpart and are as easy to implement. In this paper, we refer to MTM with the weight function $w(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$ as *globally-balanced (GB) MTM* and to MTM with a locally-balanced weight function as *locally-balanced (LB) MTM*. It will be highlighted in the following that LB MTM is closely related to gradient-based MCMC methods such as the *Metropolis-adjusted Langevin algorithm* (MALA, Roberts and Tweedie (1996)) and the *Barker proposal scheme* of Livingstone and Zanella (2022).

1.3 Organization of the paper

We now describe how the rest of the paper is organized. In Section 2, we establish a weak convergence of the Markov chains simulated by MTM towards those simulated by an *ideal* sampling scheme, as $N \to \infty$ with the dimension d fixed. This ideal sampling scheme is the continuous state-space counterpart of the class of informed samplers studied in Zanella (2020). In Section 3, we study the convergence phase of MTM algorithms through both theoretical and empirical results. The analysis shows that GB MTM indeed often has issues of convergence, while LB MTM does not. In fact, LB MTM provides drastic convergence speed-ups and improved robustness compared to GB MTM. In Section 4, we study MTM properties in stationarity. Theorem 2 provides high-dimensional scaling-limit results (as $d \to \infty$) for the ideal scheme with $w(\mathbf{x}, \mathbf{y}) \propto \pi(\mathbf{y})$ and that with a LB weight function, namely $w(\mathbf{x}, \mathbf{y}) \propto \sqrt{\pi(\mathbf{y})}$. Theorem 3 then shows that these ideal schemes are approached by MTM under conditions on the rate at which the number of candidates increases with the dimension. Combining Theorems 2 and 3 leads to an informative but somewhat negative conclusion: the ideal scheme with the LB weight function scales better with dimension, however, the (sufficient) condition under which the corresponding MTM algorithm reaches its full potential by approaching the ideal scheme is that N scales exponentially with the dimension. As a result, in common situations, LB MTM provides only mild performance

improvement compared to GB MTM in stationarity. The theoretical results of Sections 3 and 4 are derived in simple scenarios and numerical results are provided to illustrate them, including adaptive MTM implementations. To explore whether the scope of those theoretical results extends beyond simple scenarios, we study in Section 5 the application of GB and LB MTM to a real-world inference problem of immunotherapy in precision medicine where the likelihood is expensive to compute and its gradient is not available in closed form. The empirical results are consistent with the theoretical ones, including observing a drastically reduced burn-in time for LB MTM. We finish the manuscript with a discussion in Section 6. The proofs of all theoretical results are deferred to Section A. The code to reproduce all numerical results is available online.[4]

While working on our manuscript, it came to our attention that Chang et al. (2022) independently and concurrently propose to use LB weight functions within MTM and study the resulting samplers, but the context and focus are quite different. The context and focus are that of sampling from target distributions defined on discrete state-spaces and more precisely from target distributions resulting from model-selection problems. Our contributions and theirs thus have virtually no overlap. That being said, the conclusions in Chang et al. (2022) and ours are consistent, and in this sense, the two studies are complementary.

2. Ideal schemes and locally-balanced weight functions

We start in Section 2.1 with the identification of LB weight functions as effective weight functions. The arguments motivating the use of such weight functions rest upon a theoretical result for which a sketch of a proof was presented in Liu et al. (2000). The result is stated informally in Section 2.1, while a formal statement is presented in Section 2.2. In Section 2.3, a connection between LB MTM and gradient-based methods is highlighted.

2.1 Locally-balanced weight functions: A motivation

Liu et al. (2000) presented a sketch of a proof of the convergence of the Markov kernel of MTM as $N \to \infty$ to that of a MH algorithm using a proposal distribution with a PDF defined as

$$Q_{w,\sigma}(\mathbf{x}, \mathbf{y}) := \frac{w(\mathbf{x}, \mathbf{y})\, q_\sigma(\mathbf{x}, \mathbf{y})}{\int w(\mathbf{x}, \mathbf{y})\, q_\sigma(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{y}},$$

assuming that the integral in the denominator exists and is finite. We will refer to the MH algorithm using the proposal distribution $Q_{w,\sigma}$ as an *ideal* scheme as it cannot in general be implemented and it can be thought of as an MTM algorithm sampling an infinite number of candidates N at each iteration.[5]

Whenever there exists a positive continuous function g such that $w(\mathbf{x}, \mathbf{y})$ is of the form $w(\mathbf{x}, \mathbf{y}) = g(\pi(\mathbf{y})/\pi(\mathbf{x}))$, the ideal scheme corresponds to the continuous version of the class of informed samplers studied in Zanella (2020). All the weight functions considered in our paper are of this form. Even though the work of Zanella (2020) is in the context of discrete state-spaces, a part of the analysis conducted therein is applicable to the continuous

4. See ancillary files on arXiv:2211.11613.
5. For consistency, we will use the terminology "ideal scheme" also for the globally-balanced version even if using a large number of candidates N is not effective in that case.

case as well. In particular, it indicates that the choice $w(\mathbf{x}, \mathbf{y}) \propto \pi(\mathbf{y})$ yields a proposal distribution which leaves the target distribution invariant (without the MH correction) in the situation where $\sigma \to \infty$, the approach being, in that sense, global; in fact, the limiting case corresponds to independent sampling. This justifies the fact that, MTM using the weight function $w(\mathbf{x}, \mathbf{y}) \propto \pi(\mathbf{y})$ is coined *globally-balanced* (GB) MTM. Analogously, the ideal scheme using the proposal PDF $Q_{w,\sigma}(\mathbf{x}, \mathbf{y}) \propto \pi(\mathbf{y})\, q_\sigma(\mathbf{x}, \mathbf{y})$ will be referred to as the *globally-balanced* (GB) ideal scheme.

The problem with using such a function w is that, in high dimensions, the scale parameter of typical MCMC algorithms is required to be small to avoid near-zero acceptance probabilities, as indicated by the optimal-scaling theory (see, e.g., Bédard et al. (2012) in the specific context of MTM). As $\sigma \to 0$, using $Q_{w,\sigma}(\mathbf{x}, \mathbf{y}) \propto \pi(\mathbf{y})\, q_\sigma(\mathbf{x}, \mathbf{y})$ leaves π^2 invariant (without the MH correction), instead of π. In high dimensions, when σ is small, there is thus a significant discrepancy between the proposal and target distributions, and this causes the pathological behaviours of GB MTM mentioned previously.

Locally-balanced (LB) proposal distributions aim at correcting this discrepancy. Indeed, by construction, LB proposal distributions leave π invariant (without the MH correction) in the limiting situation where $\sigma \to 0$, which is the regime in agreement with high-dimensional settings. In our context, within ideal schemes, LB proposal distributions are such that $Q_{w,\sigma}(\mathbf{x}, \mathbf{y}) \propto g(\pi(\mathbf{y})/\pi(\mathbf{x}))\, q_\sigma(\mathbf{x}, \mathbf{y})$, where the balancing function g is a positive continuous function such that $g(x)/g(1/x) = x$ for $x > 0$. We thus propose to set the weight function in MTM to $w(\mathbf{x}, \mathbf{y}) = g(\pi(\mathbf{y})/\pi(\mathbf{x}))$ with g satisfying these conditions. Several functions g satisfy these conditions (see, e.g., Zanella (2020), Sansone (2022) and Vogrinc et al. (2023)). In this paper, we focus on two of them which have been thoroughly studied in other contexts (Power and Goldman, 2019; Gagnon and Maire, 2020; Gagnon, 2021; Sun et al., 2021; Hird et al., 2022; Liang et al., 2022; Livingstone and Zanella, 2022; Sun et al., 2022b,a; Zhou and Smith, 2022), namely $g(x) = \sqrt{x}$ and $g(x) = x/(1 + x)$, the latter yielding what is called the Barker proposal distribution in reference to Barker (1965)'s acceptance-probability choice. The ideal scheme using the proposal PDF $Q_{w,\sigma}(\mathbf{x}, \mathbf{y}) \propto g(\pi(\mathbf{y})/\pi(\mathbf{x}))\, q_\sigma(\mathbf{x}, \mathbf{y})$ with g satisfying the conditions above will thus be referred to as the *locally-balanced* (LB) ideal scheme. This justifies the fact that, MTM using the weight function $w(\mathbf{x}, \mathbf{y}) \propto g(\pi(\mathbf{y})/\pi(\mathbf{x}))$ with g satisfying the conditions above is coined *locally-balanced* (LB) MTM.

LB MTM will be seen to not exhibit the pathological behaviours mentioned previously. Also, LB MTM with $g(x) = \sqrt{x}$ will be seen to have an advantage over that with the Barker weight function in terms of convergence speed-ups. This advantage has been observed in other contexts (Zhou and Smith, 2022). Significant convergence speed-ups with $g(x) = \sqrt{x}$ have also been observed for MTM in Chang et al. (2022). In stationarity, the performance of LB MTM with $g(x) = \sqrt{x}$ is similar to that with the Barker weight function. All that suggests the following practical recommendation to MTM users: *use LB weight functions, and more specifically,* $g(x) = \sqrt{x}$.

2.2 Convergence towards ideal schemes

To understand why the convergence of MTM towards the ideal scheme might hold, it is useful to have a characterization of the distribution of a proposal sampled using MTM,

that we denote by $\mathbf{Y}_J$ with a capital J to represent that the choice among the candidates for the proposal is random. This distribution is conditional on the current state of the Markov chain $\mathbf{x}$, and we use $\mathbb{E}_{\mathbf{x}}$ to denote an expectation with respect to the associated PDF that depends on $\mathbf{x}$. The PDF is based on the product measure $\prod_{i=1}^{N} q_\sigma(\mathbf{x}, \mathbf{y}_i)\,\mathrm{d}\mathbf{y}_{1:N}$, where $\mathbf{y}_{1:N} := (\mathbf{y}_1, \ldots, \mathbf{y}_N)$. That characterization uses that $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ are conditionally independent and identically distributed (IID) given $\mathbf{x}$.

Proposition 1 *Given a current state* $\mathbf{x}$ *and function* h*, a proposal* $\mathbf{Y}_J$ *sampled using MTM is such that*

$$\mathbb{E}_{\mathbf{x}}[h(\mathbf{Y}_J)] = \int h(\mathbf{y}_1)\, \frac{w(\mathbf{x}, \mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{x}, \mathbf{y}_i)} \prod_{i=1}^{N} q_\sigma(\mathbf{x}, \mathbf{y}_i)\,\mathrm{d}\mathbf{y}_{1:N}.$$

The expectation in Proposition 1 is to be compared with that of $h(\mathbf{Y})$ with $\mathbf{Y} \sim Q_{w,\sigma}(\mathbf{x}, \cdot)$, given a current state $\mathbf{x}$, that can be written as

$$\mathbb{E}_{\mathbf{x}}[h(\mathbf{Y})] = \int h(\mathbf{y}_1)\, \frac{w(\mathbf{x}, \mathbf{y}_1)}{\int w(\mathbf{x}, \mathbf{y}_1)\, q_\sigma(\mathbf{x}, \mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1} \prod_{i=1}^{N} q_\sigma(\mathbf{x}, \mathbf{y}_i)\,\mathrm{d}\mathbf{y}_{1:N}. \tag{2}$$

It is apparent that the expectation in Proposition 1 is an approximation that in (2) (and that one sampling scheme approximates the other), and that, presumably, the approximation becomes more accurate as N increases.

We now present a formal result about the weak convergence of Markov chains simulated by MTM towards those simulated by the ideal sampling scheme, under some conditions. Let us denote a Markov chain simulated by MTM as $\{\mathbf{X}_N(m) : m \in \mathbb{N}\}$ and that simulated by the corresponding ideal algorithm by $\{\mathbf{X}_{\text{ideal}}(m) : m \in \mathbb{N}\}$. Also, let us denote the Euclidean norm of a vector $\mathbf{x}$ by $\|\mathbf{x}\|$.

Theorem 1 *Assume that* $\mathbb{E}[w(\mathbf{X}, \mathbf{Y}_1)^4] < \infty$ *and* $\mathbb{E}[w(\mathbf{X}, \mathbf{Y}_1)^{-4}] < \infty$ *with* $\mathbf{X} \sim \pi$ *and* $\mathbf{Y}_1 \mid \mathbf{X} \sim q_\sigma(\mathbf{X}, \cdot)$*. As* $N \to \infty$*,*

1. *given any state* $\mathbf{x}$*, the total variation between the distribution of a proposal* $\mathbf{Y}_J$ *sampled using MTM and* $Q_{w,\sigma}(\mathbf{x}, \cdot)$ *converges to 0 at a rate of* $1/\sqrt{N}$*;*

2. *if additionally*

   (a) π*,* $Q_{w,\sigma}(\cdot\,, \mathbf{y})$ *and* $Q_{w,\sigma}(\mathbf{y}, \cdot)$ *are continuous, for any* $\mathbf{y}$*,*

   (b) *for all* $\mathbf{x} \in \mathbb{R}^d$*, there exists an* $\varepsilon > 0$ *and an integrable function* $f(\mathbf{x}, \cdot)$ *such that* $\sup_{\{\boldsymbol{\epsilon} \in \mathbb{R}^d : \|\boldsymbol{\epsilon}\| \leq \varepsilon\}} Q_{w,\sigma}(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y}) \leq f(\mathbf{x}, \mathbf{y})$ *for all* $\mathbf{y} \in \mathbb{R}^d$*,*

   then $\{\mathbf{X}_N(m) : m \in \mathbb{N}\}$ *converges weakly towards* $\{\mathbf{X}_{ideal}(m) : m \in \mathbb{N}\}$ *provided that* $\mathbf{X}_N(0) \sim \pi$ *and* $\mathbf{X}_{ideal}(0) \sim \pi$*.*

This result indicates that MTM can be seen as an approximation to the ideal MH scheme using $Q_{w,\sigma}$ as a proposal distribution. The latter will thus be considered instead of the former for theoretical analyses in the next sections. The advantage of doing so is that the ideal scheme samples only one candidate at each iteration and is thus easier to analyse.

A refined version of Theorem 1 will be provided in Section 4.2; it allows to quantitatively evaluate the discrepancy between the chains simulated by MTM and the ideal scheme.

We finish this section with a result indicating that Assumption (b) in part 2 of Theorem 1 is verified in great generality.

Proposition 2 *Assume that π is upper bounded, $q_\sigma(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$, and $w(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$ or $w(\mathbf{x}, \mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$. Then, Assumption (b) in part 2 of Theorem 1 is satisfied.*

2.3 Locally-balanced MTM as a gradient-free alternative

An indirect connection can be established between MTM using $g(x) = \sqrt{x}$ and MALA, and between MTM using $g(x) = x/(1+x)$ and the MH sampler based on the Barker proposal scheme of Livingstone and Zanella (2022). Recall that both MALA and the Barker scheme are gradient-based MCMC algorithms which require that π is differentiable and that $\nabla \log \pi$ can be evaluated pointwise. Interestingly, the MALA proposal can be viewed as an approximation to the ideal scheme using $g(x) = \sqrt{x}$, whereas the Barker proposal can be viewed as an approximation to that using $g(x) = x/(1+x)$. Indeed, they both result from an approximation of $Q_{w,\sigma}(\mathbf{x}, \cdot)$ based on a first-order Taylor series expansion of $\log \pi$:

$$Q_{w,\sigma}(\mathbf{x}, \mathbf{y}) \propto g(\mathrm{e}^{\log \pi(\mathbf{y}) - \log \pi(\mathbf{x})})\, q_\sigma(\mathbf{x}, \mathbf{y}) \approx g(\mathrm{e}^{(\nabla \log \pi(\mathbf{x}))^T (\mathbf{y} - \mathbf{x})})\, q_\sigma(\mathbf{x}, \mathbf{y}).$$

When $g(x) = \sqrt{x}$ and $q_\sigma(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$, normalizing the last expression gives exactly the MALA proposal, while, when $g(x) = x/(1+x)$, it gives the Barker scheme (see, respectively, Section 5 of Zanella (2020) and Section 3 of Livingstone and Zanella, 2022). Thus, we can see MALA and the Barker proposal scheme as gradient-based approximations of the same ideal schemes as those approximated by LB MTM samplers with $g(x) = \sqrt{x}$ and $g(x) = x/(1+x)$. The approximations used in MTM are different in several aspects. First, they are stochastic, by opposition to deterministic as in the gradient-based methods. Second, one has control over the approximations (through N). Last but not least, the algorithms do not require to evaluate the gradient of $\log \pi$, which is advantageous when evaluating the gradient is either infeasible or computationally intensive.

GB MTM approximates an ideal scheme which does not correspond to any known first-order method. A rich body of literature has shown that MALA and the MH sampler with the Barker scheme behave quantitatively and qualitatively differently (in terms of ergodicity measure or scaling-limit regime) to most zero-th order methods such as random-walk Metropolis (see, e.g., Roberts and Rosenthal (1998), Bou-Rabee and Hairer (2013), Dwivedi et al. (2018) and Livingstone and Zanella (2022)). It is thus important to study whether LB MTM inherits some of those favourable properties, which cannot be expected by GB MTM.

3. Performance during the convergence phase

In this section, we evaluate the performance during the convergence phase[6] of MTM with the different weight functions. We do this by analysing the acceptance probabilities in the tails

6. By convergence phase, often called *burn-in* in the MCMC literature, we mean the iterations until the Markov chain simulated by MTM is close in distribution to π.

in Section 3.1 and by empirically measuring the convergence time of adaptively tuned MTM in Section 3.2. The adaptive tuning aims to represent how one would use and tune MTM in practice. As mentioned previously, the acceptance probabilities are near-zero in the tails with GB MTM, unless the step size σ is made extremely small, which causes convergence issues in either case. Our analysis shows that these convergence issues do not arise with LB MTM. Also, numerical results show that MTM with the Barker weight function has higher acceptance probabilities than that with $g(x) = \sqrt{x}$. This advantage has been observed in other contexts (Zanella, 2020; Livingstone and Zanella, 2022), and is attributed to the boundedness of the function $g(x) = x/(1+x)$. In the MTM context, it yields more stability in the approximation of the ideal scheme. Even though the unboundedness of $g(x) = \sqrt{x}$ yields less stability, it leads to more persistent movement from the tails to the high-probability region, which is shown in Section 3.2 to provide better convergence performance.

3.1 Acceptance probabilities in the tails

In this section, we analyse the behaviour of MTM when initialized in the tails by evaluating the conditional expected acceptance probability, given an initial state $\mathbf{x}$ with $\|\mathbf{x}\|$ large, where the expectation is with respect to the random variables involved in the proposal mechanism. A low conditional expected acceptance probability implies that it is likely that the chain gets stuck and that an issue arises in terms of convergence to the target distribution as the algorithm progresses. The analysis rests upon a theoretical result about the ideal scheme. The result is established under a specific and simple scenario: the target density factorizes, and more precisely,

$$\pi(\mathbf{x}) = \prod_{i=1}^{d} \varphi(x_i), \quad \mathbf{x} := (x_1, \dots, x_d)^T \in \mathbb{R}^d, \tag{3}$$

and $q_\sigma(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$, where φ is the PDF of a standard normal distribution. The normal assumption and the factorization allow to make precise calculations and in particular to establish that $Q_{w,\sigma}$ is a normal distribution when the weight function factorizes as well, that is when using for instance $w(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$ or $w(\mathbf{x}, \mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$.

We acknowledge that the scenario limits the scope of the analysis. Note that, in Section A.3, we provide a result which is less precise, but which holds under weaker assumptions. With the result provided in Section A.3, we cannot conduct an analysis as thorough as that performed below. The assumptions are essentially that $U := -\log \pi$ is strongly convex and L-smooth, instead of assuming that the target density factorizes into a product of normal densities. The factorization assumption has a long history in analysis of MCMC, especially in the scaling-limit literature where it is a standard assumption (see, e.g., Roberts et al. (1997), Roberts and Rosenthal (1998), Bédard (2007), Bédard et al. (2012), Durmus et al. (2017) and Gagnon et al. (2019)). The factorization is an important structural limitation which implies independence of the random variables. The normal assumption can be justified in Bayesian large-sample regimes where the models are regular enough[7] (Schmon and Gagnon, 2022), but it is an important limitation as well. We thus expect the results to be

7. Models that are regular enough are those which satisfy regularity conditions; see Schmon and Gagnon (2022) for more details.

informative at least when MTM is used to sample from a posterior distribution resulting from a large data set ($n \gg d$) and a regular model, provided that the model parameters are *a posteriori* weakly dependent. The same scenario will be considered for the scaling-limit analysis in Section 4.

We now present the result in which we use the notation α_{ideal} for the acceptance probability in the ideal MH scheme.

Proposition 3 *Consider a current state* $\mathbf{x}$ *and that* $\mathbf{Y} \sim Q_{w,\sigma}(\mathbf{x}, \cdot)$ *with* $q_\sigma(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$ *and* $w(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$*. If the target distribution is defined as in* (3)*,*

$$\mathbb{E}_{\mathbf{x}}[\alpha_{ideal}(\mathbf{x}, \mathbf{Y})] \leq \exp\left(-\|\mathbf{x}\|^2 \frac{\sigma^2}{2((1+\sigma^2)^2 - \sigma^2)}\right)\left(1 - \frac{\sigma^2}{(1+\sigma^2)^2}\right)^{-d/2}. \tag{4}$$

In particular, for any σ *and* d*, it holds that* $\lim_{\|\mathbf{x}\| \to \infty} \mathbb{E}_{\mathbf{x}}[\alpha_{ideal}(\mathbf{x}, \mathbf{Y})] = 0$.

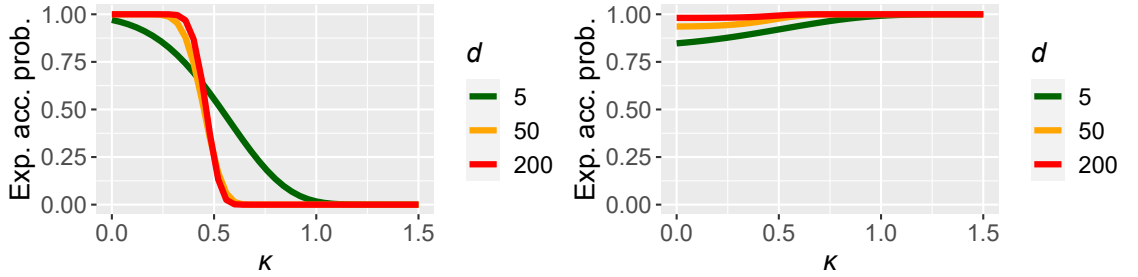
Proposition 3 highlights a pathological behaviour as most MCMC methods do not have acceptance probabilities that are near zero when the current state is in the tails of the target density. We obtained the corresponding upper bound for the ideal scheme with $w(\mathbf{x}, \mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$ and it does not converge to 0. Of course, this does not guarantee mathematically that the conditional expected acceptance probability of that scheme is not near zero in the tails, but it indicates a significant difference. We also tried deriving other upper bounds to see if they yield different results, but we did not obtain any that allows to conclude otherwise. We provide below numerical results for both the ideal scheme and MTM using $w(\mathbf{x}, \mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$ which corroborate those findings and suggest that the acceptance probabilities do not converge to 0 as the current/initial state gets further and further in the tails.

The result provided in Section A.3 essentially states that $\lim_{\|\mathbf{x}\| \to \infty} \mathbb{E}_{\mathbf{x}}[\alpha_{\text{ideal}}(\mathbf{x}, \mathbf{Y})] = 0$, for any σ and d, when $U := -\log \pi$ is strongly convex and L-smooth. While being interesting, it does not provide an explicit upper bound on the conditional expected acceptance probability as in (4) and does not allow for a precise characterization in high-dimensional regimes where $d \to \infty$. Given that we are interested in such regimes, we study the implications of (4) when $d \to \infty$, with $\mathbf{x}$ and σ functions of d. Such a study allows to characterize the relation between d and the location of $\mathbf{x}$ in the tails in the situations where there are convergence issues with the GB ideal scheme and MTM. We highlight a dependence on d of $\mathbf{x}, \mathbf{Y}, \pi$ and σ by denoting these for the rest of the section by $\mathbf{x}_d, \mathbf{Y}_d, \pi_d$ and σ_d. For the analysis, we consider that $\sigma_d^2 = \ell^2/d$ and $\|\mathbf{x}_d\| = d^\kappa$ with κ a positive constant; setting $\sigma_d^2 = \ell^2/d$ will be seen to be an effective way of scaling σ with d. With these choices, the conclusion is the following: *the conditional expected acceptance probability of the GB ideal scheme* (4) *converges to 0 when* $\kappa > 1/2$*, implying that it is sufficient for* $\|\mathbf{x}_d\|$ *to grow with* d *at any rate faster than* $\sqrt{d}$ *to lead to near-zero acceptance probabilities.* We highlight that, with a target distribution such as that defined in (3), $\|\mathbf{x}_d\| = d^\kappa$ with κ around 0.5 is not even far in the tails of the density as $\|\mathbf{X}_d\|^2$ has a chi-squared distribution with a mean of d and a standard-deviation of $\sqrt{2d}$. This implies that even a random initialization of GB MTM using a distribution slightly different from the target may lead to issues of convergence to the target distribution as the algorithm progresses.

We present in Figures 1 and 2 numerical results which complete the analysis. In both figures, we provide conditional expected acceptance probabilities as a function of κ in the situation where the target distribution is defined as in (3), $q_{\sigma_d}(\mathbf{x}_d, \cdot) = \mathcal{N}(\mathbf{x}_d, \sigma_d^2 \mathbb{I}_d)$ with $\ell = 2.38$ and the current/initial state $\mathbf{x}_d$ is set to $\mathbf{x}_d = (d^{\kappa-1/2}, \ldots, d^{\kappa-1/2})$, ensuring that $\|\mathbf{x}_d\| = d^{\kappa}$. The value $\ell = 2.38$ will be seen in Section 4.1 to be optimal for the GB ideal scheme in a high-dimensional regime. The expectations are approximated using independent Monte Carlo sampling. The approximations are based on samples of size 1,000,000.

The difference between Figure 1 and Figure 2 is that in the former the results are for the ideal schemes, whereas in the latter they are for MTM. The results in Figure 2 are for $d = 50$; we observed similar results when $d = 200$. The results for GB samplers are consistent with the theoretical result about the convergence to 0 of the expectation in (4) when $\kappa > 1/2$, with conditional expected acceptance probabilities close to 1 for κ smaller than 0.5 (for moderate to high dimensions, and moderate values of N for MTM), followed by a sharp drop around $\kappa = 0.5$. In Figure 1 (a), we notice that the conditional expected acceptance probability converges to 0 even when $d = 5$ (thus in the case where the high-dimensional regime is not attained); this is because ℓ is not large enough to yield an algorithm that performs approximately IID sampling (recall the discussion towards the end of Section 2), suggesting that the conclusion of Proposition 3 holds.

The results in Figure 1 (b) and Figure 2 (b)-(c) suggest that LB schemes do not have issues of convergence to the target distribution when initialized in the tails. They also suggest that using the Barker weight function in MTM leads to higher acceptance probabilities. As mentioned, we attribute the difference to the fact that, with MTM, the normalizing constant of $Q_{w,\sigma}(\mathbf{x}, \cdot)$ needs to be approximated and the boundedness of the function $g(x) = x/(1+x)$ yields more stability in the approximation.

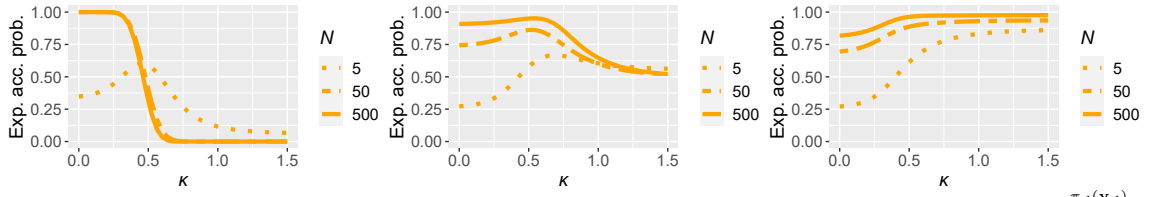


(a) Ideal scheme w. $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$ **(b) Ideal scheme w.** $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$

Figure 1: Conditional expected acceptance probability as a function of κ when $\mathbf{x}_d = (d^{\kappa-1/2}, \ldots, d^{\kappa-1/2})$ and $\ell = 2.38$, for several values of d and: (a) the ideal scheme with $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$, and (b) the ideal scheme with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$.

3.2 Convergence to stationarity: simulations with adaptive MCMC

In this section, we perform numerical simulations with adaptive MTM schemes, where the scale parameter σ is tuned on the fly while the MTM algorithm progresses. Such simulations allow to: (i) reduce the sensitivity of the simulation set-up to the choice of a specific value



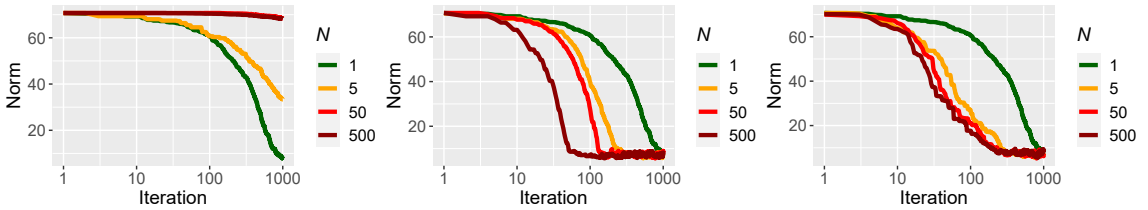
(a) MTM w. $w(\mathbf{x}_d, \mathbf{y}_d) = \frac{\pi_d(\mathbf{y}_d)}{\pi_d(\mathbf{x}_d)}$ **(b) MTM w.** $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\frac{\pi_d(\mathbf{y}_d)}{\pi_d(\mathbf{x}_d)}}$ **(c) MTM w.** $w(\mathbf{x}_d, \mathbf{y}_d) = \frac{\frac{\pi_d(\mathbf{y}_d)}{\pi_d(\mathbf{x}_d)}}{1+\frac{\pi_d(\mathbf{y}_d)}{\pi_d(\mathbf{x}_d)}}$

Figure 2: Conditional expected acceptance probability as a function of κ when $\mathbf{x}_d = (d^{\kappa-1/2}, \ldots, d^{\kappa-1/2})$, $\ell = 2.38$ and $d = 50$, for several values of N and: (a) MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$, (b) MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$, and (c) MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = (\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d))/(1 + \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d))$.

for σ; (ii) assess the impact of the choice of weight function in a more advanced and realistic MTM implementation (arguably closer to one that a careful practitioner would use).

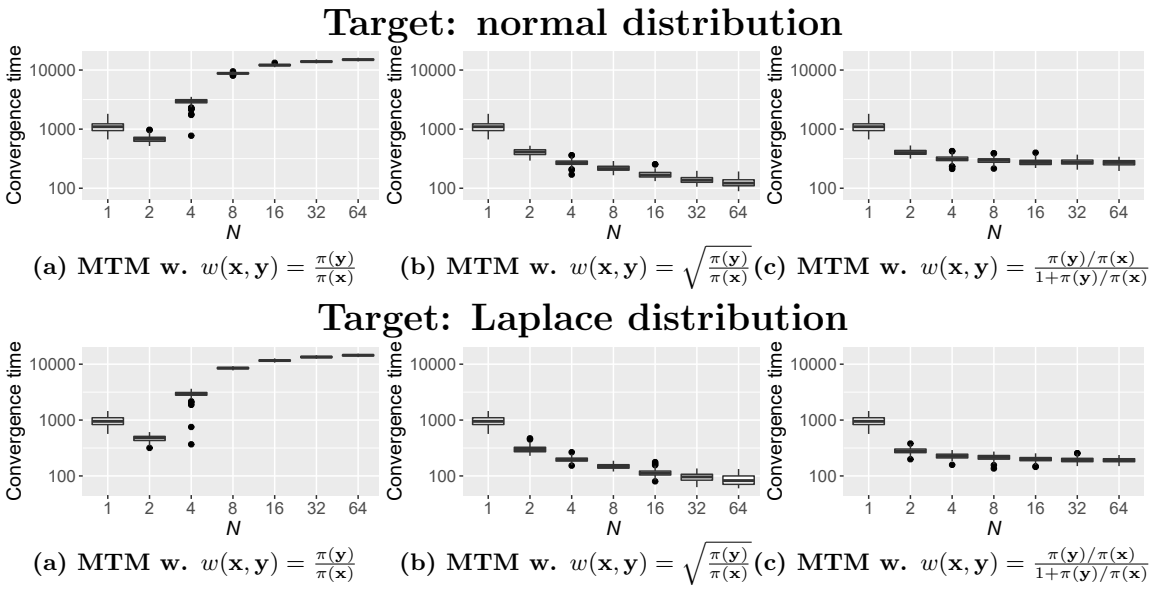
The study in this section is non-asymptotic; the mathematical objects like the target distribution, the scale parameter and the states are thus denoted without a subscript d, that is π, σ and $\mathbf{x}$. Algorithm 4 found in Section 5 of Andrieu and Thoms (2008) is used to adaptively tune σ. The algorithm targets an acceptance rate to adapt tuning parameters. The targeted acceptance rates are 25% and 50% for GB and LB MTM, respectively. These targets are chosen according to theoretical and empirical results presented in the next sections. Algorithm 4 of Andrieu and Thoms (2008) also uses a learning rate $\gamma(m)$, which here is set to $m^{-0.6}$, m representing the iteration index. A power of -0.6 allows to reach a good balance between fast adaptation and stability in this example. We experimented with different power values and obtained similar conclusions.

The results are presented in Figures 3 and 4. Figure 3 displays trace plots for GB MTM and LB MTM with $q_\sigma(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$ and target distribution as in (3) with $d = 50$. Trace plots of $\|\mathbf{X}(m)\|$ are presented, with a log-scale on the x-axis, in the situation where the algorithms are initialized from $\mathbf{x} = (10, \ldots, 10)$, which corresponds to $\mathbf{x} = (d^{\kappa-1/2}, \ldots, d^{\kappa-1/2})$ with $\kappa \approx 1.09$. We observe the pathological behaviour of GB MTM described before: increasing N deteriorates the convergence performance up to having to set σ to near-zero values to achieve non-negligible acceptance rate when N equals 50 or 500. We observe the opposite and desirable results for LB MTM, whose convergence speed increases with N. To provide a more quantitative picture, Figure 4 shows the convergence times for the same algorithms, target distribution and starting state as in Figure 3. The results are obtained from 100 independent runs for each algorithm using different values of N. Here, the convergence time is defined as the first time the chain reaches the 95th percentile of $\|\mathbf{X}\|$ under the target distribution. Figure 4 also presents analogous results for a different target distribution, namely a 50-dimensional product of standard Laplace distributions. The results are consistent with Figure 3, with LB MTM providing a smooth and regular improvement in performance with N, unlike GB MTM. Also, it is interesting to note the difference between the LB MTM with the Barker weight function and that with $g(x) = \sqrt{x}$. The performance of the former stabilizes quicker as N increases, and do not reach to same level of improvement as the latter.



(a) MTM w. $w(\mathbf{x},\mathbf{y}) = \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}$ **(b) MTM w.** $w(\mathbf{x},\mathbf{y}) = \sqrt{\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}}$ **(c) MTM w.** $w(\mathbf{x},\mathbf{y}) = \frac{\pi(\mathbf{y})/\pi(\mathbf{x})}{1+\pi(\mathbf{y})/\pi(\mathbf{x})}$

Figure 3: Trace plots of the Euclidean norm of the state when $d = 50$, the scale parameter is adaptively tuned and the initial state is $\mathbf{x} = (10, \ldots, 10)$, for several values of N and: (a) GB MTM, (b) LB MTM with $w(\mathbf{x},\mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$, and (c) LB MTM with $w(\mathbf{x},\mathbf{y}) = (\pi(\mathbf{y})/\pi(\mathbf{x}))/(1+\pi(\mathbf{y})/\pi(\mathbf{x}))$; the scale on the x-axis is logarithmic.



(a) MTM w. $w(\mathbf{x},\mathbf{y}) = \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}$ **(b) MTM w.** $w(\mathbf{x},\mathbf{y}) = \sqrt{\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}}$ **(c) MTM w.** $w(\mathbf{x},\mathbf{y}) = \frac{\pi(\mathbf{y})/\pi(\mathbf{x})}{1+\pi(\mathbf{y})/\pi(\mathbf{x})}$

(a) MTM w. $w(\mathbf{x},\mathbf{y}) = \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}$ **(b) MTM w.** $w(\mathbf{x},\mathbf{y}) = \sqrt{\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}}$ **(c) MTM w.** $w(\mathbf{x},\mathbf{y}) = \frac{\pi(\mathbf{y})/\pi(\mathbf{x})}{1+\pi(\mathbf{y})/\pi(\mathbf{x})}$

Figure 4: Convergence time as a function of N when $d = 50$, the scale parameter is adaptively tuned and the initial state is $\mathbf{x} = (10, \ldots, 10)$, for several values of N and: (a) MTM with $w(\mathbf{x},\mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$, (b) MTM with $w(\mathbf{x},\mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$, and (c) MTM with $w(\mathbf{x},\mathbf{y}) = (\pi(\mathbf{y})/\pi(\mathbf{x}))/(1+\pi(\mathbf{y})/\pi(\mathbf{x}))$; on the first row, the target is a 50-dimensional standard normal distribution, whereas it is a 50-dimensional standard Laplace distribution on the second row.

Based on all the observations made in this section, we provide a recommendation for an adaptive implementation of LB MTM.

1. Set $g(x) = \sqrt{x}$.

2. Set N equal to the number of available cores for parallel computing.

3. Initialize the step size as $\sigma = \ell/\sqrt{d}$ with $\ell = 2.38$.

4. Run MTM with adaptive tuning of σ as in Algorithm 4 of Andrieu and Thoms (2008), with $\gamma(m) = m^{-0.6}$ and a targeted acceptance rate of 50%.

4. Performance at stationarity

In this section, we characterize the high-dimensional behaviour of MTM algorithms after they have reached stationarity. To this end, we first establish in Section 4.1 the weak convergence of a transformation of the Markov chains produced by ideal MH schemes towards Langevin diffusions, each started in stationarity, as $d \to \infty$. In Section 4.2, we bridge the gap between the MTM algorithms and diffusion processes by providing conditions about the scaling of N with d ensuring the asymptotic equivalence of MTM and ideal schemes. We finish with numerical experiments in Section 4.3 that corroborate the findings of Sections 4.1 and 4.2.

It is worth mentioning that a scaling-limit analysis of MTM was conducted in Bédard et al. (2012), but the analysis in that paper is quite different from that conducted here. First, in Bédard et al. (2012), MTM is not seen as an approximation to an ideal scheme because N is considered fixed; a weak convergence of a transformation of the Markov chains produced by MTM towards Langevin diffusions is directly obtained as $d \to \infty$. The asymptotic regime considered here is that where the number of candidates N increases with d, whereas the asymptotic regime in Bédard et al. (2012) can be thought of as the situation where N is small relatively to d. Another difference is that Bédard et al. (2012) considers only GB MTM, while we study also LB MTM.

Because of the nature of the analysis conducted in Sections 4.1 and 4.2, we, as for the asymptotic analysis in Section 3.1, highlight a dependency on d of the target distribution, the scale parameter, the number of candidates, and so on, by denoting them π_d, σ_d, N_d, etc.

4.1 Scaling limits of ideal schemes

For the analysis, we consider the same scenario as in Section 3.1; in particular, we consider that the target distribution is defined as in (3). Also, for the analysis, we set $\sigma_d = \ell/d^\tau$ in $q_{\sigma_d}(\mathbf{x}_d, \cdot) = \mathcal{N}(\mathbf{x}_d, \sigma_d^2 \mathbb{I}_d)$, with ℓ being a positive tuning parameter and τ a positive constant characterizing the scalability of the algorithm with respect to the dimension (the smaller is τ, the better is the scalability with respect to d).

Before presenting the scaling-limit result, we introduce required notation. We use Φ to denote the cumulative distribution function of the standard normal distribution. We use $\{\mathbf{X}_{d,\text{ideal}}(m) : m \in \mathbb{N}\}$ to denote a Markov chain simulated by an ideal MH scheme using $Q_{w,\sigma_d}(\mathbf{x}_d, \cdot)$ for proposal distribution, and define a re-scaled continuous-time version $\{\mathbf{Z}_{d,\text{ideal}}(t) : t \geq 0\}$ using:

$$\mathbf{Z}_{d,\text{ideal}}(t) := \mathbf{X}_{d,\text{ideal}}(\lfloor d^{2\tau} t \rfloor), \tag{5}$$

with $\lfloor \cdot \rfloor$ being the floor function. A scaling limit consists in proving that the first component of $\{\mathbf{Z}_{d,\text{ideal}}(t) : t \geq 0\}$, denoted by $\{Z_{d,\text{ideal}}(t) : t \geq 0\}$, converges weakly to $\{Z(t) : t \geq 0\}$, a Langevin diffusion.

We are now ready to present the scaling-limit result. It is about the GB ideal scheme and the LB ideal scheme with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$.

Theorem 2 *Assume that π_d is as in (3) and that the proposal distribution in a MH algorithm is $Q_{w,\sigma_d}(\mathbf{x}_d, \cdot)$ with $q_{\sigma_d}(\mathbf{x}_d, \cdot) = \mathcal{N}(\mathbf{x}_d, \sigma_d^2 \mathbb{I}_d)$ and $\sigma_d = \ell/d^\tau$. Assume also that $\mathbf{X}_{d,ideal}(0) \sim \pi_d$ and that $\mathbf{Z}_{d,ideal}(t)$ for $t \geq 0$ is defined as in (7). Then, as $d \to \infty$, $\{Z_{d,ideal}(t) : t \geq 0\}$ converges weakly towards $\{Z(t) : t \geq 0\}$, a Langevin diffusion such that $Z(0) \sim \mathcal{N}(0,1)$ and*

$$\mathrm{d}Z(t) = \ell^2(\vartheta_{w,\tau}(\ell)/2)(\log \varphi(Z(t)))' \,\mathrm{d}t + \sqrt{\ell^2 \vartheta_{w,\tau}(\ell)}\,\mathrm{d}B(t),$$

with $\{B(t) : t \geq 0\}$ being a standard Brownian motion and $\vartheta_{w,\tau}$ being defined as follows: if $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$,

$$\vartheta_{w,\tau}(\ell) = \begin{cases} 2\Phi(-\ell^3/2^3) & \text{if} \quad \tau = 1/6, \\ 1 & \text{if} \quad \tau > 1/6; \end{cases}$$

if $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$,

$$\vartheta_{w,\tau}(\ell) = \begin{cases} 2\Phi(-\ell/2) & \text{if} \quad \tau = 1/2, \\ 1 & \text{if} \quad \tau > 1/2. \end{cases}$$

To establish such a result, it is crucial that the expected acceptance probability (in stationarity) $\mathbb{E}[\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d)]$ (with $\mathbf{X}_d \sim \pi_d$) converges towards a non-null function of ℓ that is independent of d; the function $\vartheta_{w,\tau}(\ell)$ in Theorem 2 is precisely this function. Theorem 2 thus indicates that $\tau \geq 1/2$ in the GB ideal scheme allows such a convergence, whereas $\tau \geq 1/6$ in the LB ideal scheme is sufficient. This implies that the LB ideal scheme has a better scaling with the dimension than the GB ideal scheme.

We present numerical results in Figure 5 of $\mathbb{E}[\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d)]$ as a function of d. These numerical results allow to show that $\mathbb{E}[\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d)]$ converges to 0 when $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ and $\tau < 1/6$ and when $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$ and $\tau < 1/2$. The expectations are approximated using independent Monte Carlo sampling; the approximations are based on samples of size 1,000,000. We stress that there is an important difference between $\mathbb{E}[\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d)]$ and what we called the *conditional expected acceptance probability* in Section 3.1: $\mathbb{E}[\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d)]$ is the unconditional expectation with $\mathbf{X}_d \sim \pi_d$, whereas the *conditional expected acceptance probability* in Section 3.1 is the conditional expectation given a current state $\mathbf{x}_d$. We highlight the difference by referring to $\mathbb{E}[\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d)]$ as (simply) the *expected acceptance probability*.

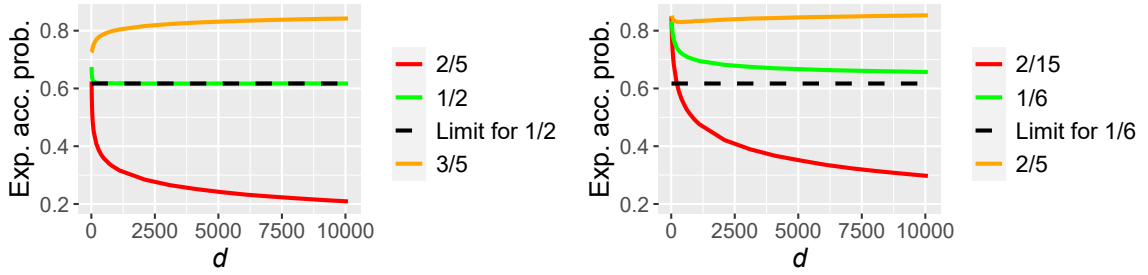
We finish this section with a discussion about the tuning of the parameter ℓ. For this part, we analyse another characteristic of the limiting stochastic process in Theorem 2. This stochastic process can be seen as a function of another process:

$$Z(t) = V(\ell^2 \vartheta_{w,\tau}(\ell) t), \tag{6}$$

where $\{V(t) : t \geq 0\}$ is the Langevin diffusion with a stochastic differential equation given by

$$\mathrm{d}V(t) = (\log \varphi(V(t)))'/2 \times \mathrm{d}t + \mathrm{d}B(t).$$

The term $\ell^2 \vartheta_{w,\tau}(\ell)$ in (6) is sometimes referred to as the *speed measure* of $\{Z(t) : t \geq 0\}$. From a MCMC perspective, the largest speed is best. Indeed, the stationary integrated autocorrelation time of *any* function h of the diffusion is proportional to the inverse of the diffusion speed. The following corollary presents the largest speeds and tuning procedures.



(a) Ideal scheme w. $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$ **(b) Ideal scheme w.** $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$

Figure 5: Expected acceptance probabilities as a function of d for: (a) the ideal scheme with $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$, $\ell = 1$, and $\tau = 2/5$, $\tau = 1/2$ and $\tau = 3/5$; (b) the ideal scheme with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$, $\ell = 2^{2/3}$, and $\tau = 2/15$, $\tau = 1/6$ and $\tau = 2/5$; the limiting expected acceptance probabilities for $\tau = 1/2$ and $\tau = 1/6$ are also presented, in the GB and LB cases, respectively; with the values used for ℓ, the limits are the same; the values for τ other than $1/2$ and $1/6$ have been obtained by increasing and decreasing these by 20%.

Corollary 1 *The speed measure when* $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ *and* $\tau = 1/6$*, given by* $2\ell^2\Phi(-\ell^3/2^3)$*, is maximized at* $\ell^* = 1.650$ *(to three decimal places), which yields a limiting expected acceptance probability of* $2\Phi(-(\ell^*)^3/2^3) = 0.574$ *(to three decimal places). The speed measure when* $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$ *and* $\tau = 1/2$*, given by* $2\ell^2\Phi(-\ell/2)$*, is maximized at* $\ell^{**} = 2.381$ *(to three decimal places), which yields a limiting expected acceptance probability of* $2\Phi(-\ell^{**}/2) = 0.234$ *(to three decimal places).*

This result highlights that: (i) the (asymptotically) optimal expected acceptance probability for the GB ideal scheme is the same as that for random-walk Metropolis, with the same maximum speed for the limiting diffusion (Roberts et al., 1997); (ii) the (asymptotically) optimal expected acceptance probability for the LB ideal scheme is the same as that for MALA, with the same maximum speed for the limiting diffusion (Roberts and Rosenthal, 1998). The latter is expected as MALA can be viewed as an approximation to the ideal scheme with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ (recall the discussion in Section 2.3), and the approximation is asymptotically exact if the step size in MALA diminishes adequately with d, as $d \to \infty$. The results in Corollary 1 can be obtained from Roberts et al. (1997) and Roberts and Rosenthal (1998).

Presenting the largest speed and a tuning procedure for the ideal scheme with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ and $\tau = 1/6$ in Corollary 1 is interesting as it allows to make that connection with MALA, but we will see in Section 4.2 that in order to take advantage of such a value for τ in MTM, one would need to use computational resource well beyond what is reasonable and realistic. We will see that, in MTM, it is more reasonable to use values around $\tau = 1/2$. With $\tau = 1/2$, the limiting diffusion of the ideal scheme with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ has a speed measure given by ℓ^2, which can be compared with that of the limiting diffusion of the ideal scheme with $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$ in Corollary 1, given by $2\ell^2\Phi(-\ell/2)$, because both schemes use the same form for the scale parameter; see Figure 6 for a comparison of the speed measures.

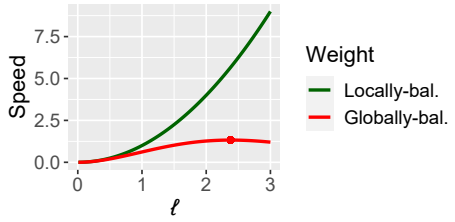


Figure 6: Speed measures as a function of ℓ of the limiting diffusions of the GB and LB ideal schemes when $\tau = 1/2$; the red point indicates the maximum speed when using the GB weight function.

Figure 6 suggests that MTM using $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ is at least as good as MTM using $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$, in high dimensions if it approximates well the ideal scheme; this is observed empirically in Section 4.3. Figure 6 also suggests to set ℓ in MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ as large as possible. However, in practice when sampling from target distributions having high but fixed dimensions d, ℓ has to be constrained to small values compared to d to reflect that it is held constant and thus does not grow with d in our asymptotic analysis. In our numerical experiments in Section 4.3, we observed that in moderate to high dimensions, with moderate to large values for N_d, optimally tuned MTM using $w(\mathbf{x}_d, \mathbf{y}_d) = g(\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d))$ with $g(x) = \sqrt{x}$ and $g(x) = x/(1+x)$ have acceptance rates in a range of 50% to 60%. We thus recommend to users to start their tuning procedures with a value of ℓ yielding an acceptance rate in that range (if they do not tune ℓ adaptively as in Section 3.2). Note that the algorithms in Section 4.3 were tuned using expected squared jumping distance (ESJD).

4.2 Characterizing the approximation of ideal schemes by MTM

In this section, we provide conditions on N_d under which MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \frac{\pi_d(\mathbf{y}_d)}{\pi_d(\mathbf{x}_d)}$ and MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ are asymptotically equivalent to their ideal counterparts as $d \to \infty$. The sense in which they are asymptotically equivalent implies a convergence of transformations of the Markov chains simulated by these MTM algorithms towards diffusions. We observed in the proof of Theorem 1 that, for an MTM algorithm to be asymptotically equivalent to its ideal counterpart, it is sufficient that: (i) the weight normalization in the proposal distribution of $\mathbf{Y}_J$ in MTM be asymptotically equivalent to the normalizing constant of Q_{w,σ_d} (recall Proposition 1 and (2)), and (ii) the acceptance probability in MTM be asymptotically equivalent to that in the ideal scheme. We present below a result stating conditions on N_d under which this holds. The result is established under the same scenario as in Section 4.1 (the target distribution is defined as in (3) and $q_{\sigma_d}(\mathbf{x}_d, \cdot) = \mathcal{N}(\mathbf{x}_d, \sigma_d^2 \mathbb{I}_d)$ with $\sigma_d = \ell/d^\tau$).

Before presenting the result, we introduce required notation. We use $\{\mathbf{X}_{d,\mathrm{MTM}}(m) : m \in \mathbb{N}\}$ to denote a Markov chain simulated by a MTM algorithm, and define a re-scaled continuous-time version $\{\mathbf{Z}_{d,\mathrm{MTM}}(t) : t \geq 0\}$ using:

$$\mathbf{Z}_{d,\mathrm{MTM}}(t) := \mathbf{X}_{d,\mathrm{MTM}}(\lfloor d^{2\tau} t \rfloor). \tag{7}$$

We use $\{Z_{d,\mathrm{MTM}}(t) : t \geq 0\}$ to denote the first component of $\{\mathbf{Z}_{d,\mathrm{MTM}}(t) : t \geq 0\}$.

We are now ready to present the result. It is about GB MTM and LB MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$.

Theorem 3 *Assume that π_d is as in (3) and that $q_{\sigma_d}(\mathbf{x}_d, \cdot) = \mathcal{N}(\mathbf{x}_d, \sigma_d^2 \mathbb{I}_d)$ with $\sigma_d = \ell/d^\tau$. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_d}$ be N_d conditionally independent random variables given $\mathbf{X}_d$, each distributed as $\mathbf{Y}_i \mid \mathbf{X}_d \sim q_{\sigma_d}(\mathbf{X}_d, \cdot)$ with $\mathbf{X}_d \sim \pi_d$. Let $\mathbf{Y}_J$ be a proposal sampled using MTM. Then, there exists a positive integer d_0 such that for any $d \geq d_0$,*

$$\mathbb{E}\left[d^{2\tau} \left| \frac{w(\mathbf{X}_d, \mathbf{Y}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d, \mathbf{Y}_i)} - \frac{w(\mathbf{X}_d, \mathbf{Y}_1)}{\mathbb{E}[w(\mathbf{X}_d, \mathbf{Y}_1) \mid \mathbf{X}_d]} \right| \right] \leq \frac{d^{2\tau}}{N_d^{1/2}}\, \varrho_1(d),$$

and

$$\mathbb{E}[d^{2\tau} |\alpha(\mathbf{X}_d, \mathbf{Y}_J) - \alpha_{ideal}(\mathbf{X}_d, \mathbf{Y}_J)|] \leq \frac{d^{2\tau}}{N_d^{1/2}}\, \varrho_2(d),$$

whenever $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ or $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$, ϱ_1 and ϱ_2 being functions of d that are explicitly defined in the proof. If $\tau \geq 1/2$, $\varrho_1(d)$ and $\varrho_2(d)$ converge to positive constants as $d \to \infty$; $N_d = d^{4\tau(1+\rho)}$ with ρ being any positive constant makes the expectations converge to 0. If $\tau < 1/2$, $\varrho_1(d)$ and $\varrho_2(d)$ grows with d and $N_d = (1+\nu)^d$ with ν being any positive constant makes the expectations converge to 0. When these expectations converge to 0, $\{Z_{d,MTM}(t) : t \geq 0\}$ converges weakly towards the same Langevin diffusion $\{Z(t) : t \geq 0\}$ as in Theorem 2, under the same conditions.

Theorem 3 indicates that, if one wanted to use a larger step size with $\tau = 1/6$, a sufficient condition for MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ to approximates well its ideal counterpart is to scale N_d exponentially with d. However, using such an enormous number of candidates is computationally prohibitive. When $\tau = 1/2$, Theorem 3 indicates that scaling N_d quadratically (essentially) with d is sufficient for both MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ and MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$ to approximate well their ideal counterparts. We note that these are only sufficient conditions, rather than necessary ones, and that the implied scalings may not be tight. In particular, taking N_d less than quadratic in d may be enough when $\tau = 1/2$ and in general more research is needed to establish optimal ways of scaling N_d with d. Nonetheless, the result suggests that, in high dimensions, LB MTM schemes will struggle to approximate their ideal counterparts when an aggressive step size with $\tau = 1/6$ is used. The numerical results in Section 4.3 complement this analysis: they show that in moderately high dimensions and with a moderately large number of candidates, optimally tuned MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ and optimally tuned MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$ have similar performance beyond the burn-in period, showing that MTM with $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ is not able to take advantage of a larger step size because with such a step size it does not approximate well the ideal scheme and does not have a good performance. This is in contrast with the results of Section 3 (e.g., Figures 3 and 4) where it is shown that even $N_d \ll d$ is sufficient to yield a significant improvement in terms of convergence speed.

4.3 Numerical experiments

In this section, we evaluate the empirical performance in stationarity of GB MTM and LB MTM under an non-asymptotic framework; the mathematical objects like the target distribution, the scale parameter and the number of candidates are thus denoted without a subscript d, that is π, σ and N. As previously, two LB MTM algorithms are evaluated: both use a weight function given by $w(\mathbf{x}, \mathbf{y}) = g(\pi(\mathbf{y})/\pi(\mathbf{x}))$; one uses $g(x) = \sqrt{x}$, and the other one, $g(x) = x/(1+x)$. The performance is evaluated using the Monte Carlo estimate of ESJD in stationarity, the latter being defined as

$$\text{ESJD} := \mathbb{E}\left[\|\mathbf{X}_{\text{MTM}}(m+1) - \mathbf{X}_{\text{MTM}}(m)\|^2\right] = \mathbb{E}\left[\|\mathbf{Y}_J - \mathbf{X}\|^2 \alpha(\mathbf{X}, \mathbf{Y}_J)\right],$$

where $\mathbf{X} \sim \pi$ and $\mathbf{Y}_J$ is sampled using the MTM mechanism. The performance evaluation is conducted under the same scenario as previously: the target distribution is defined as in (3), $d = 50$, and $q_\sigma(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$ with $\sigma = \ell/\sqrt{d}$. The results have been observed to be similar when the target is still a normal distribution but with components having different marginal variances and in higher dimensions. Note that in the scenario considered here, ESJD can be approximated efficiently using independent Monte Carlo sampling. The approximations are based on samples of size 100,000.

In Figure 7, we present results of ESJD as a function of N, for ℓ fixed, while in Figure 8, results of ESJD as a function of ℓ are presented, for several values of N. The value of ℓ used in Figure 7 corresponds to that which maximizes ESJD when $N = 5$; the same value of $\ell = 3.20$ is optimal for all algorithms (at least according to our grid search). The results in Figures 7 and 8 unveil that when N is small, all algorithms are essentially equivalent in stationarity, but they also unveil another problem with GB MTM. The algorithm behaviour changes drastically as N increases, to the extend that a performance reduction may be observed while keeping ℓ fixed. This is counter-intuitive and may be confusing to users that may diminish the value of ℓ to compensate, while the opposite is desirable. That pathological behaviour is not exhibited by LB MTM algorithms, which behave as one would expect, with a performance that increases monotonically with N, while keeping ℓ fixed. The results in Figure 8 also allow to notice that if a user manages to optimally tune GB MTM, then it is not significantly outperformed in stationarity by LB MTM for reasonable values of N, as mentioned previously.

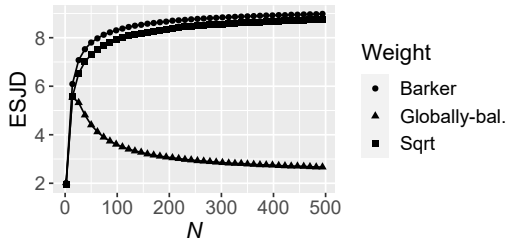
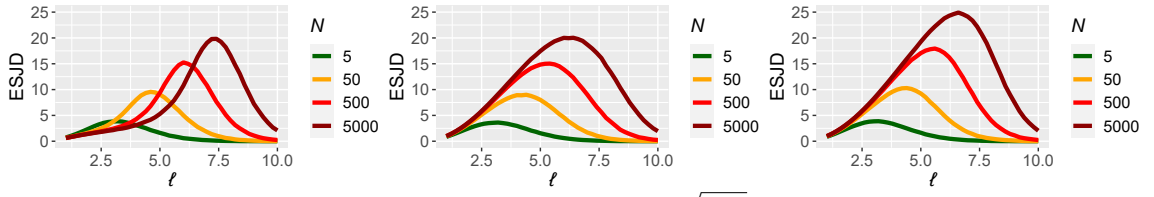


Figure 7: ESJD as a function of N when $d = 50$, $\ell = 3.20$ for MTM with $w(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$, MTM with $w(\mathbf{x}, \mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$, and MTM with $w(\mathbf{x}, \mathbf{y}) = (\pi(\mathbf{y})/\pi(\mathbf{x}))/(1 + \pi(\mathbf{y})/\pi(\mathbf{x}))$.



(a) MTM w. $w(\mathbf{x},\mathbf{y}) = \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}$ **(b) MTM w.** $w(\mathbf{x},\mathbf{y}) = \sqrt{\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}}$ **(c) MTM w.** $w(\mathbf{x},\mathbf{y}) = \frac{\pi(\mathbf{y})/\pi(\mathbf{x})}{1+\pi(\mathbf{y})/\pi(\mathbf{x})}$

Figure 8: ESJD as a function of ℓ when $d = 50$, for several values of N and: (a) MTM with $w(\mathbf{x},\mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$, (b) MTM with $w(\mathbf{x},\mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$, and (c) MTM with $w(\mathbf{x},\mathbf{y}) = (\pi(\mathbf{y})/\pi(\mathbf{x}))/(1+\pi(\mathbf{y})/\pi(\mathbf{x}))$.

5. Application of MTM for Bayesian inference in immunotherapy

We study in this section the application of MTM as an inference solution in a real-world context of immunotherapy in precision medicine involving a model with a likelihood function that is expensive to evaluate. The context is described in Section 5.1, the data and the model used to analyse them are presented in Section 5.2, and the application of MTM algorithms is studied in Section 5.3. Our study suggests that the scope of the results (empirical and theoretical) of the previous sections about GB MTM and LB MTM extends beyond the simple contexts in which they were derived. Note that in this section, we adopt a notation which is consistent with typical Bayesian-statistics contexts; it is thus different from that adopted in the other sections.

5.1 Context

Precision medicine is an innovative approach of care whereby patients are subject to personalized treatment strategies that take into account their personal data (genetics, lifestyle, health history, etc.). In Jenner et al. (2021), the authors study the effect of such personalized treatments on advanced-stage cancer patients based on cancer vaccines and oncolytic immunotherapy. Oncolytic viruses such as vesicular stomatitis virus (VSV) and vaccinia virus (VV) have the potential to destroy tumor cells and to induce a systemic anti-tumour immune response and, as such, can lead to what is commonly referred to as *virotherapy*. At the moment, several related open questions form a very active strand of research. For instance, one would like to assess the potential benefits offered by combining different oncolytic viruses or the use of virus enhancers. Even though it is not restricted to cancer treatments, being able to find the optimal treatment schedule for a given patient is a central question in precision medicine. It is believed that, provided that these questions can be addressed, oncolytic virotherapy will form a major breakthrough in cancer treatment. Statistical modelling and Bayesian inference represent a way to help answering those questions. The reliability and efficiency of the numerical methods leading to the inference is thus of crucial importance. In Section 5.3, we evaluate the performance of different MTM algorithms.

5.2 Data and model

In practice, data are collected from a cohort of K advanced-stage cancer patients that are examined at T time points. At each time point $t \in \{1, \ldots, T\}$, the state of patient $k \in \{1, \ldots, K\}$ is summarized through statistics $\mathbf{y}_k(t) \in \mathbb{R}^{m_1}$, with $m_1 \in \mathbb{N}$. These statistics represent the variables of interest. Covariate data points $\mathbf{x}_k \in \mathbb{R}^{m_2}$ with $m_2 \in \mathbb{N}$, independent of time but associated to a patient, are also collected. Each patient is assigned a personalized treatment schedule r_k which is considered as a data point from a categorical variable. The covariate data points and the personalized treatment schedules are considered to be fixed and known, in contrast to $\mathbf{y}_k(t)$ which is assumed in a statistical model to be a realization of a random variable. The assumed statistical model is a *forward model*:

$$\mathbf{Y}_k(t) = \hat{\mathbf{y}}(t, \boldsymbol{\theta}, \mathbf{x}_k, r_k) + \sigma\, \boldsymbol{\varepsilon}_k(t)\,, \tag{8}$$

where $\hat{\mathbf{y}}(t, \boldsymbol{\theta}, \mathbf{x}_k, r_k)$ is the output of a dynamical system proposed in Jenner et al. (2021) and described below, $\sigma > 0$ is a scale parameter, and $\boldsymbol{\varepsilon}_1(1), \ldots, \boldsymbol{\varepsilon}_1(T), \ldots, \boldsymbol{\varepsilon}_K(1), \ldots, \boldsymbol{\varepsilon}_K(T) \in \mathbb{R}^{m_1}$ are random standardized errors. The scale parameter is considered here fixed and known to simplify. The unknown parameter is $\boldsymbol{\theta} \in \mathbb{R}^d$ with $d = 14$. We assume that $\boldsymbol{\varepsilon}_k(t) \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_{m_1})$, $t = 1, \ldots, T$ and $k = 1, \ldots, K$, are IID random variables independent of $\boldsymbol{\theta}$, which is a common assumption in the literature.

The output $\hat{\mathbf{y}}(t, \boldsymbol{\theta}, \mathbf{x}_k, r_k)$ is produced given $(t, \boldsymbol{\theta}, \mathbf{x}_k, r_k)$ from a discretized version of the numerical solution of a system of time-delay differential equations (DDE):

$$\frac{\mathrm{d}\mathbf{y}_k(t)}{\mathrm{d}t} = \Phi_{\boldsymbol{\theta}}(t, \mathbf{y}_k(t), \mathbf{y}_k^-(t), \mathbf{x}_k, r_k)\,, \tag{9}$$

where $\Phi_{\boldsymbol{\theta}}$ is a differential operator, parameterized by $\boldsymbol{\theta}$. The notation $\mathbf{y}_k^-(t)$ refers to $\{\mathbf{y}_k(t')\,, t' < t\}$. Time-delay differential equations are commonly used to model dynamical systems of interest in fields such as epidemiology and demography. More details on the operator $\Phi_{\boldsymbol{\theta}}$ used can be found in Jenner et al. (2021). In particular, the parameter $\boldsymbol{\theta}$ (see Table TS3 in Jenner et al. (2021)) consists essentially of a logarithmic transformation of biological rates such as rates at which certain cell cycles occur, infection rates of VV and VSV, etc. It is important to stress that (9) cannot be solved exactly but that a computationally-intensive numerical solver (which requires solving a system of intermediate ordinary differential equations obtained by the so-called *linear chain technique*) exists and is made available in Jenner et al. (2021). This solver needs to run at each evaluation of the likelihood function.

The work of Jenner et al. (2021) is in a context of optimization of treatment schedule, and is based on a virtual cohort. The virtual patients are created from: 1) simulated covariate data points $\mathbf{x}_k$ with summary statistics similar to real cohorts, 2) a treatment schedule r_k that is assigned to each virtual patient, and 3) outputs $\hat{\mathbf{y}}(t, \boldsymbol{\theta}, \mathbf{x}_k, r_k)$ resulting from the numerical solution of (9). The parameter $\boldsymbol{\theta}$ used to produce the outputs $\hat{\mathbf{y}}(t, \boldsymbol{\theta}, \mathbf{x}_k, r_k)$ is set to a value $\boldsymbol{\theta}^*$ based on expert opinion. Here the problem that we consider is that of numerical estimation of the unknown parameter $\boldsymbol{\theta}$ in the model (8) based on a data set.

Our data set has been simulated from a virtual cohort (generated as in Jenner et al. (2021)) that is then fed into the model defined in (8). The simulated data set consists of a virtual cohort of $K = 10$ patients, where each patient is examined once per week for

$T = 20$ weeks. At each examination, $m_1 = 4$ statistics are measured; they are described in Figure 9. These data were simulated using the expert opinion $\boldsymbol{\theta}^*$. Having knowledge about the parameter (that is considered unknown in the Bayesian analysis) helps to evaluate the reliability of the different MTM algorithms.

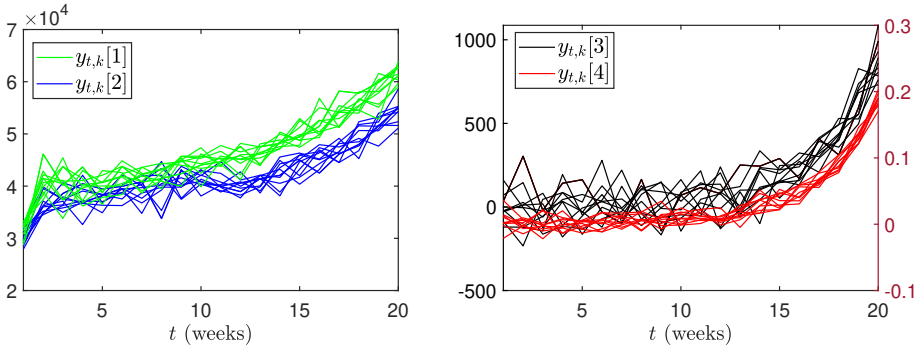


Figure 9: Data $\{\mathbf{y}_k(t)\} := \{y_{t,k}[1], y_{t,k}[2], y_{t,k}[3], y_{t,k}[4]\}$ for the cohort, simulated using (8); $y_{t,k}[1]$ is related to the number of quiescent tumor cells at time t for individual k, $y_{t,k}[2]$ is related to the number of G1-phase tumour cell population, $y_{t,k}[3]$ is related to the total infected cell population, $y_{t,k}[4]$ is related to the total virus load.

5.3 Application study of MTM for Bayesian inference

The goal in practice is to perform Bayesian inference for $\boldsymbol{\theta}$, given the data related to the cohort of patients. We here discuss how one can proceed and present a study of the reliability of MTM algorithms.

A non-informative truncated Gaussian prior distribution is assigned to $\boldsymbol{\theta}$. The truncation stems from that the DDE solver is numerically unstable when the parameter value is beyond a compact set $\boldsymbol{\Theta}$, which includes $\boldsymbol{\theta}^*$. It can be readily checked that the posterior distribution verifies

$$\pi(\boldsymbol{\theta} \,|\, \{\mathbf{y}_k(t)\}) \propto \bar{\pi}(\boldsymbol{\theta} \,|\, \{\mathbf{y}_k(t)\}) := \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t,k} (\mathbf{y}_k(t) - \hat{\mathbf{y}}(t, \boldsymbol{\theta}, \mathbf{x}_k, r_k))^2 - \frac{1}{2} \|\boldsymbol{\theta} - \boldsymbol{\mu}\|^2 \right\} \mathbb{1}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \,,$$

where $\bar{\pi}$ is the unnormalized version of the posterior density and $\boldsymbol{\mu} \in \mathbb{R}^d$ is a prior hyper-parameter. Because of the terms $\hat{\mathbf{y}}(t, \boldsymbol{\theta}, \mathbf{x}_k, r_k)$, posterior expectations cannot be derived in closed form for non-trivial observables and IID sampling from π is virtually impossible. Moreover, assuming that π is differentiable, the gradient of π is not available explicitly and is computationally expensive to approximate. This makes standard gradient-based methods, such as MALA or Hamiltonian Monte Carlo, computationally intensive and not straightforward to implement. Also, the posterior distribution exhibits an irregular and complex behaviour; see, e.g., the strong correlations and truncation effect in Figure 10. This makes posterior-approximation methods requiring a tractable approximation to the posterior distribution, such as importance sampling, independent MH, as well as default variational methods, highly non-trivial to apply successfully.

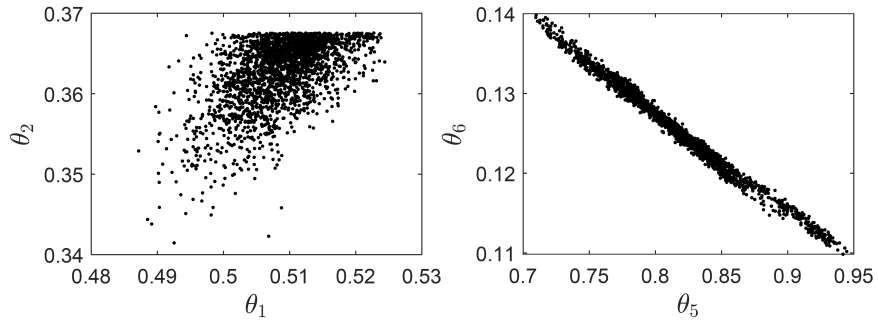


Figure 10: Pairwise marginal samples from π.

To perform Bayesian inference in such a situation, one may thus naturally turn to a random-walk Metropolis algorithm. If parallel computing is available, LB MTM with $w(\mathbf{x}, \mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$ is an appealing alternative as it can exploit parallel computing to speed up convergence. The non-trivial shape of the posterior distribution provides an interesting test case for the results derived in the previous sections with regular and isotropic target distributions. Here, we compare MH with LB MTM and GB MTM using different values of N. All samplers use the proposal $q_\sigma(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$, with $\sigma = \ell/\sqrt{d}$, and the parameter ℓ is adaptively tuned as in Section 3.2.

Figure 11 shows the convergence to stationarity of the different algorithms, monitored through the log-posterior-density value (up to a normalizing constant). The right panel of Figure 11 displays the evolution of the step-size parameter $\{\ell_m\}$ across MCMC iterations. We clearly observe the pathology of GB MTM described previously, with a performance that deteriorates as N increases. In particular, it can be seen that GB MTM requires extremely small step sizes to navigate the low density regions (see Figure 11, right panel).

LB MTM instead exhibits a convergence speed that increases with N, as one expects. In particular, in this example, LB MTM with $N = 20$ converges significantly faster compared to $N = 1$, roughly by a factor of 7 based on a quantitative comparison of the trace plots in Figure 11, left panel. With our parallel implementation[8], the computational cost per iteration of MTM with $N = 20$ is about 3.5 times higher than MH, resulting in an effective reduction of wall-clock time required for burn-in by a factor of roughly 2. Note that the actual improvement ratio can depend heavily on the model and parallel implementation used, with the improvement typically larger for higher-dimensional problems.

6. Discussion

In this paper, we revisited the promises and pitfalls of a popular MCMC method, namely MTM, through several new theoretical and empirical results. We proposed to use a novel class of weight functions, based on the so-called locally-balanced proposal distributions, that can be employed instead of the classical GB weight function without impacting neither the computational cost nor the coding complexity of MTM. The resulting LB MTM scheme

8. In our experiment, we used a desktop computer with 32 cores (thus larger than the number of candidates N), AMD Ryzen 9 5950X processor, Alma Linux 8.5 operating system, 64 GB of RAM, and off-the-shelf high-level MATLAB parallelization.

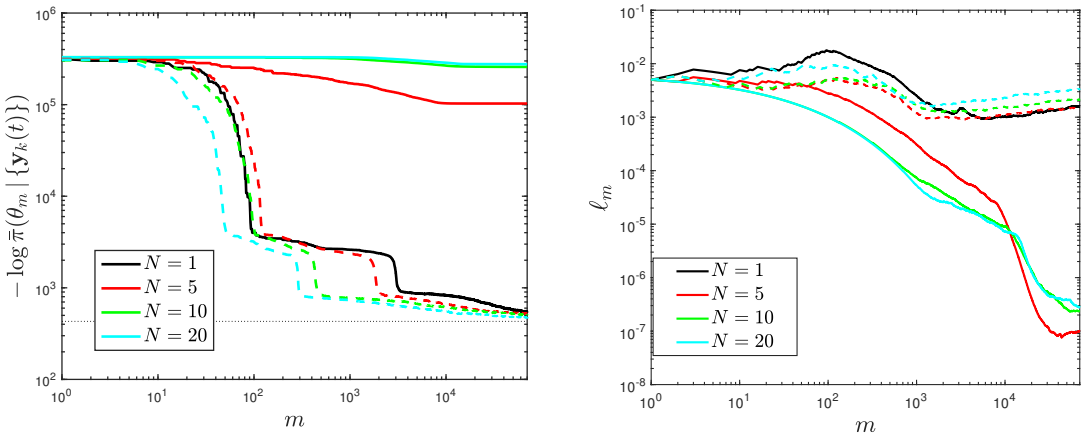


Figure 11: Convergence of GB MTM (solid lines) and LB MTM (dashed lines), with a log-scale on the x-axis; all chains are started from the same state $\boldsymbol{\theta}_0$ belonging to the tails of π; left panel: trace plots of $-\log \bar{\pi}(\boldsymbol{\theta}_m \mid \{\mathbf{y}_k(t)\})$ indicating the progress towards the high-probability region (designated by the dotted black line); right panel: adaptation of the step-size parameter ℓ as the algorithms progress.

is remarkably and positively different compared to the GB MTM one, especially regarding the behaviour of the induced Markov chains during the convergence phase. This difference is associated with substantially reduced burn-in time and an element of stability that, together with an easier tuning of the method, make LB MTM an appealing and competitive algorithm for high-dimensional Bayesian inference problems for which nothing except the unnormalized posterior density is known and when the latter is computationally expensive to evaluate, which makes in-step parallel computing beneficial.

Part of the research effort conducted in this paper can also be cast as an attempt to generalize the use of LB samplers beyond the discrete-state-space scenario. In particular, LB MTM is similar in spirit to exact-approximation schemes, in the sense of Andrieu and Vihola (2016), given that it can be thought of as approximations of ideal LB samplers that use a noisy version of the acceptance probability without affecting the invariant distribution. Further work in this direction is needed to explore these connections to strands of literature. Also, the results in this paper motivate further non-asymptotic (in N and d) theoretical analyses to identify other quantitative or qualitative differences between LB MTM algorithms and their GB counterpart. The rich literature on non-asymptotic analysis of MH exact-approximations (see, e.g., Andrieu and Vihola (2016), Andrieu et al. (2020), and Wang (2022)) may prove useful in this direction.

While we have focused mainly on MTM and the impact of the weight function, it would be interesting to provide a systematic comparison of LB MTM with other multiple-proposal MCMC schemes (e.g., Neal (2003); Tjelmeland (2004); Frenkel (2004); Calderhead (2014); Holbrook (2023) and references therein) and more generally to other approaches that could in principle exploit in-step parallelization to speed-up convergence of MCMC. One such approach would be to exploit parallel target-density evaluations to derive numerical (e.g.

finite-difference) approximations to the gradient, $\nabla \log \pi$, and then employ off-the-shelf gradient-based MCMC methods. We leave the exploration of such comparisons for future work, noting that it is likely that the optimal scheme among the ones mentioned above will depend on features of the target distribution (such as dimensionality, smoothness and regularity), and on the parallel-computing environment available (see, e.g., Glatt-Holtz et al. (2022) for examples of GPU implementations of multiple-proposal MCMC methods with large values of N) as well as potentially other aspects.

Acknowledgments

Philippe Gagnon acknowledges support from NSERC (Natural Sciences and Engineering Research Council of Canada) and FRQNT (Fonds de recherche du Québec – Nature et technologies). Florian Maire acknowledges support from NSERC. Giacomo Zanella acknowledges support from the European Research Council (ERC), through StG "PrSc-HDBayLe" grant ID 101076564. Also, the authors thank three anonymous referees for helpful suggestions that led to an improved manuscript. The authors additionally thank Morgan Craig, professor in the Department of Mathematics and Statistics of Université de Montréal, for an introduction to the use of dynamical systems in precision medicine and for computer code to numerically solve the system of intermediate ordinary differential equations in the application of Section 5.

References

Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Stat. Comput.*, 18(4):343–373, 2008.

Christophe Andrieu and Matti Vihola. Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.*, 26(5):2661–2696, 2016.

Christophe Andrieu, Sinan Yıldırım, Arnaud Doucet, and Nicolas Chopin. Metropolis–Hastings with averaged acceptance ratios. *arXiv:2101.01253*, 2020.

Av A Barker. Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Austral. J. Phys.*, 18(2):119–134, 1965.

Mylene Bédard. Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.*, 17(4):1222–1244, 2007.

Mylène Bédard, Randal Douc, and Eric Moulines. Scaling analysis of multiple-try MCMC methods. *Stochastic Process. Appl.*, 122(3):758–786, 2012.

Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA J. Numer. Anal.*, 33(1):80–110, 2013.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Ben Calderhead. A general construction for parallelizing Metropolis–Hastings algorithms. *Proceedings of the National Academy of Sciences*, 111(49):17408–17413, 2014.

Hyunwoong Chang, Changwoo J Lee, Zhao Tang Luo, Huiyan Sang, and Quan Zhou. Rapidly mixing multiple-try Metropolis algorithms for model selection problems. *Advances in Neural Information Processing Systems*, 36, 2022.

Alain Durmus, Sylvain Le Corff, Eric Moulines, and Gareth O. Roberts. Optimal scaling of the random walk Metropolis algorithm under l^p mean differentiability. *J. Appl. Probab.*, 54(4):1233–1260, 2017.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis–Hastings algorithms are fast! In *Conference on learning theory*, pages 793–797. PMLR, 2018.

Stewart N Ethier and Thomas G Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, 1986.

Daan Frenkel. Speed-up of Monte Carlo simulations by sampling of rejected states. *Proceedings of the National Academy of Sciences*, 101(51):17571–17575, 2004.

Philippe Gagnon. Informed reversible jump algorithms. *Electron. J. Stat.*, 15(2):3951–3995, 2021.

Philippe Gagnon and Florian Maire. An asymptotic Peskun ordering and its application to lifted samplers. *arXiv:2003.05492*, 2020.

Philippe Gagnon, Mylène Bédard, and Alain Desgagné. Weak convergence and optimal tuning of the reversible jump algorithm. *Math. Comput. Simulation*, 161:32–51, 2019.

Nathan E Glatt-Holtz, Andrew J Holbrook, Justin A Krometis, and Cecilia F Mondaini. Parallel MCMC algorithms: Theoretical foundations, algorithm design, case studies. *arXiv:2209.04750*, 2022.

W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Max Hird, Samuel Livingstone, and Giacomo Zanella. A fresh take on 'Barker dynamics' for MCMC. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 169–184. Springer, 2022.

Andrew J Holbrook. Generating MCMC proposals by randomly rotating the regular simplex. *J. Multivariate Anal.*, 194:1–15, 2023.

Jonathan Hunter Huggins. An information-theoretic analysis of resampling in sequential Monte Carlo. Master's thesis, Massachusetts Institute of Technology, 2014.

Pierre E Jacob, John O'Leary, and Yves F Atchadé. Unbiased Markov chain Monte Carlo methods with couplings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 82(3):543–600, 2020.

Adrianne L Jenner, Tyler Cassidy, Katia Belaid, Marie-Claude Bourgeois-Daigneault, and Morgan Craig. In silico trials predict that combination strategies for enhancing vesicular stomatitis oncolytic virus are determined by tumor aggressivity. *Journal for immunotherapy of cancer*, 9(2):1–13, 2021.

Alan F Karr. Weak convergence of a sequence of Markov chains. *Z. Wahrsch. Verw. Gebiete*, 33(1):41–48, 1975.

Xitong Liang, Samuel Livingstone, and Jim Griffin. Adaptive random neighbourhood informed Markov chain Monte Carlo for high-dimensional Bayesian variable selection. *Stat. Comput.*, 32(5):1–52, 2022.

Jun S Liu, Faming Liang, and Wing Hung Wong. The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Assoc.*, 95(449):121–134, 2000.

Samuel Livingstone and Giacomo Zanella. The Barker proposal: combining robustness and efficiency in gradient-based MCMC. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 84(2): 496–523, 2022.

L Martino and F Louzada. Issues in the multiple try Metropolis mixing. *Comput. Statist.*, 32(1):239–252, 2017.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

Alfred Müller and Dietrich Stoyan. *Comparison methods for stochastic models and risks*. John Wiley & Sons Ltd., Chichester, 2002.

Radford M Neal. Markov chain sampling for non-linear state space models using embedded hidden markov models. *arXiv preprint math/0305039*, 2003.

Silvia Pandolfi, Francesco Bartolucci, and Nial Friel. A generalization of the Multiple-try Metropolis algorithm for Bayesian estimation and model selection. In *International Conference on Artificial Intelligence and Statistics*, pages 581–588, 2010.

Samuel Power and Jacob Vorstrup Goldman. Accelerated sampling on discrete spaces with non-reversible markov processes. *arXiv preprint arXiv:1912.04681*, 2019.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 60(1):255–268, 1998.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Gareth O Roberts, Andrew Gelman, and Walter R Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 1997.

Jeffrey S Rosenthal. Parallel computing and Monte Carlo algorithms. *Far East J. Theor. Stat.*, 4(2):207–236, 2000.

Emanuele Sansone. LSB: Local self-balancing MCMC in discrete spaces. In *International Conference on Machine Learning*, pages 19205–19220. PMLR, 2022.

S M Schmon and P Gagnon. Optimal scaling of random walk Metropolis algorithms using Bayesian large-sample asymptotics. *Stat. Comput.*, 32(2):1–16, 2022.

S M Schmon, G Deligiannidis, A Doucet, and M K Pitt. Large-sample asymptotics of the pseudo-marginal method. *Biometrika*, 108(1):37–51, 2021.

Haoran Sun, Hanjun Dai, Wei Xia, and Arun Ramamurthy. Path auxiliary proposal for MCMC in discrete space. In *International Conference on Learning Representations*, 2021.

Haoran Sun, Hanjun Dai, Bo Dai, Haomin Zhou, and Dale Schuurmans. Discrete Langevin sampler via Wasserstein gradient flow. *arXiv:2206.14897*, 2022a.

Haoran Sun, Hanjun Dai, and Dale Schuurmans. Optimal scaling for locally balanced proposals in discrete spaces. *arXiv:2209.08183*, 2022b.

Hakon Tjelmeland. Using all Metropolis–Hastings proposals to estimate mean values. Technical report, 2004.

Jure Vogrinc, Samuel Livingstone, and Giacomo Zanella. Optimal design of the Barker proposal and other locally-balanced Metropolis–Hastings algorithms. *To appear in Biometrika*, 2023.

Guanyang Wang. On the theoretical properties of the Exchange algorithm. *Bernoulli*, 28 (3):1935–1960, 2022.

Giacomo Zanella. Informed proposals for local MCMC in discrete spaces. *J. Amer. Statist. Assoc.*, 115(530):852–865, 2020.

Quan Zhou and Aaron Smith. Rapid convergence of informed importance tempering. In *International Conference on Artificial Intelligence and Statistics*, pages 10939–10965. PMLR, 2022.

Appendix A. Proofs

Proof [Proposition 1] First, let us look at

$$\mathbb{P}_{\mathbf{x}}(J = j \mid \mathbf{Y}_1, \dots, \mathbf{Y}_N) = \mathbb{E}_{\mathbf{x}}[\mathbb{1}_{J=j} \mid \mathbf{Y}_1, \dots, \mathbf{Y}_N] = \frac{w(\mathbf{x}, \mathbf{Y}_j)}{\sum_{i=1}^N w(\mathbf{x}, \mathbf{Y}_i)}.$$

Using this and that $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ are conditionally IID given $\mathbf{x}$, we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}[h(\mathbf{Y}_J)] = \sum_{j=1}^N \mathbb{E}_{\mathbf{x}}[h(\mathbf{Y}_j)\, \mathbb{1}_{J=j}] &= \sum_{j=1}^N \mathbb{E}_{\mathbf{x}}[h(\mathbf{Y}_j)\, \mathbb{E}[\mathbb{1}_{J=j} \mid \mathbf{Y}_1, \dots, \mathbf{Y}_N]] \\
&= \sum_{j=1}^N \mathbb{E}_{\mathbf{x}}\left[h(\mathbf{Y}_j)\, \frac{w(\mathbf{x}, \mathbf{Y}_j)}{\sum_{i=1}^N w(\mathbf{x}, \mathbf{Y}_i)}\right] \\
&= \sum_{j=1}^N \int h(\mathbf{y}_j)\, \frac{w(\mathbf{x}, \mathbf{y}_j)}{\sum_{i=1}^N w(\mathbf{x}, \mathbf{y}_i)} \prod_{i=1}^N q_{\sigma_d}(\mathbf{x}, \mathbf{y}_i)\, \mathrm{d}\mathbf{y}_{1:N} \\
&= \int h(\mathbf{y}_1)\, \frac{w(\mathbf{x}, \mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{x}, \mathbf{y}_i)} \prod_{i=1}^N q_{\sigma_d}(\mathbf{x}, \mathbf{y}_i)\, \mathrm{d}\mathbf{y}_{1:N}.
\end{aligned}$$

■

A.1 Proof of Theorem 1

In order to prove Theorem 1, we provide a general convergence result.

Theorem 4 *Let $\{P_N : N \geq 1\}$ be a collection of Markov transition kernels and P a Markov transition kernel such that P_N (for any N) and P admit π as invariant distribution. If the following assumptions hold:*

(a) *the Markov transition kernels $\{P_N : N \geq 1\}$ satisfy*

$$\int |P_N h(\mathbf{x}) - Ph(\mathbf{x})|\, \pi(\mathbf{x})\, \mathrm{d}\mathbf{x} \to 0$$

as $N \to \infty$ for all $h \in \mathcal{C}_b$, where

$$P_N h(\mathbf{x}) := \int P_N(\mathbf{x}, \mathbf{y})\, h(\mathbf{y})\, \mathrm{d}\mathbf{y},$$

with an analogous definition for Ph, $\mathcal{C}_b$ being the space of bounded continuous functions;

(b) *Ph is continuous for any $h \in \mathcal{C}_b$,*

then $\{\mathbf{X}_N(m) : m \in \mathbb{N}\}$ converges weakly to $\{\mathbf{X}(m) : m \in \mathbb{N}\}$ provided that $\mathbf{X}_N(0) \sim \pi$ and $\mathbf{X}(0) \sim \pi$, $\{\mathbf{X}_N(m) : m \in \mathbb{N}\}$ and $\{\mathbf{X}(m) : m \in \mathbb{N}\}$ being the Markov chains with transition kernels P_N and P, respectively.

We present a proof here for self-containedness, but do not claim the originality of the result. The proof is strongly inspired by that of Theorem 2 in Schmon et al. (2021).

Proof To prove that $\{\mathbf{X}_N(m) : m \in \mathbb{N}\}$ converges weakly to $\{\mathbf{X}(m) : m \in \mathbb{N}\}$, we only need to prove the convergence of finite-dimensional distributions (Karr, 1975). It thus suffices to prove that for any positive integer k

$$|\mathbb{E}[f_0(\mathbf{X}_N(0)) \dots f_k(\mathbf{X}_N(k))] - \mathbb{E}[f_0(\mathbf{X}(0)) \dots f_k(\mathbf{X}(k))]| \to 0,$$

as $N \to \infty$, for all $f_0, \dots, f_k \in \mathcal{C}_\mathrm{b}$ given that $(\mathbf{x}(0), \dots, \mathbf{x}(k)) \mapsto \prod_{i=0}^{k} f_i(\mathbf{x}(i))$ is measure determining by Proposition 4.6 in Chapter 2 of Ethier and Kurtz (1986).

We prove this by induction. For $k = 0$, the convergence is trivial because $\mathbf{X}_N(0) \sim \pi$ and $\mathbf{X}(0) \sim \pi$. Now assume that it is true for $k \geq 0$, and let us verify that it is true for $k + 1$.

$$\begin{aligned}
&|\mathbb{E}[f_0(\mathbf{X}_N(0)) \dots f_k(\mathbf{X}_N(k)) f_{k+1}(\mathbf{X}_N(k+1))] - \mathbb{E}[f_0(\mathbf{X}(0)) \dots f_k(\mathbf{X}(k)) f_{k+1}(\mathbf{X}(k+1))]| \\
&= |\mathbb{E}[f_0(\mathbf{X}_N(0)) \dots f_k(\mathbf{X}_N(k)) P_N f_{k+1}(\mathbf{X}_N(k))] - \mathbb{E}[f_0(\mathbf{X}(0)) \dots f_k(\mathbf{X}(k)) P f_{k+1}(\mathbf{X}(k))]| \\
&\leq |\mathbb{E}[f_0(\mathbf{X}_N(0)) \dots f_k(\mathbf{X}_N(k)) P_N f_{k+1}(\mathbf{X}_N(k)) - f_0(\mathbf{X}_N(0)) \dots f_k(\mathbf{X}_N(k)) P f_{k+1}(\mathbf{X}_N(k))]| \\
&\quad + |\mathbb{E}[f_0(\mathbf{X}_N(0)) \dots f_k(\mathbf{X}_N(k)) P f_{k+1}(\mathbf{X}_N(k))] - \mathbb{E}[f_0(\mathbf{X}(0)) \dots f_k(\mathbf{X}(k)) P f_{k+1}(\mathbf{X}(k))]| \\
&\leq M^k \mathbb{E}[|P_N f_{k+1}(\mathbf{X}_N(k)) - P f_{k+1}(\mathbf{X}_N(k))|] \\
&\quad + |\mathbb{E}[f_0(\mathbf{X}_N(0)) \dots f_k(\mathbf{X}_N(k)) P f_{k+1}(\mathbf{X}_N(k))] - \mathbb{E}[f_0(\mathbf{X}(0)) \dots f_k(\mathbf{X}(k)) P f_{k+1}(\mathbf{X}(k))]|,
\end{aligned}$$

using that there exists a positive constant M such that $f_i \leq M$ for all i. The term on the penultimate line vanishes as a consequence of Assumption (a). That on the last line vanishes because Pf_{k+1} is bounded and continuous by Assumption (b). ■

Before presenting the proof of Theorem 1, we present a result that will be used in it and in other proofs.

Proposition 4 *(Müller and Stoyan, 2002, Corollary 1.5.24) For any $N \geq 2$ exchangeable random variables $X_1, \ldots, X_N$ and any convex function ϕ, we have*

$$\mathbb{E}\left[\phi\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)\right] \leq \mathbb{E}\left[\phi\left(\frac{1}{N-1}\sum_{i=1}^{N-1} X_i\right)\right],$$

whenever the expectations exist.

Proof [Theorem 1] We start with Result 1. The strategy to prove the result is the same as that to prove Theorem 2.1.1 in Huggins (2014). The probability that $\mathbf{Y}_J$ belongs to a set A, for fixed $\mathbf{x}$, as a function of N is

$$Q^N_{w,\sigma}(\mathbf{x}, A) := \mathbb{P}_{\mathbf{x},N}(\mathbf{Y}_J \in A) = \mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_{\mathbf{Y}_1 \in A}\, \frac{w(\mathbf{x}, \mathbf{Y}_1)}{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{x}, \mathbf{Y}_i)}\right],$$

using Proposition 1. To simplify the notation for this part of the proof, we omit the dependence on w, σ and $\mathbf{x}$ that are fixed throughout, and use $Q^N(A) := Q^N_{w,\sigma}(\mathbf{x}, A)$ and $Q(A) := Q_{w,\sigma}(\mathbf{x}, A)$.

We have that

$$Q^N(A) = \mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_{\mathbf{Y}_1 \in A}\, w(\mathbf{x}, \mathbf{Y}_1)\, \mathbb{E}_{\mathbf{x}}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{x}, \mathbf{Y}_i)} \mid \mathbf{Y}_1\right]\right].$$

With

$$h(\mathbf{Y}_1) := \mathbb{E}_{\mathbf{x}}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{x}, \mathbf{Y}_i)} \mid \mathbf{Y}_1\right],$$

we have that

$$\begin{aligned} Q^N(A) = \mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_{\mathbf{Y}_1 \in A}\, w(\mathbf{x}, \mathbf{Y}_1)\, h(\mathbf{Y}_1)\right] &= \int_A w(\mathbf{x}, \mathbf{y}_1)\, h(\mathbf{y}_1)\, q_\sigma(\mathbf{x}, \mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1 \\ &= \int_A \mathbb{E}_{\mathbf{x}}[w(\mathbf{x}, \mathbf{Y}_1)]\, h(\mathbf{y}_1)\, Q_{w,\sigma}(\mathbf{x}, \mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1. \end{aligned}$$

This implies that we have an expression for the following Radon–Nikodym derivative:

$$\begin{aligned} \frac{\mathrm{d}Q^N}{\mathrm{d}Q}(\mathbf{y}_1) &= \mathbb{E}_{\mathbf{x}}[w(\mathbf{x}, \mathbf{Y}_1)]\mathbb{E}_{\mathbf{x}}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{x}, \mathbf{Y}_i)} \mid \mathbf{Y}_1 = \mathbf{y}_1\right] \\ &\geq \frac{\mathbb{E}_{\mathbf{x}}[w(\mathbf{x}, \mathbf{Y}_1)]}{\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{x}}[w(\mathbf{x}, \mathbf{Y}_i) \mid \mathbf{Y}_1 = \mathbf{y}_1]} \\ &= \frac{N\mathbb{E}_{\mathbf{x}}[w(\mathbf{x}, \mathbf{Y}_1)]}{(N-1)\mathbb{E}_{\mathbf{x}}[w(\mathbf{x}, \mathbf{Y}_1)] + w(\mathbf{x}, \mathbf{y}_1)}, \end{aligned}$$

using Jensen's inequality.

To prove Result 1, we use that the total variation between $Q^N(A)$ and $Q(A)$ can be bounded above using the Kullback–Leibler divergence, and more precisely, by

$$\sqrt{\frac{1}{2}\int Q(\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1 \log\frac{\mathrm{d}Q}{\mathrm{d}Q^N}(\mathbf{y}_1)}.$$

We have that

$$\begin{aligned}
\int Q(\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1 \log\frac{\mathrm{d}Q(\mathbf{y}_1)}{\mathrm{d}Q^N(\mathbf{y}_1)} &\leq \int Q(\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1 \log\frac{(N-1)\mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]+w(\mathbf{x},\mathbf{y}_1)}{N\mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]}\\
&\leq \log\int Q(\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1 \frac{(N-1)\mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]+w(\mathbf{x},\mathbf{y}_1)}{N\mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]}\\
&= \log\left(1+\frac{\int w(\mathbf{x},\mathbf{y}_1)Q(\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1 - \mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]}{N\mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]}\right)\\
&= \log\left(1+\frac{\int w(\mathbf{x},\mathbf{y}_1)^2 q_\sigma(\mathbf{x},\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1 - \mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]^2}{N\mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]^2}\right)\\
&\leq \frac{\mathrm{var}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]}{N\mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]^2},
\end{aligned}$$

using Jensen's inequality and that $\log(1+x)\leq x$ for all $x>-1$. This concludes the proof of Result 1 as, by assumption $\mathbb{E}_{\mathbf{x}}\left[w(\mathbf{x},\mathbf{Y}_1)^4\right]<\infty$, and $\mathbb{E}_{\mathbf{x}}[w(\mathbf{x},\mathbf{Y}_1)]>0$ given that w is a strictly positive function.

We now prove Result 2. To achieve this, we use Theorem 4, which requires that:

(a) for every $h\in\mathcal{C}_{\mathrm{b}}$ (the space of bounded continuous functions),

$$\int |P_N h(\mathbf{x}) - P_{\mathrm{ideal}}h(\mathbf{x})|\,\pi(\mathbf{x})\,\mathrm{d}\mathbf{x}\to 0,$$

where P_N is the Markov kernel simulated by MTM and P_{ideal} is the Markov kernel simulated by the ideal scheme;

(b) $P_{\mathrm{ideal}}h$ is continuous for any $h\in\mathcal{C}_{\mathrm{b}}$.

Let us first define P_N and P_{ideal}:

$$\begin{aligned}
P_N(\mathbf{x},A) := &\sum_{j=1}^N \int_{\mathbf{y}_j\in A} \frac{w(\mathbf{x},\mathbf{y}_j)}{\sum_{i=1}^N w(\mathbf{x},\mathbf{y}_i)} \prod_{i=1}^N q_\sigma(\mathbf{x},\mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_j,\mathbf{z}_i)\,\alpha(\mathbf{x},\mathbf{y}_j)\,\mathrm{d}(\mathbf{y}_{1:N},\mathbf{z}_{1:N-1})\\
&+ \mathbb{1}_{\mathbf{x}\in A}\sum_{j=1}^N \int \frac{w(\mathbf{x},\mathbf{y}_j)}{\sum_{i=1}^N w(\mathbf{x},\mathbf{y}_i)} \prod_{i=1}^N q_\sigma(\mathbf{x},\mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_j,\mathbf{z}_i)\,(1-\alpha(\mathbf{x},\mathbf{y}_j))\,\mathrm{d}(\mathbf{y}_{1:N},\mathbf{z}_{1:N-1}),
\end{aligned}$$

and

$$P_{\mathrm{ideal}}(\mathbf{x},A) := \int_{\mathbf{y}\in A} Q_{w,\sigma}(\mathbf{x},\mathbf{y})\,\alpha_{\mathrm{ideal}}(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y} + \mathbb{1}_{\mathbf{x}\in A}\int Q_{w,\sigma}(\mathbf{x},\mathbf{y})\,(1-\alpha_{\mathrm{ideal}}(\mathbf{x},\mathbf{y}))\,\mathrm{d}\mathbf{y},$$

where

$$\alpha_{\text{ideal}}(\mathbf{x}, \mathbf{y}) := 1 \wedge \frac{\pi(\mathbf{y})\, Q_{w,\sigma}(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})\, Q_{w,\sigma}(\mathbf{x}, \mathbf{y})}.$$

In $P_N(\mathbf{x}, A)$, the integrals are the same for all j. Therefore,

$$\begin{aligned} P_N(\mathbf{x}, A) := &\int_{\mathbf{y}_1 \in A} \frac{w(\mathbf{x}, \mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{x}, \mathbf{y}_i)} \prod_{i=1}^N q_\sigma(\mathbf{x}, \mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_1, \mathbf{z}_i)\, \alpha(\mathbf{x}, \mathbf{y}_1)\, \mathrm{d}(\mathbf{y}_{1:N}, \mathbf{z}_{1:N-1}) \\ &+ \mathbb{1}_{\mathbf{x} \in A} \int \frac{w(\mathbf{x}, \mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{x}, \mathbf{y}_i)} \prod_{i=1}^N q_\sigma(\mathbf{x}, \mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_1, \mathbf{z}_i)\, (1 - \alpha(\mathbf{x}, \mathbf{y}_1))\, \mathrm{d}(\mathbf{y}_{1:N}, \mathbf{z}_{1:N-1}). \end{aligned}$$

Consequently,

$$\begin{aligned} P_N h(\mathbf{x}) = &\int \frac{w(\mathbf{x}, \mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{x}, \mathbf{y}_i)} \prod_{i=1}^N q_\sigma(\mathbf{x}, \mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_1, \mathbf{z}_i)\, \alpha(\mathbf{x}, \mathbf{y}_1)\, h(\mathbf{y}_1)\, \mathrm{d}(\mathbf{y}_{1:N}, \mathbf{z}_{1:N-1}) \\ &+ h(\mathbf{x}) \int \frac{w(\mathbf{x}, \mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{x}, \mathbf{y}_i)} \prod_{i=1}^N q_\sigma(\mathbf{x}, \mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_1, \mathbf{z}_i)\, (1 - \alpha(\mathbf{x}, \mathbf{y}_1))\, \mathrm{d}(\mathbf{y}_{1:N}, \mathbf{z}_{1:N-1}), \end{aligned}$$

and

$$P_{\text{ideal}} h(\mathbf{x}) := \int Q_{w,\sigma}(\mathbf{x}, \mathbf{y})\, \alpha_{\text{ideal}}(\mathbf{x}, \mathbf{y})\, h(\mathbf{y})\, \mathrm{d}\mathbf{y} + h(\mathbf{x}) \int Q_{w,\sigma}(\mathbf{x}, \mathbf{y})\, (1 - \alpha_{\text{ideal}}(\mathbf{x}, \mathbf{y}))\, \mathrm{d}\mathbf{y}. \tag{10}$$

We first prove that

$$\int |P_N h(\mathbf{x}) - P_{\text{ideal}} h(\mathbf{x})|\, \pi(\mathbf{x})\, \mathrm{d}\mathbf{x} \to 0.$$

Using the triangle inequality, it suffices to prove that

$$\begin{aligned} \int \Bigg| &\int \frac{w(\mathbf{x}, \mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{x}, \mathbf{y}_i)} \prod_{i=1}^N q_\sigma(\mathbf{x}, \mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_1, \mathbf{z}_i)\, \alpha(\mathbf{x}, \mathbf{y}_1)\, h(\mathbf{y}_1)\, \mathrm{d}(\mathbf{y}_{1:N}, \mathbf{z}_{1:N-1}) \\ &- \int Q_{w,\sigma}(\mathbf{x}, \mathbf{y})\, \alpha_{\text{ideal}}(\mathbf{x}, \mathbf{y})\, h(\mathbf{y})\, \mathrm{d}\mathbf{y} \Bigg|\, \pi(\mathbf{x})\, \mathrm{d}\mathbf{x} \to 0, \end{aligned}$$

and

$$\begin{aligned} \int \Bigg| h(\mathbf{x}) &\int \frac{w(\mathbf{x}, \mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{x}, \mathbf{y}_i)} \prod_{i=1}^N q_\sigma(\mathbf{x}, \mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_1, \mathbf{z}_i)\, (1 - \alpha(\mathbf{x}, \mathbf{y}_1))\, \mathrm{d}(\mathbf{y}_{1:N}, \mathbf{z}_{1:N-1}) \\ &- h(\mathbf{x}) \int Q_{w,\sigma}(\mathbf{x}, \mathbf{y})\, (1 - \alpha_{\text{ideal}}(\mathbf{x}, \mathbf{y}))\, \mathrm{d}\mathbf{y} \Bigg|\, \pi(\mathbf{x})\, \mathrm{d}\mathbf{x} \to 0. \end{aligned}$$

We prove the first convergence and the other one follows using the same arguments. Using that we can rewrite

$$\begin{aligned} &\int Q_{w,\sigma}(\mathbf{x}, \mathbf{y})\, \alpha_{\text{ideal}}(\mathbf{x}, \mathbf{y})\, h(\mathbf{y})\, \mathrm{d}\mathbf{y} \\ &\qquad = \int \frac{w(\mathbf{x}, \mathbf{y}_1)}{\int w(\mathbf{x}, \mathbf{y}_1)\, q_\sigma(\mathbf{x}, \mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1}\, \alpha_{\text{ideal}}(\mathbf{x}, \mathbf{y}_1)\, h(\mathbf{y}_1)\, \mathrm{d}(\mathbf{y}_{1:N}, \mathbf{z}_{1:N-1}), \end{aligned}$$

Jensen's inequality and that h is bounded, let us say by a positive constant M,

$$\begin{aligned}
&\int \left| \int \frac{w(\mathbf{x},\mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{x},\mathbf{y}_i)} \prod_{i=1}^N q_\sigma(\mathbf{x},\mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_1,\mathbf{z}_i)\, \alpha(\mathbf{x},\mathbf{y}_1)\, h(\mathbf{y}_1)\, \mathrm{d}(\mathbf{y}_{1:N},\mathbf{z}_{1:N-1}) \right. \\
&\qquad \left. - \int Q_{w,\sigma}(\mathbf{x},\mathbf{y})\, \alpha_{\text{ideal}}(\mathbf{x},\mathbf{y})\, h(\mathbf{y})\, \mathrm{d}\mathbf{y} \right| \pi(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
&\leq M \iint \left| \frac{w(\mathbf{x},\mathbf{y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{x},\mathbf{y}_i)}\, \alpha(\mathbf{x},\mathbf{y}_1) - \frac{w(\mathbf{x},\mathbf{y}_1)}{\int w(\mathbf{x},\mathbf{y}_1)\, q_\sigma(\mathbf{x},\mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1}\, \alpha_{\text{ideal}}(\mathbf{x},\mathbf{y}_1) \right| \\
&\qquad \times \prod_{i=1}^N q_\sigma(\mathbf{x},\mathbf{y}_i) \prod_{i=1}^{N-1} q_\sigma(\mathbf{y}_1,\mathbf{z}_i)\, \mathrm{d}(\mathbf{y}_{1:N},\mathbf{z}_{1:N-1})\, \pi(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
&= M\mathbb{E}\left[\left| \frac{w(\mathbf{X},\mathbf{Y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{X},\mathbf{Y}_i)}\, \alpha(\mathbf{X},\mathbf{Y}_1) - \frac{w(\mathbf{X},\mathbf{Y}_1)}{\int w(\mathbf{X},\mathbf{y}_1)\, q_\sigma(\mathbf{X},\mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1}\, \alpha_{\text{ideal}}(\mathbf{X},\mathbf{Y}_1) \right| \right].
\end{aligned}$$

From the strong law of large numbers, we have that with probability 1

$$\frac{1}{N}\sum_{i=1}^N w(\mathbf{x},\mathbf{Y}_i) \to \int w(\mathbf{x},\mathbf{y}_1)\, q_\sigma(\mathbf{x},\mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1, \quad \text{as } N \to \infty.$$

for all $\mathbf{x}$. Therefore, with probability 1,

$$\frac{1}{N}\sum_{i=1}^N w(\mathbf{X},\mathbf{Y}_i) \to \int w(\mathbf{X},\mathbf{y}_1)\, q_\sigma(\mathbf{X},\mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1.$$

Also, with probability 1,

$$\begin{aligned}
\alpha(\mathbf{X},\mathbf{Y}_1) &= 1 \wedge \frac{\pi(\mathbf{Y}_1)\, q_\sigma(\mathbf{Y}_1,\mathbf{X})\, w(\mathbf{Y}_1,\mathbf{X}) \Big/ \left(\frac{1}{N}\left(\sum_{i=1}^{N-1} w(\mathbf{Y}_1,\mathbf{Z}_i) + w(\mathbf{Y}_1,\mathbf{X}) \right) \right)}{\pi(\mathbf{X})\, q_\sigma(\mathbf{X},\mathbf{Y}_1)\, w(\mathbf{X},\mathbf{Y}_1) \Big/ \left(\frac{1}{N}\left(\sum_{i=2}^{N} w(\mathbf{X},\mathbf{Y}_i) + w(\mathbf{X},\mathbf{Y}_1) \right) \right)} \\
&\to 1 \wedge \frac{\pi(\mathbf{Y}_1)\, q_\sigma(\mathbf{Y}_1,\mathbf{X})\, w(\mathbf{Y}_1,\mathbf{X}) \Big/ \int w(\mathbf{Y}_1,\mathbf{z}_1)\, q_\sigma(\mathbf{Y}_1,\mathbf{z}_1)\, \mathrm{d}\mathbf{z}_1}{\pi(\mathbf{X})\, q_\sigma(\mathbf{X},\mathbf{Y}_1)\, w(\mathbf{X},\mathbf{Y}_1) \Big/ \int w(\mathbf{X},\mathbf{y}_1)\, q_\sigma(\mathbf{X}_1,\mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1} \\
&= 1 \wedge \frac{\pi(\mathbf{Y}_1)\, Q_{w,\sigma}(\mathbf{Y}_1,\mathbf{X})}{\pi(\mathbf{X})\, Q_{w,\sigma}(\mathbf{X},\mathbf{Y}_1)} = \alpha_{\text{ideal}}(\mathbf{X},\mathbf{Y}_1).
\end{aligned}$$

Therefore, with probability 1,

$$\left| \frac{w(\mathbf{X},\mathbf{Y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{X},\mathbf{Y}_i)}\, \alpha(\mathbf{X},\mathbf{Y}_1) - \frac{w(\mathbf{X},\mathbf{Y}_1)}{\int w(\mathbf{X},\mathbf{y}_1)\, q_\sigma(\mathbf{X},\mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1}\, \alpha_{\text{ideal}}(\mathbf{X},\mathbf{Y}_1) \right| \to 0.$$

To be able to conclude that the expectation converges, we prove that the random variable is uniformly integrable. We more specifically prove that

$$\sup_N \mathbb{E}\left[\left(\frac{w(\mathbf{X},\mathbf{Y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{X},\mathbf{Y}_i)}\, \alpha(\mathbf{X},\mathbf{Y}_1) - \frac{w(\mathbf{X},\mathbf{Y}_1)}{\int w(\mathbf{X},\mathbf{y}_1)\, q_\sigma(\mathbf{X},\mathbf{y}_1)\, \mathrm{d}\mathbf{y}_1}\, \alpha_{\text{ideal}}(\mathbf{X},\mathbf{Y}_1) \right)^2 \right] < \infty,$$

which implies that the random variable is uniformly integrable. We have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{w(\mathbf{X},\mathbf{Y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{X},\mathbf{Y}_i)}\,\alpha(\mathbf{X},\mathbf{Y}_1) - \frac{w(\mathbf{X},\mathbf{Y}_1)}{\int w(\mathbf{X},\mathbf{y}_1)\,q_\sigma(\mathbf{X},\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1}\,\alpha_{\text{ideal}}(\mathbf{X},\mathbf{Y}_1)\right)^2\right] \\
&\leq 2\mathbb{E}\left[\left(\frac{w(\mathbf{X},\mathbf{Y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{X},\mathbf{Y}_i)}\,\alpha(\mathbf{X},\mathbf{Y}_1)\right)^2\right] \\
&\qquad + 2\mathbb{E}\left[\left(\frac{w(\mathbf{X},\mathbf{Y}_1)}{\int w(\mathbf{X},\mathbf{y}_1)\,q_\sigma(\mathbf{X},\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1}\,\alpha_{\text{ideal}}(\mathbf{X},\mathbf{Y}_1)\right)^2\right] \\
&\leq 2\mathbb{E}\left[\left(\frac{w(\mathbf{X},\mathbf{Y}_1)}{\frac{1}{N}\sum_{i=1}^N w(\mathbf{X},\mathbf{Y}_i)}\right)^2\right] + 2\mathbb{E}\left[\left(\frac{w(\mathbf{X},\mathbf{Y}_1)}{\int w(\mathbf{X},\mathbf{y}_1)\,q_\sigma(\mathbf{X},\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1}\right)^2\right] \\
&\leq 2\mathbb{E}\left[w(\mathbf{X},\mathbf{Y}_1)^4\right]^{1/2}\mathbb{E}\left[w(\mathbf{X},\mathbf{Y}_1)^{-4}\right]^{1/2} + 2\mathbb{E}\left[w(\mathbf{X},\mathbf{Y}_1)^2\right]\left[\int w(\mathbf{X},\mathbf{y}_1)\,q_\sigma(\mathbf{X},\mathbf{y}_1)\,\mathrm{d}\mathbf{y}_1\right]^{-2},
\end{aligned}$$

which is finite. We used that for any real numbers a, b, $(a+b)^2 \leq 2a^2 + 2b^2$, that $0 \leq \alpha, \alpha_{\text{ideal}} \leq 1$, Cauchy–Schwarz inequality, and Proposition 4 for

$$\mathbb{E}_{\mathbf{x}}\left[\left(\frac{1}{N}\sum_{i=1}^N w(\mathbf{x},\mathbf{Y}_i)\right)^{-4}\right] \leq \mathbb{E}_{\mathbf{x}}\left[w(\mathbf{x},\mathbf{Y}_1)^{-4}\right],$$

with $\mathbb{E}\left[w(\mathbf{X},\mathbf{Y}_1)^{-4}\right] < \infty$.

We now show that $P_{\text{ideal}}h$ is continuous. Define $\mathbf{x}_{\boldsymbol{\epsilon}} := \mathbf{x} + \boldsymbol{\epsilon}$ with $\|\boldsymbol{\epsilon}\| \leq \varepsilon$, where $\varepsilon > 0$ can be chosen to be arbitrarily small. We want to prove that

$$\lim_{\varepsilon \to 0} P_{\text{ideal}}h(\mathbf{x}_{\boldsymbol{\epsilon}}) = P_{\text{ideal}}h(\mathbf{x}).$$

This is true under the assumptions if we can interchange the limit and the integral in (10). We are allowed to do it using the dominated convergence theorem because h is bounded, $0 \leq \alpha_{\text{ideal}} \leq 1$ and $Q_{w,\sigma}(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}) \leq f(\mathbf{x},\mathbf{y})$, an integrable (in $\mathbf{y}$) function, for all $\mathbf{x}$. ■

A.2 Proof of Proposition 2

Proof [Proposition 2] We prove the result for the case where $w(\mathbf{x},\mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$. The proof is analogous for the case where $w(\mathbf{x},\mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$.

We have that

$$Q_{w,\sigma}(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}) = \frac{w(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})\,q_\sigma(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})}{\int w(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})\,q_\sigma(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})\,\mathrm{d}\mathbf{y}} = \frac{\pi(\mathbf{y})\,q_\sigma(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})}{\int \pi(\mathbf{y})\,q_\sigma(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})\,\mathrm{d}\mathbf{y}}.$$

Also,

$$\begin{aligned}
q_\sigma(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}) &= \frac{1}{(2\pi\sigma^2)^{d/2}}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-(\mathbf{x}+\boldsymbol{\epsilon}))^T(\mathbf{y}-(\mathbf{x}+\boldsymbol{\epsilon}))\right) \\
&= \frac{1}{(2\pi\sigma^2)^{d/2}}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{x})^T(\mathbf{y}-\mathbf{x})\right)\exp\left(\frac{1}{2\sigma^2}\boldsymbol{\epsilon}^T(\mathbf{y}-\mathbf{x})\right)\exp\left(-\frac{1}{2\sigma^2}\,\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\right).
\end{aligned}$$

Therefore,

$$q_\sigma(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}) \leq q_\sigma(\mathbf{x},\mathbf{y}) \exp\left(\frac{\varepsilon}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|\right),$$

given that

$$\exp\left(-\frac{1}{2\sigma^2}\,\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\right) \leq 1,$$

and

$$\boldsymbol{\epsilon}^T(\mathbf{y}-\mathbf{x}) \leq |\boldsymbol{\epsilon}^T(\mathbf{y}-\mathbf{x})| \leq \|\boldsymbol{\epsilon}\|\|\mathbf{y}-\mathbf{x}\| \leq \varepsilon\|\mathbf{y}-\mathbf{x}\|,$$

by Cauchy–Schwarz inequality.

We thus have an upper bound of the numerator of $Q_{w,\sigma}(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})$ that does not depend on $\boldsymbol{\epsilon}$. We now find a lower bound of the denominator that does not depend on $\boldsymbol{\epsilon}$. By Fatou's lemma, we have that

$$\liminf_{\varepsilon\to 0} \int \pi(\mathbf{y})\, q_\sigma(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})\,\mathrm{d}\mathbf{y} \geq \int \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}.$$

We have that $\|\boldsymbol{\epsilon}\| \leq \varepsilon$ with $\varepsilon > 0$ an arbitrarily small value. We thus know that we can choose ε such that

$$\int \pi(\mathbf{y})\, q_\sigma(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})\,\mathrm{d}\mathbf{y} \geq \int \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y} - \xi,$$

with ξ an arbitrarily small value. In particular, we can choose ε such that

$$\xi \leq \int \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y},$$

implying that

$$\int \pi(\mathbf{y})\, q_\sigma(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y})\,\mathrm{d}\mathbf{y} \geq \frac{1}{2}\int \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}.$$

Note that $\int \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y} > 0$ because $\pi(\mathbf{y})$ is assumed to be strictly positive. Note also that $\int \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y} < \infty$. Indeed,

$$\int \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y} \leq \frac{1}{(2\pi\sigma^2)^{d/2}}\int \pi(\mathbf{y})\,\mathrm{d}\mathbf{y} < \infty.$$

In the case where $w(\mathbf{x},\mathbf{y}) = \sqrt{\pi(\mathbf{y})/\pi(\mathbf{x})}$, we use Cauchy–Schwarz inequality instead of the boundedness of q_σ to reach the same conclusion.

To summarize, we know that we can choose ε so that

$$Q_{w,\sigma}(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}) \leq \frac{\pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y}) \exp\left(\frac{\varepsilon}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|\right)}{\frac{1}{2}\int \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbf{y}},$$

which is independent of $\boldsymbol{\epsilon}$. We know that the denominator on the right-hand side (RHS) is strictly positive and finite. To conclude the proof, we need to show that the numerator is integrable.

We have that

$$\int \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y}) \exp\left(\frac{\varepsilon}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|\right) \mathrm{d}\mathbf{y} = \int_{A_\mathbf{x}} \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y}) \exp\left(\frac{\varepsilon}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|\right) \mathrm{d}\mathbf{y} + \int_{A_\mathbf{x}^\complement} \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y}) \exp\left(\frac{\varepsilon}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|\right) \mathrm{d}\mathbf{y},$$

where $A_\mathbf{x}$ is the set of values of $\mathbf{y}$, with $\mathbf{x}$ fixed, such that $\|\mathbf{y}-\mathbf{x}\| \leq 1$. Given that this set is compact and that the integrand is upper bounded, we know that

$$\int_{A_\mathbf{x}} \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y}) \exp\left(\frac{\varepsilon}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|\right) \mathrm{d}\mathbf{y} < \infty.$$

Now, let us analyse the other integral, we have that

$$\begin{aligned}\int_{A_\mathbf{x}^\complement} \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y}) \exp\left(\frac{\varepsilon}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|\right) \mathrm{d}\mathbf{y} &\leq \int_{A_\mathbf{x}^\complement} \pi(\mathbf{y})\, q_\sigma(\mathbf{x},\mathbf{y}) \exp\left(\frac{\varepsilon}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|^2\right) \mathrm{d}\mathbf{y} \\ &\leq \int \pi(\mathbf{y})\, \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1-\varepsilon}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\|^2\right) \mathrm{d}\mathbf{y} \\ &\leq \frac{1}{(2\pi\sigma^2)^{d/2}} \int \pi(\mathbf{y})\, \mathrm{d}\mathbf{y} < \infty,\end{aligned}$$

which concludes the proof. ■

A.3 Proof of Proposition 3

Before presenting the proof of Proposition 3, we provide two lemmas which are used in it and in several other proofs. After presenting the proof of Proposition 3, we present a result stating that $\lim_{\|\mathbf{x}\|\to\infty} \mathbb{E}_\mathbf{x}[\alpha_{\text{ideal}}(\mathbf{x},\mathbf{Y})] = 0$ for any σ and d under weaker assumptions than those supposed in Proposition 3. As mentioned in Section 3.1, the assumptions are essentially that $U := -\log\pi$ is strongly convex and L-smooth. In the proof of that result, it will be noticed that L-smoothness is not necessary but makes the proof simpler. It will also be noticed in the statement of the result that, instead of strong convexity, we assume that $\|\nabla U(\mathbf{x})\| \to \infty$ as $\|\mathbf{x}\| \to \infty$, which is weaker, but is in fact the crucial assumption for having $\lim_{\|\mathbf{x}\|\to\infty} \mathbb{E}_\mathbf{x}[\alpha_{\text{ideal}}(\mathbf{x},\mathbf{Y})] = 0$.

Lemma 1 *When the target density is defined as in* (3) *and the proposal distribution is* $Q_{w,\sigma}$ *with* $q_\sigma(\mathbf{x},\cdot) = \mathcal{N}(\mathbf{x},\sigma^2\mathbb{I}_d)$ *and* $w(\mathbf{x},\mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$*, we have that*

$$\alpha_{ideal}(\mathbf{x},\mathbf{y}) = 1 \wedge \exp\left(\frac{1}{2(1+\sigma^2)}\sum_{i=1}^{d}(y_i^2 - x_i^2)\right).$$

Proof [Lemma 1] We have that

$$\begin{aligned}
\alpha_{\text{ideal}}(\mathbf{x},\mathbf{y}) &= 1 \wedge \frac{\pi(\mathbf{y})\, Q_{w,\sigma}(\mathbf{y},\mathbf{x})}{\pi(\mathbf{x})\, Q_{w,\sigma}(\mathbf{x},\mathbf{y})} \\
&= 1 \wedge \prod_{i=1}^{d} \frac{\int \varphi(z_i)\,(1/\sigma)\varphi((z_i - x_i)/\sigma)\,\mathrm{d}z_i}{\int \varphi(z_i)\,(1/\sigma)\varphi((z_i - y_i)/\sigma)\,\mathrm{d}z_i} \\
&= 1 \wedge \prod_{i=1}^{d} \frac{\int \varphi(x_i + \sigma u_i)\,\varphi(u_i)\,\mathrm{d}u_i}{\int \varphi(y_i + \sigma u_i)\,\varphi(u_i)\,\mathrm{d}u_i} \\
&= 1 \wedge \prod_{i=1}^{d} \frac{\int \exp\left(-\frac{\sigma^2}{2}(u_i + x_i/\sigma)^2\right)\,\varphi(u_i)\,\mathrm{d}u_i}{\int \exp\left(-\frac{\sigma^2}{2}(u_i + y_i/\sigma)^2\right)\,\varphi(u_i)\,\mathrm{d}u_i} \\
&= 1 \wedge \exp\left(\frac{1}{2(1+\sigma^2)} \sum_{i=1}^{d} (y_i^2 - x_i^2)\right),
\end{aligned}$$

using the definition of $Q_{w,\sigma}$, the factorization of the target and proposal densities, a change of variable and the equality

$$\int \exp\left(-\frac{\sigma^2}{2}(u_i + x_i/\sigma_d)^2\right)\,\varphi(u_i)\,\mathrm{d}u_i = \mathbb{E}\left[\exp\left(-\frac{\sigma^2}{2} Z_i\right)\right] = \frac{\exp\left(-\frac{x_i^2}{2(1+\sigma^2)}\right)}{(1+\sigma^2)^{1/2}},$$

with Z_i that follows a non-central chi-squared distribution. ■

Lemma 2 *When the target density is defined as in* (3) *and the proposal distribution is* $Q_{w,\sigma}$ *with* $q_\sigma(\mathbf{x},\cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$ *and* $w(\mathbf{x},\mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$*, we have that* $\mathbf{Y} \sim Q_{w,\sigma}(\mathbf{x},\cdot)$ *is equal in distribution to*

$$\frac{\mathbf{x}}{1+\sigma^2} + \sqrt{\frac{\sigma^2}{1+\sigma^2}}\,\mathbf{U},$$

where the components of $\mathbf{U} := (U_1, \ldots, U_d)^T$ *are* d *independent standard-normal random variables.*

Proof [Lemma 2] The PDF $Q_{w,\sigma}(\mathbf{x},\cdot)$ is such that

$$Q_{w,\sigma}(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{d} \varphi(y_i)\,(1/\sigma)\varphi((y_i - x_i)/\sigma) \propto \prod_{i=1}^{d} \exp\left(-\frac{1}{2}\frac{1+\sigma^2}{\sigma^2}\left(y_i - \frac{x_i}{1+\sigma^2}\right)^2\right),$$

implying that the components in $\mathbf{Y} := (Y_1, \ldots, Y_d)^T$ are d conditionally independent random variables with

$$Y_i \sim \mathcal{N}\left(\frac{X_i}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right),$$

for all i. This concludes the proof. ■

Proof [Proposition 3] Using Lemma 1,

$$\mathbb{E}[\alpha_{\text{ideal}}(\mathbf{x}, \mathbf{Y})] = \mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma^2)} \sum_{i=1}^{d} (Y_i^2 - x_i^2)\right)\right].$$

We have that

$$\begin{aligned}
\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma^2)} \sum_{i=1}^{d} (Y_i^2 - x_i^2)\right)\right] &\leq \mathbb{E}\left[\exp\left(\frac{1}{2(1+\sigma^2)} \sum_{i=1}^{d} (Y_i^2 - x_i^2)\right)\right] \\
&= \exp\left(-\frac{1}{2(1+\sigma^2)} \|\mathbf{x}\|^2\right) \mathbb{E}\left[\exp\left(\frac{1}{2(1+\sigma^2)} \sum_{i=1}^{d} \left(\frac{x_i}{1+\sigma^2} + \sqrt{\frac{\sigma^2}{1+\sigma^2}} U_i\right)^2\right)\right] \\
&= \exp\left(-\frac{1}{2(1+\sigma^2)} \|\mathbf{x}\|^2\right) \mathbb{E}\left[\exp\left(\frac{\sigma^2}{2(1+\sigma^2)^2} \sum_{i=1}^{d} \left(\frac{x_i}{\sigma\sqrt{1+\sigma^2}} + U_i\right)^2\right)\right] \\
&= \exp\left(-\frac{1}{2(1+\sigma^2)} \|\mathbf{x}\|^2\right) \exp\left(\frac{\frac{1}{2(1+\sigma^2)^3} \|\mathbf{x}\|^2}{1 - \frac{\sigma^2}{(1+\sigma^2)^2}}\right) \left(1 - \frac{\sigma^2}{(1+\sigma^2)^2}\right)^{-d/2},
\end{aligned}$$

using Lemma 2 and the explicit expression of the moment generating function of a non-central chi-squared distribution. Note that the latter can be used because $\sigma^2/(1+\sigma^2)^2 < 1$. The first two terms on the RHS above can be combined and simplified; their product is equal to

$$\exp\left(-\|\mathbf{x}\|^2 \frac{\sigma^2}{2((1+\sigma^2)^2 - \sigma^2)}\right).$$

■

Proposition 5 *Consider a current state* $\mathbf{x}$ *and that* $\mathbf{Y} \sim Q_{w,\sigma}(\mathbf{x}, \cdot)$ *with* $q_\sigma(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \mathbb{I}_d)$ *and* $w(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$. *Assume that* $U := -\log \pi$ *is continuously differentiable. Assume that* U *is* L*-smooth, meaning that its gradient,* ∇U*, is* L*-Lipschitz. Finally, assume that* $\|\nabla U(\mathbf{x})\| \to \infty$ *as* $\|\mathbf{x}\| \to \infty$. *Then, for any* σ *and* d*, it holds that*

$$\lim_{\|\mathbf{x}\| \to \infty} \mathbb{E}_{\mathbf{x}}[\alpha_{ideal}(\mathbf{x}, \mathbf{Y})] = 0.$$

Proof [Proposition 5] Let us define $\mathcal{Z}(\mathbf{x}) := \int \pi(\mathbf{y})\, q_\sigma(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{y}$. We first provide a lower bound on $\mathcal{Z}(\mathbf{x})/\pi(\mathbf{x})$ which will be useful to prove our result. To provide this lower bound, we will use that

$$|U(\mathbf{y}) - U(\mathbf{x}) - \nabla U(\mathbf{x})^T(\mathbf{y} - \mathbf{x})| \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2,$$

given that U is L-smooth; see (Bubeck, 2015, Lemma 3.4). Using this, we have that

$$\begin{aligned}
\frac{\mathcal{Z}(\mathbf{x})}{\pi(\mathbf{x})} &= \int \exp(-(U(\mathbf{y}) - U(\mathbf{x}))) \, q_\sigma(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{y} \\
&\geq \int \exp\left(-\nabla U(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) - \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2\right) \, q_\sigma(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{y} \\
&= \int \exp\left(-\nabla U(\mathbf{x})^T\mathbf{z} - \frac{L}{2}\|\mathbf{z}\|^2\right) \, \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{z}\|^2\right) \, \mathrm{d}\mathbf{z} \\
&= \left(\frac{\sigma^{-2}}{L + \sigma^{-2}}\right)^{d/2} \int \exp\left(-\nabla U(\mathbf{x})^T\mathbf{z}\right) \, \frac{1}{(2\pi(L + \sigma^{-2}))^{d/2}} \exp\left(-\frac{L + \sigma^{-2}}{2}\|\mathbf{z}\|^2\right) \, \mathrm{d}\mathbf{z} \\
&= \left(\frac{\sigma^{-2}}{L + \sigma^{-2}}\right)^{d/2} \exp\left(\frac{1}{2(L + \sigma^{-2})}\|\nabla U(\mathbf{x})\|^2\right),
\end{aligned}$$

where the third line follows from a change of variables $\mathbf{z} = \mathbf{y} - \mathbf{x}$, and the last line follows from the explicit expression of the moment generating function of a multivariate normal distribution.

Now, we make use of that bound. When $\mathbf{Y} \sim Q_{w,\sigma}(\mathbf{x}, \cdot)$ with $q_\sigma(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2\mathbb{I}_d)$ and $w(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y})/\pi(\mathbf{x})$,

$$\alpha_{\text{ideal}}(\mathbf{x}, \mathbf{y}) = 1 \wedge \frac{\mathcal{Z}(\mathbf{x})}{\mathcal{Z}(\mathbf{y})} \leq \frac{\mathcal{Z}(\mathbf{x})}{\mathcal{Z}(\mathbf{y})}.$$

Therefore, we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}[\alpha_{\text{ideal}}(\mathbf{x}, \mathbf{Y})] &\leq \int \frac{\pi(\mathbf{y}) \, q_\sigma(\mathbf{x}, \mathbf{y})}{\mathcal{Z}(\mathbf{y})} \, \mathrm{d}\mathbf{y} \\
&= \int \frac{\pi(\mathbf{x} + \mathbf{z}) \, q_\sigma(\mathbf{0}, \mathbf{z})}{\mathcal{Z}(\mathbf{x} + \mathbf{z})} \, \mathrm{d}\mathbf{z} \\
&\leq \int \left(\frac{\sigma^{-2}}{L + \sigma^{-2}}\right)^{-d/2} \exp\left(-\frac{1}{2(L + \sigma^{-2})}\|\nabla U(\mathbf{x} + \mathbf{z})\|^2\right) \, q_\sigma(\mathbf{0}, \mathbf{z}) \, \mathrm{d}\mathbf{z}
\end{aligned}$$

after a change of variables $\mathbf{z} = \mathbf{y} - \mathbf{x}$. We conclude the proof using the bounded convergence theorem: for any σ, d, L and $\mathbf{z}$,

$$\left(\frac{\sigma^{-2}}{L + \sigma^{-2}}\right)^{-d/2} \exp\left(-\frac{1}{2(L + \sigma^{-2})}\|\nabla U(\mathbf{x} + \mathbf{z})\|^2\right) \to 0,$$

as $\|\mathbf{x}\| \to \infty$ and

$$\left(\frac{\sigma^{-2}}{L + \sigma^{-2}}\right)^{-d/2} \exp\left(-\frac{1}{2(L + \sigma^{-2})}\|\nabla U(\mathbf{x} + \mathbf{z})\|^2\right) \leq \left(\frac{\sigma^{-2}}{L + \sigma^{-2}}\right)^{-d/2}.$$

■

A.4 Proof of Theorem 2

Proof [Theorem 2] We prove a weak convergence towards a diffusion, denoted by $\{Z(t) : t \geq 0\}$, in the Skorokhod topology (for more details about this type of convergence, see Chapter 3 of Ethier and Kurtz (1986)). In order to prove the result, we demonstrate the convergence of the finite-dimensional distributions of $\{Z_{d,\text{ideal}}(t) : t \geq 0\}$ to those of $\{Z(t) : t \geq 0\}$. To achieve this, we verify Condition (c) of Theorem 8.2 from Chapter 4 of Ethier and Kurtz (1986). The weak convergence then follows from Corollary 8.6 of Chapter 4 of Ethier and Kurtz (1986). The remaining conditions of Theorem 8.2 and the conditions of Corollary 8.6 are either straightforward or easily derived from the proof given here.

The proof of the convergence of the finite-dimensional distributions relies on the convergence of (what we call) the *pseudo-generator* of $\{Z_{d,\text{ideal}}(t) : t \geq 0\}$, an operator that we now introduce. The proof follows.

Pseudo-generator. The process $\{Z_{d,\text{ideal}}(t) : t \geq 0\}$ is a jump process for which the time in between the (possible) jumps is deterministic: we know that every $1/d^{2\tau}$ unit of time, the process jumps if the proposal is accepted. The pseudo-generator is a discrete version of infinitesimal generators of stochastic processes. It is defined as follows:

$$\phi_{d,\text{ideal}}(t) := d^{2\tau}\mathbb{E}[h(Z_{d,\text{ideal}}(t + 1/d^{2\tau})) - h(Z_{d,\text{ideal}}(t)) \mid \mathcal{F}_{\mathbf{Z}_{d,\text{ideal}}}(t)],$$

where h is a test function and $\mathcal{F}_{\mathbf{Z}_{d,\text{ideal}}}(t)$ is the natural filtration associated to $\{\mathbf{Z}_{d,\text{ideal}}(t) : t \geq 0\}$. The Markov property, the fact that $\mathbf{Z}_{d,\text{ideal}}(0) \sim \pi_d$ and that $\{\mathbf{X}_{d,\text{ideal}}(m) : m \in \mathbb{N}\}$ is time-homogeneous imply that for any t,

$$\begin{aligned}\phi_{d,\text{ideal}}(t) &= d^{2\tau}\mathbb{E}[h(Z_{d,\text{ideal}}(t + 1/d^{2\tau})) - h(Z_{d,\text{ideal}}(t)) \mid \mathbf{Z}_{d,\text{ideal}}(t)] \\ &\overset{\text{dist.}}{=} d^{2\tau}\mathbb{E}[(h(Y_1) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d],\end{aligned}$$

where " $\overset{\text{dist.}}{=}$ " denotes an equality in distribution, $\mathbf{X}_d \sim \pi_d$ and Y_1 is the first coordinate of $\mathbf{Y}_d \sim Q_{w,\sigma}(\mathbf{X}_d, \cdot\,)$.

We prove the convergence of $\phi_{d,\text{ideal}}(t)$ towards $Gh(Z_{d,\text{ideal}}(t))$ in some sense, where G is the generator of the diffusion. The form of G allows to restrict our attention to test functions $h \in \mathcal{C}_c^\infty(\mathbb{R})$, the space of infinitely differentiable functions on $\mathbb{R}$ with compact support (Theorem 2.5 from Chapter 8 of Ethier and Kurtz (1986)).

Proof of the convergence of the finite-dimensional distributions. Condition (c) of Theorem 8.2 from Chapter 4 of Ethier and Kurtz (1986) essentially reduces to the following convergence:

$$\mathbb{E}|\phi_{d,\text{ideal}}(t) - Gh(Z_{d,\text{ideal}}(t))| \to 0 \quad \text{as} \quad d \to \infty,$$

for all t. The generator is such that

$$\begin{aligned}Gh(Z_{d,\text{ideal}}(t)) &= \ell^2(\vartheta_{w,\tau}(\ell)/2)(\log\varphi(Z_{d,\text{ideal}}(t)))'h'(Z_{d,\text{ideal}}(t)) + \ell^2(\vartheta_{w,\tau}(\ell)/2)h''(Z_{d,\text{ideal}}(t)) \\ &\overset{\text{dist.}}{=} \ell^2(\vartheta_{w,\tau}(\ell)/2)(\log\varphi(X_1))'h'(X_1) + \ell^2(\vartheta_{w,\tau}(\ell)/2)h''(X_1),\end{aligned}$$

where the equality in distribution follows from the fact that the process starts in stationarity, that is $\mathbf{Z}_{d,\text{ideal}}(0) \sim \pi_d$. We can thus see $\phi_{d,\text{ideal}}(t) - Gh(Z_{d,\text{ideal}}(t))$ in the expectation above

as a difference of two functions of $\mathbf{X}_d \sim \pi_d$, and will write $Gh(X_1)$ instead of $Gh(Z_{d,\text{ideal}}(t))$ in the expectation. Note that the form of the generator indicates that φ is the unique invariant PDF of the diffusion.

We prove that

$$\mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_1) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d] - Gh(X_1)| \to 0.$$

The key here is to use a Taylor expansion in $(h(Y_1) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d)$ to obtain derivatives of h as in $Gh(X_1)$. Specifically, we write

$$h(Y_1) - h(X_1) = h'(X_1)(Y_1 - X_1) + h''(X_1)\frac{(Y_1 - X_1)^2}{2} + h'''(W)\frac{(Y_1 - X_1)^3}{6},$$

where W belongs to (X_1, Y_1) or (Y_1, X_1) (depending which one of X_1 and Y_1 is smaller). Therefore, using the triangle inequality,

$$\begin{aligned}
&\mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_1) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d] - Gh(X_1)| \\
&\quad = \mathbb{E}|d^{2\tau}\mathbb{E}[h'(X_1)(Y_1 - X_1)\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d] - \ell^2(\vartheta_{w,\tau}(\ell)/2)(\log\varphi(X_1))'h'(X_1)| \\
&\qquad + \mathbb{E}\left|d^{2\tau}\mathbb{E}\left[h''(X_1)\frac{(Y_1 - X_1)^2}{2}\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d\right] - \ell^2(\vartheta_{w,\tau}(\ell)/2)h''(X_1)\right| \\
&\qquad + \mathbb{E}\left|d^{2\tau}\mathbb{E}\left[h'''(W)\frac{(Y_1 - X_1)^3}{6}\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d\right]\right|. \qquad (11)
\end{aligned}$$

We now prove that each term on the RHS converges to 0. We prove this for the case $w(\mathbf{x}_d, \mathbf{y}_d) = \pi(\mathbf{y}_d)/\pi(\mathbf{x}_d)$; the case $w(\mathbf{x}_d, \mathbf{y}_d) = \sqrt{\pi(\mathbf{y}_d)/\pi(\mathbf{x}_d)}$ is proved similarly.

We have that

$$\begin{aligned}
&\mathbb{E}\left|d^{2\tau}\mathbb{E}\left[h'''(W)\frac{(Y_1 - X_1)^3}{6}\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d\right]\right| \le \frac{M}{6}d^{2\tau}\mathbb{E}[\mathbb{E}[|Y_1 - X_1|^3 \mid \mathbf{X}_d]] \\
&\le \frac{M}{6}d^{2\tau}\left(\mathbb{E}\left|\frac{\sigma_d^2 X_1}{1+\sigma_d^2}\right|^3 + 3\mathbb{E}\left[\left(\frac{\sigma_d^2 X_1}{1+\sigma_d^2}\right)^2\right]\sqrt{\frac{\sigma_d^2}{1+\sigma_d^2}}\mathbb{E}|U_1| + 3\mathbb{E}\left[\left|\frac{\sigma_d^2 X_1}{1+\sigma_d^2}\right|\right]\frac{\sigma_d^2}{1+\sigma_d^2}\mathbb{E}[U_1^2]\right. \\
&\qquad\qquad \left. + \left(\frac{\sigma_d^2}{1+\sigma_d^2}\right)^{3/2}\mathbb{E}|U_1|^3\right),
\end{aligned}$$

using Jensen's inequality, that $0 \le \alpha_{\text{ideal}} \le 1$, that there exists a positive constant M such that $|h'''| \le M$, Lemma 2 and the triangle inequality. The random variables X_1 and U_1 are independent and both follow a standard normal distribution, implying that $\mathbb{E}|X_1|^p\mathbb{E}|U_1|^q$ is finite and independent of d for any p and q. Recall that $\sigma_d = \ell/d^\tau$. The sum above thus converges to 0.

For the other terms in (11), we view the function $\alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d)$ (for any realization of $\mathbf{X}_d$ and $\mathbf{Y}_d$) as a function of y_1 (while keeping the other variables fixed) and we use a Taylor expansion around x_1 to obtain a function independent of x_1 and y_1. To see this, we recall that

$$\alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d) = 1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=1}^{d}(y_i^2 - x_i^2)\right),$$

using Lemma 1. We thus write

$$\alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d) = \alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d^*) + \left(\frac{\partial}{\partial y_1} \alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d) \Big|_{y_1 = x_1} \right) (y_1 - x_1) + \left(\frac{\partial^2}{\partial y_1^2} \alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d) \Big|_{y_1 = w} \right) \frac{(y_1 - x_1)^2}{2},$$

where $\mathbf{y}_d^* := (x_1, y_2, \ldots, y_d)$ and w belongs to (x_1, y_1) or (y_1, x_1). We have that

$$\alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d^*) = 1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} \sum_{i=2}^{d} (y_i^2 - x_i^2) \right),$$

$$\frac{\partial}{\partial y_1} \alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d) \Big|_{y_1 = x_1} = \frac{x_1}{1+\sigma_d^2} \exp\left(\frac{1}{2(1+\sigma_d^2)} \sum_{i=2}^{d} (y_i^2 - x_i^2) \right) \mathbb{1}\left(\sum_{i=2}^{d} (y_i^2 - x_i^2) < 0 \right), \tag{12}$$

$$\frac{\partial^2}{\partial y_1^2} \alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d) \Big|_{y_1 = w} = \left(\frac{1}{1+\sigma_d^2} \exp\left(\frac{1}{2(1+\sigma_d^2)} \sum_{i=2}^{d} (y_i^2 - x_i^2) + w^2 - x_1^2 \right) + \frac{w^2}{1+\sigma_d^2} \exp\left(\frac{1}{2(1+\sigma_d^2)} \sum_{i=2}^{d} (y_i^2 - x_i^2) + w^2 - x_1^2 \right) \right) \mathbb{1}\left(\sum_{i=2}^{d} (y_i^2 - x_i^2) + w^2 - x_1^2 < 0 \right).$$

We replace $\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d)$ in

$$\mathbb{E}| d^{2\tau} \mathbb{E}[h'(X_1)(Y_1 - X_1)\, \alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d] - \ell^2 (\vartheta_{w,\tau}(\ell)/2)(\log \varphi(X_1))' h'(X_1)|$$

in (11) by the sum above. We want to prove that the expectation converges to 0. Using the triangle inequality and that M can be chosen such that $|h'| \leq M$, it is sufficient to show that

$$M\mathbb{E}\Bigg| d^{2\tau} \mathbb{E}\Bigg[(Y_1 - X_1) \left(1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} \sum_{i=2}^{d} (Y_i^2 - X_i^2) \right) \right) + (Y_1 - X_1)^2 \frac{X_1}{1+\sigma_d^2} \exp\left(\frac{1}{2(1+\sigma_d^2)} \sum_{i=2}^{d} (Y_i^2 - X_i^2) \right) \mathbb{1}\left(\sum_{i=2}^{d} (Y_i^2 - X_i^2) < 0 \right) \mid \mathbf{X}_d \Bigg] - \ell^2 (\vartheta_{w,\tau}(\ell)/2)(\log \varphi(X_1))' \mid \to 0,$$

and that the following expectation converges to 0:

$$M d^{2\tau} \mathbb{E}\left| \frac{(Y_1 - X_1)^3}{2} \left(\frac{1}{1+\sigma_d^2} \exp\left(\frac{1}{2(1+\sigma_d^2)} \sum_{i=2}^{d} (Y_i^2 - X_i^2) + W^2 - X_1^2 \right) + \frac{W^2}{1+\sigma_d^2} \exp\left(\frac{1}{2(1+\sigma_d^2)} \sum_{i=2}^{d} (Y_i^2 - X_i^2) + W^2 - X_1^2 \right) \right) \mathbb{1}\left(\sum_{i=2}^{d} (Y_i^2 - X_i^2) + W^2 - X_1^2 < 0 \right) \right|.$$

We start with the last term. Using the triangle inequality, it is lesser than or equal to

$$\frac{Md^{2\tau}}{2(1+\sigma_d^2)}\mathbb{E}|Y_1 - X_1|^3 + \frac{Md^{2\tau}}{2(1+\sigma_d^2)}\mathbb{E}[|Y_1 - X_1|^3 W^2].$$

We have seen before that $d^{2\tau}\mathbb{E}|Y_1 - X_1|^3 \to 0$. Also, given X_1 and writing $Y_1 = \frac{X_1}{1+\sigma_d^2} + \sqrt{\frac{\sigma_d^2}{1+\sigma_d^2}}U_1$ under the conditional expectation, we can show that $|W| \leq |X_1| + \sqrt{\frac{\sigma_d^2}{1+\sigma_d^2}}|U_1|$, and consequently, that $d^{2\tau}\mathbb{E}[|Y_1 - X_1|^3 W^2] \to 0$ in the same way we proved that $d^{2\tau}\mathbb{E}|Y_1 - X_1|^3 \to 0$.

For the other term, we first note that $(\log \varphi(X_1))' = -X_1$. We simplify the notation by defining

$$f_1(\mathbf{X}_d) := \mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2 - X_i^2)\right) \mid \mathbf{X}_d\right],$$

and

$$f_2(\mathbf{X}_d) := \mathbb{E}\left[\exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2 - X_i^2)\right)\mathbb{1}\left(\sum_{i=2}^{d}(Y_i^2 - X_i^2) < 0\right) \mid \mathbf{X}_d\right].$$

Using the conditional independence among $Y_1, \ldots, Y_d$ given $\mathbf{X}_d$,

$$\begin{aligned}
&\mathbb{E}\left[(Y_1 - X_1)\left(1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2 - X_i^2)\right)\right)\right. \\
&\quad \left. + (Y_1 - X_1)^2 \frac{X_1}{1+\sigma_d^2}\exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2 - X_i^2)\right)\mathbb{1}\left(\sum_{i=2}^{d}(Y_i^2 - X_i^2) < 0\right) \mid \mathbf{X}_d\right] \\
&= \mathbb{E}\left[Y_1 - X_1 \mid \mathbf{X}_d\right] f_1(\mathbf{X}_d) + \frac{X_1}{1+\sigma_d^2}\mathbb{E}\left[(Y_1 - X_1)^2 \mid \mathbf{X}_d\right] f_2(\mathbf{X}_d) \\
&= \mathbb{E}\left[-\frac{\sigma_d^2 X_1}{1+\sigma_d^2} + \sqrt{\frac{\sigma_d^2}{1+\sigma_d^2}}U_1 \mid \mathbf{X}_d\right] f_1(\mathbf{X}_d) \\
&\qquad + \frac{X_1}{1+\sigma_d^2}\mathbb{E}\left[\frac{\sigma_d^4 X_1^2}{(1+\sigma_d^2)^2} - 2\frac{\sigma_d^2 X_1}{1+\sigma_d^2}\sqrt{\frac{\sigma_d^2}{1+\sigma_d^2}}U_1 + \frac{\sigma_d^2}{1+\sigma_d^2}U_1^2 \mid \mathbf{X}_d\right] f_2(\mathbf{X}_d) \\
&= -\frac{\sigma_d^2 X_1}{1+\sigma_d^2} f_1(\mathbf{X}_d) + \frac{\sigma_d^4 X_1^3}{(1+\sigma_d^2)^3} f_2(\mathbf{X}_d) + \frac{\sigma_d^2 X_1}{(1+\sigma_d^2)^2} f_2(\mathbf{X}_d) \\
&= -\sigma_d^2 X_1\left(\frac{f_1(\mathbf{X}_d)}{1+\sigma_d^2} - \frac{f_2(\mathbf{X}_d)}{(1+\sigma_d^2)^2}\right) + \frac{\sigma_d^4 X_1^3}{(1+\sigma_d^2)^3} f_2(\mathbf{X}_d).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left| d^{2\tau}\mathbb{E}\left[(Y_1-X_1)\left(1\wedge\exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2-X_i^2)\right)\right)\right.\right.\\
&\quad\left.+(Y_1-X_1)^2\frac{X_1}{1+\sigma_d^2}\exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2-X_i^2)\right)\mathbb{1}\left(\sum_{i=2}^{d}(Y_i^2-X_i^2)<0\right)\mid\mathbf{X}_d\right]\\
&\quad\left.-\ell^2(\vartheta_{w,\tau}(\ell)/2)(\log\varphi(X_1))'\mid\right.\\
&\leq\mathbb{E}\left|-X_1\ell^2\left(\left(\frac{f_1(\mathbf{X}_d)}{1+\sigma_d^2}-\frac{f_2(\mathbf{X}_d)}{(1+\sigma_d^2)^2}\right)-\frac{\vartheta_{w,\tau}(\ell)}{2}\right)+\frac{\sigma_d^4X_1^3}{(1+\sigma_d^2)^3}f_2(\mathbf{X}_d)\right|\\
&\leq\mathbb{E}\left|-X_1\ell^2\left(\left(\frac{f_1(\mathbf{X}_d)}{1+\sigma_d^2}-\frac{f_2(\mathbf{X}_d)}{(1+\sigma_d^2)^2}\right)-\frac{\vartheta_{w,\tau}(\ell)}{2}\right)\right|+d^{2\tau}\mathbb{E}\left|\frac{\sigma_d^4X_1^3}{(1+\sigma_d^2)^3}\right|,
\end{aligned}$$

using the triangle inequality and that $0\leq f_2(\mathbf{X}_d)\leq 1$. As previously,

$$d^{2\tau}\mathbb{E}\left|\frac{\sigma_d^4X_1^3}{(1+\sigma_d^2)^3}\right|\to 0.$$

We also have that

$$\frac{f_1(\mathbf{X}_d)}{1+\sigma_d^2}-\frac{f_2(\mathbf{X}_d)}{(1+\sigma_d^2)^2}=\frac{\mathbb{E}[\mathbb{1}\left(\sum_{i=2}^{d}(Y_i^2-X_i^2)\geq 0\right)\mid\mathbf{X}_d]}{(1+\sigma_d^2)^2}+\frac{\sigma_d^2f_1(\mathbf{X}_d)}{(1+\sigma_d^2)^2}.$$

Using that $0\leq f_1(\mathbf{X}_d)\leq 1$, the triangle inequality and because

$$\frac{\ell^2\sigma_d^2}{(1+\sigma_d^2)^2}\mathbb{E}|X_1|\to 0,$$

we can now focus on

$$\mathbb{E}\left|-X_1\ell^2\left(\frac{\mathbb{E}\left[\mathbb{1}\left(\sum_{i=2}^{d}(Y_i^2-X_i^2)\geq 0\right)\mid\mathbf{X}_d\right]}{(1+\sigma_d^2)^2}-\frac{\vartheta_{w,\tau}(\ell)}{2}\right)\right|.$$

We have that $Y_i^2-X_i^2=(Y_i-X_i)(Y_i+X_i)$, and as previously, we use that given $\mathbf{X}_d$, we can write $Y_i=\frac{X_i}{1+\sigma_d^2}+\sqrt{\frac{\sigma_d^2}{1+\sigma_d^2}}U_i$, and consequently,

$$\begin{aligned}
Y_i-X_i&=-\frac{\sigma_d^2}{1+\sigma_d^2}X_i+\frac{1}{\sqrt{1+\sigma_d^2}}\sigma_dU_i,\\
Y_i+X_i&=\frac{2+\sigma_d^2}{1+\sigma_d^2}X_i+\frac{1}{\sqrt{1+\sigma_d^2}}\sigma_dU_i,
\end{aligned}$$

and

$$\begin{aligned}
(Y_i-X_i)(Y_i+X_i)=&-\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2X_i^2-\frac{1}{(1+\sigma_d^2)^{3/2}}\sigma_d^3X_iU_i\\
&+\frac{(2+\sigma_d^2)}{(1+\sigma_d^2)^{3/2}}\sigma_dU_iX_i+\frac{1}{1+\sigma_d^2}\sigma_d^2U_i^2.
\end{aligned}$$

We define $S_d := \sum_{i=2}^d Y_i^2 - X_i^2$ and $W_d := -\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2 + \frac{(2+\sigma_d^2)}{(1+\sigma_d^2)^{3/2}}\sigma_d \sum_{i=2}^d U_i X_i + \frac{\ell^2 d}{(1+\sigma_d^2)(d-1)}$. If $\tau = 1/2$, in S_d, we notice that $\frac{1}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d U_i^2 \to \ell^2$ with probability 1 as a result of the strong law of large numbers. Also, $\frac{1}{(1+\sigma_d^2)^{3/2}}\sigma_d^3 \sum_{i=2}^d X_i U_i \to 0$ with probability 1 for the same reason. Therefore, given $\mathbf{X}_d$, S_d follows essentially a normal distribution with mean $-\frac{2+\sigma_d^2}{1+\sigma_d^2}\sum_{i=2}^d \sigma_d^2 X_i^2 + \ell^2$ and variance $\frac{(2+\sigma_d^2)^2}{(1+\sigma_d^2)^3}\sigma_d^2 \sum_{i=2}^d X_i^2$. We want to use this and that is why we will prove that S_d and W_d are asymptotically equivalent; W_d has a conditional normal distribution. We have an explicit expression for $\mathbb{E}\left[\mathbb{1}\left(W_d \geq 0\right) \mid \mathbf{X}_d\right]$ and we can use it.

For the rest of the proof, we consider that $\tau = 1/2$. If $\tau > 1/2$, we can use the same strategy as below, but with $W_d := -\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2 + \frac{(2+\sigma_d^2)}{(1+\sigma_d^2)^{3/2}}\sigma_d \sum_{i=2}^d U_i X_i$ because $\frac{1}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d U_i^2 \to 0$. In this case, W_d has a conditional normal distribution whose mean is $-\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2$ and variance $\frac{(2+\sigma_d^2)^2}{(1+\sigma_d^2)^3}\sigma_d^2 \sum_{i=2}^d X_i^2$. Both converge to 0 with probability 1, but the mean converges quicker than the standard deviation, implying that the limit of the explicit expression for $\mathbb{E}\left[\mathbb{1}\left(W_d \geq 0\right) \mid \mathbf{X}_d\right]$ is $\Phi(0) = 1/2$, which allows to conclude.

Let us now return to the case $\tau = 1/2$. Using the triangle inequality,

$$\begin{aligned}\mathbb{E}\left|-X_1\ell^2\left(\frac{\mathbb{E}\left[\mathbb{1}\left(S_d \geq 0\right) \mid \mathbf{X}_d\right]}{(1+\sigma_d^2)^2} - \frac{\vartheta_{w,\tau}(\ell)}{2}\right)\right| &\leq \mathbb{E}\left|-X_1\ell^2\left(\frac{\mathbb{E}\left[\mathbb{1}\left(W_d \geq 0\right) \mid \mathbf{X}_d\right]}{(1+\sigma_d^2)^2} - \frac{\vartheta_{w,\tau}(\ell)}{2}\right)\right| \\ &\quad + \mathbb{E}\left|-\frac{X_1\ell^2}{(1+\sigma_d^2)^2}\left(\mathbb{E}\left[\mathbb{1}\left(W_d \geq 0\right) \mid \mathbf{X}\right] - \mathbb{E}\left[\mathbb{1}\left(S_d \geq 0\right) \mid \mathbf{X}\right]\right)\right|. \end{aligned} \tag{13}$$

We now prove that the last expectation converges to 0. Using the Cauchy–Schwarz inequality and Jensen's inequality, it is sufficient to prove that $\mathbb{E}[(\mathbb{1}\left(W_d \geq 0\right) - \mathbb{1}\left(S_d \geq 0\right))^2] \to 0$ given that

$$\frac{\ell^2}{(1+\sigma_d^2)^2}\mathbb{E}[X_1^2]^{1/2} = \frac{\ell^2}{(1+\sigma_d^2)^2} \to \ell^2.$$

We have that

$$\begin{aligned}\mathbb{E}[(\mathbb{1}\left(W_d \geq 0\right) - \mathbb{1}\left(S_d \geq 0\right))^2] &= \mathbb{P}(W_d \geq 0, S_d < 0) + \mathbb{P}(W_d < 0, S_d \geq 0) \\ &= \mathbb{P}(W_d \geq 0, S_d < 0, |W_d - S_d| > d^{-1/4}) \\ &\quad + \mathbb{P}(W_d \geq 0, S_d < 0, |W_d - S_d| \leq d^{-1/4}) \\ &\quad + \mathbb{P}(W_d < 0, S_d \geq 0, |W_d - S_d| > d^{-1/4}) \\ &\quad + \mathbb{P}(W_d < 0, S_d \geq 0, |W_d - S_d| \leq d^{-1/4}) \\ &\leq 2\mathbb{P}(|W_d - S_d| > d^{-1/4}) \\ &\quad + \mathbb{P}(W_d \geq 0, S_d < 0, W_d - S_d \leq d^{-1/4}) \\ &\quad + \mathbb{P}(W_d < 0, S_d \geq 0, S_d - W_d \leq d^{-1/4}) \\ &\leq 2\mathbb{P}(|W_d - S_d| > d^{-1/4}) + \mathbb{P}(-d^{-1/4} \leq W_d \leq d^{-1/4}).\end{aligned}$$

Using Markov's inequality,

$$\begin{aligned}\mathbb{P}(|W_d - S_d| > d^{-1/4}) &\leq \frac{\mathbb{E}|W_d - S_d|}{d^{-1/4}} \\ &= \frac{\mathbb{E}\left|\frac{\ell^2 d}{(1+\sigma_d^2)(d-1)} - \frac{1}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d U_i^2 + \frac{1}{(1+\sigma_d^2)^{3/2}}\sigma_d^3 \sum_{i=2}^d X_i U_i\right|}{d^{-1/4}}\end{aligned}$$

Also,

$$\begin{aligned}&d^{1/4}\mathbb{E}\left|\frac{\ell^2 d}{(1+\sigma_d^2)(d-1)} - \frac{1}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d U_i^2\right| \\ &\quad\leq d^{1/4}\mathbb{E}\left[\left(\frac{\ell^2 d}{(1+\sigma_d^2)(d-1)} - \frac{1}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d U_i^2\right)^2\right]^{1/2} \\ &\quad= d^{1/4}\mathrm{var}\left[\frac{1}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d U_i^2\right]^{1/2} = \frac{\ell^2 d^{1/4}\sqrt{d-1}}{(1+\sigma_d^2)d} \to 0,\end{aligned}$$

and

$$d^{1/4}\mathbb{E}\left|\frac{1}{(1+\sigma_d^2)^{3/2}}\sigma_d^3 \sum_{i=2}^d X_i U_i\right| \leq \frac{\ell^3 (d-1) d^{1/4}}{(1+\sigma_d^2)^{3/2} d^{3/2}} \frac{2}{\pi} \to 0.$$

To compute $\mathbb{P}(-d^{-1/4} \leq W_d \leq d^{-1/4})$, we use that given $\mathbf{X}_d$,

$$W_d \sim \mathcal{N}\left(-\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2 + \frac{\ell^2 d}{(1+\sigma_d^2)(d-1)}, \frac{(2+\sigma_d^2)^2}{(1+\sigma_d^2)^3}\sigma_d^2 \sum_{i=2}^d X_i^2\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}(-d^{-1/4} \leq W_d \leq d^{-1/4}) = \mathbb{E}&\left[\Phi\left(\frac{d^{-1/4} + \frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2 - \frac{\ell^2 d}{(1+\sigma_d^2)(d-1)}}{\sqrt{\frac{(2+\sigma_d^2)^2}{(1+\sigma_d^2)^3}\sigma_d^2 \sum_{i=2}^d X_i^2}}\right)\right] \\ &- \mathbb{E}\left[\Phi\left(\frac{-d^{-1/4} + \frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2 - \frac{\ell^2 d}{(1+\sigma_d^2)(d-1)}}{\sqrt{\frac{(2+\sigma_d^2)^2}{(1+\sigma_d^2)^3}\sigma_d^2 \sum_{i=2}^d X_i^2}}\right)\right].\end{aligned}$$

As $d \to \infty$,

$$\frac{\pm d^{-1/4} + \frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2 - \frac{\ell^2 d}{(1+\sigma_d^2)(d-1)}}{\sqrt{\frac{(2+\sigma_d^2)^2}{(1+\sigma_d^2)^3}\sigma_d^2 \sum_{i=2}^d X_i^2}} \longrightarrow \frac{\ell}{2} \quad \text{with probability 1.}$$

So, using Lebesgue's dominated convergence theorem, we know that $\mathbb{P}(-d^{-1/4} \leq W_d \leq d^{-1/4}) \to 0$.

We return to the other term in (13):

$$\mathbb{E}\left|-X_1\ell^2\left(\frac{\mathbb{E}\left[\mathbb{1}\left(W_d \geq 0\right) \mid \mathbf{X}_d\right]}{(1+\sigma_d^2)^2} - \frac{\vartheta_{w,\tau}(\ell)}{2}\right)\right| \\ \leq \mathbb{E}[X_1^2]^{1/2}\ell^2\mathbb{E}\left[\left(\frac{\mathbb{E}\left[\mathbb{1}\left(-W_d \leq 0\right) \mid \mathbf{X}_d\right]}{(1+\sigma_d^2)^2} - \frac{\vartheta_{w,\tau}(\ell)}{2}\right)^2\right]^{1/2},$$

using the Cauchy–Schwarz inequality. We have that

$$\mathbb{E}\left[\mathbb{1}\left(-W_d \leq 0\right) \mid \mathbf{X}_d\right] = \Phi\left(\frac{-\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2\sum_{i=2}^d X_i^2 + \frac{\ell^2 d}{(1+\sigma_d^2)(d-1)}}{\sqrt{\frac{(2+\sigma_d^2)^2}{(1+\sigma_d^2)^3}\sigma_d^2\sum_{i=2}^d X_i^2}}\right) \to \Phi\left(-\frac{\ell}{2}\right),$$

with probability 1. Therefore,

$$\mathbb{E}\left[\left(\frac{\mathbb{E}\left[\mathbb{1}\left(-W_d \leq 0\right) \mid \mathbf{X}_d\right]}{(1+\sigma_d^2)^2} - \frac{\vartheta_{w,\tau}(\ell)}{2}\right)^2\right]^{1/2} \to 0,$$

using Lebesgue's dominated convergence theorem, which concludes this part given that $\mathbb{E}[X_1^2] = 1$.

There remains to prove that

$$\mathbb{E}\left|d^{2\tau}\mathbb{E}\left[h''(X_1)\frac{(Y_1-X_1)^2}{2}\,\alpha_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_d) \mid \mathbf{X}_d\right] - \ell^2\frac{\vartheta_{w,\tau}(\ell)}{2}h''(X_1)\right| \to 0,$$

in (11). We proceed as before with a Taylor expansion around x_1 of $\alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_d)$, viewed as function of y_1. This time it is less complicated because we write

$$\alpha_{\text{ideal}}(\mathbf{x}_d,\mathbf{y}_d) = \alpha_{\text{ideal}}(\mathbf{x}_d,\mathbf{y}_d^*) + \left(\frac{\partial}{\partial y_1}\alpha_{\text{ideal}}(\mathbf{x}_d,\mathbf{y}_d)\Big|_{y_1=w}\right)(y_1-x_1).$$

Using that M can be chosen such that $|h''| \leq M$, the triangle inequality, and that $0 \leq \exp(x)\mathbb{1}(x<0) \leq 1$ (see the partial derivative $\frac{\partial}{\partial y_1}\alpha_{\text{ideal}}(\mathbf{x}_d,\mathbf{y}_d)$ (12)),

$$\mathbb{E}\left|d^{2\tau}\mathbb{E}\left[h''(X_1)\frac{(Y_1-X_1)^2}{2}\,\alpha_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_d) \mid \mathbf{X}_d\right] - \ell^2\frac{\vartheta_{w,\tau}(\ell)}{2}h''(X_1)\right| \\ \leq M\mathbb{E}\left|d^{2\tau}\mathbb{E}\left[\frac{(Y_1-X_1)^2}{2} \mid \mathbf{X}_d\right]f_1(\mathbf{X}_d) - \ell^2\frac{\vartheta_{w,\tau}(\ell)}{2}\right| + \frac{d^{2\tau}}{2(1+\sigma_d^2)}\mathbb{E}[|W||Y_1-X_1|^3].$$

From what we have seen before, we know that

$$\frac{d^{2\tau}}{2(1+\sigma_d^2)}\mathbb{E}[|W||Y_1-X_1|^3] \to 0.$$

We also know that

$$\mathbb{E}\left[\frac{(Y_1 - X_1)^2}{2} \mid \mathbf{X}_d\right] = \mathbb{E}\left[\frac{\sigma_d^4 X_1^2}{2(1+\sigma_d^2)^2} - \frac{\sigma_d^2 X_1}{1+\sigma_d^2}\sqrt{\frac{\sigma_d^2}{1+\sigma_d^2}}U_1 + \frac{\sigma_d^2}{2(1+\sigma_d^2)}U_1^2 \mid \mathbf{X}_d\right].$$

Therefore, using the triangle inequality and that $0 \leq f_1(\mathbf{X}_d) \leq 1$,

$$\begin{aligned}\mathbb{E}\left|d^{2\tau}\mathbb{E}\left[\frac{(Y_1 - X_1)^2}{2} \mid \mathbf{X}_d\right] f_1(\mathbf{X}_d) - \ell^2\frac{\vartheta_{w,\tau}(\ell)}{2}\right| &\leq \ell^2\mathbb{E}\left|\frac{1}{2(1+\sigma_d^2)}f_1(\mathbf{X}_d) - \frac{\vartheta_{w,\tau}(\ell)}{2}\right| \\ &\quad + d^{2\tau}\mathbb{E}\left[\frac{\sigma_d^4 X_1^2}{2(1+\sigma_d^2)^2}\right].\end{aligned}$$

We have that

$$d^{2\tau}\mathbb{E}\left[\frac{\sigma_d^4 X_1^2}{2(1+\sigma_d^2)^2}\right] = \frac{\ell^4}{2d(1+\sigma_d^2)^2} \to 0.$$

To conclude the proof, there thus remains to show that

$$\mathbb{E}\left|\frac{1}{1+\sigma_d^2}\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2 - X_i^2)\right) \mid \mathbf{X}_d\right] - \vartheta_{w,\tau}(\ell)\right| \to 0.$$

We proceed similarly as before when we proved that

$$\mathbb{E}\left|\mathbb{E}\left[\mathbb{1}\left(\sum_{i=2}^{d}(Y_i^2 - X_i^2) \geq 0\right) \mid \mathbf{X}_d\right] - \Phi\left(-\frac{\ell}{2}\right)\right| \longrightarrow 0,$$

when $\tau = 1/2$.

Using the triangle inequality,

$$\begin{aligned}&\mathbb{E}\left|\frac{1}{1+\sigma_d^2}\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2 - X_i^2)\right) \mid \mathbf{X}_d\right] - \vartheta_{w,\tau}(\ell)\right| \\ &\quad\leq \mathbb{E}\left|\frac{1}{1+\sigma_d^2}\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2 - X_i^2)\right) \mid \mathbf{X}_d\right]\right. \\ &\qquad\left.-\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2 - X_i^2)\right) \mid \mathbf{X}_d\right]\right| \\ &\qquad+ \mathbb{E}\left|\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)}\sum_{i=2}^{d}(Y_i^2 - X_i^2)\right) \mid \mathbf{X}_d\right] - \vartheta_{w,\tau}(\ell)\right|.\end{aligned}$$

The first term on the RHS converges to 0 given that $0 \leq 1 \wedge \exp(x) \leq 1$.

Given $\mathbf{X}_d$, we saw that we can write

$$\begin{aligned}(Y_i - X_i)(Y_i + X_i) = &-\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 X_i^2 - \frac{1}{(1+\sigma_d^2)^{3/2}}\sigma_d^3 X_i U_i \\ &+ \frac{(2+\sigma_d^2)}{(1+\sigma_d^2)^{3/2}}\sigma U_i X_i + \frac{1}{1+\sigma_d^2}\sigma_d^2 U_i^2.\end{aligned}$$

We define $S_d := \sum_{i=2}^d Y_i^2 - X_i^2$ and $W_d := -\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2 + \frac{(2+\sigma_d^2)}{(1+\sigma_d^2)^{3/2}}\sigma_d \sum_{i=2}^d U_i X_i + \frac{\ell^2 d}{(1+\sigma_d^2)(d-1)}$. For the rest of the proof, we consider that $\tau = 1/2$. If $\tau > 1/2$, we can use the same strategy as below, but with $W_d := -\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2 + \frac{(2+\sigma_d^2)}{(1+\sigma_d^2)^{3/2}}\sigma_d \sum_{i=2}^d U_i X_i$ because $\frac{1}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d U_i^2 \to 0$. In this case, W_d has a conditional normal distribution whose mean is $-\frac{2+\sigma_d^2}{1+\sigma_d^2}\sigma_d^2 \sum_{i=2}^d X_i^2$ and variance $\frac{(2+\sigma_d^2)^2}{(1+\sigma_d^2)^3}\sigma_d^2 \sum_{i=2}^d X_i^2$. Both converge to 0 with probability 1, but the mean converges quicker than the standard deviation, implying that the limit of the explicit expression for $\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} W_d\right) \mid \mathbf{X}_d\right]$ is $2\Phi(0) = 1$, which allows to conclude.

Let us return to the case $\tau = 1/2$. Using the triangle inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} S_d\right) \mid \mathbf{X}_d\right] - \vartheta_{w,\tau}(\ell)\right|\right] \\
&\quad \leq \mathbb{E}\left[\left|\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} S_d\right) \mid \mathbf{X}_d\right] - \mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} W_d\right) \mid \mathbf{X}_d\right]\right|\right] \\
&\qquad + \mathbb{E}\left[\left|\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} W_d\right) \mid \mathbf{X}_d\right] - \vartheta_{w,\tau}(\ell)\right|\right].
\end{aligned}$$

We now show that each term vanishes. We start with the first one:

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} S_d\right) \mid \mathbf{X}_d\right] - \mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} W_d\right) \mid \mathbf{X}_d\right]\right|\right] \\
&\quad \leq \mathbb{E}\left[\left|1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} S_d\right) - 1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} W_d\right)\right|\right] \\
&\quad \leq \frac{1}{2(1+\sigma_d^2)} \mathbb{E}|S_d - W_d|,
\end{aligned}$$

using Jensen's inequality and that the function $1 \wedge \exp(x)$ is 1-Lipschitz continuous. It has been proved previously that $\mathbb{E}|S_d - W_d| \to 0$.

Given $\mathbf{X}_d$,

$$\frac{1}{2(1+\sigma_d^2)} W_d \sim \mathcal{N}\left(\mu_d, s_d^2\right),$$

with

$$\mu_d := -\frac{2+\sigma_d^2}{2(1+\sigma_d^2)^2}\sigma_d^2 \sum_{i=1}^d X_i^2 + \frac{\ell^2 d}{2(1+\sigma_d^2)^2(d-1)},$$

and

$$s_d^2 := \frac{(2+\sigma_d^2)^2}{4(1+\sigma_d^2)^5}\sigma_d^2 \sum_{i=1}^d X_i^2.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} W_d\right) \mid \mathbf{X}_d\right] &= \Phi\left(\frac{\mu_d}{s_d}\right) + \exp\left(\mu_d + \frac{s_d^2}{2}\right) \Phi\left(-s_d - \frac{\mu_d}{s_d}\right) \\
&\to 2\Phi\left(-\frac{\ell}{2}\right),
\end{aligned}$$

with probability 1. Lebesgue's dominated convergence theorem allows to establish that

$$\mathbb{E}\left[\left|\mathbb{E}\left[1 \wedge \exp\left(\frac{1}{2(1+\sigma_d^2)} W_d\right) \mid \mathbf{X}_d\right] - \vartheta_{w,\tau}(\ell)\right|\right] \to 0,$$

which concludes the proof. ■

A.5 Proof of Theorem 3

Proof [Theorem 3] We first prove that if

$$\mathbb{E}\left[d^{2\tau}\left|\frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i)} - \frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\mathbb{E}[w(\mathbf{X}_d,\mathbf{Y}_1)\mid \mathbf{X}_d]}\right|\right] \leq \frac{d^{2\tau}}{N_d^{1/2}}\,\varrho_1(d) \to 0,$$

and

$$\mathbb{E}[d^{2\tau}|\alpha(\mathbf{X}_d,\mathbf{Y}_J) - \alpha_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_J)|] \leq \frac{d^{2\tau}}{N_d^{1/2}}\,\varrho_2(d) \to 0,$$

then $\{Z_{d,\text{MTM}}(t) : t \geq 0\}$ converges weakly towards the same Langevin diffusion $\{Z(t) : t \geq 0\}$ as in Theorem 2, where $\varrho_1(d)$ and $\varrho_2(d)$ are explicitly defined below.

We saw in the proof of Theorem 2 that to prove a weak convergence towards a diffusion, it is essentially sufficient to prove that the pseudo-generator of $\{Z_{d,\text{MTM}}(t) : t \geq 0\}$ converges towards the generator of the diffusion in the 1-norm. We thus first derive the pseudo-generator of $\{Z_{d,\text{MTM}}(t) : t \geq 0\}$. It is defined as follows:

$$\phi_{d,\text{MTM}}(t) := d^{2\tau}\mathbb{E}[h(Z_{d,\text{MTM}}(t+1/d^{2\tau})) - h(Z_{d,\text{MTM}}(t)) \mid \mathcal{F}_{\mathbf{Z}_{d,\text{MTM}}}(t)],$$

where h is a test function and $\mathcal{F}_{\mathbf{Z}_{d,\text{MTM}}}(t)$ is the natural filtration associated to $\{\mathbf{Z}_{d,\text{MTM}}(t) : t \geq 0\}$. The Markov property, the fact that $\mathbf{Z}_{d,\text{MTM}}(0) \sim \pi_d$ and that $\{\mathbf{X}_{d,\text{MTM}}(m) : m \in \mathbb{N}\}$ is time-homogeneous imply that for any t,

$$\begin{aligned}\phi_{d,\text{MTM}}(t) &= d^{2\tau}\mathbb{E}[h(Z_{d,\text{MTM}}(t+1/d^{2\tau})) - h(Z_{d,\text{MTM}}(t)) \mid \mathbf{Z}_{d,\text{MTM}}(t)] \\ &\overset{\text{dist.}}{=} d^{2\tau}\mathbb{E}[(h(Y_{J,1}) - h(X_1))\,\alpha(\mathbf{X}_d,\mathbf{Y}_J) \mid \mathbf{X}_d],\end{aligned}$$

where the last equality is in distribution, $\mathbf{X}_d \sim \pi_d$ and $Y_{J,1}$ is the first coordinate of $\mathbf{Y}_J$, a proposal generated by MTM.

The convergence in the 1-norm of the pseudo-generator of $\{Z_{d,\text{MTM}}(t) : t \geq 0\}$ towards the generator of the diffusion thus corresponds to

$$\begin{aligned}&\mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_{J,1}) - h(X_1))\,\alpha(\mathbf{X}_d,\mathbf{Y}_J) \mid \mathbf{X}_d] - Gh(X_1)| \\ &\leq \mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_{J,1}) - h(X_1))\,\alpha(\mathbf{X}_d,\mathbf{Y}_J) \mid \mathbf{X}_d] - d^{2\tau}\mathbb{E}[(h(Y_1) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_d) \mid \mathbf{X}_d]| \\ &\quad + \mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_1) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_d) \mid \mathbf{X}_d] - Gh(X_1)|,\end{aligned}$$

using the triangle inequality, where Y_1 is the first coordinate of $\mathbf{Y}_d \sim Q_{w,\sigma}(\mathbf{X}_d,\cdot)$. We saw in the proof of Theorem 2 that

$$\mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_1) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_d) \mid \mathbf{X}_d] - Gh(X_1)| \to 0.$$

Using the triangle inequality,

$$\begin{aligned}
&\mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_{J,1}) - h(X_1))\,\alpha(\mathbf{X}_d, \mathbf{Y}_J) \mid \mathbf{X}_d] - d^{2\tau}\mathbb{E}[(h(Y_1) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d]| \\
&\quad \leq \mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_{J,1}) - h(X_1))\,\alpha(\mathbf{X}_d, \mathbf{Y}_J) \mid \mathbf{X}_d] \\
&\qquad\quad - d^{2\tau}\mathbb{E}[(h(Y_{J,1}) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_J) \mid \mathbf{X}_d]| \\
&\qquad + \mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_{J,1}) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_J) \mid \mathbf{X}_d] \\
&\qquad\quad - d^{2\tau}\mathbb{E}[(h(Y_1) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_d) \mid \mathbf{X}_d]|.
\end{aligned}$$

We now prove that each of the two expectations on the RHS converges to 0 if

$$\mathbb{E}\left[d^{2\tau}\left|\frac{w(\mathbf{X}_d, \mathbf{Y}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d, \mathbf{Y}_i)} - \frac{w(\mathbf{X}_d, \mathbf{Y}_1)}{\mathbb{E}[w(\mathbf{X}_d, \mathbf{Y}_1) \mid \mathbf{X}_d]}\right|\right] \leq \frac{d^{2\tau}}{N_d^{1/2}}\,\varrho_1(d) \to 0,$$

and

$$\mathbb{E}[d^{2\tau}|\alpha(\mathbf{X}_d, \mathbf{Y}_J) - \alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_J)|] \leq \frac{d^{2\tau}}{N_d^{1/2}}\,\varrho_2(d) \to 0.$$

We start with the second one. Given any realisation $\mathbf{x}_d$, we have an explicit expression for the conditional expressions and use them; we thus use the notation $\mathbb{E}_{\mathbf{x}_d}$. Using Proposition 1,

$$\begin{aligned}
&|\mathbb{E}_{\mathbf{x}_d}[(h(Y_{J,1}) - h(x_1))\,\alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{Y}_J)] - \mathbb{E}_{\mathbf{x}_d}[(h(Y_1) - h(x_1))\,\alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{Y}_d)]| \\
&\leq \int |h(y_{1,1}) - h(x_1)|\,\alpha_{\text{ideal}}(\mathbf{x}_d, \mathbf{y}_1)\left|\frac{w(\mathbf{x}_d, \mathbf{y}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{x}_d, \mathbf{y}_i)} - \frac{w(\mathbf{x}_d, \mathbf{y}_1)}{\int w(\mathbf{x}_d, \mathbf{y}_1)\,q_{\sigma_d}(\mathbf{x}_d, \mathbf{y}_1)\,d\mathbf{y}_1}\right| \\
&\qquad \times \prod_{i=1}^{N_d} q_{\sigma_d}(\mathbf{x}_d, \mathbf{y}_i)\,d\mathbf{y}_{1:N_d} \\
&\leq 2M\mathbb{E}_{\mathbf{x}_d}\left|\frac{w(\mathbf{x}_d, \mathbf{Y}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{x}_d, \mathbf{Y}_i)} - \frac{w(\mathbf{x}_d, \mathbf{Y}_1)}{\mathbb{E}_{\mathbf{x}_d}[w(\mathbf{x}_d, \mathbf{Y}_1)]}\right|,
\end{aligned}$$

using Jensen's inequality and the triangle inequality, along with the fact that there exists a positive constant M such that $|h| \leq M$, where $y_{1,1}$ is the first coordinate of $\mathbf{y}_1$.

Regarding the first one,

$$\begin{aligned}
&\mathbb{E}|d^{2\tau}\mathbb{E}[(h(Y_{J,1}) - h(X_1))\,\alpha(\mathbf{X}_d, \mathbf{Y}_J) \mid \mathbf{X}_d] - d^{2\tau}\mathbb{E}[(h(Y_{J,1}) - h(X_1))\,\alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_J) \mid \mathbf{X}_d]| \\
&\quad \leq 2M\mathbb{E}[d^{2\tau}|\alpha(\mathbf{X}_d, \mathbf{Y}_J) - \alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_J)|],
\end{aligned}$$

using Jensen's inequality and the triangle inequality, along with the fact that $|h| \leq M$.

Therefore, if

$$\mathbb{E}\left[d^{2\tau}\left|\frac{w(\mathbf{X}_d, \mathbf{Y}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d, \mathbf{Y}_i)} - \frac{w(\mathbf{X}_d, \mathbf{Y}_1)}{\mathbb{E}[w(\mathbf{X}_d, \mathbf{Y}_1) \mid \mathbf{X}_d]}\right|\right] \leq \frac{d^{2\tau}}{N_d^{1/2}}\,\varrho_1(d) \to 0,$$

and

$$\mathbb{E}[d^{2\tau}|\alpha(\mathbf{X}_d, \mathbf{Y}_J) - \alpha_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_J)|] \leq \frac{d^{2\tau}}{N_d^{1/2}}\,\varrho_2(d) \to 0,$$

then $\{Z_{d,\mathrm{MTM}}(t) : t \geq 0\}$ converges weakly towards $\{Z(t) : t \geq 0\}$.

We now prove that each of these two expectations converges to 0. We first prove that

$$\mathbb{E}\left[d^{2\tau}\left|\frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i)} - \frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\mathbb{E}[w(\mathbf{X}_d,\mathbf{Y}_1)\mid \mathbf{X}_d]}\right|\right] \leq \frac{d^{2\tau}}{N_d^{1/2}}\,\varrho_1(d), \tag{14}$$

and next we prove that

$$\mathbb{E}[d^{2\tau}|\alpha(\mathbf{X}_d,\mathbf{Y}_J) - \alpha_{\mathrm{ideal}}(\mathbf{X}_d,\mathbf{Y}_J)|] \leq \frac{d^{2\tau}}{N_d^{1/2}}\,\varrho_2(d). \tag{15}$$

We have that

$$\begin{aligned}
&\mathbb{E}\left[d^{2\tau}\left|\frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i)} - \frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\mathbb{E}[w(\mathbf{X}_d,\mathbf{Y}_1)\mid \mathbf{X}_d]}\right|\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[d^{2\tau}\left|\frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i)} - \frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\mathbb{E}[w(\mathbf{X}_d,\mathbf{Y}_1)\mid \mathbf{X}_d]}\right| \mid \mathbf{X}_d\right]\right].
\end{aligned}$$

For any realisation $\mathbf{x}_d$, we have an explicit expression for the conditional expectation and therefore write it as follows:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_d}\left[d^{2\tau}\left|\frac{w(\mathbf{x}_d,\mathbf{Y}_1)}{\sum_{i=1}^{N_d} w(\mathbf{x}_d,\mathbf{Y}_i)} - \frac{w(\mathbf{x}_d,\mathbf{Y}_1)}{\mathbb{E}_{\mathbf{x}_d}[w(\mathbf{x}_d,\mathbf{Y}_1)]}\right|\right] \\
&\quad = \mathbb{E}_{\mathbf{x}_d}\left[d^{2\tau}\left|\frac{w(\mathbf{x}_d,\mathbf{x}_d+\sigma_d\mathbf{U}_1)}{\sum_{i=1}^{N_d} w(\mathbf{x}_d,\mathbf{x}_d+\sigma_d\mathbf{U}_i)} - \frac{w(\mathbf{x}_d,\mathbf{x}_d+\sigma_d\mathbf{U}_1)}{\mathbb{E}_{\mathbf{x}_d}[w(\mathbf{x}_d,\mathbf{x}_d+\sigma_d\mathbf{U}_1)]}\right|\right],
\end{aligned}$$

using that $\mathbf{Y}_i \sim q_{\sigma_d}(\mathbf{x}_d,\cdot\,)$ is equal in distribution to $\mathbf{x}_d+\sigma_d\mathbf{U}_i$ with $\mathbf{U}_i := (U_{i,1},\ldots,U_{i,d})$, $U_{i,1},\ldots,U_{i,d}$ being d (conditionally) independent standard normal random variables. We prove the result for the case $w(\mathbf{x}_d,\mathbf{y}_d) = \pi(\mathbf{y}_d)/\pi(\mathbf{x}_d)$; the case $w(\mathbf{x}_d,\mathbf{y}_d) = \sqrt{\pi(\mathbf{y}_d)/\pi(\mathbf{x}_d)}$ is proved similarly.

Using the definition of the GB weight function and the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_d}\left[d^{2\tau}\left|\frac{w(\mathbf{x}_d,\mathbf{x}_d+\sigma_d\mathbf{U}_1)}{\sum_{i=1}^{N_d} w(\mathbf{x}_d,\mathbf{x}_d+\sigma_d\mathbf{U}_i)} - \frac{w(\mathbf{x}_d,\mathbf{x}_d+\sigma_d\mathbf{U}_1)}{\mathbb{E}_{\mathbf{x}_d}[w(\mathbf{x}_d,\mathbf{x}_d+\sigma_d\mathbf{U}_1)]}\right|\right] \\
&\quad = \mathbb{E}_{\mathbf{x}_d}\left[d^{2\tau}\left|\frac{\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_1)\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_i) - \mathbb{E}_{\mathbf{x}_d}[\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_1)]\right)}{\frac{1}{N_d}\sum_{i=1}^{N_d}\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_i)\mathbb{E}_{\mathbf{x}_d}[\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_1)]}\right|\right] \\
&\quad \leq d^{2\tau}\mathbb{E}_{\mathbf{x}_d}\left[\left(\frac{\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d}\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_i)}\right)^2\right]^{1/2} \\
&\qquad \times \mathbb{E}_{\mathbf{x}_d}\left[\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\frac{\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_i)}{\mathbb{E}_{\mathbf{x}_d}[\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_1)]} - 1\right)^2\right]^{1/2}.
\end{aligned}$$

We analyse these two terms separately. First,

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}_d}\left[\left(\frac{\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d}\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_i)}\right)^2\right]^{1/2}\\
&\quad=\mathbb{E}_{\mathbf{x}_d}\left[\left(\frac{\exp\left(-\frac{1}{2}\sum_{j=1}^{d}(x_j+\sigma_d U_{1j})^2\right)}{\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^{d}(x_j+\sigma_d U_{ij})^2\right)}\right)^2\right]^{1/2}\\
&\quad\le\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-2\sigma_d^2\sum_{j=1}^{d}(U_{1j}+x_j/\sigma_d)^2\right)\right]^{1/4}\\
&\qquad\times\mathbb{E}_{\mathbf{x}_d}\left[\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{\sigma_d^2}{2}\sum_{j=1}^{d}(U_{ij}+x_j/\sigma_d)^2\right)\right)^{-4}\right]^{1/4}\\
&\quad\le\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-2\sigma_d^2\sum_{j=1}^{d}(U_{1j}+x_j/\sigma_d)^2\right)\right]^{1/4}\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(2\sigma_d^2\sum_{j=1}^{d}(U_{1j}+x_j/\sigma_d)^2\right)\right]^{1/4}\\
&\quad=\frac{\exp\left(-\frac{\|\mathbf{x}_d\|^2}{2(1+4\sigma_d^2)}\right)}{(1+4\sigma_d^2)^{d/8}}\frac{\exp\left(\frac{\|\mathbf{x}_d\|^2}{2(1-4\sigma_d^2)}\right)}{(1-4\sigma_d^2)^{d/8}}=\frac{\exp\left(\frac{4\sigma_d^2\|\mathbf{x}_d\|^2}{1-16\sigma_d^4}\right)}{(1+4\sigma_d^2)^{d/8}(1-4\sigma_d^2)^{d/8}},
\end{aligned}
$$

using the Cauchy–Schwarz inequality, Proposition 4, and the explicit expression of the moment generating function of a non-central chi-squared distribution. Note that

$$
\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(2\sigma_d^2\sum_{j=1}^{d}(U_{1j}+x_j/\sigma_d)^2\right)\right]
$$

exists for large enough d; it more precisely exists when $4\sigma_d^2<1$.

We now turn to the other term:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}_d}\left[\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\frac{\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_i)}{\mathbb{E}_{\mathbf{x}_d}[\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_1)]}-1\right)^2\right]^{1/2}&=\frac{1}{N_d^{1/2}}\mathrm{var}_{\mathbf{x}_d}\left[\frac{\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_1)}{\mathbb{E}_{\mathbf{x}_d}[\pi_d(\mathbf{x}_d+\sigma_d\mathbf{U}_1)]}\right]^{1/2}\\
&\le\frac{1}{N_d^{1/2}}\frac{\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\sum_{j=1}^{d}(x_j+\sigma_d U_{1j})^2\right)\right]^{1/2}}{\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\frac{1}{2}\sum_{j=1}^{d}(x_j+\sigma_d U_{1j})^2\right)\right]}\\
&=\frac{1}{N_d^{1/2}}\frac{(1+\sigma_d^2)^{d/2}}{(1+2\sigma_d^2)^{d/4}}\exp\left(\frac{\sigma_d^2\|\mathbf{x}_d\|^2}{2(1+2\sigma_d^2)(1+\sigma_d^2)}\right).
\end{aligned}
$$

Putting all that together yields

$$\begin{aligned}
&\mathbb{E}\left[d^{2\tau}\left|\frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i)} - \frac{w(\mathbf{X}_d,\mathbf{Y}_1)}{\mathbb{E}[w(\mathbf{X}_d,\mathbf{Y}_1)\mid \mathbf{X}_d]}\right|\right] \\
&\quad\le \frac{d^{2\tau}}{N_d^{1/2}}\frac{(1+\sigma_d^2)^{d/2}}{(1+2\sigma_d^2)^{d/4}(1-16\sigma_d^4)^{d/8}}\,\mathbb{E}\left[\exp\left(\frac{4\sigma_d^2\|\mathbf{X}_d\|^2}{(1-16\sigma_d^4)}\right)\exp\left(\frac{\sigma_d^2\|\mathbf{X}\|^2}{2(1+2\sigma_d^2)(1+\sigma_d^2)}\right)\right] \\
&\quad= \frac{d^{2\tau}}{N_d^{1/2}}\frac{(1+\sigma_d^2)^{d/2}}{(1+2\sigma_d^2)^{d/4}(1-16\sigma_d^4)^{d/8}}\,\mathbb{E}\left[\exp\left(\frac{\sigma_d^2(9+24\sigma_d^2)\|\mathbf{X}\|^2}{2(1-16\sigma_d^4)(1+2\sigma_d^2)(1+\sigma_d^2)}\right)\right] \\
&\quad= \frac{d^{2\tau}}{N_d^{1/2}}\frac{(1+\sigma_d^2)^{d/2}}{(1+2\sigma_d^2)^{d/4}(1-16\sigma_d^4)^{d/8}}\left(1-\frac{\sigma_d^2(9+24\sigma_d^2)}{(1-16\sigma_d^4)(1+2\sigma_d^2)(1+\sigma_d^2)}\right)^{-d/2},
\end{aligned}$$

using the explicit expression of the moment generating function of a chi-squared distribution. Note that the expectation in the penultimate line exists for large enough d.

When $\tau \ge 1/2$,

$$\varrho_1(d) := \frac{(1+\sigma_d^2)^{d/2}}{(1+2\sigma_d^2)^{d/4}(1-16\sigma_d^4)^{d/8}}\left(1-\frac{\sigma_d^2(9+24\sigma_d^2)}{(1-16\sigma_d^4)(1+2\sigma_d^2)(1+\sigma_d^2)}\right)^{-d/2}$$

converges to a constant. So the expectation converges to 0 if

$$N_d = d^{4\tau(1+\rho)},$$

with any $\rho > 0$.

When $w(\mathbf{x}_d,\mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$, the terms are different, but the speed is the same. Therefore having $\tau = 1/2$ with $N_d = d^{4\tau(1+\rho)}$ also makes the expectation vanish, but we can also use $\tau = 1/6$ with $N_d = (1+\nu)^d$ to make the expectation vanish, ν being any positive constant.

There remains to prove the bound in (15), that is

$$\mathbb{E}[d^{2\tau}|\alpha(\mathbf{X}_d,\mathbf{Y}_J) - \alpha_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_J)|] \le \frac{d^{2\tau}}{N_d^{1/2}}\,\varrho_2(d).$$

We first use that the function $1 \wedge x$ is 1-Lipschitz continuous:

$$\mathbb{E}[d^{2\tau}|\alpha(\mathbf{X}_d,\mathbf{Y}_J) - \alpha_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_J)|] \le \mathbb{E}[d^{2\tau}|r(\mathbf{X}_d,\mathbf{Y}_J) - r_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_J)|],$$

where

$$r(\mathbf{X}_d,\mathbf{Y}_J) := \frac{\pi_d(\mathbf{Y}_J)\,q_{\sigma_d}(\mathbf{Y}_J,\mathbf{X}_d)\,w(\mathbf{Y}_J,\mathbf{X}_d)\Big/\left(\sum_{i=1}^{N_d-1} w(\mathbf{Y}_J,\mathbf{Z}_i) + w(\mathbf{Y}_J,\mathbf{X}_d)\right)}{\pi_d(\mathbf{X}_d)\,q_{\sigma_d}(\mathbf{X}_d,\mathbf{Y}_J)\,w(\mathbf{X}_d,\mathbf{Y}_J)\Big/\left(\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i)\right)},$$

and

$$r_{\text{ideal}}(\mathbf{X}_d,\mathbf{Y}_J) := \frac{\pi_d(\mathbf{Y}_J)\,Q_{w,\sigma_d}(\mathbf{Y}_J,\mathbf{X}_d)}{\pi_d(\mathbf{X}_d)\,Q_{w,\sigma_d}(\mathbf{X}_d,\mathbf{Y}_J)}.$$

Recall that

$$Q_{w,\sigma_d}(\mathbf{x}_d, \mathbf{y}_d) := \frac{w(\mathbf{x}_d, \mathbf{y}_d)\, q_{\sigma_d}(\mathbf{x}_d, \mathbf{y}_d)}{\int w(\mathbf{x}_d, \mathbf{y}_d)\, q_{\sigma_d}(\mathbf{x}_d, \mathbf{y}_d)\, \mathrm{d}\mathbf{y}_d}.$$

Using Proposition 1, we can write $\mathbb{E}[d^{2\tau}|r(\mathbf{X}_d, \mathbf{Y}_J) - r_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_J)|]$ as $\mathbb{E}[d^{2\tau}|r(\mathbf{X}_d, \mathbf{Y}_1) - r_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_1)|]$, where the latter expectation is computed with respect to the following PDF:

$$\pi_d(\mathbf{x}_d) \frac{w(\mathbf{x}_d, \mathbf{y}_1)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{x}_d, \mathbf{y}_i)} \prod_{i=1}^{N_d} q_{\sigma_d}(\mathbf{x}_d, \mathbf{y}_i) \prod_{i=1}^{N_d-1} q_{\sigma_d}(\mathbf{y}_1, \mathbf{z}_i).$$

This means that the expectation $\mathbb{E}[d^{2\tau}|r(\mathbf{X}_d, \mathbf{Y}_1) - r_{\text{ideal}}(\mathbf{X}_d, \mathbf{Y}_1)|]$ can be written as

$$\tilde{\mathbb{E}}\left[d^{2\tau}\left|\frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\frac{1}{N_d}\sum_{i=1}^{N_d-1} w(\mathbf{Y}_1, \mathbf{Z}_i) + w(\mathbf{Y}_1, \mathbf{X}_d)} - \frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1, \mathbf{X}_d)]} \frac{\mathbb{E}_{\mathbf{X}_d}[w(\mathbf{X}_d, \mathbf{Y}_1)]}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d, \mathbf{Y}_i)}\right|\right],$$

with respect to a PDF given by

$$\pi_d(\mathbf{y}_1) \prod_{i=2}^{N_d} q_{\sigma_d}(\mathbf{x}_d, \mathbf{y}_i) \prod_{i=1}^{N_d-1} q_{\sigma_d}(\mathbf{y}_1, \mathbf{z}_i)\, q_{\sigma_d}(\mathbf{y}_1, \mathbf{x}_d),$$

where $\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1, \mathbf{X}_d)]$ is a fonction of the random variable $\mathbf{Y}_1$ for which any realisation $\mathbf{y}_1$ is mapped to

$$\int w(\mathbf{y}_1, \mathbf{x}_d)\, q_{\sigma_d}(\mathbf{y}_1, \mathbf{x}_d)\, \mathrm{d}\mathbf{x}_d;$$

$\mathbb{E}_{\mathbf{X}_d}[w(\mathbf{X}_d, \mathbf{Y}_1)]$ is defined analogously. We noted a change of PDF in the expectation by using the notation " $\tilde{\mathbb{E}}$ ".

Using the triangle inequality,

$$\begin{aligned}
&\tilde{\mathbb{E}}\left[d^{2\tau}\left|\frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\frac{1}{N_d}\sum_{i=1}^{N_d-1} w(\mathbf{Y}_1, \mathbf{Z}_i) + w(\mathbf{Y}_1, \mathbf{X}_d)} - \frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1, \mathbf{X}_d)]} \frac{\mathbb{E}_{\mathbf{X}_d}[w(\mathbf{X}_d, \mathbf{Y}_1)]}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d, \mathbf{Y}_i)}\right|\right] \\
&\quad \le \tilde{\mathbb{E}}\left[d^{2\tau}\left|\frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\frac{1}{N_d}\sum_{i=1}^{N_d-1} w(\mathbf{Y}_1, \mathbf{Z}_i) + w(\mathbf{Y}_1, \mathbf{X}_d)} - \frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1, \mathbf{X}_d)]}\right|\right] \\
&\qquad + \tilde{\mathbb{E}}\left[d^{2\tau}\left|\frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1, \mathbf{X}_d)]} - \frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1, \mathbf{X}_d)]} \frac{\mathbb{E}_{\mathbf{X}_d}[w(\mathbf{X}_d, \mathbf{Y}_1)]}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d, \mathbf{Y}_i)}\right|\right].
\end{aligned}$$

The first expectation on the RHS can be seen as an expectation with respect to the PDF $\pi_d(\mathbf{y}_1)\prod_{i=1}^{N_d-1} q_{\sigma_d}(\mathbf{y}_1, \mathbf{z}_i)\, q_{\sigma_d}(\mathbf{y}_1, \mathbf{x}_d)$; so this expectation is equal to that in (14) and it thus converges to 0 under the same conditions.

For the other one, we have that

$$\begin{aligned}
&\tilde{\mathbb{E}}\left[d^{2\tau}\left|\frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1, \mathbf{X}_d)]} - \frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1, \mathbf{X}_d)]} \frac{\mathbb{E}_{\mathbf{X}_d}[w(\mathbf{X}_d, \mathbf{Y}_1)]}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d, \mathbf{Y}_i)}\right|\right] \\
&= \tilde{\mathbb{E}}\left[d^{2\tau} \frac{w(\mathbf{Y}_1, \mathbf{X}_d)}{\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1, \mathbf{X}_d)]} \frac{1}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d, \mathbf{Y}_i)} \left|\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d, \mathbf{Y}_i) - \mathbb{E}_{\mathbf{X}_d}[w(\mathbf{X}_d, \mathbf{Y}_1)]\right|\right].
\end{aligned}$$

This expectation can be seen as an expectation with respect to the PDF

$$\pi_d(\mathbf{y}_1)\prod_{i=2}^{N_d} q_{\sigma_d}(\mathbf{x}_d,\mathbf{y}_i)\, q_{\sigma_d}(\mathbf{y}_1,\mathbf{x}_d) = \pi(\mathbf{y}_1)\prod_{i=1}^{N_d} q_{\sigma_d}(\mathbf{x}_d,\mathbf{y}_i),$$

using that q_{σ_d} is symmetric. We prove the result for the case $w(\mathbf{x}_d,\mathbf{y}_d) = \pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)$; the case $w(\mathbf{x}_d,\mathbf{y}_d) = \sqrt{\pi_d(\mathbf{y}_d)/\pi_d(\mathbf{x}_d)}$ is proved similarly. For any realisation $\mathbf{y}_1$, we have that

$$\begin{aligned}\frac{w(\mathbf{y}_1,\mathbf{X}_d)}{\mathbb{E}_{\mathbf{y}_1}[w(\mathbf{y}_1,\mathbf{X}_d)]} &= \frac{\pi_d(\mathbf{X}_d)}{\frac{1}{(2\pi)^{d/2}}\mathbb{E}_{\mathbf{y}_1}\left[\exp\left(-\frac{\sigma_d^2}{2}\sum_{j=1}^d (U_{1j}+y_{1j}/\sigma_d)^2\right)\right]}\\ &= \frac{\pi_d(\mathbf{X}_d)}{(2\pi)^{-d/2}(1+\sigma_d^2)^{-d/2}\exp\left(-\frac{1}{2(1+\sigma_d^2)}\sum_{j=1}^d y_{1j}^2\right)},\end{aligned}$$

using that $q_{\sigma_d}(\mathbf{x}_d,\cdot\,)$ is a normal distribution, where, as previously, $U_{11},\ldots,U_{1N_d}$ are (conditionally) independent standard normal random variables. We can thus rewrite the expectation

$$\tilde{\mathbb{E}}\left[d^{2\tau}\frac{w(\mathbf{Y}_1,\mathbf{X}_d)}{\mathbb{E}_{\mathbf{Y}_1}[w(\mathbf{Y}_1,\mathbf{X}_d)]}\frac{1}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i)}\left|\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i) - \mathbb{E}_{\mathbf{X}_d}[w(\mathbf{X}_d,\mathbf{Y}_1)]\right|\right]$$

as

$$\overset{\approx}{\mathbb{E}}\left[d^{2\tau}(1+\sigma_d^2)^{d/2}\frac{\exp\left(-\frac{\sigma_d^2}{2(1+\sigma_d^2)}\|\mathbf{Y}_1\|^2\right)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i)}\left|\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i) - \mathbb{E}_{\mathbf{X}_d}[w(\mathbf{X}_d,\mathbf{Y}_1)]\right|\right],$$

where the expectation is with respect to the PDF $\pi_d(\mathbf{x}_d)\prod_{i=1}^{N_d} q_{\sigma_d}(\mathbf{x}_d,\mathbf{y}_i)$. Using the definition of the GB weight function,

$$\begin{aligned}&\overset{\approx}{\mathbb{E}}\left[d^{2\tau}(1+\sigma_d^2)^{d/2}\frac{\exp\left(-\frac{\sigma_d^2}{2(1+\sigma_d^2)}\sum_{j=1}^d Y_{1j}^2\right)}{\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i)}\left|\frac{1}{N_d}\sum_{i=1}^{N_d} w(\mathbf{X}_d,\mathbf{Y}_i) - \mathbb{E}_{\mathbf{X}_d}[w(\mathbf{X}_d,\mathbf{Y}_1)]\right|\right]\\ &= \overset{\approx}{\mathbb{E}}\left[d^{2\tau}(1+\sigma_d^2)^{d/2}\frac{\exp\left(-\frac{\sigma_d^2}{2(1+\sigma_d^2)}\sum_{j=1}^d Y_{1j}^2\right)}{\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d Y_{ij}^2\right)}\right.\\ &\quad\left.\times\left|\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d Y_{ij}^2\right) - \mathbb{E}_{\mathbf{X}_d}\left[\exp\left(-\frac{1}{2}\sum_{j=1}^d Y_{1j}^2\right)\right]\right|\right].\end{aligned}$$

We omit the $\approx$ above $\overset{\approx}{\mathbb{E}}$ for the rest of the proof to simplify the notation and note that for the rest of the proof the random variables are such that $\mathbf{X}_d \sim \pi_d$ and $\mathbf{Y}_1,\ldots,\mathbf{Y}_{N_d}$ are conditionally independent given $\mathbf{X}_d$ with a distribution given by $\mathbf{Y}_i \sim \mathcal{N}(\mathbf{X}_d,\sigma_d^2\mathbb{I}_d)$.

We have that

$$\begin{aligned}
&\mathbb{E}\left[\frac{\exp\left(-\frac{\sigma_d^2}{2(1+\sigma_d^2)}\sum_{j=1}^d Y_{1j}^2\right)}{\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d Y_{ij}^2\right)}\right. \\
&\qquad\left.\times\left|\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d Y_{ij}^2\right)-\mathbb{E}_{\mathbf{X}_d}\left[\exp\left(-\frac{1}{2}\sum_{j=1}^d Y_{1j}^2\right)\right]\right|\right] \\
&=\mathbb{E}\left[\mathbb{E}\left[\frac{\exp\left(-\frac{\sigma_d^2}{2(1+\sigma_d^2)}\sum_{j=1}^d Y_{1j}^2\right)}{\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d Y_{ij}^2\right)}\right.\right. \\
&\qquad\left.\left.\times\left|\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d Y_{ij}^2\right)-\mathbb{E}_{\mathbf{X}_d}\left[\exp\left(-\frac{1}{2}\sum_{j=1}^d Y_{1j}^2\right)\right]\right|\mid\mathbf{X}_d\right]\right].
\end{aligned}$$

Given any realisation $\mathbf{x}_d$, we have an explicit expression for the conditional expectation and therefore write it as follows:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_d}\left[\frac{\exp\left(-\frac{\sigma_d^2}{2(1+\sigma_d^2)}\sum_{j=1}^d (x_j+\sigma_d U_{1j})^2\right)}{\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d (x_j+\sigma_d U_{ij})^2\right)}\right. \\
&\qquad\left.\times\left|\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d (x_j+\sigma_d U_{ij})^2\right)-\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\frac{1}{2}\sum_{j=1}^d (x_j+\sigma_d U_{1j})^2\right)\right]\right|\right] \\
&\leq\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\frac{\sigma_d^2}{1+\sigma_d^2}\sum_{j=1}^d (x_j+\sigma_d U_{1j})^2\right)\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d (x_j+\sigma_d U_{ij})^2\right)\right)^{-2}\right]^{1/2} \\
&\qquad\times\mathbb{E}_{\mathbf{x}_d}\left[\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^d (x_j+\sigma_d U_{ij})^2\right)-\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\frac{1}{2}\sum_{j=1}^d (x_j+\sigma_d U_{1j})^2\right)\right]\right)^2\right]^{1/2}
\end{aligned}$$

using that $\mathbf{Y}_i\sim q_{\sigma_d}(\mathbf{x}_d,\cdot\,)$ is equal in distribution to $\mathbf{x}_d+\sigma_d\mathbf{U}_i$ and the Cauchy–Schwarz inequality.

We analyse the two terms on the RHS separately. First,

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\frac{\sigma_d^2}{1+\sigma_d^2}\sum_{j=1}^{d}(x_j+\sigma_d U_{1j})^2\right)\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^{d}(x_j+\sigma_d U_{ij})^2\right)\right)^{-2}\right]^{1/2}\\
&\quad\leq \mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\frac{4\sigma_d^2}{1+\sigma_d^2}\sum_{j=1}^{d}(x_j+\sigma_d U_{1j})^2\right)\right]^{1/4}\\
&\qquad\times \mathbb{E}_{\mathbf{x}_d}\left[\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^{d}(x_j+\sigma_d U_{ij})^2\right)\right)^{-4}\right]^{1/4}\\
&\quad\leq \mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\frac{4\sigma_d^2}{1+\sigma_d^2}\sum_{j=1}^{d}(x_j+\sigma_d U_{1j})^2\right)\right]^{1/4}\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(2\sum_{j=1}^{d}(x_j+\sigma_d U_{1j})^2\right)\right]^{1/4}\\
&\quad= \mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\frac{4\sigma_d^4}{1+\sigma_d^2}\sum_{j=1}^{d}(x_j/\sigma_d+U_{1j})^2\right)\right]^{1/4}\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(2\sigma_d^2\sum_{j=1}^{d}(x_j/\sigma_d+U_{1j})^2\right)\right]^{1/4}\\
&\quad= \frac{\exp\left(-\frac{\sigma_d^2\|\mathbf{x}_d\|^2}{1+\sigma_d^2+8\sigma_d^4}\right)}{\left(1+\frac{8\sigma_d^4}{1+\sigma_d^2}\right)^{d/8}}\frac{\exp\left(\frac{\|\mathbf{x}_d\|^2}{2(1-4\sigma_d^2)}\right)}{\left(1-4\sigma_d^2\right)^{d/8}}\\
&\quad\leq \frac{1}{\left(1+\frac{8\sigma_d^4}{1+\sigma_d^2}\right)^{d/8}}\frac{\exp\left(\frac{\|\mathbf{x}_d\|^2}{2(1-4\sigma_d^2)}\right)}{\left(1-4\sigma_d^2\right)^{d/8}}
\end{aligned}
$$

using the Cauchy–Schwarz inequality, Proposition 4, and the expression for the moment generating function of a non-central chi-squared distribution. Note that

$$
\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(2\sigma_d^2\sum_{j=1}^{d}(x_j/\sigma_d+U_{1j})^2\right)\right]
$$

exists for large enough d; it more precisely exists when $4\sigma_d^2<1$.

Second,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_d}\left[\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^{d}(x_j+\sigma_d U_{ij})^2\right)-\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\frac{1}{2}\sum_{j=1}^{d}(x_j+\sigma_d U_{1j})^2\right)\right]\right)^2\right]^{1/2}\\
&=\frac{1}{N_d^{1/2}}\mathrm{var}_{\mathbf{x}_d}\left[\exp\left(-\frac{1}{2}\sum_{j=1}^{d}(x_j+\sigma_d U_{1j})^2\right)\right]^{1/2}\\
&\leq\frac{1}{N_d^{1/2}}\mathbb{E}_{\mathbf{x}_d}\left[\exp\left(-\sigma_d^2\sum_{j=1}^{d}(x_j/\sigma_d+U_{1j})^2\right)\right]^{1/2}\\
&=\frac{1}{N_d^{1/2}}\frac{\exp\left(\frac{-\|\mathbf{x}_d\|^2}{2(1+2\sigma_d^2)}\right)}{(1+2\sigma_d^2)^{d/4}}.
\end{aligned}$$

Putting all the results above together yields

$$\begin{aligned}
&\mathbb{E}\left[d^{2\tau}(1+\sigma_d^2)^{d/2}\frac{\exp\left(-\frac{\sigma_d^2}{2(1+\sigma_d^2)}\sum_{j=1}^{d}Y_{1j}^2\right)}{\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^{d}Y_{ij}^2\right)}\right.\\
&\quad\left.\times\left|\frac{1}{N_d}\sum_{i=1}^{N_d}\exp\left(-\frac{1}{2}\sum_{j=1}^{d}Y_{ij}^2\right)-\mathbb{E}_{\mathbf{X}_d}\left[\exp\left(-\frac{1}{2}\sum_{j=1}^{d}Y_{1j}^2\right)\right]\right|\right]\\
&\leq\frac{d^{2\tau}}{N_d^{1/2}}\frac{(1+\sigma_d^2)^{d/2}}{\left(1+\frac{8\sigma_d^4}{1+\sigma_d^2}\right)^{d/8}\left(1-4\sigma_d^2\right)^{d/8}(1+2\sigma_d^2)^{d/4}}\mathbb{E}\left[\exp\left(\frac{3\sigma_d^2\|\mathbf{X}_d\|^2}{(1-4\sigma_d^2)(1+2\sigma_d^2)}\right)\right]\\
&=\frac{d^{2\tau}}{N_d^{1/2}}\frac{(1+\sigma_d^2)^{d/2}}{\left(1+\frac{8\sigma_d^4}{1+\sigma_d^2}\right)^{d/8}\left(1-4\sigma_d^2\right)^{d/8}(1+2\sigma_d^2)^{d/4}\left(1-\frac{6\sigma_d^2}{(1-4\sigma_d^2)(1+2\sigma_d^2)}\right)^{d/2}}
\end{aligned}$$

using the explicit expression for the moment generating function of a chi-squared distribution. Note that the expectation in the penultimate line exist for large enough d. The last term, seen as a function of d, behaves as that in the bound (14), and thus converges to 0 under the same conditions. Note that

$$\begin{aligned}
\varrho_2(d):=&\frac{(1+\sigma_d^2)^{d/2}}{(1+2\sigma_d^2)^{d/4}(1-16\sigma_d^4)^{d/8}}\left(1-\frac{\sigma_d^2(9+24\sigma_d^2)}{(1-16\sigma_d^4)(1+2\sigma_d^2)(1+\sigma_d^2)}\right)^{-d/2}\\
&+\frac{(1+\sigma_d^2)^{d/2}}{\left(1+\frac{8\sigma_d^4}{1+\sigma_d^2}\right)^{d/8}\left(1-4\sigma_d^2\right)^{d/8}(1+2\sigma_d^2)^{d/4}\left(1-\frac{6\sigma_d^2}{(1-4\sigma_d^2)(1+2\sigma_d^2)}\right)^{d/2}}.
\end{aligned}$$

■