Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

An Eulerian approach to the regularized JKO scheme with low-rank tensor decompositions for Bayesian inversion

Vitalii Aksenov, Martin Eigel

submitted: November 28, 2024

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: vitalii.aksenov@wias-berlin.de
martin.eigel@wias-berlin.de

No. 3143
Berlin 2024

2020 *Mathematics Subject Classification.* 46E27, 49Q22, 62F15, 68W25.

Key words and phrases. Wasserstein distance, JKO scheme, low-rank tensor decompositions, Bayesian inverse problems.

VA acknowledges the support of the EMPIR project 20IND04-ATMOC. This project (20IND04 ATMOC) has received funding from the EMPIR programme co-financed by the Participating States and from the European Union's Horizon 2020 research and innovation programme. ME was supported by ANR-DFG project "COFNET" and DFG SPP 2298 "Theoretical Foundations of Deep Learning".

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

An Eulerian approach to the regularized JKO scheme with low-rank tensor decompositions for Bayesian inversion

Vitalii Aksenov, Martin Eigel

Abstract

The possibility of using the Eulerian discretization for the problem of modelling high-dimensional distributions and sampling, is studied. The problem is posed as a minimization problem over the space of probability measures with respect to the Wasserstein distance and solved with the entropy-regularized JKO scheme. Each proximal step can be formulated as a fixed-point equation and solved with accelerated methods, such as Anderson's. The usage of the low-rank Tensor Train format allows to overcome the *curse of dimensionality*, i.e. the exponential growth of degrees of freedom with dimension, inherent to Eulerian approaches. The resulting method requires only pointwise computations of the unnormalized posterior and is, in particular, gradient-free. Fixed Eulerian grid allows to employ a caching strategy, significally reducing the expensive evaluations of the posterior. When the Eulerian model of the target distribution is fitted, the passage back to the Lagrangian perspective can also be made, allowing to approximately sample from the distribution. We test our method both for synthetic target distributions and particular Bayesian inverse problems and report comparable or better performance than the baseline Metropolis-Hastings MCMC with the same amount of resources. Finally, the fitted model can be modified to facilitate the solution of certain associated problems, which we demonstrate by fitting an importance distribution for a particular quantity of interest.

We release our code at https://github.com/viviaxenov/rJKOtt

1 Introduction

The present paper focus on acquiring a tractable model of some complicated probability distribution ρ_∞ such that an efficient sampling process is achieved. We assume that this distribution is absolutely continuous on some $X \subseteq \mathbb{R}^d$ and in the notation do not distinguish between the distribution and its probability density function. We focus on the setting where the target distribution is accessible via pointwise evaluation of the unnormalized probability density function. One of the most important practical tasks falling into this setting is the Bayesian inverse problem, which we consider as an application of the proposed method. Oftentimes, parameters of a physical system need to be estimated, along with the quantification of the uncertainty of the estimation, from indirect measurements. In the Bayesian setting, this is done by studying the Bayesian posterior distribution

$$\rho_\infty(x|y) \propto \rho_0(x)\ell(F(x), y),$$

where x stands for the unknown parameters, the *forward operator* F describes the response of the system with given parameters x, y is the noisy measurement of the said response, and ρ_0 and l describe the belief on the prior distribution of the parameters and the distribution of noise. For example, if an inverse problem for a heat equation is considered, x could be the parameters, describing the initial distribution of the temperature, and the operator F would map these parameters to temperatures at a set of fixed temporal and spatial points, in which the measurements were performed.

DOI 10.20347/WIAS.PREPRINT.3143 Berlin 2024

The evaluation of F quite often requires computationally intense calls to numerical solvers for the underlying PDE- or ODE-based models. In application with more complex measurement systems such as astrophysical ones, the computation of the likelihood function can also be an expensive subroutine [1]. Although either uniform or Gaussian priors are usually assumed, it can be advantageous to use more involved priors, e.g. derived from data in medical imaging problems [2]. Thus, the computation of the posterior distribution easily becomes the bottleneck in the numerical algorithms for such problems. Hence, the number of these evaluations required to produce an sufficiently accurate sampling algorithm is a suitable measure to compare the performance of different approaches.

Although in principle the formula above contains all information about the posterior, actual uncertainty quantification tasks such as evaluation of expectations and credible intervals of quantities of interest requires numerical integration. In general, its complexity scales exponentially with the dimension d of the problem (i.e. the number of model parameters to be determined), rendering this approach impractical even for moderate dimensions.

1.1 Sampling approaches

One of the possible approaches is to produce an approximate sample from the target, i.e. $\{x_i\}_{i=1}^N$: $x_i \sim \rho_\infty$i.i.d.. A large class of sampling methods, called *Markov Chain Monte-Carlo*, iterate through a Markov Chain, which has the target as its invariant distribution. In the Metropolis-Hastings algorithm, a new iterate is proposed based on the previous one, and then the proposal is accepted or rejected, with the criterion being designed to satisfy the *detailed balance* condition, which, in turn, guarantees that the unique stationary distribution is ρ_∞. If the gradient information can be utilised, one can consider the *Langevin dynamics* method [3]. Noticing that the stochastic differential equation (SDE)

$$\mathrm{d}X_t = \frac{1}{2}\nabla \log \rho_\infty(X_t)\mathrm{d}t + \mathrm{d}W_t \tag{1.1.1}$$

with W_t is the standard Brownian motion, has ρ_∞ as a stationary distribution and converges in probability and its law converges in distribution to it, one can consider a discretization of this equation, for example, with Euler-Maruyama method. Since the time discretization converges to the target distribution only when time step converges to zero, it can be combined with a Metropolis-Hastings rejection step, yielding the *Metropolis-adjusted Langevin Algorithm* (MALA). It is a good practice to run multiple chains in parallel, for example, to reduce variance in the estimation of the autocorrelation time of the chain or track convergence with Gelman-Rubin criterion [4]. The next step in this direction is to consider the iterates in the parallel chains as an ensemble of particles, and use the information from the whole ensemble to further improve the properties of the algorithm. In [5], a set of proposal distributions, defined on the whole ensemble, is proposed, which ensures the affine invariance property of the algorithm. In practice, this property means that convergence rate is invariant under the affine transformation in the state space. Combining the idea of the particle ensemble and gradient information yields the *Affine-Invariant Langevin Dynamic* (ALDI) sampler [6], which is a time discretization of the SDE

$$\mathrm{d}X_t^{(i)} = C(X_t)\nabla_{X_t^{(i)}} \log \rho_\infty(X_t^{(i)})\mathrm{d}t + \frac{d+1}{N}\left(X_t^{(i)} - m(X_t)\right)\mathrm{d}t + \sqrt{2}C^{\frac{1}{2}}(X_t)\mathrm{d}W_t^{(i)} \tag{1.1.2}$$

with $X_t = \left(X_t^{(1}, \ldots, X_t^{(N)}\right)$ being the ensemble of N particles, and $m(X_t),\ C(X_t)$ are respectively the empirical mean and covariance of the ensemble. The efficiency and robustness of this method can be further improved with such numeric techniques as ensemble enrichment and homotopy [7].

The number of posterior calls in a Metropolis-Hastings MCMC can be estimated as $O(N_{\text{chains}} \cdot N_{\text{steps}})$. Usually, a (possibly substantial) fraction of the initial steps, referred to as *burn-in stage*, is discarded since the distribution of the chains first has to converge to the invariant distribution to be admissible. The consequent draws are not independent with the chain requiring multiple steps to «forget» the influence of the predecessors, which can be estimated by the *autocorrelation time* τ_{ac}. Because of this, only a $\frac{1}{\tau_{ac}}$ fraction of the performed steps yields independent samples. The computation of the gradient, e.g. required by MALA and ALDI, in principle can be achieved in $O(1)$ time compared to the computation time of the posterior if the numerical implementation of ρ_∞ supports *fast automatic differentiation* [8]. Practically, this is not the case if a rather complicated numerical solver for the forward model is used. The gradient hence might have to be estimated via finite-difference formulas, adding an $O(d)$ factor to the complexity. All these considerations render MCMC approaches time-consuming. Additionally, the extensive amount of the posterior calls cannot be reused in any way.

1.2 Gradient flow structure of sampling approaches

The seminal paper of Jordan, Kinderlehrer and Otto [9], the Langevin process (1.1.1) is viewed from the perspective of the geometry of the space of probability distribution endowed with the Wasserstein distance. To be more precise, the process is proved to be the gradient flow of the *Kullback-Leibler divergence*

$$\mathrm{KL}(\rho|\rho_\infty) = \int \log\left(\frac{\rho}{\rho_\infty}\right) \mathrm{d}\rho, \tag{1.2.1}$$

which qualitatively means that the distribution is gradually changing in the direction of the steepest descent of the functional. In general, a gradient flow of a functional $\mathcal{E}$ starting from some initial distribution ρ^0 can be defined as a continuous time limit (in case one exists) of the proximal scheme

$$\rho^{k+1} = \arg\min_\rho \mathcal{E}(\rho) + \frac{1}{2T_k} W_2^2(\rho, \rho^k), \quad k = 0, 1, \dots \tag{1.2.2}$$

as $\max_k T_k \to 0$. The scheme is referred to as *minimizing movements* (MM) or, in case that $\mathcal{E} = \mathrm{KL}(\cdot|\rho_\infty)$, the *Jordan, Kinderlehrer and Otto* (JKO) scheme. The mathematical background regarding existence, uniqueness and rates of convergence to the optimum for the gradient flows is comprehensively presented in [10]. The theory provides a unified framework for mathematical modelling of a multitude of natural processes, which are characterized by conservation of mass and monotone decrease of a certain functional. Evolutions governed by diffusion, porous media [11], quantum drift-diffusion [12] or Keller-Segel [13] equations all exhibit the structure of a gradient flow with respect to the Wasserstein distance. The approach can be extended to models with a discontinuous functional, where the PDE model does not exist, e.g. crowd motion with congestion [14].

From a numerical perspective, using the Wasserstein framework allows to construct novel algorithms for the evolution of distributions based on the various reformulations and regularization of the optimization problem (1.2.2). For instance in [15, 16] the primal-dual reformulation is studied. Another important variation is based on the entropic regularization of the problem. With this, the respective entropic JKO scheme is studied in [17]. It is shown that under a certain coupling law between the time step and the amount of the entropic regularization, the limit of the entropic scheme is the original gradient flow. Numerical examples for this scheme are presented, for example, in [18]. The approach presented relies on using the proximal stepping in the Wasserstein directly, instead of considering the PDE model of its continuous time limit, the latter either not existing, or being excessively complex. This idea is demonstrated by tackling such problems as crowd motion with congestion and a gradient flow on a Riemannian manifold.

In cases when gradient flows are viewed in the context of modelling natural processes, the dimension of the problem is typically not higher than $d = 3$ and Eulerian approaches with spatial discretization of densities and other appropriate functions are frequently used. There are examples of finite difference [19], finite volume [20] and finite element [15] discretizations. However, as the amount of degrees of freedom required for the spatial discretization normally scales exponentially with the dimension, Lagrangian techniques are preferred in higher dimensions. These approaches consider either the behaviour of a particle ensemble or parametrize a flow transporting particles in such a way that they have the target distribution. A notable example is *Stein variational gradient descent* (SVGD) introduced in [21] and analyzed in the framework of gradient flows in [22]. In this method one assumes the intermediate transports appearing in the gradient flow to lie in a certain *reproducing kernel Hilbert space* with which a tractable deterministic update formula to compute the evolution of the ensemble is obtained. The approach proves to be quite powerful and can for instance be extended to a gradient-free setting [23] or to gradient flows with respect to different distances [24].

1.3 Low-rank Tensor decompositions

A central motivation for this work is that – in our opinion – the explicit discretization as in Eulerian methods is a more structured and organized way to model posterior distributions than using Langrangian approaches as mentioned above. For the latter, the position of the next particle is not known in advance and each step requires the renewed computation of the movement toward the posterior distribution. The new position might in fact be close to one of the previously encountered positions and thus not bear a lot of additional information to improve the solution. In the Eulerian approach, the posterior is evaluated on a prescribed grid. We hypothesize that if the posterior has some additional structure, the values on the entire grid can be approximated within reasonable tolerance by computing only the values at a few nodes. Additionally, since the nodes are known in advance, a caching strategy can be implemented, rendering the posterior calls reusable. Additionally, the Eulerian discretization can be adapted to the particular problem at hand and the required precision of the solution.

The goal of the current work is to explore how Eulerian approaches can be efficiently adapted for higher dimensions ($d \geq 4$). To alleviate the exponential growth of the number of degrees of freedom in terms of the dimension, a compression format for multidimensional arrays (*tensors*) is required. One has to represent the initial data in the compressed format and be able to perform all required operations of the algorithm exactly or approximately without leaving the format. In the current work we focus of the popular *Tensor-Train* (TT) decomposition [25]. A tensor $A \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ is in the TT format if it has the form

$$A_{i_1 \dots i_d} = A^1_{i_1 j_1} A^2_{j_1 i_2 j_2} \cdot \dots \cdot A^d_{j_{d-1} i_d}, \tag{1.3.1}$$

where for each $k \in \overline{1,\ d}$, $A^k_{j_{k-1} i_k j_k} \in \mathbb{R}^{r_{k-1} \times N_k \times r_k}$ is a 3-dimensional tensor, $r_0 = r_k = 1$ and Einstein summation convention is assumed. Values r_k are called *TT-ranks*. If $\max_k r_k = r,\ \max_k N_k = N$, the storage complexity of the tensor is only $\mathcal{O}(dNr^2)$, which in case of low-rank r can be drastically less than $\mathcal{O}(N^d)$ required for the storage of the uncompressed full tensor.

For a given tensor in format (1.3.1), its approximation with lower rank and controlled error can be computed by the TT-SVD algorithm [**?**]. Linear algebra operations such as linear combinations, index contractions, inner products or norms can also be efficiently computed. If the TT decomposition is used for the discretization of a function, index contraction corresponds to integrating with respect to a subset of variables. This means that some typically complex tasks such as the computation of marginal distributions become tractable. A plethora of approximate methods have been adapted for the TT format. For example, a general minimization problem with respect to an argument in TT format can

be solved with the *alternating linear scheme* (ALS) or Modified ALS (MALS), also known as *density matrix renormalization group* (DMRG) [26]. In the setting where one strives to obtain a low-rank TT approximation of a tensor, each individual entry can be computed by *tensor completion*, e.g. by using a *cross approximation* (TT-cross) [25].

With various algorithms at hand, the TT format has been applied to a variety of practical problems. Of relevance to the current topic would be, firstly, density estimation where TT are either used directly [27] or in conjunction with Normalizing Flows (NF) [28]. The approximation of distributions, accessible only via pointwise evaluations of unnormalized densities, was studied in [29]. In this paper, the strategy is to construct a series of intermediate densities. A transport between them is approximated via the *inverse Rosenblatt transport*. Additionally, tensor trains are reported to be used for PDEs in genereal (see review [30]), and, in particular, parabolic PDEs, similar to the ones appearing in the current work [31, 32].

1.4 Contribution and paper organization

In this paper, we present a novel method for approximating probability distributions. It aims at applications in Bayesian inversion [33]. The underlying idea is to minimize a functional $\mathcal{E} = \mathrm{KL}(\cdot|\rho_\infty)$, which quantifies the dissimilarity to the posterior distribution. The entropic regularization of the scheme (1.2.2) is used to produce a discrete sequence of successive approximations of the target distribution in a tractable compressed representation. The optimality condition for the proximal step is reformulated in terms of entropic potentials as a system of nonlinearily coupled heat equations. The system is explicitly discretized in an Eulerian way. Subsequently, the heat equations are solved by means of finite difference methods on a high-dimensional grid. The coupling is then considered a fixed-point problem, which is tackled by accelerated fixed-point methods. To the authors' knowledge, this is the first use of fixed-point methods together with the low-rank tensor-train format. As it will be revealed below, the underlying dynamical formulation of the proximal steps provides dynamics that interpolate between the approximate measures in the form of an ODE or an SDE. Numeric solution of these equations provides a way to efficiently draw samples from the model distributions.

The presented method has the same inputs as the common sampling method such as Metropolis-Hastings MCMC or SVGD, namely the unnormalized posterior density and general a priori parameter bounds. We implement the caching of the posterior calls during the solution process. In the Bayesian setting, these contain the expensive solution operator of the model. Together with the accelerated fixed-point scheme, the caching noticeably increases the efficiency of the method. Once the model is fitted, certain computations can be performed without additional posterior calls. For example, a normalized approximation of the posterior or its marginal distributions are accessible. This is important for Bayesian inversion since any statistical quantification of the determined parameters can be carried out efficiently. The approximation of the dynamics, which evolves the particles to the target distribution, is also possible. This means that an arbitrary amount of the samples from the approximate posterior can be drawn without additional (costly) posterior calls. Finally, the acquired model can be modified to facilitate the solution of a sampling problem for a distribution, similar to the Bayesian posterior. We demonstrate it by approximating the optimal importance distribution for a certain quantity of interest.

Previous research closest to the present one can be found in [32] and [34]. In [32], the Fokker-Planck equation with a general, possibly time-dependent drift term is considered. The solution is determined by splitting the operator, describing either the drift or diffusion process. The diffusion term is quite well suited for approximation in the TT format due to its intrinsic low-rank structure. We utilize a similar approach to solve the heat equation that arises in our formulation. The solution of the drift part can be

obtained pointwise by solving the characteristic ODEs. The TT representation for the whole solution is obtained by means of a tensor reconstruction with the TT-cross algorithm. Formally, this method can be applied for the approximation of a target density by setting the drift term f equal to the score of the distribution: $f(x,t) = \nabla \log \rho_\infty(x)$. However, the implementation would require the gradient of the posterior, which to be avoided for the sake of computational efficacy. Although quite similar to our method in the range of applied numerical techniques, the other method comes from a different context for simulating an actual physical system. Thus, it is not clear if the method would be numerically efficient in a setting where only approaching the invariant distribution is of importance. In [34], the same entropy-regularized proximal steps are applied to a linearized version of the target functional. The authors argue that under a certain choice of coupling between the time step and the regularization factor, the limit of the regularized scheme for the linearized functional is again the desired distribution. With a TT-cross approximation of a certain intermediate quantity, an efficient deterministic update formula is acquired and a particle ensemble can be evolved towards the target in a fashion somewhat similar to SVGD. We demonstrate that the regularized step for the original method, despite requiring the solution of a nonlinear coupling problem, can still be tackled with appropriate tools. Moreover, we argue that the direct approximation of the potentials and densities opens up possibilities to extension of the method. An example is importance sampling for the expectations of quantities of interest with respect to the posterior. Additionally, the two aforementioned papers report experiments in dimensions up to $d = 6$. While it is already impossible to tackle the considered problem numerically without any compression such as tensor trains, in our opinion this dimension is still at the lower end for reasonable practical applications. Thus, we verify the performance of our method in much higher dimension up to $d = 30$.

The remaining paper is structured as follows. In the second section, reference information on Wasserstein spaces is provided. In the third section, the PDE system underlying the method is derived. The fourth section contains details of the implementation. The subsequent two sections contain the numerical experiments. Firstly, the method is verified on trial distributions. We then present the use cases of a Bayesian inverse problem with inference of an initial condition with a parabolic and a hyperbolic PDE. The manuscript is concluded with a short outlook on possible further developments.

2 Metric space of measures with respect to the Wasserstein distance

The overall idea of the current approach is to treat the task of sampling from the target distribution or to approximate it with a tractable model as a minimization of a certain functional, that quantifies the dissimilarity to the target. We denote the functional $\mathcal{E}$ and in what follows fix it to be the KL divergence:

$$\mathcal{E}(\rho) = \mathrm{KL}(\rho|\rho_\infty) = \int \log\left(\frac{\rho}{\rho_\infty}\right) \mathrm{d}\rho.$$

This functional is of particular importance because the unnormalized density is used in its computation instead of the true one. Since the value of the functional would only change by a constant the minimizer stays unchanged. In the sequel, however, we would use the $\mathcal{E}$ notation for the functional, except for explicitly mentioned parts, to signify that the proximal method can be used for a variety of functionals (as long as the aforementioned parts can be efficiently tackled numerically).

Probability measures lack a natural linear space structure. However, if a metric structure is imposed, an analog of a steepest descent can be defined, either as a discrete sequence of proximal steps

(minimizing movements) or as a continuous process (a *gradient flow*). The Wasserstein distance is based on the solution of an optimal transport problem. Unlike other metrics defined on probability measures such as Hellinger or total variance (TV), it is takes into account the metric of the underlyig space, thus being «aware» of its the topology. Additionally, transport maps that are optimal with respect to this distance can be shown to exist under certain conditions. This motivates numerous applications of the Wasserstein distance, both as an element of the analysis in PDE theory (e.g. [9, 11, 35, 36] and references therein) and as a practical tool in statistics [37], machine learning [38], image processing [39] and others. [10] provides a detailed reference on the gradient flows both in general metric spaces and in the Wasserstein space. In the rest of the section we recall the concepts required for the definition of our method.

2.1 Wasserstein distance and its dynamic reformulation

In the most generality, the Wasserstein distance can be defined on any separable metric space X with the Radon property given by

$$W_p^p(\rho_0, \rho_1) = \min_{\pi \in \Pi(\rho_0, \rho_1)} \int_{X \times X} d^p(x, y) \mathrm{d}\pi(x, y) \tag{2.1.1}$$

with Borel probability measures $\rho_0,\ \rho_1$ on X, the set of probability measures on $X \times X$ with marginals ρ_0 and ρ_1 denoted by $\Pi(\mu, \nu)$ metric d on X. The set of probability measures with finite p-moments defined by

$$\mathscr{P}_p(X) = \left\{ \rho \in \mathscr{P}(X) : \int_X d^p(x, x^0) \mathrm{d}\rho(x) \right\} \tag{2.1.2}$$

is a separable metric space with respect to the p-Wasserstein distance. It is complete if X is complete. Henceforth we consider $X = \mathbb{R}^d$ with Euclidean metric $d(x, y) = \|x - y\|_2^2$ and the $2-$Wasserstein.

One can study a notion of curves in the Wasserstein space to define the dynamics, that smoothly interpolates between the initial and terminal distributions.

Definition 2.1. *Absolutely continuous curve A curve* $\rho(t) : [a, b] \mapsto \mathscr{P}_2(X)$ *is called* absolutely continuous *if* $\exists m \in L^1([a, b])$ *such that*

$$W_2^2(\rho(s), \rho(t)) \leq \int_s^t m(r) dr, \quad \forall a \leq s < t \leq b.$$

In $\mathscr{P}_2$ there is an important characterization of the absolutely continuous curves with the continuity equation.

Theorem 2.1. *Continuity equation [10, Theorem 8.3.1] There exists a Borel vector field* $v_t \in L^2(\rho_t, X) : \|v_t\|_{L^2(\rho_t, X)} \leq |\rho_t'|$ *and the continuity equation*

$$\partial_t \rho_t + \nabla(\rho_t v_t) = 0$$

holds in the sense of distributions, i.e.

$$\int_{t_1}^{t_2} \int_X \left(\partial_t \varphi(x, t) + \langle v_t(x), \nabla_x \varphi(x, t) \rangle \right) \mathrm{d}\rho_t \mathrm{d}t = 0, \quad \varphi \in \mathrm{Cyl}(X \times [t_1, t_2]).$$

Under additional regularity assumptions on the vector field v_t (cf. [40] or [10, Lemma 8.1.6]), there exists a diffeomorphism X_t such that the solution admits the following characterization:

$$\begin{aligned} \rho_t &= (X_t)_\sharp \rho_0, \\ \frac{\mathrm{d}}{\mathrm{d}t} X_t(x,t) &= v_t(X_t(x,t),t), \\ X_t(x,0) &= x \end{aligned} \tag{2.1.3}$$

In other words, a passage to the Lagrangian point of view can be made, in which the dynamics of individual particles is described. From the numerical point of view this means that given samples from ρ_0 a numerical model (approximating the interpolation of measures in the Eulerian sense) can be used to provide samples from the distribution ρ_1.

Based on the interpretation of absolutely continuous curves via the continuity equation, one can reformulate the optimal transport problem as a problem of finding the geodesic curve in the Wasserstein space. This leads to the *dynamic optimal transport* problem due to Benamou and Brenier [41]:

$$\begin{aligned} W_2^2(\rho_0,\rho_1) = \min_{v_t} & \int_0^1 \int_X \|v_t(x)\|^2 \rho_t(x) \mathrm{d}x\mathrm{d}t, \\ & \partial_t \rho_t + \nabla(\rho_t v_t) = 0, \\ \text{s. t.} \quad & \rho_t(0) = \rho_0 \\ & \rho_t(1) = \rho_1, \end{aligned} \tag{2.1.4}$$

complemented by a no-outflow boundary condition in case X has a boundary.

2.2 Minimizing movements scheme

The minimization of the target functional $\mathcal{E}$ can be performed iteratively via proximal steps,

$$\rho^{k+1} = \arg\min_\rho \mathcal{E}(\rho) + \frac{1}{2T_k} W_2^2(\rho,\rho^k). \tag{2.2.1}$$

This is the counterpart of the implicit gradient descent in the Wasserstein space. If one is explicitly interested in the curve interpolating between successive steps is of interest, the dynamic formulation of the optimal transport problem can be used with (2.2.1). One then arrives at the PDE-constrained minimization problem

$$\begin{aligned} \rho^{k+1} = \arg\min_{\rho_t, v_t} & \int_0^1 \int_X \|v_t(x)\|^2 \rho_t(x) \mathrm{d}x\mathrm{d}t + 2T_k \mathcal{E}(\rho_1), \\ & \partial_t \rho_t + \nabla(\rho_t v_t) = 0, \\ \text{s. t.} \quad & \rho_t(0) = \rho^k. \end{aligned} \tag{2.2.2}$$

Just as in the case of minimization over a linear space, convexity of the target functional allows to determine the convergence rate of the method. Unfortunately, it is known that W_2^2 is not convex along geodesics, which complicates the analysis. Thus, a generalization is required leading to the notion of *convexity along generalized geodesics*.

Definition 2.2. *Convexity along generalized geodesics [10, Definition 9.2.4] The functional* $\mathcal{E} : \mathscr{P}_2(X) \to \mathbb{R}$ *is said to be* λ-convex along generalized geodesics *for some* $\lambda \in \mathbb{R}$ *if for any* $\rho_1,\ \rho_2,\ \rho_3 \in D(\mathcal{E})$ *there exists a plan*

$$\boldsymbol{\rho} \in \Pi(\rho_1, \rho_2, \rho_3) : p^{1,2}_\sharp \boldsymbol{\rho} \in \Pi_{opt}(\rho_1, \rho_2),\ p^{1,3}_\sharp \boldsymbol{\rho} \in \Pi_{opt}(\rho_1, \rho_3)$$

and a generalized geodesic, which is a curve of the form

$$\mu_t^{2\to 3} = \left(t\pi^2 + (1-t)\pi^3\right)_\sharp \boldsymbol{\rho}$$

such that

$$\mathcal{E}(\rho_t^{2\to 3}) \le (1-t)\mathcal{E}(\rho_2) + t\mathcal{E}(\rho_3) - \frac{1}{2}\lambda t(1-t)W^2_{\boldsymbol{\rho}}(\rho_2, \rho_3). \tag{2.2.3}$$

Here, D *is the domain of the functional,* $\Pi(\rho_1, \rho_2, \rho_3)$ *are probability distributions over* X^3 *with one-dimensional marginals* ρ_i, $i \in \{1, 2, 3\}$, $\Pi_{opt}(\rho_i, \rho_j)$ *is the set of optimal plans between* ρ_i *and* ρ_j *(in the sense of (2.1.1)),* p^i *is the projection onto the* i*-th variable, i.e.* $p^i(x_1, \dots, x_K) = x_i$*, and* $W^2_{\boldsymbol{\rho}}(\rho_2, \rho_3)$ *is defined as*

$$W^2_{\boldsymbol{\rho}}(\rho_2, \rho_3) := \int_{X^3} |x_2 - x_3|^2 d\boldsymbol{\rho}(x_1, x_2, x_2) \ge W_2^2(\rho_2, \rho_3).$$

The convexity of $\mathrm{KL}(\cdot|\rho_\infty)$ depends on the properties of the target distribution. A characterization can be obtained as a straightforward corollary of Propositions 9.3.2 and 9.3.9 of [10]:

Corollary 2.1.1. *Let* $X = \mathbb{R}^d$*. Let the target density have the form* $\rho_\infty = e^{-V}\,\mathrm{d}\,\mathcal{L}^d$ *with the Lebesgue measure* L^d *and* $V : \mathbb{R}^d \to \mathbb{R}$ *is* λ*-strongly convex for some* $\lambda \ge 0$*. Then the functional* $\mathrm{KL}(\cdot|\rho_\infty)$ *is* λ*-convex along generalized geodesics.*

Proof. The $\mathrm{KL}(\cdot|\rho_\infty)$ functional decomposes into two terms:

$$\mathrm{KL}(\rho|\rho_\infty) = \int \log\left(\frac{\rho}{\rho_\infty}\right)\mathrm{d}\rho = \int \log\rho\mathrm{d}\rho + \int V\mathrm{d}\rho = \mathcal{E}_{\text{int}}(\rho) + \mathcal{E}_{\text{pot}}(\rho).$$

Due to Proposition 9.3.9, the internal energy term $\mathcal{E}_{\text{int}}$ is geodesically convex (λ = 0). Due to Proposition 9.3.2, the potential energy $\mathcal{E}_{\text{pot}}$ is λ-convex along any interpolating curve. Thus, the inequalities of type (2.2.3) can be summed, leading to λ-convexity of $\mathrm{KL}(\cdot|\rho_\infty)$. □

If the target functional is λ-convex along generalized geodesics, the sequence generated by the Minimizing Movement scheme can be characterized by *Evolution Variational Inequalities*.

Theorem 2.2. *Variational inequalities, see [10, Theorem 4.1.2] Let* $\mathcal{E}$ *be* λ*-convex along generalized geodesics. For any* $\rho \in \overline{D(\mathcal{E})},\ T : \lambda T > -1$

- *the proximal minimization problem (1.2.2) has a unique solution*

$$\rho_T = \arg\min_{\tilde\rho} \frac{1}{2T}W_2^2(\tilde\rho, \rho) + \mathcal{E}(\tilde\rho)$$

- *for any other* $\tilde\rho \in D(\mathcal{E})$*,*

$$\frac{1}{2T}W_2^2(\rho_T, \tilde\rho) - \frac{1}{2T}W_2^2(\rho, \tilde\rho) + \frac{1}{2}\lambda W_2^2(\rho_T, \tilde\rho) \le \mathcal{E}(\tilde\rho) - \mathcal{E}_T(\rho), \tag{2.2.4}$$

  where

$$\mathcal{E}_T(\rho) = \frac{1}{2T}W_2^2(\rho_T, \rho) + \mathcal{E}(\rho_T) \ge \mathcal{E}(\rho_T).$$

Corollary 2.2.1. *Let $\mathcal{E} = \mathrm{KL}(\cdot|\rho_\infty)$ be λ-convex along generalized geodesics. Then ρ_∞ is a global minimum of the functional. For any ρ and denoting $\tilde{\rho} = \rho_\infty$, one can see that*

$$W_2(\rho_T, \rho_\infty) \leq \frac{1}{\sqrt{1+\lambda T}} W_2(\rho, \rho_\infty).$$

For the sequence generated by the minimizing movement scheme $\rho^{k+1} = \left(\rho^k\right)_{T_k}$ a convergence rate can be estimated by

$$W_2(\rho^N, \rho_\infty) \leq \prod_{k=1}^{N} \frac{1}{\sqrt{1+\lambda T_k}} W_2(\rho^0, \rho_\infty).$$

We have to admit that the standard analysis presented above is rather incomplete because, for example, the convexity property for the KL does not hold in case of a multimodal target measure. One also has to take the suboptimality of the solution into account.

3 Entropic regularization of the Minimizing Movements scheme

To obtain a numerical solution it is quite common to regularize the computation of the Wasserstein distance ((2.1.1)) by adding a term, which makes the problem strongly convex. Although other approaches for example based on quadratic regularization [42] and Rényi divergences [43] exist, the most prominent approach is *entropic optimal transport* [44]. At the cost of introducing a small bias controlled by the regularization parameter, one gets a strongly convex problem. When $\rho_{0,1}$ are two empirical measures (e.g. in a setting of comparing two point clouds as is popular in data science and machine learning), this enjoys better asymptotic complexity than the linear program solution required for the unregularized problem [45]. Additionally, the regularized problem allows for efficient parallel implementation with the Sinkhorn algorithm [46].

Inspired by the previous works [47, 48] we introduce the regularization into the dynamic reformulation of the problem (2.2.2),

$$\begin{aligned} \rho^{k+1} = \arg\min_{v,\rho} & \int_0^{T_k} \int_X \frac{1}{2} \|v(t,x)\|^2 \rho(t,x) \mathrm{d}x \mathrm{d}t + \mathcal{E}(\rho(T_k,\cdot)), \\ \text{s.t.} \quad & \partial_t \rho + \nabla(\rho v) = \beta_k \Delta \rho, \\ & \rho(0,x) = \rho^k(x). \end{aligned} \tag{3.0.1}$$

Essentially a diffusion term is added to the flow constraint. The parameter $\beta_k > 0$ controls the strength of the regularization. In [47] it is shown that this is in fact equivalent to adding entropic and Fisher information regularizing terms to the target functional. The equivalence between the dynamic and static regularized OT formulations is explained in [49], which can be obviously extended to the Minimizing Movements scheme. As will be evident from the derivation below, this particular form of regularization allows to reformulate the problem as a coupled PDE system. Using a fixed-point method is an obvious approach for the solution, similar to the well-known Sinkhorn algorithm and well suited for the implementation with Tensor Train decompositions.

3.1 Derivation of the method

We introduce a dual variable (as Lagrange multiplier) for the continuity equation constraint $\Phi(x,t)$. Writing down the Lagrangian and the optimality conditions one can see that $v = \nabla\Phi$ and it holds the

system

$$\partial_t \rho + \nabla(\rho \nabla \Phi) = \beta_k \Delta \rho, \tag{3.1.1}$$

$$\partial_t \Phi + \frac{1}{2}\|\nabla \Phi\|^2 = -\beta_k \Delta \Phi, \tag{3.1.2}$$

$$\rho(0,x) = \rho^k(x), \quad \Phi(T_k, x) = -\delta\mathcal{E}(\rho(T_k,\cdot),x). \tag{3.1.3}$$

Here, $\delta\mathcal{E}$ is the first variation of the functional $\mathcal{E}$. This can be seen as a primal-dual formulation of the problem. For a general energy we see that the terminal condition has nonlinear dependence on $\Phi,\ \rho$.

With a change of variables in terms of the (*Hopf-Cole transform*)

$$\eta(t,x) = e^{\frac{\Phi(t,x)}{2\beta_k}}, \quad \hat{\eta}(t,x) = \rho(t,x)e^{\frac{-\Phi(t,x)}{2\beta_k}} \tag{3.1.4}$$

yielding

$$\Phi = 2\beta_k \log \eta, \quad \rho = \eta\hat{\eta}, \tag{3.1.5}$$

the system of a forward-in-time Fokker-Planck (3.1.1) and a backward-in-time Hamilton-Jacobi (3.1.2) transforms into a system of two heat equations, one forward and one backward in time. It is given by

$$\partial_t \hat{\eta} = \beta_k \Delta \hat{\eta} \tag{3.1.6}$$

$$\partial_t \eta = -\beta_k \Delta \eta \tag{3.1.7}$$

$$\hat{\eta}(0,x) = \frac{\rho^k(x)}{\eta(0,x)} \tag{3.1.8}$$

$$\eta(T_k,x) = e^{-\frac{1}{2\beta_k}\delta\mathcal{E}(\eta(T_k,\cdot),\hat{\eta}(T_k,\cdot),x)}. \tag{3.1.9}$$

3.2 Fixed-point formulation of the proximal step

We argue that system (3.1.6)-(3.1.9) is more suitable for the numeric solution than (3.1.1)-(3.1.3). Firstly, now two similar linear PDEs are considered. The solution of the general initial-value problem for the heat equation can be obtained by a convolution with the fundamental solution, namely

$$\hat{\eta}(T_k,\cdot) = H_d(\beta_k T_k)\hat{\eta}(0,\cdot) \tag{3.2.1}$$

$$\eta(0,\cdot) = H_d(\beta_k T_k)\eta(T_k,\cdot), \tag{3.2.2}$$

where $H_d : L_1(\mathbb{R}^d) \times \mathbb{R}^+ \to L_1(\mathbb{R})$ is the integral operator such that

$$\left(H_d(s)u\right)(x) = \int_{\mathbb{R}^d} G_d(s, y-x)u(y) \tag{3.2.3}$$

$$G_d(s,\xi) = \frac{1}{(4\pi s)^{\frac{d}{2}}} e^{-\frac{1}{4s}\|\xi\|^2}. \tag{3.2.4}$$

We suggest that the problem of satisfying the whole system together with the initial and the terminal conditions lends itself better for a reformulation as a fixed-point iteration. Consider the following steps for some η_m:

$$\eta_{m,0} = H_d(\beta_k T_k)\eta_m, \tag{3.2.5}$$

$$\hat{\eta}_{m,0} = \frac{\rho^k(x)}{\eta_{m,0}}, \tag{3.2.6}$$

$$\hat{\eta}_m = H_d(\beta_k T_k)\hat{\eta}_{m,0}, \tag{3.2.7}$$

$$\tilde{\eta}_m = e^{-\frac{1}{2\beta_k}\delta\mathcal{E}(\tilde{\eta}_m,\hat{\eta}_m,\cdot)}. \tag{3.2.8}$$

If $(\eta^*, \hat{\eta}^*)$ were the solution of (3.1.6)-(3.1.9) with $\eta_m = \eta^*(T_k, \cdot)$, then $\eta_{m,0} = \eta^*(0, \cdot)$, $\hat{\eta}_{m,0} = \hat{\eta}^*(0, \cdot)$, $\hat{\eta}_m = \hat{\eta}^*(T_k, \cdot)$ and η_m is the fixed point of the operator G, implicitly defined by (3.2.5)-(3.2.8)

$$\tilde{\eta}_m = G(\eta_m) = \eta_m.$$

We propose to reformulate the coupling problem as a classical Picard iteration $\eta_{m+1} = \tilde{\eta}_m$, $m = 0, 1, \ldots$, starting from some initial guess η_0 and computing successive iterates until η_m and $\tilde{\eta}_m$ are sufficiently close in some metric. The condition (3.2.8) is itself a nonlinear equation since the variation of energy depends on $\tilde{\eta}_m$. For general energy functional this can be relaxed by taking the value of η from the previous iteration, i.e.,

$$\tilde{\eta}_m = e^{-\frac{1}{2\beta_k}\delta\mathcal{E}(\eta_m, \hat{\eta}_m, x)}.$$

However, in the case of KL divergence as considered in the current work, one can see that

$$\delta\,\mathrm{KL}(\rho|\rho_\infty)(x) = \log\rho - \log\rho_\infty + 1 + \text{const.}.$$

Here, the constant appears because the target density ρ_∞ is only known up to a multiplicative constant. We note that adding a constant to the target functional does not change the $\arg\min$ in the minimization problem. Thus we assume that we can minimize the functional $\mathrm{KL}(\cdot|\rho_\infty) + C$ with C chosen in such a way that the constant in the $\delta\,\mathrm{KL}$ term cancels out. In conclusion, one can explicitly rewrite $\tilde{\eta}_m$ in (3.2.8) as a function of already known variables by

$$\tilde{\eta}_m = \left(\frac{\rho_\infty}{\hat{\eta}_m}\right)^{\frac{1}{1+2\beta_k}}. \tag{3.2.9}$$

The fixed-point cycle defined above is schematically depicted in Figure 1. We point out that formally replacing $\beta_k T_k$ with some positive ϵ and $\beta_k = 0$ yields the Sinkhorn algorithm [50].

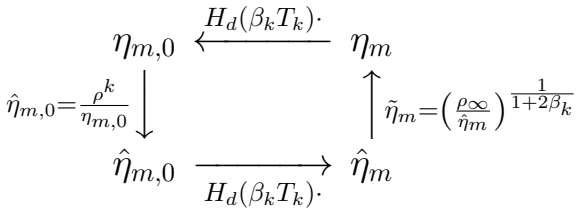


Figure 1: Schematic of proposed fixed-point algorithm.

For a system that converged (to a prescribed accuracy) after M fixed-point iterations, the next density is given by

$$\rho^{k+1}(x) \approx \eta_M(x) \cdot \hat{\eta}_M(x). \tag{3.2.10}$$

Additionally, one can determine the drift term of the Fokker-Planck equation describing the flow in (3.0.1) from the optimality condition $v = \nabla_x \Phi$ by

$$\begin{aligned} v(x,t) &= 2\beta_k \nabla_x \log\eta(t,x), \\ \eta(t,x) &= (H^d(\beta_k(T-t))\eta_M)(x). \end{aligned} \tag{3.2.11}$$

Alternatively, using that

$$\begin{aligned} \partial_t\rho + \nabla(\rho\nabla\Phi) - \beta_k\Delta\rho = \\ = \partial_t\rho + \nabla(\rho\nabla\Phi) - \beta_k\nabla\cdot(\nabla\rho) = \partial_t\rho + \nabla(\rho\nabla\Phi) - \beta_k\nabla\cdot(\rho\nabla\log\rho) = \\ = \partial_t\rho + \nabla\left(\rho\nabla(\Phi - \beta_k\log\rho)\right) \end{aligned}$$

and recalling the Hopf-Cole transform (3.1.5), one can get an equivalent deterministic dynamics of the form

$$\partial_t \rho + \nabla(\rho \tilde{v}) = 0$$

with the drift term $\tilde{v}$ given by

$$\begin{aligned} \tilde{v}(t,x) &= \beta_k \nabla_x \left(\log \eta(t,x) - \log \hat{\eta}(t,x)\right), \\ \hat{\eta}(t,x) &= \left(H_d(\beta_k t)\hat{\eta}_{0,M}\right)(x), \quad \eta(t,x) = \left(H_d(\beta_k(T_k - t))\eta_M\right)(x). \end{aligned} \tag{3.2.12}$$

3.3 Convergence of the Minimizing Movements with regularized proximal

Convergence of the scheme (3.0.1) is – to the knowledge of the authors – predominantly studied in the context of convergence of the discrete sequence $\{\rho^k\}_{k=1}^{\infty}$ to the continuous solution of the gradient flow for the original energy functional $\mathcal{E}$ as both the time step and regularization parameter approach zero. In [17] it is shown that in order to approximate the original gradient flow, the regularization has to decay sufficiently faster than the time step size. However, we are more interested in the rate of convergence of the discrete sequence to the minimizer of the energy functional, given a predefined sequence of steps and regularizations and possibly also making an informed choice of such a sequence. This problem can be attacked in the following way. In [51] the EVI similar to (2.2.4) is derived for a sequence $\{\rho^k\}$, where each term is not an exact solution of (1.2.2) but a suboptimal one. To make this precise, we assume that the Wasserstein subgradient of the proximal functional at ρ_{k+1} is bounded by some predefined sequence Ξ_k, i.e.,

$$\begin{aligned} \mathcal{F}_{k+1}(\rho) &:= \mathcal{E}(\rho) + \frac{1}{2\tau_k} W_2^2(\rho, \rho_k), \\ \rho_{k+1} &: \exists\, \xi_{k+1} \in \partial_{W_2} \mathcal{F}_{k+1}(\rho_{k+1}) \quad \text{s.t.} \quad \|\xi_{k+1}\|_{L_2(\rho_{k+1})} \leq \Xi_k. \end{aligned}$$

Then (again in case of a functional λ-convex along generalized geodesics) we deduce that

$$\left(1 + \frac{T_k \lambda}{2}\right) W_2^2(\rho_{k+1}, \rho_\infty) + 2T_k \left(\mathcal{E}(\rho_{k+1}) - \mathcal{E}(\rho_\infty)\right) \leq W_2^2(\rho_k, \rho_\infty) + \frac{2T_k}{\lambda} \Xi_k^2. \tag{3.3.1}$$

If applied by induction for instance for uniformly bounded $\Xi_k \leq \Xi$ and constant timestep $T_k = T$, it yields the linear convergence rate ([51, Theorem 4.3])

$$W_2^2(\rho_{k+1}, \rho_\infty) \leq \left(1 + \frac{T\lambda}{2}\right)^{-k} W_2^2(\rho_0, \rho_\infty) + \frac{4\Xi^2}{\lambda^2}.$$

We hypothesize that the subgradient norm Ξ_k can be controlled by a proper choice of (T_k, β_k) but leave this to future studies.

4 Numerical solution

The main challenge for a numerical implementation of the method in the Eulerian approach is the number of degrees of freedom in the discretization, which is growing exponentially with the dimension d of the problem. We argue that this can be overcome by utilizing low-rank tensor methods, assuming that the problem exhibits some low-rank structure. This section provides the details on the low-rank methods used for the implementation.

4.1 Discretization and Tensor Train compression

For a proof-of-concept implementation we consider a «sufficiently large» cube $\bigotimes_{k=1}^{d}[L_k; R_k]$ and a finite difference approximation on a regular grid with N_k nodes in each dimension, namely

$$h_k = \frac{R_k - L_k}{N_k}, \qquad x_{k,i} = -L_k + h_k \cdot i.$$

A tensor with function values g_α with multiindex α is defined by

$$g_{i_1 \dots i_d} = g(x_{1,i_1}, \dots, x_{d,i_d}),$$

where g is one of $\eta_m,\ \eta_{0,m},\ \hat{\eta}_{0,m},\ \hat{\eta}_m$. If the dimension d of the problem is high, the amount of memory required to store all the values of the tensor increases exponentially, rendering it prohibitively complex in practice. Hence, some compression approach has to be used and in the present work we focus on the Tensor Train format

Definition 4.1. *Tensor Train format A tensor $g_\alpha \in \mathbb{R}^{N_1 \times \dots \times N_d}$ is in the TT format if it has the form*

$$g_{i_1 \dots i_d} = G^1_{i_1 l_1} G^2_{l_1 i_2 l_2} \dots G^d_{l_{d-1} i_d}, \tag{4.1.1}$$

where for each $n \in \overline{1,\ d}$, $G^n_{l_{n-1} i_n l_n} \in \mathbb{R}^{r_{n-1} \times N_n \times r_n}$ is a 3-dimensional tensor with $r_0 = r_d = 1$ and Einstein summation convention is assumed. The values r_k are called TT-ranks*.*

If $N = \max_n N_n$ and TT-rank $r = \max_n r_n$, accessing a component requires $\mathcal{O}(dr^2)$ operations and the storage complexity is $\mathcal{O}(dNr^2)$ compared to $\mathcal{O}(1)$ and $\mathcal{O}(N^d)$ in the case of storing all the values in a full tensor. Compression is achieved if the rank is significantly lower than the full rank.

We assume that for each proximal step an adequate initial guess for the fixed-point $\eta_{0,i_1 \dots i_d}$ in the TT format can be obtained. In the sequel it is explained how to approximate all the solution steps while staying in the TT format.

4.2 Solution of the heat equation in TT format

The solution of the heat equation amounts to applying the operator $H_d(s)$. It is required during the fixed-point iteration (see (3.2.1)-(3.2.2)) and later to compute the drift terms for the sampling dynamics ((3.2.11), (3.2.12)). We first replace the Laplace operator Δ with its second-order in space finite difference approximation

$$L_h = D_{1,h} \otimes I_2 \otimes \dots \otimes I_d + I_1 \otimes D_{2,h} \otimes \dots \otimes I_d + \dots + I_1 \otimes I_2 \otimes \dots \otimes D_{d,h}, \tag{4.2.1}$$

where $\otimes$ denotes the tensor product, $I_k \in \mathbb{R}^{N_k \times N_k}$ is an identity matrix and

$$D_{n,h} = \frac{1}{h_n^2} \begin{pmatrix} -1 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -1 \end{pmatrix}$$

for $n \in \overline{1,\ d}$ is a $N_n \times N_n$ matrix, which represents the second-order finite difference approximation of $\frac{\partial^2}{\partial x_n^2}$. Then, for a time-dependent tensor $g(t)$ the following ODE holds true

$$\frac{\mathrm{d}}{\mathrm{dt}} g(t) = \beta L_h g(t).$$

Its solution has the form

$$g(t) = e^{\beta_k t L_h} g(0). \tag{4.2.2}$$

Due to the structure of the discrete Laplacian (4.2.1), the matrix exponential of (4.2.2) has the form (see [52, Section 2.4])

$$e^{\beta_k t L_h} = e^{\beta_k t D_{1,h}} \otimes \cdots \otimes e^{\beta_k t D_{d,h}}. \tag{4.2.3}$$

Finally, if the initial tensor $g(0)$ is in TT format, the solution $g(t)$ is given by

$$\begin{aligned} g(t)_{i_1 \dots i_d} &= \left(e^{\beta_k t L_h}\right)_{i_1 \dots i_d j_1 \dots j_d} g(0)_{j_1 \dots j_d} = \\ &= \left(e^{\beta_k t L_h}\right)_{i_1 \dots i_d j_1 \dots j_d} G^1_{j_1 k_1} G^2_{k_1 j_2 k_2} \dots G^d_{k_{d-1} j_d} = \left(e^{\beta_k t D_{1,h}}_{i_1 j_1} \dots e^{\beta_k t D_{d,h}}_{i_d j_d}\right) G^1_{j_1 l_1} G^2_{l_1 j_2 l_2} \dots G^d_{l_{d-1} j_d} = \\ &= \left(e^{\beta_k t D_{1,h}}_{i_1 j_1} G^1_{j_1 l_1}\right) \left(e^{\beta_k t D_{2,h}}_{i_2 j_2} G^2_{l_1 j_2 l_2}\right) \dots \left(e^{\beta_k t D_{d,h}}_{i_d j_d} G^d_{l_{d-1} j_d}\right). \end{aligned}$$

This means that the approximate solution can also be represented in TT format with new cores

$$e^{\beta_k t D_{n,h}}_{i_n j_n} G^n_{l_{n-1} j_n l_n} \in \mathbb{R}^{r_{n-1} \times N_n \times r_n}, \tag{4.2.4}$$

evidently having the same TT rank as the initial condition. The update of the TT cores by (4.2.4) requires $\mathcal{O}(dN^2 r^2)$ operations compared to $\mathcal{O}(N^{2d})$ operations required do compute a general convolution of type $A_{i_1 \dots i_d j_1 \dots j_d} b_{j_1 \dots j_d}$. For given (β_k, T_k), d matrix exponentials (or, in case $N_n,\ h_n$ in every direction are the same, only one) can be precomputed, for example by means of a Padé approximation [53]. Thus we argue that the heat equation required during the fixed-point iteration can be efficiently solved in TT format without explicit time stepping.

4.3 Solution of initial and terminal conditions

Tensors $\hat{\eta}_{m,0},\ \tilde{\eta}_m$ are defined by (3.2.6) and (3.2.9) in the sense that their value can be computed for any index given index. We aim to acquire a TT representation of these tensors by computing the values at a much smaller subset of indices. Note that it is crucial to control the amount of evaluations of the right-hand side of (3.2.9). This is because in this step the posterior evaluations take place, making it a computational bottleneck of the whole algorithm. Such a task to construct a tensor from relatively few measurements is called *tensor completion* or *reconstruction*. A common algorithm is the well-known TT-cross approximation [25], which we would like to briefly explain. Given a multi-dimensional tensor $T \in \mathbb{R}^{n_1 \times \dots \times n_d}$, it may be of use to view it as a 2-dimensional matrix, produced by grouping indices of several dimensions $(i_1, i_2, \dots i_k)$ into a joint index $i_1 i_2 \dots i_k \in \overline{1, n_1 \cdots n_k}$, by some one-to-one mapping. This matrix is then given by

$$\left[T^{(k)}\right]_{i_1 i_2 \dots i_k, i_{k+1} \dots i_d} = T_{i_1, \dots i_d}$$

and called the *unfolding* of the tensor. If a tensor has a best TT approximation error ϵ with ranks r_k, the method reconstructs it by recursively building skeleton decompositions of the aforementioned unfoldings. For a given sequence of indices for the skeleton decomposition $\left\{(I^{\leq k}, I^{>k})\right\}_{k=\overline{1,d-1}}$, where each $I^{\leq k} = \{(i_1^j, i_2^j, \dots i_k^j)\}$, the error can be estimated as follows.

Theorem 4.1 ([54, Theorem 2]). *For a given tensor $\mathcal{T} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$, assume there is a tensor $\mathcal{T}_{r_k}$ in TT format with ranks r_k such that*

$$r = \max_{k=\overline{1,d-1}} r_k, \quad \epsilon = \max_{k=\overline{1,d-1}} \|T^{\langle k \rangle} - T^{\langle k \rangle}_{r_k}\|_F,$$
$$\kappa = \max_{k=\overline{1,d-1}} \left\{ \|T^{\langle k \rangle}(:, I^{>k}) \cdot (T^{\langle k \rangle}(I^{\leq k}, I^{>k}))^{-1}\|, \|(T^{\langle k \rangle}(I^{\leq k}, I^{>k}))^{-1} \cdot T^{\langle k \rangle}(I^{\leq k}, :)\| \right\}.$$

Then, for sufficiently small $\varepsilon > 0$ there exists a TT cross approximation $\mathcal{T}_{\text{TT-cross}}$ such that

$$\|\mathcal{T} - \mathcal{T}_{\text{TT-cross}}\|_F \leq \frac{(3\kappa)^{\lceil \log_2 N \rceil} - 1}{3\kappa - 1}(r+1)\varepsilon. \tag{4.3.1}$$

One pass of computing the TT cores for each dimension by building skeleton decompositions is referred to as a *sweep*. It requires $\mathcal{O}(dNr^2)$ evaluations of the right-hand side tensor elements. Practically, the optimal indices are not known, thus, several sweeps are performed, using the approximation from the previous ones and various heuristics to improve the selection of the indices. The description of these approachse can be found in [55] and references therein. For an arbitrary tensor, the low-rank approximation is never guaranteed, but the maximum rank can be estimated according to the knowledge of the problem and numerical resources available. The rank of the cross-approximation can be dynamically adapted with the DMRG technique [56].

A hallmark of the Eulerian approach is that by using a fixed grid, the evaluations of the posterior are carried out in a comprehensive and “organized” way. More specifically, since in our problem the posterior can be evaluated only at a finite (although extremely large even for moderated dimensions $d > 4$) number of points determined in advance by the grid, we can implement a caching strategy in practice to reduce the number of evaluations drastically. In the current implementation, we simply store the first N_{cache} calls to the posterior. The cache is shared between inner fixed-point iterations and outer proximal steps. In the numerical experiments below we demonstrate that this feature indeed significantly decreases the amount of the posterior calls required to solve the problem.

4.4 Fixed-point iteration

As explained in Section 3.2, the PDE coupling problem of one proximal step is recast as a fixed-point problem

$$\eta^* = G(\eta^*)$$

with TT representation η of the potential $\eta(T_k, x)$ and operator G is the approximation of the cycle defined in (3.2.5)-(3.2.8) and depicted in Figure 1. The numerical approximations are described in the preceding sections. For an iterative solution procedure, consider a sequence

$$x_m = \eta_m, \quad g_m = G(x_m), \quad r_m = g_m - x_m.$$

The simplest approach is the relaxation method

$$x_{m+1} = q_m g_m + (1 - q_m) x_m,$$

where $q_m \in (0;, 1]$ is called *relaxation factor*. It can either be constant or selected adaptively. If constant, setting smaller values of q_m improves the stability of the method but deteriorates its speed. For $q_m \equiv 1$, one recovers the Picard method. Aitken’s scheme [57] updates q_m adaptively depending on two previous residuals. However, it is evident that in the region where the operator G is a contraction,

this scheme cannot converge faster than the Picard method with $q = 1$. In our preliminary experiments, the fixed-point problems appear to always converge for the Picard method with $q = 1$ and hence the application of Aitken's scheme is not reported.

A more involved acceleration method called *Anderson Acceleration* (AA) relies on the minimization in the affine hull of P previous residuals. It is described by

$$(\alpha_0, \ldots, \alpha_{P-1}) = \arg \min_{\alpha_0+\cdots+\alpha_{P-1}=1} \left\| \sum_{i=0}^{P-1} \alpha_i r_{m-i} \right\|_2^2, \tag{4.4.1}$$
$$x_{m+1} = q \sum_{i=0}^{P-1} \alpha_i g_{m-i} + (1-q) \sum_{i=0}^{P-1} \alpha_i x_{m-i}.$$

The theoretical results on the convergence of accelerated FP in general can be found in [58] and for AA in particular in [59]. AA can be related to multisecant methods [60]. In fact, for a linear problem it is equivalent to GMRES [61]. Practically, AA has been reported to improve both speed and robustness of fixed-point iteration compared to the Picard iteration (see [62, 63] and references therein). For the application with TT representations of the iterates, each fixed-point update is followed by a TT truncation up to prescribed rank $r_{\max}$ and tolerance ε.

4.5 Sampling the approximate solution

The PDE-based approach of our method combines the Lagrangian and Eulerean viewpoint since the approximate solution of the PDE system defines a dynamics that describes the evolution of particles, moving from a start towards an updated distribution. The representation (2.1) provides a deterministic description of the dynamics, resulting in a deterministic sampling method. It requires solving an ODE with right-hand side defined by (3.2.12), i.e.,

$$\dot{X}(t) = \beta_k \nabla_x \big(\log \eta(t, X(t)) - \log \hat{\eta}(t, X(t)) \big), \qquad X(0) \sim \rho_0.$$

Similar methods have recently become popular, for instance in the generative modelling community [64]. The approach only requires random samples of the initial density, which can be made tractable by construction. Note that the ODE can be solved in parallel with high precision and with an adaptive choice of time steps for each trajectory. Alternatively, from the Fokker-Planck equation (3.1.1) and the fact that $\Phi = 2\beta_k \log \eta$, an SDE can be defined by

$$\mathrm{d}X_t = 2\beta_k \nabla \log \eta(X_t, t)\mathrm{d}t + \sqrt{2\beta_k}\mathrm{d}W_t, \qquad X_0 \sim \rho_0,$$

with standard Brownian motion W_t and the law of X_t denoted by ρ_t. Any suitable numerical method of the SDE solution as for example Euler-Maruyama provides an algorithm for approximate sampling based on this dynamics.

Remark. *Our preliminary computations show that it might be beneficial to employ a combination of both dynamics. The ODE dynamic can be approximately integrated with high-order adaptive methods such as* $\mathrm{RK45}$*. In our sampling tasks, the steps taken are quite large in the beginning of the time interval. However, towards the terminal time for a small subset of the trajectories the step size estimated by the adaptive solver becomes extremely small. Additionally, due to the discretization error in the regions of low probability and approximate computation of the gradient, the latter can become too small and particles might get stuck in such regions. This leads to samples produced solely with the ODE scheme*

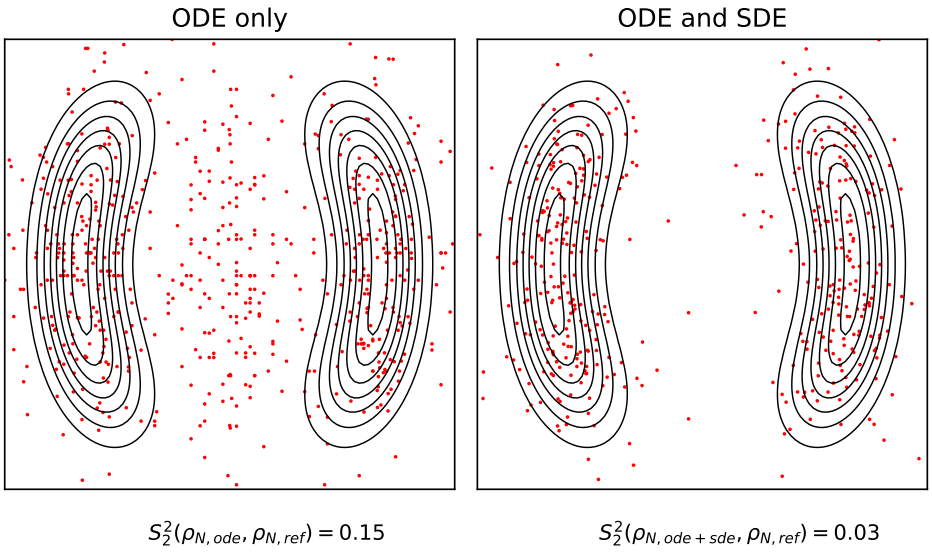


Figure 2: Contour lines of the test distribution and the samples generated with the deterministic method (left) and the combination of deterministic and stochastic method (right).

to exhibit certain unwanted artifacts. Heuristically, some scheme with addition of noise would push the particles away from the regions with zero gradient. Thus, we suggest to combine the deterministic and the stochastic dynamics. More precisely, for some predefined value ϵ_{SDE} we solve the ODE system (4.5) on the time interval $[0; (1-\epsilon_{SDE})T_k]$ (with outer step size T_k in the proximal scheme). Then, the final $\epsilon_{SDE}T_k$ time is integrated with the Euler-Maruyama method with a predefined number of steps. The SDE time fraction ϵ_{SDE} is chosen heuristically and takes values $0.01 \sim 0.001$

Figure 2 illustrates the effect discussed above. The plot on the left depicts the sample produced with the ODE dynamics only. It clearly exhibits a "line-like" artefact in the center, where the actual probability density of the target is rather low. The sample generated with the combination of ODE and SDE dynamics (right) fits the reference distribution better (for this particular example, the Sinkhorn distance is roughly 5 times lower). In fact, it also requires less computational time because the region where the ODE solver requires small time steps is avoided.

Post-processing procedures for deterministic sampling methods that involve the addition of noise are quite common. For example, [64] suggests using a predictor-corrector scheme, where the solution of the probability flow ODE is corrected at each step by a step of Langevin dynamics towards the current distribution. A Langevin post-processing with multiple steps is also reported in [65]. The joint dynamics presented in the current work provides a novel viewpoint on this established computational know-how.

5 Numerical experiments

This section is concerned with the numerical assessment of the method proposed above. The method's properties are first explored on a trivial example of fitting a Gaussian target with diagonal covariance matrix. The dependence of the convergence properties of the fixed-point iteration on the stepsize and

regularization parameter is studied. We demonstrate that, in principle, a single proximal step can be sufficient to provide a good approximation. This comes at the cost of increasing complexity of the fixed-point iteration, which, in turn, can be mitigated by implementing accelerated fixed-point schemes.

The method is then tested against the baseline Metropolis-Hastings MCMC method on several synthetic problems. We highlight the method's performance in case of multimodal distributions and distributions with nonconvex potentials, which are of particular complexity for the baseline. We carefully estimate the «computational budget» of our method, expressed as the number of the calls to the posterior, and demonstrate, that, given the same amount of the posterior calls, the approximation by the baseline is less precise.

For the Python implementation the `teneva` package [66], which contains all basic tensor-train operation, as well as a cross-approximation solver, was employed.

5.1 Approach to performance assessment

Before providing the numerical applications of the described method, we briefly discuss the approaches used to comprehensively assess its practical performance. The method itself allows to track the KL divergence to the posterior by carrying out an additional cross-approximation of the quantity

$$\log\left(\frac{\rho_{\mathsf{TT}}}{\rho_\infty}\right)\rho_{\mathsf{TT}} \tag{5.1.1}$$

on a grid and a successive approximation of the integral by contracting the low-rank TT approximation. Note that $\rho_{\mathsf{TT}} = \eta_M \cdot \hat{\eta}_M$ and for the converged solution, from (3.2.9)

$$\log\left(\frac{\rho_{\mathsf{TT}}}{\rho_\infty}\right) = -2\beta_k \log \eta_M.$$

We hence suggest that the divergence can actually be evaluated without additional posterior calls. The KL divergence estimated in this fashion can be utilized for the assessment of the convergence of the method.

Considering the targeted practical application in Bayesian inversion and related applications, estimating the expectations of various quantities of interest is of importance. Certain functions such as x_i and $x_i x_j$ do not depend on the solution of the forward model and have a straightforward low-rank representation. Thus, their moments (e.g. means and covariances of the parameters) can be estimated with negligible computational costs with TT contractions. For more general quantities, an approximation similar to the one discussed above in the context of estimating the KL divergence is possible. However, this possible approach is left for the future studies. Instead, we focus on using the acquired model to sample from the approximate posterior in order to relate and compare the method to «Eulerian» methods that appear to be more established in the area.

We aim to compare the produced samples either to the samples from the posterior (for tractable ones such as Gaussian mixtures) or to a baseline sampling method (such as Langevin dynamics or MCMC). Several issues are to be dealt with in order to do so. Firstly, the empirical measure associated with a sample,

$$\rho_N = \sum_{i=1}^{N} \delta_{x_i},\ x_i \sim \rho$$

is obviously different from the original measure ρ. Thus, there is some nonzero Wasserstein distance between them. It can be observed that two different samples from the same distribution can have quite

a large Wasserstein distance, especially for higher dimensions of the problem. This suggests that when the approximate sample is sufficiently close to the target, the error is dominated by the finite-sample effect and not with the approximation error. This effect can be somewhat mitigated by using a larger sample size. Secondly, the computational complexity of the original linear OT problem (2.1.1) does not scale well with the number of samples, requiring alternatives to be considered. One way to tackle with these issues, is to solve the *entropic OT* distance given by

$$W_{2,\varepsilon}^2(\rho_1,\ \rho_2) = \min_{\pi\in\Pi(\rho_1,\rho_2)} \int \|x-y\|^2 + \varepsilon\,\mathrm{KL}\left(\pi|\rho_1\otimes\rho_2\right).$$

which can be done efficiently. The entropic regularizer introduces a «bias», i.e., $W_{2,\varepsilon}^2(\rho,\rho)\neq 0$ for general ρ. Introducing «debiasing» terms, one gets a so-called *Sinkhorn divergence* by

$$S_{2,\varepsilon}^2(\rho_1,\rho_2) = W_{2,\varepsilon}^2(\rho_1,\ \rho_2) - \frac{1}{2}\left(W_{2,\varepsilon}^2(\rho_1,\ \rho_1) + W_{2,\varepsilon}^2(\rho_2,\ \rho_2)\right).$$

This divergence is an approximation of W_2^2 up to $O(\varepsilon^2)$ [67] and can be efficiently computed with the Sinkhorn algorithm [46] or by means of stochastic optimization [68] in higher dimensions. In our numerical experiments, the implementation of the Sinkhorn algorithm from the `geomloss` package [69] is used.

An alternative approach to measure the OT distance betwen samples would be to use the *sliced Wasserstein* distance given by

$$\overline{W}_2^2(\rho_1,\rho_2) = \sup_{\theta\in\mathbb{S}^{d-1}} W_2^2(\theta_\sharp\rho_1,\theta_\sharp\rho_2),$$
$$\theta_\sharp = (\langle\theta,\cdot\rangle)_\sharp,$$

where $\theta\in\mathbb{S}^{d-1}$ is a direction and $\theta_\sharp$ denotes a pushforward of the measure with the map that projects vectors to the direction θ. It is shown in [70] that while $\overline{W}_2^2$ induces a metric equivalent to W_2^2, the computation of it relies on solving unidimensional OT problems for which a closed-form solution exists. The implementation of the sliced Wasserstein distance from `POT` package [71] was employed, where $\sup$ is estimated by taking a finite number of random projections. Except for the numerical efficiency, the sliced OT metric does not show any qualitative benefits compared to the Sinkhorn distance in our test. Hence, in the sequel the results are only reported in the latter distance.

We aim to show that our method produces samples that are consistent with the target measure by comparing them to the reference samples from the latter. Due to the finite-sample effect discussed in the beginning of this section, optimal transport distances between the samples even from the same distribution are in fact random variables not identical to zero. Following [7], we suggest comparing the distributions of this random variables. More specifically, we consider the distribution of $S_{2,\varepsilon}^2(\rho_{N_1}^1,\ \rho_{N_2}^2)$, where $\rho_N^i = \frac{1}{N}\sum_{k=1}^N \delta_{x_k}$ and $x_k\sim\rho^1$i.i.d., i.e. the empirical measure defined by an i.i.d. sample of size N from the measure ρ^1. If the measures ρ^1 and ρ^2 are actually equal, the distributions of $S_{2,\varepsilon}^2(\rho_{N_1}^1,\ \rho_{N_2}^2)$ and $S_{2,\varepsilon}^2(\rho_{N_1}^1,\ \rho_{N_2}^1)$ should coincide. Based on this, we acknowledge that our method generates samples from the target measure if the distribution of the Sinkhorn distances to the reference samples is close as a distribution to the distribution of Sinkhorn distances between independent reference samples.

The rest of the section is organized as follows. The first part deals with verification computations for one step of the method with a Gaussian distribution with diagonal covariance matrix as posterior. The goal is to analyze the behavior of the method for different combinations of values of the parameters $T,\ \beta$ and the acceleration of the fixed-point scheme. In the second part, the method is compared to a baseline Metropolis-Hastings MCMC method for the task of sampling. Test distributions include multimodal distributions and a distribution with nonconvex potential.

5.2 Verification

We perform verification computations for a simple setting where both ρ_0 and ρ_∞ are Gaussian distributions. For this, assume a uniform grid on the hypercube $[-L, L]^d,\ L = 3$ with $N = 30$ nodes in each dimension. Moreover, $\rho_0 = \mathcal{N}(0, \mathrm{I}_d)$ and $\rho_\infty = \mathcal{N}(m, \sigma\,\mathrm{I}_d)$ with m generated randomly with $m_i \sim U([-L/2,\ L/2])$ and $\sigma = 0.5$. The dimension of the problem is $d = 16$. We track the KL divergence as discussed above.

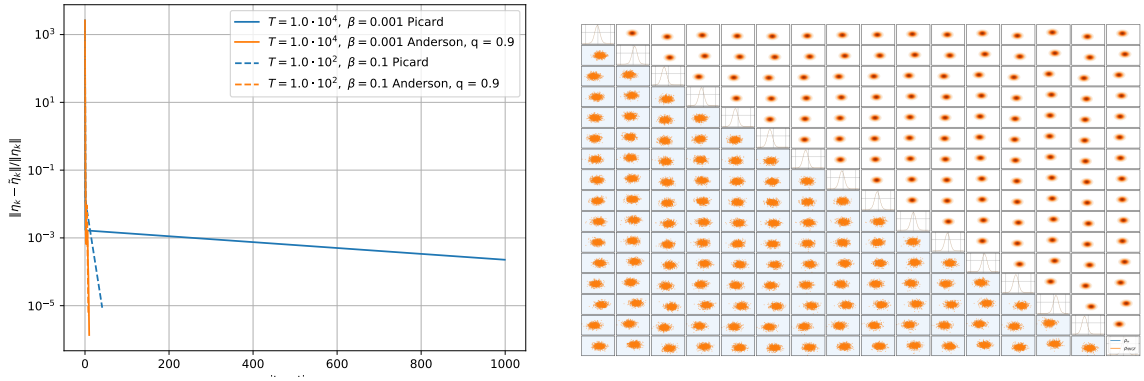
A single step of the method is performed for a broad range of parameters T and β. The results are presented in Table 1 One can see that with small values of βT, which plays the role of the effective

	$\mathrm{KL}(\rho_{\mathrm{TT}}\|\rho_\infty)$					N_{fp}				
β / T	1	0.1	0.01	0.001	0.0001	1	0.1	0.01	0.001	0.0001
0.1	9.1	-	-	-	-	13	-	-	-	-
1	5.7	-	-	-	-	13	-	-	-	-
10	6.7	0.13	-	-	-	13	58	-	-	-
100	6.8	0.14	0.0014	-	-	13	42	$4 \cdot 10^2$	-	-
1000	6.7	0.14	0.0016	$1.7 \cdot 10^{-5}$	-	13	42	$3.7 \cdot 10^2$	$1 \cdot 10^3$	-
10000	6.6	0.14	0.0016	$1.8 \cdot 10^{-5}$	$2.6 \cdot 10^{-6}$	13	42	$3.7 \cdot 10^2$	$1 \cdot 10^3$	$1 \cdot 10^3$
100000	6.7	0.14	0.0016	$1.8 \cdot 10^{-5}$	$2.6 \cdot 10^{-6}$	13	42	$3.7 \cdot 10^2$	$1 \cdot 10^3$	$1 \cdot 10^3$

Table 1: Approximated KL and fixed-point complexity with Picard iteration for one step of the method.

time horizon in the heat equations, the method does not converge. When this value is sufficiently large, a rather good reconstruction can be achieved in one step for small β. The minimal value of $\mathrm{KL}(\rho_{\mathrm{TT}}|\rho_\infty) \sim 10^{-6}$ is obtained for $\beta = 10^{-4},\ T = 10^5$. The fixed point iterations are performed until the relative change of the discretized variables $\frac{\|\eta_m - \tilde{\eta}_m\|_F}{\|\eta_m\|_F}$ is less than 10^{-5}. The convergence of this quantity is linear with the rate decreasing when β decreases. The amount of fixed-point iterations can quickly become large, motivating the introduction of acceleration methods.

In our case, for acceleration the Anderson method based on $P = 2$ iterates is chosen because it admits a closed form solution to the minimization subproblem.



(a) Relative error during fixed-point iterations for one step of the method.

(b) TT approximation after one step with $T = 10^5,\ \beta = 10^{-4}$.

Figure 3: Reconstruction for a Gaussian target distribution.

The example of the fixed-point convergence plot is presented in Figure 3a. For illustrative purposes, we present the convergence results for two problems with the same value of $\beta T = 10$ and higher and lower regularization, $\beta = 10^{-1}$ (dashed lines) and $\beta = 10^{-3}$ (solid lines) correspondingly in the same plot. For the Picard iteration, one can clearly see that the linear convergence rate (slope of the line on the plot) is larger for larger regularization. Accelerated fixed-point iteration (orange) provides superlinear convergence and requires at maximum tens of iterations, giving a speed-up of up to 3 orders of magnitude while potentially reaching higher fixed-point tolerance.

A downside of the Anderson acceleration is the presence of the relaxation parameter q. From our numerical experiments, it cannot always be set to 1 in this problem since the computation may diverge. It is known that small q reduce convergence speed and to the authors' knowledge, there is no way to select it a priori or adjust it adaptively. Despite of this, in the further experiments Anderson acceleration with $P = 2$ and q around $0.8 - 0.9$ is used.

Figure 3b depicts the approximate distributions. Plots on the diagonal represent the marginal distributions of each variable x_i. The upper right triangle shows the contour lines of joint marginal distributions of variables (x_i, x_j) of the approximate and the true posterior. In the bottom left triangle, the contour lines belong to the true posterior and the scatter plot represents the sample from the approximate posterior acquired with the deterministic ODE method as described in Section 4.5.

5.3 Sampling from test distributions

The main focus of this part is to study the sampling capabilities of our method and compare the samples to baseline sampling methods using metrics based on optimal transport. The choice of the test distributions is partially inspired by [34]. In addition to the Gaussian mixture family, we consider distributions described by their potential $V : \rho \propto e^{-V}$. Multimodal distributions and distributions with nonconvex V can be particularly challenging for classical sampling methods such as MCMC or Langevin dynamics.

The chosen test posteriors are given as follows:

Gaussian mixture

$$\rho_{\mathsf{GM}} \propto \sum_{k=1}^{K} w_k \rho_{m_k, C_k},$$

where ρ_{m_k,C_k} is a normalized density of a d-dimensional Gaussian distribution with mean m_k and covariance C_k, w_k are positive weights and K is the number of components. In our test, we set $d = 30,\ K = 5, C_k = \sigma^2 I_d,\ \sigma^2 = 0.5$, weights are identical and $m_{i,j}$ is generated randomly from $U([-L/2, L/2])$.

The usage of the next two distributions is inspired by [34]. Due to multimodality (in the former case) and the non-convexity of the potential (in the latter), they are known to be hard to sample from with MCMC methods.

Double-moon potential distribution

$$\rho_{\mathsf{DM}}(x) \propto \exp\left(-2(\|x\|_2 - a)^2\right)\left(\exp(-2(x_1 - a)^2) + \exp(-2(x_1 + a)^2)\right)$$

with scalar parameter $a = 2$

Nonconvex potential distribution

$$V_{\text{NC}}(x) = \left(\sum_{i=1}^{d} \sqrt{|x_i - a_i|} \right)^2$$

with means vector $a = (a_i, \ldots, a_d)$ taken as $a_i = (-1)^i$.

The experiments are organized as follows. The Tensor-Train approximation is computed with one regularized JKO step. During the fixed-point iterations, the calls to the posterior are cached and the number of actual calls to the posterior and those taken from the cache is kept track of. The sample of size $n_s = 400$ is then generated with the composition of ODE and SDE dynamics as described in Section 4.5.

For comparison, reference samples are generated. In case of a Gaussian mixture, the target is sampled exactly. For the two other distributions, the reference is generated with Metropolis-Hastings MCMC (`emcee` package[72] is chosen as a particular implementation). The autocorrelation time τ_{ac} of the parameters can be estimated within the package, following the approach in [5]. We run $n_s = 400$ independent chains with starting value distributed as $\mathcal{N}(0, I_d)$ for a sufficiently large number of iterations ($\gtrsim 10000$) so that $N_{\text{iter}} > 50\tau_{ac}$. The covariance parameter in the Gaussian step is chosen so that the acceptance rate is close to 25%.

Finally, to have some perspective on the computational efficiency of the method, we compare it to the baseline method. In order to do this, MCMC is run again with the same parameters but with a shorter chain and the number of iteration chosen so that the number of the posterior calls is equal to the number of *unique* calls carried out with the TT method.

The procedure is repeated 20 times for the TT method and the short MCMC chain. We also generate 20 reference samples (in case of the long MCMC chain, 20 states separated by t_{ac} iterations are taken from the end of the chain). We then compute $S_2^2(\rho_{n_s}^{\text{TT}}, \rho_{n_s}^{\text{ref}})$, $S_2^2(\rho_{n_s}^{\text{MCMC}}, \rho_{n_s}^{\text{ref}})$, $S_2^2(\rho_{n_s}^{\text{ref}}, \rho_{n_s}^{\text{ref}})$, for each pair of different samples.

Distribution	d	r_{max}	N_∞		S_ε^2			Double OT	
			Unique	Total	Ref. to ref.	Ref. to TT	Ref. to MCMC	TT	MCMC
Mixture	30	5	$1290K$	$2561K$	9.04 ± 0.10	9.05 ± 0.14	9.77 ± 0.33	$2.1 \cdot 10^{-3}$	$5.8 \cdot 10^{-1}$
Double-Moon	6	3	$111K$	$164K$	0.65 ± 0.02	0.67 ± 0.03	0.66 ± 0.03	$3.4 \cdot 10^{-4}$	$1.2 \cdot 10^{-4}$
Nonconvex	6	2	$25K$	$183K$	0.06 ± 0.01	0.08 ± 0.01	1.75 ± 0.11	$3.9 \cdot 10^{-4}$	$2.9 \cdot 10^{0}$

Table 2: Summary of results for each distribution, showing problem dimension, maximal TT rank, posterior call counts, average Sinkhorn distance with standard deviation for each sampling method, and the Double OT estimate between distributions.

The results are presented in Table 2. For each distribution, the dimension of the problem and the maximal TT rank of the solution r_{max} are presented. N_∞ denotes the number of the posterior calls carried out by the TT method, including the unique calls (when the posterior is actually called) and the total number (i.e. unique and re-used from cache). The Sinkhorn distance S_ε^2 averaged with respect to every different pair of samples generated by the TT method, by the short MCMC chain and between the reference itself, is presented alongside its standard deviation. Finally, the OT distance between these distributions is denoted by Double OT. The results are illustrated in Figure 4. The contour lines of a two-dimensional marginal of the test distribution are shown alongside with the reference sample, a

sample generated with the TT method and the MCMC sample with a shorter chain generated with the same amount of posterior calls as for the TT method.

Interestingly, the test runs show that approximation of multimodal and concentrated distributions is in fact possible with quite low TT ranks. One can see that the cache plays a significant role in the computation, increasing the effective amount of the posterior calls (compared to the scheme with no cache) by a factor ranging from ~ 1.5 times in the case of the Double-Moon problem up to ~ 7.3 times in the case of the Nonconvex problem. In general, the method performs at least as good as the baseline with the same computational effort. We note that the Nonconvex problem was particularly difficulty for the MCMC method, where the difference between the short chain MCMC sample and the reference can be seen immediately in Figure 4.

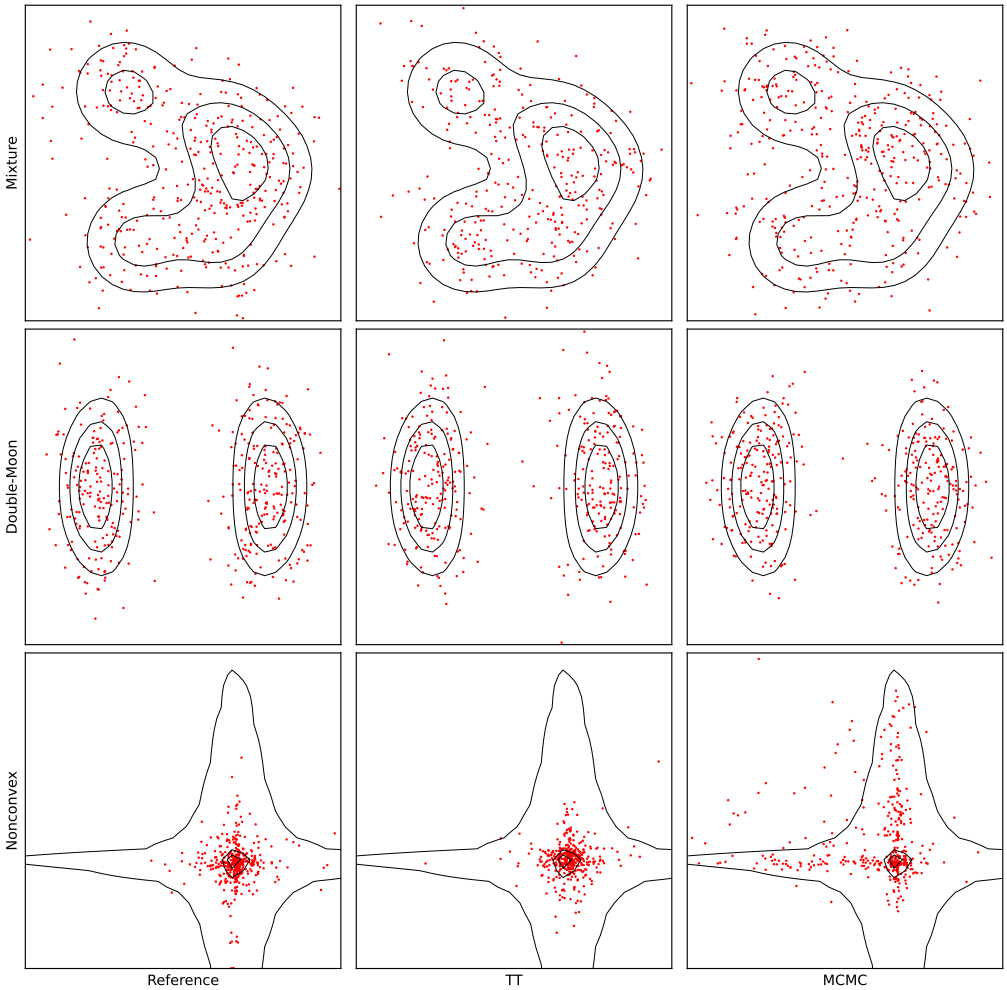


Figure 4: Contour lines of two-dimensional marginals of the test distributions, shown with the reference sample, the TT-generated sample, and an MCMC sample using a shorter chain with an equivalent number of posterior calls as the TT method.

6 Application to Bayesian inverse problems

The present section consists of case studies for two certain PDE-constrained Bayesian inverse problem. We consider the application of our method for Bayesian inversion. Here our method is reviewed in the context of estimating the unknown parameters while providing uncertainty quantification. We also demonstrate how the fitted Tensor-Train model for the Bayesian posterior can be re-used to efficiently solve an associated problem, in our case, importance sampling for some quantity of interest, that non-trivially depends on the unknown parameters.

Specifically, we aim to reconstruct the initial condition in an initial value problems for PDEs. Our two test cases are described in the following.

Hyperbolic equation The initial value problem for the wave equation is considered:

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial t^2} = 0, \tag{6.0.1}$$

$$u(0,x) = h_\theta(x),\ \frac{\partial u}{\partial t}(0,x) = 0 \tag{6.0.2}$$

where the initial condition depends on the unknown parameters θ, the position of «peaks» in the initial distribution

$$h_\theta(x) = \sum_{i=1}^{d} e^{-(x-\theta_i)^2}.$$

The solution to this problem is given by

$$u(\theta;t,x) = \frac{1}{2}(h_\theta(x-t) + h_\theta(x+t)).$$

Parabolic equation Here the initial-boundary value problem for the heat equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \tag{6.0.3}$$

$$\left.\frac{\partial u}{\partial x}\right|_{x=\{-1,1\}} = 0, \tag{6.0.4}$$

$$u(0,x) = h_\theta(x) \tag{6.0.5}$$

with the initial condition parametrized by a truncated Fourier series in the cosine basis

$$h_\theta(x) = \sum_{j=0}^{d-1} \theta_i \cos(j\pi x).$$

is considered. The solution to this problem is given by

$$u(\theta;t,x) = \sum_{j=0}^{d-1} \theta_j e^{-(\pi j)^2 t} \cos(j\pi x).$$

For both forward process, the inverse problem is formulated in the same way. Firstly, a Gaussian prior on the parameters is chosen by

$$\theta_k \sim \mathcal{N}(0, \sigma_{0,k}^2) \qquad \text{independent of each other.}$$

For the hyperbolic problem, these distributions have the same standard deviation $\sigma_{0,k} = \sigma_0$. For the parabolic ones, the standard deviation decays like $\sigma_{0,k} = \frac{\sigma_0}{k+1}$, corresponding to a truncation of a trace-class covariance operator in a Karhunen-Loève expansion. A set of parameters θ^* from the prior distribution is sampled and fixed as «true» parameters. Afterwards, noisy measurements of the solution

$$\tilde{u}_{ij} = u(\theta^*; t_i, x_j) + \xi_{ij}, \quad \xi_{ij} \sim \mathcal{N}(0, \sigma^2_{meas}\,\mathrm{I}_d) \quad \text{i.i.d.}$$

are generated. The space and time discretization points are located equidistantly. The goal is to reconstruct the parametrized initial condition, given these measurements, their likelihood function and the prior distribution of the parameters. The potential of the posterior distribution is defined by

$$V_\infty(\theta) = \frac{1}{2\sigma^2_{meas}} \sum_{i,j} |u(\theta; t_i, x_j) - \tilde{u}_{ij}|^2 + \frac{1}{2} \sum_{k=1}^{d} \frac{1}{\sigma^2_{0,k}} |\theta_k|^2. \tag{6.0.6}$$

We opt to sample from the posterior distribution. In addition to comparing the sample to the reference, we provide an estimate to the parameters along with some uncertainty quantification. Since from the fitted TT approximation a discretized version of the marginal probability density of each of the parameters can be obtained, various statistical quantities such as means, variances, quantiles, etc. can be estimated. We choose to provide a *maximal a posteriori* (MAP) estimate and an 89%-highest density credible interval, i.e. the minimal width interval containing 89% of the marginal posterior.

One can notice that in its original form the forward problem of the hyperbolic equation is invariant under permutation of the parameters (positions of the «peaks» in the initial distribution). In the preliminary computations we noticed the said invariance leads to a posterior distributions with very many modes, which then hinders the convergence of all the methods and renders the interpretation of the results difficult. Thus, we choose to enforce the condition of the parameters being in increasing order by sampling with the perturbed potential

$$V_{\infty,s}(\theta) = V_\infty(\theta) + \begin{cases} 0, \text{ if } \theta_1 \le \theta_2 \le \cdots \le \theta_d \\ +\infty, \text{ otherwise} \end{cases}.$$

First, the approximation to the posterior with the TT method is constructed. Next, samples with the TT model and MCMC with the same amount of the posterior calls are generated and compared in the same fashion as in the previous section. The results are presented in Table 3, with the same metrics

Distribution	d	r_{max}	N_∞		$S^2_\varepsilon \times 10^3$			Double OT	
			Unique	Total	Ref. to ref.	Ref. to TT	Ref. to MCMC	TT	MCMC
Hyperbolic	6	1	$22K$	$92K$	0.56 ± 0.21	5.23 ± 0.84	277.37 ± 31.34	$2.2 \cdot 10^{-5}$	$7.8 \cdot 10^{-2}$
Parabolic	10	1	$15K$	$60K$	1.61 ± 0.28	2.09 ± 0.30	781.28 ± 55.29	$2.4 \cdot 10^{-7}$	$6.1 \cdot 10^{-1}$

Table 3: Summary of results for each problem, showing its dimension, maximal TT rank, posterior call counts, average Sinkhorn distance with standard deviation for each sampling method, and the Double OT estimate between distributions.

tracked as in the previous section. One can observe that quite a good approximation superior to the baseline method can be achieved by our TT representation. The caching provides at least 4 times acceleration in terms of the number of posterior calls.

The results of the parameter estimation is presented in Table 4. The leftmost column provides the true parameters θ^*. Next, minimal and maximal value of a *Highest Density Interval* (HDI) along with the

θ^*	θ_{MCMC}			θ_{TT}			ϵ_{rel}
	min	MAP	max	min	MAP	max	
-2.30	-2.31	-2.13	-1.90	-2.33	-2.15	-1.97	0.11
-1.07	-1.24	-1.15	-0.82	-1.24	-1.06	-0.86	0.04
-0.61	-0.84	-0.42	-0.39	-0.81	-0.64	-0.49	0.14
-0.53	-0.44	-0.37	-0.04	-0.36	-0.21	-0.07	0.13
0.87	0.74	0.96	1.12	0.73	0.88	1.04	0.12
1.62	1.69	1.91	2.13	1.70	1.91	2.07	0.08

(a) Hyperbolic

θ^*	θ_{MCMC}			θ_{TT}			ϵ_{rel}
	min	MAP	max	min	MAP	max	
1.62	1.60	1.64	1.69	1.58	1.67	1.70	0.17
0.31	0.13	0.24	0.34	0.12	0.23	0.33	0.03
0.18	0.03	0.24	0.45	0.05	0.25	0.47	0.05
0.27	-0.35	-0.24	0.27	-0.36	-0.05	0.24	0.03
-0.17	-0.34	-0.01	0.26	-0.35	-0.04	0.25	0.00
0.38	-0.28	0.02	0.21	-0.26	-0.01	0.26	0.07
-0.25	-0.23	-0.01	0.22	-0.22	0.00	0.23	0.02
0.10	-0.19	0.03	0.21	-0.19	0.00	0.20	0.02
-0.04	-0.15	-0.02	0.20	-0.17	-0.00	0.17	0.08
0.02	-0.16	-0.00	0.15	-0.16	0.00	0.16	0.02

(b) Parabolic

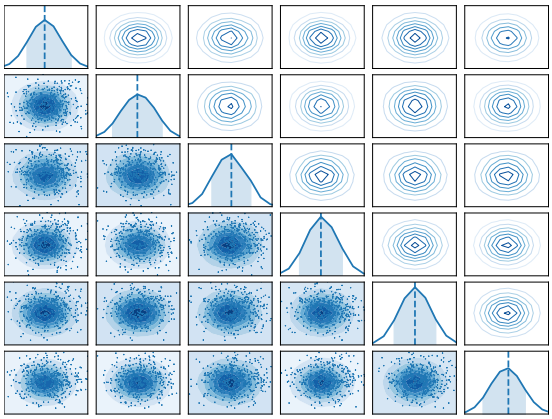
Table 4: Parameters and their credible intervals, estimated by the TT method and the baseline MCMC. Low relative error ϵ_{rel} shows good agreement between the methods.

MAP estimate are presented. This is first for the reference MCMC chain with $\gtrsim 10^5$ iterations and then for the TT method. The error of determination of the intervals relative to their respective lengths (as uncertainty quantification) is provided in the rightmost column.

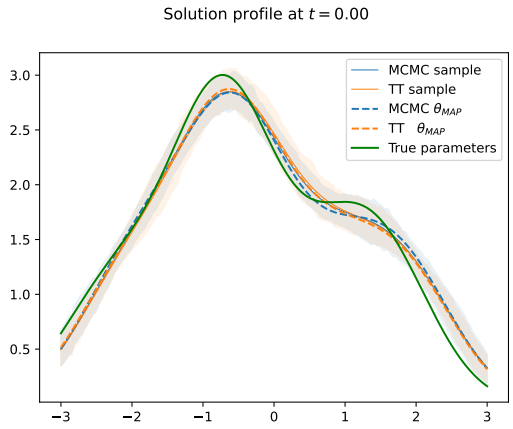
6.1 Modelling of the importance distribution

We argue that the structured representation of the posterior density as a discretized function provides the opportunity to modify the solution while reusing the computationally expensive posterior calls. For example, new variables can be added to the distribution by concatenating new cores to the Tensor Train representing the posterior density. In case when the forward operator calls are the bottleneck in computation of the posterior density, these can be cached (instead of the posterior calls) and reused in posterior fitting for multiple experiments with different measurements but within the same parameter bounds. Leaving a thorough examination of these extensions to further work, we provide a single example, namely demonstrating that approximating the importance distribution for a certain quantity of interest can be carried out without any additional calls to the posterior density.

To briefly summarize the idea of importance sampling, given a certain quantity of interest F for which $\mathbb{E}_{\rho_\infty}[F]$ has to be estimated, one designs a distribution ρ_F and a weighting function w such that



(a) Contour lines and samples of 1- and 2-d marginals of the TT approximation of the posterior. The 89% credible interval for each parameter is filled.



(b) Reconstruction of the initial condition by the TT method and MCMC reference. The filled patch represents a 89% credible interval for the solution $u(x)|_{t=0}$ estimated via samples.

Figure 5: Posterior distribution and reconstruction of the initial solution for the hyperbolic problem.

$\mathbb{E}_{\rho_F}[wF] = \mathbb{E}_{\rho_\infty}[F]$ and the variance is minimized. The *importance weight* takes the form

$$w(x) = \frac{\rho_\infty(x)}{\rho_F(x)} \tag{6.1.1}$$

and in case of the optimal importance distribution [73, Chapter 5.5],

$$\rho_{F,\mathsf{opt}}(x) \propto |F(x)|\rho_\infty(x) \quad \text{and} \quad w_{\mathsf{opt}}(x) = \frac{1}{|F(x)|}. \tag{6.1.2}$$

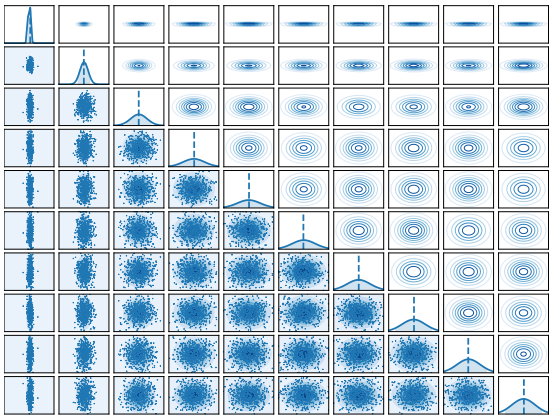
The estimator for $\mathbb{E}_{\rho_\infty}[F]$ is then given by

$$\hat{F}_N = \frac{\sum_{i=1}^N w(x_i)F(x_i)}{\sum_{i=1}^N w(x_i)} \qquad \text{with} \quad x_i \sim \rho_F \text{ i.i.d.} \tag{6.1.3}$$

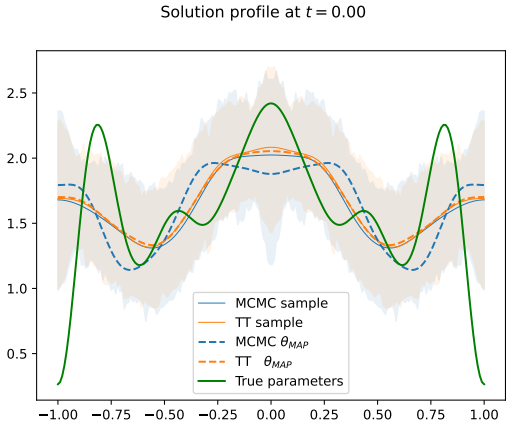
Although ρ_∞ and ρ_F are both known only up to multiplicative factors, these factors cancel out in (6.1.3). We note that this estimate is biased.

To demonstrate importance sampling with the TT model, we modify the setting used in the inverse problem for the parabolic equation. Mimicking certain real-world settings, we assume that the solution in the inside area is of practical interest but not accessible for direct measurements [74, 75, 76]. We use the same geometry and the same priors on the parameters, as in the parabolic problem in the previous section, but assume there are no measurement points in $x \in [-\frac{1}{2}, \frac{1}{2}]$ and the 10 measurements are taken outside of the interval with equal spacing. We suppose that a reliable measurement of $\mathbb{E}_{\theta\sim\rho_\infty} u(\theta; 0, 0)$ is required. Note that the computation of the function $F(\theta) := u(\theta; 0, 0)$ only requires the evaluation of the parametrized initial condition but not the expensive forward model.

We fit one step of the method with initial step ρ_{TT} (the TT approximation of the posterior) and the target $\propto \|F\|\rho_{\mathsf{TT}}$, assuming the resulting approximation $\rho_{F,\mathsf{TT}}$ to be an adequate approximation to the importance distribution. Then, 10 samples of size $N_{\mathsf{samples}} = 400$ each are generated. And for



(a) Contour lines and samples of 1- and 2-d marginals of the TT approximation of the posterior. The 89% credible interval for each parameter is filled.



(b) Reconstruction of the initial condition by the TT method and MCMC reference. The filled patch represents a 89% credible interval for the solution $u(x)|_{t=0}$ estimated via samples.

Figure 6: Posterior distribution and reconstruction of the initial solution for the parabolic problem.

each of them, the expectation of F is estimated via the estimator (6.1.3). The computation of the true importance weight (6.1.1) still requires N_{samples} posterior calls. We have noted however that the resulting distribution of the estimates for F does not change significantly if the optimal weight (6.1.2) is used instead. Thus, the estimate can be computed without any posterior calls at all. This approach is compared to estimating the mean directly from the samples $F(x),\ x \sim \rho_\infty$. The results are depicted in Figure 7. One can see that the estimates without importance sampling (blue) are quite spread out.

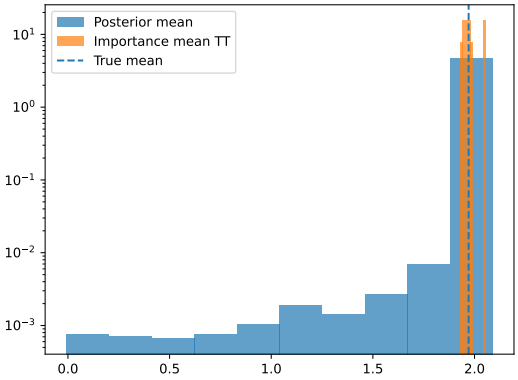


Figure 7: Distribution of the estimate of $\mathbb{E}_{\rho_\infty} F$ via the posterior mean (blue) and the importance estimator (6.1.3), along with the true mean (vertical blue line)

The estimates via the importance distribution are not centered around the actual mean due to the bias in the estimate, but in general are much closer to it. Thus, the importance estimator visibly reduces variance of the estimation, and is a useful tool, provided that with our method it can be acquired without any additional posterior calls.

7 Discussion

With the current implementation using a finite difference approximation of the functions, we are already able to demonstrate the main ideas of the proposed new method, namely the advantages of the Eulerian approach, the benefit of caching and the re-usability of the results for similar problems. There is however still a lot of room for improvements. The interpolation can be improved by using the *functional Tensor-Train* format [77]. A careful choice of the basis function should be dictated by the properties of the problem. To allow for even more re-using of the posterior calls, one can consider building a model for the joint distribution $\rho(x, y)$ and conditioning it to certain values of the measurement y. In this approach, the model is re-used to estimate multiple parameter sets from multiple measurements carried out with the same experimental setup. Numerically, this can be achieved either by directly conditioning the TT model for $\rho(x, y)$ by means of a multidimensional integration or by considering proximal steps with respect to the conditional Wasserstein distance [78].

The regularized JKO scheme presented in the current paper can in principle be viewed as a general-purpose minimization algorithm in Wasserstein space. This can be achieved as long as an efficient computation of the energy first variation of a general energy by (3.2.8) can be implemented. Energies different from KL divergence may for example arise in case when in addition to approximating the target distribution the model has to meet certain constraints such as fairness, safety or interpretability (see [79] and references therein).

Acknowledgements

The authors would like to express their gratitude to Mathias Oster, David Sommer, Robert Gruhlke and Reinhold Schneider for helpful discussions.

References

[1] Nabila Aghanim, Yashar Akrami, Mark Ashdown, J Aumont, Carlo Baccigalupi, M Ballardini, Anthony J Banday, RB Barreiro, N Bartolo, S Basak, et al. Planck 2018 results-V. CMB power spectra and likelihoods. *Astronomy & Astrophysics*, 641:A5, 2020.

[2] Qiong Liu, Yu-Jung Tsai, Jean-Dominique Gallezot, Xueqi Guo, Ming-Kai Chen, Darko Pucar, Colin Young, Vladimir Panin, Michael Casey, Tianshun Miao, Huidong Xie, Xiongchao Chen, Bo Zhou, Richard Carson, and Chi Liu. Population-based deep image prior for dynamic PET denoising: A data-driven approach to improve parametric quantification. *Medical Image Analysis*, 95:103180, 2024.

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

[4] Andrew Gelman and Donald B Rubin. A single series from the Gibbs sampler provides a false sense of security. *Bayesian statistics*, 4(1):625–631, 1992.

[5] Jonathan Goodman and Jonathan Weare. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, 5(1):65–80, 2010.

[6] Alfredo Garbuno-Inigo, Nikolas Nüsken, and Sebastian Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. *SIAM Journal on Applied Dynamical Systems*, 19(3):1633–1658, 2020.

[7] Martin Eigel, Robert Gruhlke, and David Sommer. Less interaction with forward models in Langevin dynamics. *arXiv preprint arXiv:2212.11528*, 2022.

[8] Yuri Evtushenko. Computation of exact gradients in distributed dynamic systems. *Optimization Methods and Software*, 9(1-3):45–75, 1998.

[9] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[10] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[11] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[12] Ugo Gianazza, Giuseppe Savaré, and Giuseppe Toscani. The Wasserstein gradient flow of the fisher information and the quantum drift-diffusion equation. *Archive for rational mechanics and analysis*, 194(1):133–220, 2009.

[13] Adrien Blanchet. A gradient flow approach to the Keller-Segel systems (progress in variational problems : Variational problems interacting with probability theories). 2013.

[14] Hugo Leclerc, Quentin M'erigot, Filippo Santambrogio, and Federico Stra. Lagrangian discretization of crowd motion and linear diffusion. *SIAM J. Numer. Anal.*, 58:2093–2118, 2019.

[15] Jean-David Benamou, Guillaume Carlier, and Maxime Laborde. An augmented lagrangian approach to Wasserstein gradient flows and applications. *ESAIM: Proceedings and surveys*, 54:1–17, 2016.

[16] Jose A Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for Wasserstein gradient flows. *Foundations of Computational Mathematics*, pages 1–55, 2022.

[17] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

[18] Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[19] Rafael Bailo, José A. Carrillo, and Jingwei Hu. Fully Discrete Positivity-Preserving and Energy-Dissipating Schemes for Aggregation-Diffusion Equations with a Gradient-Flow Structure. *Communications in Mathematical Sciences*, 18(5):1259–1303, 2020.

[20] Clément Cances, Thomas O Gallouët, and Gabriele Todeschi. A variational finite volume scheme for Wasserstein gradient flows. *Numerische Mathematik*, 146:437–480, 2020.

[21] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

[22] Qiang Liu. Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, 30, 2017.

[23] Jun Han and Qiang Liu. Stein variational gradient descent without gradient. In *International Conference on Machine Learning*, pages 1900–1908. PMLR, 2018.

[24] Aimee Maurais and Youssef Marzouk. Sampling in unit time with kernel Fisher-Rao flow. *arXiv preprint arXiv:2401.03892*, 2024.

[25] Ivan Oseledets and Eugene Tyrtyshnikov. TT-cross approximation for multidimensional arrays. *Linear Algebra and its Applications*, 432(1):70–88, 2010.

[26] Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. The alternating linear scheme for tensor optimization in the tensor train format. *SIAM Journal on Scientific Computing*, 34(2):A683–A713, 2012.

[27] Georgii S Novikov, Maxim E Panov, and Ivan V Oseledets. Tensor-train density estimation. In *Uncertainty in artificial intelligence*, pages 1321–1331. PMLR, 2021.

[28] Yinuo Ren, Hongli Zhao, Yuehaw Khoo, and Lexing Ying. High-dimensional density estimation with tensorizing flow. *Research in the Mathematical Sciences*, 10(3):30, 2023.

[29] Tiangang Cui and Sergey Dolgov. Deep composition of tensor-trains using squared inverse Rosenblatt transports. *Foundations of Computational Mathematics*, 22(6):1863–1922, 2022.

[30] Markus Bachmayr. Low-rank tensor methods for partial differential equations. *Acta Numerica*, 32:1–121, 2023.

[31] Lorenz Richter, Leon Sallandt, and Nikolas Nüsken. Solving high-dimensional parabolic PDEs using the tensor train format. In *International Conference on Machine Learning*, pages 8998–9009. PMLR, 2021.

[32] Andrei Chertkov and Ivan Oseledets. Solution of the Fokker–Planck equation by cross approximation method in the tensor train format. *Frontiers in Artificial Intelligence*, 4:668215, 2021.

[33] Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

[34] Fuqun Han, Stanley Osher, and Wuchen Li. Tensor train based sampling algorithms for approximating regularized Wasserstein proximal operators. *arXiv preprint arXiv:2401.13125*, 2024.

[35] Matthias Liero and Alexander Mielke. Gradient structures and geodesic convexity for reaction–diffusion systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(2005):20120346, 2013.

[36] Katy Craig. Nonconvex gradient flow in the Wasserstein metric and applications to constrained nonlocal interactions. *Proceedings of the London Mathematical Society*, 114(1):60–102, 2017.

[37] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 10 2019.

[38] Luis Caicedo Torres, Luiz Manella Pereira, and M Hadi Amini. A survey on optimal transport for machine learning: Theory and applications. *arXiv preprint arXiv:2106.01963*, 2021.

[39] Nicolas Bonneel and Julie Digne. A survey of optimal transport for computer graphics and computer vision. In *Computer Graphics Forum*, volume 42, pages 439–460. Wiley Online Library, 2023.

[40] Luigi Ambrosio, Stefano Lisini, and Giuseppe Savaré. Stability of flows associated to gradient vector fields and convergence of iterated transport maps. *manuscripta mathematica*, 121(1):1–50, 2006.

[41] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[42] Dirk A Lorenz, Paul Manns, and Christian Meyer. Quadratically regularized optimal transport. *Applied Mathematics & Optimization*, 83(3):1919–1949, 2021.

[43] Jonas Bresch and Viktor Stein. Interpolating between optimal transport and KL regularized optimal transport using Rényi divergences. *arXiv preprint arXiv:2404.18834*, 2024.

[44] Marcel Nutz and Johannes Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184(1):401–424, 2022.

[45] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.

[46] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[47] Wuchen Li, Jianfeng Lu, and Li Wang. Fisher information regularization schemes for Wasserstein gradient flows. *Journal of Computational Physics*, 416:109449, 2020.

[48] Wuchen Li, Siting Liu, and Stanley Osher. A kernel formula for regularized Wasserstein proximal operators. *arXiv preprint arXiv:2301.10301*, 2023.

[49] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. On the relation between optimal transport and schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169:671–691, 2016.

[50] Yongxin Chen, Tryphon Georgiou, and Michele Pavon. Entropic and displacement interpolation: a computational approach using the hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396, 2016.

[51] Xiuyuan Cheng, Jianfeng Lu, Yixin Tan, and Yao Xie. Convergence of flow-based generative models via proximal gradient descent in Wasserstein space. *arXiv preprint arXiv:2310.17582*, 2023.

[52] Alexander Graham. *Kronecker products and matrix calculus with applications*. Courier Dover Publications, 2018.

[53] Awad H. Al-Mohy and Nicholas J. Higham. A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications*, 31(3):970–989, 2010.

[54] Zhen Qin, Alexander Lidiak, Zhexuan Gong, Gongguo Tang, Michael B Wakin, and Zhihui Zhu. Error analysis of tensor-train cross approximation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 14236–14249. Curran Associates, Inc., 2022.

[55] Sergey Dolgov and Dmitry Savostyanov. Parallel cross interpolation for high-precision calculation of high-dimensional integrals. *Computer Physics Communications*, 246:106869, 2020.

[56] Dmitry Savostyanov and Ivan Oseledets. Fast adaptive interpolation of multi-dimensional arrays in tensor train format. In *The 2011 International Workshop on Multidimensional (nD) Systems*, pages 1–8. IEEE, 2011.

[57] Philipp Birken, Tobias Gleim, Detlef Kuhl, and Andreas Meister. Fast solvers for unsteady thermal fluid structure interaction. *International Journal for Numerical Methods in Fluids*, 79(1):16–29, 2015.

[58] Jisun Park and Ernest K Ryu. Exact optimal accelerated complexity for fixed-point iterations. In *International Conference on Machine Learning*, pages 17420–17457. PMLR, 2022.

[59] Alex Toth and Carl T Kelley. Convergence analysis for Anderson acceleration. *SIAM Journal on Numerical Analysis*, 53(2):805–819, 2015.

[60] Haw-ren Fang and Yousef Saad. Two classes of multisecant methods for nonlinear acceleration. *Numerical linear algebra with applications*, 16(3):197–221, 2009.

[61] Homer F Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

[62] Vitalii Aksenov, Maxim Chertov, and Konstantin Sinkov. Application of accelerated fixed-point algorithms to hydrodynamic well-fracture coupling. *Computers and Geotechnics*, 129:103783, 2021.

[63] Bohao Tang, Nicholas C. Henderson, and Ravi Varadhan. Accelerating fixed-point algorithms in statistics and data science: A state-of-art review. *Journal of Data Science*, 21(1):1–26, 2022.

[64] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[65] David Sommer, Robert Gruhlke, Max Kirstein, Martin Eigel, and Claudia Schillings. Generative modelling with tensor train approximations of Hamilton–Jacobi–Bellman equations. *arXiv preprint arXiv:2402.15285*, 2024.

[66] Andrei Chertkov, Gleb Ryzhakov, and Ivan Oseledets. Black box approximation in the tensor train format initialized by ANOVA decomposition. *SIAM Journal on Scientific Computing*, 45(4):A2101–A2118, 2023.

[67] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

[68] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

[69] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

[70] Erhan Bayraktar and Gaoyue Guo. Strong equivalence between metrics of Wasserstein type. 2021.

[71] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[72] Daniel Foreman-Mackey, David W Hogg, Dustin Lang, and Jonathan Goodman. emcee: the MCMC hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306, 2013.

[73] Robert Scheichl and Jakob Zech. Numerical methods for Bayesian inverse problems. 2021.

[74] TJ Martin and GS Dulikravich. Inverse determination of boundary conditions and sources in steady heat conduction with heat generation. 1996.

[75] TJ Martin and GS Dulikravich. Inverse determination of temperatures and heat fluxes on inaccessible surfaces. *WIT Transactions on Modelling and Simulation*, 8, 2024.

[76] AS Carasso. Space marching difference schemes in the nonlinear inverse heat conduction problem. *Inverse Problems*, 8(1):25, 1992.

[77] Ivan V Oseledets. Constructive representation of functions in low-rank tensor formats. *Constructive Approximation*, 37:1–18, 2013.

[78] Jannis Chemseddine, Paul Hagemann, Christian Wald, and Gabriele Steidl. Conditional wasserstein distances with applications in bayesian ot flow matching. *arXiv preprint arXiv:2403.18705*, 2024.

[79] Xingchao Liu, Xin Tong, and Qiang Liu. Sampling with trusthworthy constraints: A variational gradient framework. *Advances in Neural Information Processing Systems*, 34:23557–23568, 2021.