Journal of Machine Learning Research 27 (2026) 1-38 Submitted 1/26; Revised 7/26; Published 9/26

Anderson Mixing in Bures–Wasserstein Space of Gaussian Measures

Vitalii Aksenov VITALII.AKSENOV@WIAS-BERLIN.DE
RG4
WIAS
Berlin, Germany

Martin Eigel MARTIN.EIGEL@WIAS-BERLIN.DE
RG4
WIAS
Berlin, Germany

Mathias Oster OSTER@IGPM.RWTH-AACHEN.DE
IGPM
RWTH
Aachen, Germany

Editor: Jianfeng Lu

Abstract

Various statistical tasks, including sampling or computing Wasserstein barycenters, can be reformulated as fixed-point problems for operators on probability distributions. Accelerating standard fixed-point iteration schemes provides a promising novel approach to the design of efficient numerical methods for these problems. The Wasserstein geometry on the space of probability measures, although not precisely Riemannian, allows us to define various useful Riemannian notions, such as tangent spaces, exponential maps and parallel transport, motivating the adaptation of Riemannian numerical methods. We demonstrate this by developing and implementing the Riemannian Anderson mixing (RAM) method for Gaussian distributions. The method reuses the history of the residuals and improves the iteration complexity, and we argue that the additional costs, compared to the Picard method, are negligible. We show that certain balls in the Bures–Wasserstein manifold satisfy the requirements for convergence of RAM. The numerical experiments show a significant acceleration compared to the Picard iteration, and performance on par with Riemannian gradient descent and conjugate gradient methods.

1. Introduction

There is a plethora of tasks in statistics, ranging from sampling given a Bayesian posterior to calculating Wasserstein barycenters and medians, that can be reformulated as a fixed-point problem in Wasserstein space.

For example, the dynamics of the Wasserstein gradient flow of certain functionals is contractive and, thus, induces fixed-point schemes. To make this more precise, let $\mathcal{P}_2(\mathbb{R}^d)$ be the space of probability measures with finite second moments

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int \|x\|_2^2 \, \mathrm{d}\mu(x) < \infty \right\}$$

©2026 Vitalii Aksenov, Martin Eigel, Mathias Oster.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/26-0075.html.

endowed with the 2-Wasserstein distance

$$W_2^2(\mu_1, \mu_2) := \inf_{\pi \in \Pi(\mu_1, \mu_2)} \int \|x - y\|_2^2 \, \mathrm{d}\pi(x, y),$$

where $\Pi(\mu_1, \mu_2)$ denotes the set of probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ_1 and μ_2.

As shown in Ambrosio et al. (2005), if some functional $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ is λ-convex along generalized geodesics for $\lambda > 0$ (Ambrosio et al., 2005, Definition 9.2.4), then for any $\mu \in \overline{D(\mathcal{E})} := \overline{\{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mathcal{E}(\mu) < \infty\}}$ there exists a unique locally Lipschitz curve $S[\mu](t) : (0, +\infty) \to \mathcal{P}_2(\mathbb{R}^d)$, which is a gradient flow of $\mathcal{E}$ with initial value μ. For a fixed time τ, the operator defined by $G_\tau(\mu) = S[\mu](\tau)$ satisfies the following contraction estimate:

$$W_2(G_\tau(\mu_1), G_\tau(\mu_2)) \le e^{-\tau\lambda} W_2(\mu_1, \mu_2).$$

If $\mathcal{E} = \mathrm{KL}(\cdot|\rho_*)$ is the Kullback–Leibler divergence and $\rho_* \propto e^{-V}$ for some coercive potential V, then λ-convexity along generalized geodesics is equivalent to λ-convexity of V and ρ_* is a fixed point of G_τ. The case of KL divergence is of particular interest, as a multitude of numerical algorithms for sampling can be interpreted as relaxations and approximations of this gradient flow. Important examples are Langevin dynamics, see, e.g., Wibisono (2018), and Stein variational gradient descent, see, e.g., Liu and Wang (2016); Duncan et al. (2023).

Furthermore, MCMC methods such as Metropolis–Hastings (Hastings, 1970) or Metropolis-adjusted Langevin dynamics (Roberts and Tweedie, 1996) have by design the target distribution as an invariant distribution of the Markov chain. Thus, they can also be seen as a fixed-point iteration. In addition, fixed-point iterations can be used to find the Wasserstein barycenter as noted in Álvarez-Esteban et al. (2016).

Finally, it is known that for a convex function $f : \mathbb{R}^d \to \mathbb{R}$, a fixed point of $\mathrm{Id} - h\nabla f$ is the root of the gradient and, therefore, the global minimizer. Thus, a fixed-point method can be used as a first-order optimization method and can be applied to a variety of optimization problems.

However, in various instances fixed-point schemes based on a contraction principle suffer from a slow convergence rate, for example, if the contraction constant is close to unity. In order to speed up the calculations one can generalize Euclidean or Riemannian acceleration schemes to Wasserstein space by exploiting the manifold-like structure. Previously, similar ideas have been applied in gradient-based optimization. In Kent et al. (2021), the Frank–Wolfe method is adapted to the Wasserstein setting. A Riemannian Frank–Wolfe method was applied to the Wasserstein barycenter problem in Weber and Sra (2023). A counterpart of Nesterov acceleration in Wasserstein space is considered in Liu et al. (2018); Chen et al. (2025). Riemannian minimization on the Bures–Wasserstein manifold is studied in Han et al. (2021).

It is known that the space of probability measures on $\mathbb{R}^d$ can be endowed with the 2-Wasserstein metric and then forms a complete separable metric space (Ambrosio et al., 2005, Proposition 7.1.5). This space is not exactly Riemannian (i.e., there is no atlas of maps to subsets of a fixed linear space), but nevertheless all the important geometric notions, such as geodesics, tangent spaces, exponential map and parallel transport, can be defined. However, the deviations of the geometry from standard Riemannian assumptions, e.g. the exponential map having a vanishing injectivity radius (Gigli, 2008), pose hurdles in designing schemes in Wasserstein space beyond the Gaussian case based on already known algorithms.

Contributions: We adapt the Riemannian Anderson mixing (RAM) algorithm introduced in Li and Bao (2023) to Gaussian measures on the Bures–Wasserstein manifold (referred to as BWRAM). While following the proof strategy of Li and Bao (2023), we use the explicit structure of the Bures–Wasserstein manifold to verify some assumptions for the convergence analysis of RAM and to identify the limits of generalizing the methods to general measures in Wasserstein space. Furthermore, by revisiting the proof and simplifying certain steps for this special case, we obtain improved constants that are explicitly related to spectral assumptions on the fixed point. The

statement of the convergence theorem for BWRAM with all required assumptions can be found in Section 4.1. Informally, it can be summarized as follows:

Theorem 1 (Convergence of BWRAM (informal statement)) *Let $G(\Sigma) = \mathrm{Exp}_\Sigma(-F(\Sigma))$ be a κ–contractive mapping on the Bures–Wasserstein manifold and $\Sigma_* \succeq \lambda \mathrm{Id}$ its fixed point with $\lambda > 0$. If the initial iterate $\Sigma_0 \in B_r(\Sigma_*)$ for a sufficiently small r, and under additional smoothness assumptions on F, then the following convergence estimate holds*

$$W_2(\Sigma_{k+1}, G(\Sigma_{k+1})) \leq [(1-\beta_k) + \kappa\beta_k] \|r_k\| + \sum_{i=0}^{\min\{m,k\}} \mathcal{O}(\|r_{k-i}\|^2), \tag{1}$$

where Exp *denotes the Riemannian exponential map, $r_k = -F(\Sigma_k)$ the fixed-point residual, m the maximal number of history vectors and $\beta_k > 0$ the relaxation parameter.*

We point out that the asymptotic behavior is exactly as in the Euclidean and RAM cases, but with an improved pre-asymptotic regime. The theoretical estimate says that, up to higher-order terms, the method is guaranteed to converge at least as fast as Picard iteration. We argue that the additional computational costs incurred by the method are negligible in comparison with the operator evaluation, and thus an overall speedup of the solution is provided. The improved convergence behavior of BWRAM in comparison to Picard is illustrated with various numerical examples such as estimation of the steady state of the Ornstein–Uhlenbeck process, minimization of the Kullback–Leibler divergence, and accelerated solution of averaging problems, such as Wasserstein barycenters and geometric medians. We also demonstrate that our method outperforms other well-known Riemannian minimization methods, such as Riemannian gradient descent and conjugate gradient.

Limitations: Empirically, the method achieves convergence speed superior to Picard iteration at the expense of additional memory costs and solving a small-scale minimization problem. The performance of the method is sensitive to the choice of hyperparameters, such as the number of history vectors. Although the Picard method is outperformed robustly for an arbitrary selection, the maximal performance is only achieved for particular values, which cannot be known in advance. In $\mathbb{R}^d$, there exist strategies to mitigate the issue, for example, adaptive restarting, see, e.g., Wei et al. (2023), and regularization, see, e.g., Wei et al. (2021). Their adaptation is left for future work.

Outline: The paper is structured as follows: In Section 2 we recall the Anderson mixing algorithm in Euclidean and Riemannian spaces. Thereafter, we analyze the convergence behavior of Anderson mixing on Bures–Wasserstein space based on Li and Bao (2023), with the proofs deferred to the Appendix. Finally, in Section 5 numerical experiments are shown.

2. Anderson mixing in Euclidean space and for Riemannian manifolds with bounded sectional curvature

Anderson mixing (AM) (Fang and Saad, 2009) reuses previous calculations in order to accelerate fixed-point iterations. Given the problem of finding x such that $x = G(x)$ for some operator $G : \mathbb{R}^d \to \mathbb{R}^d$, at each iteration k, AM uses the $m+1$ most recent iterates $\{x_{k-m+i}\}_{i=0}^m$ and the values of the operator $g_i = G(x_i)$ to choose the next iterate as a linear combination of the history and the current update candidate. The coefficients are determined by finding weights $\{v_i\}_{i=0}^m$ with $\sum_{i=0}^m v_i = 1$ such that $\|\sum_{i=0}^m v_i r_{k-m+i}\|_2$ is minimized. In a linear space $\mathbb{R}^d$, it takes the form of Algorithm 1. Note that the method can be reformulated using residuals $r_k = g_k - x_k$, such that finding the fixed point is equivalent to setting the residual to zero. Introducing a shorthand notation for the forward finite difference

$$\Delta r_k = r_{k+1} - r_k, \ \Delta x_k = x_{k+1} - x_k, \tag{4}$$

Algorithm 1 Anderson mixing in $\mathbb{R}^d$

Parameters: $x_0 \in \mathbb{R}^d$, relaxation parameters $0 \leq \beta_k \leq 1$, and memory parameter $m \geq 1$.
Returns: A sequence $x_0, x_1, \ldots$, intended to converge to a fixed point of $G : \mathbb{R}^d \to \mathbb{R}^d$.

1: **for** $k = 0, 1, \ldots$ until convergence **do**
2: Compute $\tilde{R}_k = (r_{k-m}, \ldots, r_k)$, where $r_i = G(x_i) - x_i$.
3: Solve

$$\alpha^{(k)} = \arg\min_{v \in \mathbb{R}^{m+1}} \{\|\tilde{R}_k v\|_2 \quad \text{s.t.} \sum_{i=0}^{m} v_i = 1\}. \tag{2}$$

4: Set

$$x_{k+1} = (1 - \beta_k) \sum_{i=0}^{m} \alpha_i^{(k)} x_{k-m+i} + \beta_k \sum_{i=0}^{m} \alpha_i^{(k)} G(x_{k-m+i}). \tag{3}$$

5: **end for**

and the matrices

$$X_k = [\Delta x_{k-m}, \Delta x_{k-m+1}, \ldots, \Delta x_{k-1}], \tag{5}$$
$$R_k = [\Delta r_{k-m}, \Delta r_{k-m+1}, \ldots, \Delta r_{k-1}], \tag{6}$$
$$\Gamma_k = \left(\gamma_{k-m}^{(k)}, \ldots, \gamma_{k-1}^{(k)}\right)^T, \tag{7}$$

the Anderson mixing in $\mathbb{R}^d$ can be rewritten as

$$\begin{cases} \Gamma_k = \arg\min_{\Gamma} \|r_k - R_k \Gamma\|_2, \\ \bar{r}_k = r_k - R_k \Gamma_k, \\ x_{k+1} = x_k - X_k \Gamma_k + \beta_k \bar{r}_k. \end{cases} \tag{8}$$

In the Euclidean setting, AM is reported to provide significant numerical advantage in comparison to Picard iteration, see Aksenov et al. (2021); Walker and Ni (2011) and references therein. Observe that the convergence analysis (Evans et al., 2020) in the Euclidean setting yields the estimate

$$\|r_{k+1}\| \leq \frac{\|\bar{r}_k\|}{\|r_k\|}(1 - \beta + \beta\kappa)\|r_k\| + \sum_{i=0}^{m_k} \mathcal{O}(\|r_{k-i}\|^2),$$

where $\frac{\|\bar{r}_k\|}{\|r_k\|} \leq 1$. In the Riemannian setting we cannot expect better results. As the manifold has no linear structure, differences of the residuals are not well defined. In order to generalize AM to Riemannian manifolds, the definitions of Δr_k and Δx_k need to be adapted. One needs to transport tangent vectors from the tangent space of previous iterates to the tangent spaces at the current iterate. Therefore, a suitable vector transport has to be defined on the tangent bundle in order to allow linear combinations. Additionally, the update needs to be approximated by a retraction mapping to stay on the manifold.

Incorporating these ideas, in Li and Bao (2023), the extension of Anderson mixing to Riemannian manifolds (RAM) was introduced. The authors work in the setting where the operator takes the form

$$G(x) := \mathrm{Exp}_x(-F(x)) \tag{9}$$

for some vector field F on $\mathcal{M}$, i.e., $F(x) \in \mathrm{Tan}_x \mathcal{M}$. In this setting, x_* is a fixed point if and only if $F(x_*) = 0$. The method is summarized in Algorithm 2 with the vector transport $\mathcal{T}_y^x : \mathrm{Tan}_y \mathcal{M} \to \mathrm{Tan}_x \mathcal{M}$ and the retraction mapping $\mathcal{R}_x : \mathrm{Tan}_x \mathcal{M} \to \mathcal{M}$.

Algorithm 2 Riemannian Anderson mixing method

Parameters: $x_0 \in \mathcal{M}, \epsilon, \beta_k > 0, m \in \mathbb{N}^*, k = 1.$
$r_0 = -F(x_0)$ and $x_1 = \mathcal{R}_{x_0}(r_0)$.
while $\|F(x_k)\| \geq \epsilon$ **do**
$m_k = \min\{m, k\}$.
$\Delta x_{k-i}^{(k)} = \mathcal{T}_{x_{k-1}}^{x_k} \Delta x_{k-i}^{(k-1)} \in \mathrm{Tan}_{x_k} \mathcal{M}, i = 1, \ldots, m_k.$
$\Delta r_{k-1}^{(k)} = r_k - \mathcal{T}_{x_{k-1}}^{x_k} r_{k-1} \in \mathrm{Tan}_{x_k} \mathcal{M}.$
if $k \geq 2$ **then**
$\Delta r_{k-i}^{(k)} = \mathcal{T}_{x_{k-1}}^{x_k} \Delta r_{k-i}^{(k-1)} \in \mathrm{Tan}_{x_k} \mathcal{M}, i = 2, \ldots, m_k.$
end if
$X_k = [\Delta x_{k-m_k}^{(k)}, \ldots, \Delta x_{k-1}^{(k)}], R_k = [\Delta r_{k-m_k}^{(k)}, \ldots, \Delta r_{k-1}^{(k)}].$
$r_k = -F(x_k)$.
$\Gamma_k = \arg\min_{\Gamma \in \mathbb{R}^{m_k}} \|r_k - R_k \Gamma\|.$
$\bar{r}_k = r_k - R_k \Gamma_k \in \mathrm{Tan}_{x_k} \mathcal{M}, \Delta x_k^{(k)} = -X_k \Gamma_k + \beta_k \bar{r}_k \in \mathrm{Tan}_{x_k} \mathcal{M}.$
$x_{k+1} = \mathcal{R}_{x_k}(\Delta x_k^{(k)})$, $k = k + 1$.
end while

3. The Bures–Wasserstein space of Gaussians

In this section, we recall the geometric structure of the Wasserstein space and of its restriction to centered Gaussian measures, the Bures–Wasserstein manifold $\mathcal{N}_0^d$. We introduce the Riemannian notions required by RAM, namely tangent spaces, the exponential map and parallel transport, and establish local properties of $\mathcal{N}_0^d$ around a nondegenerate covariance matrix that are needed for the convergence analysis.

3.1 Optimal transport and the Wasserstein space

The Wasserstein distance between two probability measures ρ_0, ρ_1 is defined as the solution of the optimal transport problem with cost equal to the underlying space's metric:

$$W_2^2(\rho_0, \rho_1) := \inf_{\gamma \in \Pi(\rho_0, \rho_1)} \int \|x - y\|_2^2 \mathrm{d}\gamma(x, y), \tag{10}$$

where $\Pi(\rho_0, \rho_1)$ is the set of all couplings with first and second marginals ρ_0 and ρ_1, respectively. If both measures have finite second moments, the Wasserstein distance is finite. We thus define the Wasserstein space as the set of all probability measures with finite second moments

$$\mathcal{P}_2(\mathbb{R}^d) := \left\{ \rho \in \mathcal{P}(\mathbb{R}^d) \;\middle|\; \int \|x\|_2^2 \mathrm{d}\rho(x) < \infty \right\} \tag{11}$$

One can show that $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a metric space.

Perhaps the most important feature of $\mathcal{P}_2(\mathbb{R}^d)$ is that for every *absolutely continuous* curve $[0, 1] \to \mathcal{P}_2(\mathbb{R}^d)$ there exists a vector field $v_t \in L^2_{\rho_t}$ such that the curve can be described as a solution of the continuity equation

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0 \tag{12}$$

One can see that for a given ρ_t, the vector field v_t satisfying (12) is not unique. The solution with minimal $L^2_{\rho_t}$-norm is unique and belongs to the space

$$\mathrm{Tan}_\rho \mathcal{P}_2(\mathbb{R}^d) := \overline{\{\nabla \phi : \phi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2_\rho} \tag{13}$$

which is called the *tangent* space to $\mathcal{P}_2(\mathbb{R}^d)$ at ρ. The continuity equation allows one to use a dynamic optimal transport formulation, proven in Benamou and Brenier (2000)

$$W_2^2(\rho_0,\rho_1) = \min_{\rho_t,\ v_t} \int_0^1 \|v_t\|^2_{L^2_{\rho_t}}\,\mathrm{d}t, \quad \text{s.t.} \begin{array}{c} \partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0, \\ \rho_t|_{t=0} = \rho_0, \quad \rho_t|_{t=1} = \rho_1. \end{array}$$

This formula establishes the analogy between the Wasserstein space and a finite-dimensional Riemannian manifold, where the geodesics are defined as the minimizers of the energy functional $\int g_{x_t}(\dot{x}_t,\dot{x}_t)\mathrm{d}t$. In this interpretation, vector fields v_t correspond to tangent vectors and the scalar product of $L^2_{\rho_t}$ plays the role of the Riemannian metric. One has to tread carefully though, since the analogy cannot be made strict and the Wasserstein space cannot be identified with some Hilbert manifold. For example, the tangent spaces to measures supported on finite sets have finite dimensionality, but for an arbitrary ρ the tangent space (13) is infinite-dimensional.

The parallel transport on $\mathcal{P}_2(\mathbb{R}^d)$ has been considered in Ambrosio and Gigli (2008). The authors argue that parallel transport can be defined along curves, along which the topology of the measure's support is preserved. They construct a family of *regular* curves $\rho_t : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$, generated by some Lipschitz vector field v_t, such that there exists a family of mappings τ_s^t with for arbitrary $s,\ t \in [0,1]$

$$\rho_t = (\tau_s^t)_\sharp \rho_s$$

and the mappings satisfy the group property, i.e., $\tau_s^p \circ \tau_t^s = \tau_t^p$ and $\tau_s^t = (\tau_t^s)^{-1}$. Then, an isometric map called *vector translation* between $L^2_{\rho_t}$ and $L^2_{\rho_s}$ is defined as $u \mapsto u\circ\tau_s^t$ and the parallel transport is constructed as a limit of the approximation

$$\mathscr{P}_{t_k}^{t_{k+1}} u := \Pi_{\mathrm{Tan}_{\rho_{t_{k+1}}}\mathcal{P}_2(\mathbb{R}^d)}\left(u\circ\tau_{t_{k+1}}^{t_k}\right), \tag{14}$$

$$\mathcal{P}[t_0,\dots,t_N]u := \mathscr{P}_{t_{N-1}}^{t_N} \circ \mathscr{P}_{t_{N-2}}^{t_{N-1}} \circ \cdots \circ \mathscr{P}_{t_0}^{t_1} u \tag{15}$$

with respect to partitions $0 = t_0 < t_1 < \cdots < t_N = 1$ as $\max_k |t_{k+1} - t_k| \to 0$.

3.2 Bures–Wasserstein distance and geodesic

The set of Gaussian measures forms a closed subset of the space of probability measures with finite second moments $\mathcal{P}_2(\mathbb{R}^d)$ with respect to the Wasserstein metric, see, e.g., Takatsu (2010, 2011); Malagò et al. (2018), and can be given a Riemannian manifold structure. The advantage of this restriction is that many geometric notions such as parallel transport have a more computationally feasible formulation in terms of matrix equations in the Gaussian setting than in the general case.

We define the set of all Gaussians with zero mean as $\mathcal{N}_0^d = \{\mathcal{N}(0,\Sigma), 0 \prec \Sigma \in \mathbb{R}^{d\times d}\}$. The set $\mathcal{N}_0^d$ can be identified with the set $\mathrm{Sym}^{++}(d)$ of symmetric positive-definite matrices (each measure is identified with its covariance matrix Σ). Then the Wasserstein distance between two Gaussian measures can be computed as

$$W_2^2\left(\mathcal{N}(0,\Sigma_0),\mathcal{N}(0,\Sigma_1)\right) = \mathrm{Tr}(\Sigma_0) + \mathrm{Tr}(\Sigma_1) - 2\,\mathrm{Tr}(\left(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\right)^{1/2}). \tag{16}$$

We identify $\mathcal{N}(0,\Sigma)$ with Σ and write $W_2(\Sigma_0,\Sigma_1) = W_2(\mathcal{N}(0,\Sigma_0),\mathcal{N}(0,\Sigma_1))$. Given a linear map with matrix T, a Gaussian measure with covariance Σ_0 can be transformed to a new Gaussian with covariance $\Sigma_1 = T\Sigma_0 T^T$ by $x \mapsto Tx$. Conversely, given two *centered* Gaussians with covariance matrices $\Sigma_{0,1}$ there is an optimal linear map that transforms the distributions into each other. Its matrix T can be found as the unique positive definite solution of the Riccati equation

$$\Sigma_1 = T\Sigma_0 T^T, \qquad T = \Sigma_0^{-1/2}\left(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\right)^{1/2}\Sigma_0^{-1/2}. \tag{17}$$

For two measures with parameters $(0, \Sigma_{0,1})$ consider the convex function

$$\mathcal{W}(x) = \frac{1}{2}\langle x, Tx\rangle \tag{18}$$

where T is the matrix defining the optimal linear map interrelating Σ_0 and Σ_1. One can see that $(\mathrm{id}, \nabla\mathcal{W})_\sharp\mathcal{N}(0,\Sigma_0)$ is the optimal transport coupling between $\mathcal{N}(0,\Sigma_0)$ and $\mathcal{N}(0,\Sigma_1)$. Furthermore, for the curve $\Sigma(t) : [0,\ 1] \to \mathcal{N}_0^d$ defined by

$$\Sigma(t) := ((1-t)\,\mathrm{Id} + tT)\,\Sigma_0\,((1-t)\,\mathrm{Id} + tT)\,, \tag{19}$$

one can show that it is a geodesic in the metric sense, i.e.,

$$W_2(\Sigma(t), \Sigma(s)) = |t-s| W_2(\Sigma_0, \Sigma_1), \quad \forall t, s \in [0,1] \tag{20}$$

It was shown in Takatsu (2011); Malagò et al. (2018); Bhatia et al. (2019) that the structure of a Riemannian manifold can be imposed on $\mathcal{N}_0^d$, with metric that induces the Wasserstein distance. We summarize these results in Section 3.3.

Remark 2 *For non-centered Gaussians with means $m_0,\ m_1$, the squared distance is augmented by adding $\|m_0 - m_1\|^2$. Then the linear map T becomes affine and $\mathcal{W}(x) = \frac{1}{2}\langle x - m_0, T(x-m_0)\rangle + \langle x, m_1\rangle$. The geodesic is augmented by $m_t = (1-t)m_0 + tm_1$.*

3.3 Riemannian structure on $\mathcal{N}_0^d$

We briefly summarize the results of Takatsu (2011), see Malagò et al. (2018); Bhatia et al. (2019) for additional details. Consider the manifold $\mathcal{G}(d)$ of all invertible matrices in $\mathbb{R}^{d\times d}$, equipped with the Frobenius metric $\langle Z, W\rangle_F := \mathrm{Tr}(Z^T W)$. The mapping $\Pi : A \mapsto AA^T$ is a surjection from $\mathcal{G}(d)$ to $\mathcal{N}_0^d$ whose differential is given by $\mathrm{d}\Pi_A(Z) = ZA^T + AZ^T$. For each A, $\mathrm{Tan}_A(\mathcal{G}(d))$ splits into $\mathrm{Ker}\,\mathrm{d}\Pi_A$ and its orthogonal complement, called *vertical* and *horizontal* subspaces. The horizontal subspace consists exactly of those $Z \in \mathbb{R}^{d\times d}$ such that ZA^{-1} is symmetric. Let us denote by $Z_\mathcal{H}$ the orthogonal projection of Z onto the horizontal subspace $\mathcal{H}_A$. If the metric g_{BW} on $\mathcal{N}_0^d$ is defined by

$$g_{\mathrm{BW}}(\mathrm{d}\Pi_A(Z), \mathrm{d}\Pi_A(W)) = \langle Z_\mathcal{H}, W_\mathcal{H}\rangle_F$$

then, as shown in Takatsu (2011), Π is a Riemannian submersion. We identify $\mathrm{Tan}_\Sigma\mathcal{N}_0^d$ with the horizontal subspace $\mathcal{H}_A$ for some $A : AA^T = \Sigma$, and $\mathcal{H}_A$ is identified with $\mathrm{Sym}(d)$ via

$$U^T = U = Z_\mathcal{H} A^{-1}. \tag{21}$$

Thus, $\mathrm{Tan}_\Sigma\mathcal{N}_0^d \simeq \mathrm{Sym}(d)$, and

$$\langle Z_\mathcal{H}, W_\mathcal{H}\rangle_F = \mathrm{Tr}(Z_\mathcal{H}^T W_\mathcal{H}) = \mathrm{Tr}\left((UA)^T VA\right) = \mathrm{Tr}(A^T UVA) = \mathrm{Tr}(VAA^T U) = \mathrm{Tr}(V\Sigma U).$$

This is indeed a symmetric bilinear form that does not depend on the particular choice of A. We thus denote the scalar product on the tangent space as

$$\langle U, V\rangle_\Sigma := \mathrm{Tr}(U\Sigma V), \quad \|V\|_\Sigma^2 = \mathrm{Tr}(V\Sigma V). \tag{22}$$

The form of the optimal transport map (18) suggests that the tangent vectors in $\mathrm{Tan}_\Sigma\mathcal{N}_0^d$ can be isometrically identified with tangent vector fields in $\mathrm{Tan}_{\mathcal{N}(0,\Sigma)}\mathcal{P}_2(\mathbb{R}^d)$ that are linear maps:

$$v(x) = Vx \in L^2_{\mathcal{N}(0,\Sigma)}(\mathbb{R}^d) \quad \mapsto \quad V \in \mathrm{Sym}(d)$$

and, writing $\rho_\Sigma := \mathcal{N}(0,\Sigma)$, it holds that

$$\langle u, v\rangle_{L^2_{\rho_\Sigma}} = \int \langle Ux, Vx\rangle \mathrm{d}\rho_\Sigma(x) = \langle U, V\rangle_\Sigma$$

The exponential map takes the form

$$\mathrm{Exp}_\Sigma(V) := (\mathrm{Id}+V)\Sigma(\mathrm{Id}+V). \tag{23}$$

This definition corresponds to the definition of the exponential from Ambrosio and Gigli (2008)

$$\mathrm{Exp}_\rho(v) := (\mathrm{id}+v)_\sharp\rho \tag{24}$$

where id is the identity map on $\mathbb{R}^d$. Due to the pushforward formula, the covariance Σ' of the exponential of $\mathcal{N}(0,\Sigma)$ can be computed as

$$\int x'(x')^T \mathrm{d}\rho_{\mathrm{Exp}_\Sigma(V)}(x') = \int (x+Vx)(x+Vx)^T \mathrm{d}\rho_\Sigma(x) = (\mathrm{Id}+V)\left[\int xx^T \mathrm{d}\rho_\Sigma(x)\right](\mathrm{Id}+V) = (\mathrm{Id}+V)\Sigma(\mathrm{Id}+V)$$

Since $(\mathrm{id}, \mathrm{id}+v)_\sharp\mathcal{N}(0,\Sigma)$ is a valid coupling between $\mathcal{N}(0,\Sigma)$ and $\mathcal{N}(0,\mathrm{Exp}_\Sigma(V))$, we can see that

$$W_2(\Sigma, \mathrm{Exp}_\Sigma(V)) \le \|V\|_\Sigma \tag{25}$$

The following statement allows one to explicitly compute the tangent vector field $V(t)$ for a given curve $\Sigma(t)$ on the Bures–Wasserstein space.

Proposition 3 *For $\Sigma \in \mathcal{N}_0^d$, consider an operator P_Σ*

$$P_\Sigma : (\mathrm{Sym}(d),\ \|\cdot\|_F) \to (\mathrm{Sym}(d),\ \|\cdot\|_F)$$
$$P_\Sigma[V] := \Sigma V + V\Sigma$$

P_Σ is symmetric and invertible. Denote $L_\Sigma := P_\Sigma^{-1}$.

Let $\Sigma(t) :\ [0,1] \to \mathcal{N}_0^d$ be a smooth curve, and $V(t) \in \mathrm{Tan}_{\Sigma(t)}\mathcal{N}_0^d$ its tangent vector. Then it holds

$$\dot\Sigma(t) = P_{\Sigma(t)}[V(t)],$$
$$V(t) = L_{\Sigma(t)}[\dot\Sigma(t)],$$

where $\dot\Sigma(t)$ stands for the time derivative of $\Sigma(t)$. The vector field $v_t(x) := V(t)x$ is the unique tangent vector field to the curve $\rho_t := \mathcal{N}(0,\Sigma(t))$ in the Wasserstein space. In particular, the continuity equation is satisfied

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

Remark 4 *In Malagò et al. (2018), the representation of the tangent vectors in* $\mathrm{Sym}(d)$ *is related to the representation* (21)*, used in the present paper as well as in Takatsu (2010), by means of the operator L_Σ. For example, the exponential map in this representation is given by* $(\mathrm{Id}+L_\Sigma[Z])\Sigma(\mathrm{Id}+L_\Sigma[Z])$ *for some $Z \in \mathrm{Sym}(d)$ instead of* (23)*. We resort to Takatsu's notation since it simplifies the presentation. Moreover, it allows one to intuitively identify the Wasserstein gradients as $L^2_{\mathcal{N}(0,\Sigma)}$ vector fields (formulae for which are known from the infinite-dimensional theory) with the corresponding tangent vectors.*

As for the parallel transport, the geodesics on $\mathcal{N}_0^d$ are regular curves in the sense of Ambrosio and Gigli (2008). As shown in Malagò et al. (2018), the notion of parallel transport in the sense of Ambrosio and Gigli (2008) coincides with standard Riemannian parallel transport (assuming the initial vector field is linear in the spatial coordinate, thus being able to be identified with a tangent vector to $\mathcal{N}_0^d$).

3.4 Properties in the neighborhood of a nondegenerate distribution

In order to apply the local convergence theory of RAM from Li and Bao (2023), one has to know certain properties of the manifold in the neighborhood of the fixed point. In that regard, we have proven several novel statements that allow us to relate the Bures–Wasserstein distance between the distributions and norms of tangent vectors to Frobenius norms of the symmetric matrices that represent them. This significantly simplifies the subsequent convergence analysis. We summarize our findings in the following theorem:

Theorem 5 *Let Σ_* be a nondegenerate covariance matrix, and*

$$0 < \lambda_d^* \leq \cdots \leq \lambda_1^*$$

be the eigenvalues of Σ_*. Define the Bures–Wasserstein ball as

$$B_r(\Sigma_*) := \{\Sigma \in \mathcal{N}_0^d : W_2(\Sigma, \Sigma_*) \leq r\} \tag{26}$$

and consider $r < \frac{1}{2}\sqrt{\lambda_d^*}$.

Then, for $\Sigma, \Sigma_1, \Sigma_2 \in B_r(\Sigma_*)$:

- *$B_r(\Sigma_*)$ is compact in $\mathcal{N}_0^d$.*
- *Its sectional curvature is uniformly bounded.*
- *The following estimates hold:*

$$(\sqrt{\lambda_d^*} - r)\|V\|_F \leq \|V\|_\Sigma \leq (\sqrt{\lambda_1^*} + r)\|V\|_F \tag{27}$$

$$\|V\|_{\Sigma_2} \leq \frac{\sqrt{\lambda_1^*} + r}{\sqrt{\lambda_d^*} - r}\|V\|_{\Sigma_1} \tag{28}$$

$$\|\Sigma_1 - \Sigma_2\|_F \leq 2(\sqrt{\lambda_1^*} + \sqrt{\lambda_d^*})W_2(\Sigma_1,\ \Sigma_2) \tag{29}$$

$$2(\sqrt{\lambda_d^*} - r)W_2(\Sigma_1,\ \Sigma_2) \leq \|\Sigma_1 - \Sigma_2\|_F \tag{30}$$

- *For every $V \in \mathrm{Tan}_\Sigma \mathcal{N}_0^d$ such that $\|V\|_\Sigma \leq \frac{1}{2}r$, it holds that*

$$W_2(\Sigma, \mathrm{Exp}_\Sigma(V)) = \|V\|_\Sigma \quad \forall \Sigma \in B_r(\Sigma_*) \tag{31}$$

The proof is split into multiple lemmata, which can be found in Section A.1. The overall idea is to show that a perturbation of the covariance matrix, which is small in terms of the Bures–Wasserstein distance, leads to a small perturbation of the eigenvalues. A uniform bound on the eigenvalues leads to bounded sectional curvature due to the explicit formulas derived in Takatsu (2010). Additionally, it allows us to compare the tangent norm to the Frobenius norm by (27). As the Frobenius norm does not depend on the current distribution, the dynamic optimal transport can be used to estimate the length of the Bures–Wasserstein geodesic by a Euclidean geodesic with the same endpoints, leading to (29).

Remark 6 *The property* (31) *establishes the injectivity of the exponential map in a neighborhood of zero in the tangent space. Together with the sectional curvature bound, these properties are a cornerstone of the convergence analysis of RAM. These properties do not hold for the whole Wasserstein space, rendering the convergence analysis for general measures difficult.*

4. Riemannian Anderson mixing on Bures–Wasserstein space

We now apply Riemannian Anderson mixing (Algorithm 2) to fixed-point problems on $\mathcal{N}_0^d$, which we refer to as BWRAM. We first state its local convergence result, then construct computationally tractable vector transport mappings, give examples of contractive operators, and discuss the computational complexity of the method.

4.1 Convergence

In order to show the convergence of Anderson mixing in Bures–Wasserstein space we follow the lines of Li and Bao (2023). As we have explicit formulas for the parallel transport $\mathcal{P}$ as well as the exponential map, some arguments of Li and Bao (2023) can be simplified. Hence, we repeat the proof in the appendix. As in the Euclidean case in Toth and Kelley (2015); Wei et al. (2022), we impose the "coercive" and "locally Lipschitz continuous" properties on the vector field F as well as on its Jacobian H. Define the fixed-point map $G(\Sigma) := \mathrm{Exp}_\Sigma(-F(\Sigma))$.

Assumption 4.1 *Assume $\Sigma_* \in \mathcal{N}_0^d$ with $\Sigma_* = G(\Sigma_*)$ and let $\lambda_i^* > 0$ be its eigenvalues in decreasing order. Assume there exists a ball $B_r(\Sigma_*)$ with $r < \frac{1}{2}\sqrt{\lambda_d^*}$ and constants $0 < L_1 < L_2$ and $L_H > 0$ such that for all $\Sigma_1,\ \Sigma_2 \in B_r(\Sigma_*)$ the following hold:*

$$L_1 W_2(\Sigma_1, \Sigma_2) \le \|F(\Sigma_1) - \mathcal{P}_{\Sigma_2}^{\Sigma_1} F(\Sigma_2)\|_{\Sigma_1} \le L_2 W_2(\Sigma_1, \Sigma_2), \tag{32}$$

$$\left\| \mathcal{P}_{\Sigma_2}^{\Sigma_1} H(\Sigma_2) \mathcal{P}_{\Sigma_1}^{\Sigma_2} - H(\Sigma_1) \right\|_{\Sigma_1} \le L_H W_2(\Sigma_1, \Sigma_2), \tag{33}$$

$$W_2(G(\Sigma_1), G(\Sigma_2)) \le \kappa W_2(\Sigma_1, \Sigma_2) \text{ for some } \kappa < 1. \tag{34}$$

Assumption 4.2 *The vector transport $\mathcal{T}$ is continuously differentiable and there exists some $M > 0$ such that*

$$\left\| \mathcal{T}_{\Sigma_1}^{\Sigma_2} V \right\|_{\Sigma_2} \le M \|V\|_{\Sigma_1}$$

for all $\Sigma_1, \Sigma_2 \in B_r(\Sigma_*),\ \forall V \in \mathrm{Tan}_{\Sigma_1} \mathcal{N}_0^d$.

Observe that for the two vector transports introduced in Section 4.2, we have $M \le 1$ and $M \le \frac{\sqrt{\lambda_1^*}+r}{\sqrt{\lambda_d^*}-r}$, respectively.

The final assumption concerns the uniform boundedness of the extrapolation coefficients Γ_k in the RAM method. This assumption is common for analyzing convergence in the Euclidean setting, see Toth and Kelley (2015).

Assumption 4.3 *There exists a positive constant M_Γ such that $\|\Gamma_k\|_\infty \le M_\Gamma$ for all $k \in \mathbb{N}$.*

Following Li and Bao (2023), we can formulate the following theorem.

Theorem 7 *Assume $\{\Sigma_n\}_{n\ge1}$ is generated by Algorithm 2 and suppose Assumptions 4.1–4.3 hold. Let $\beta := \inf_k \beta_k$. Then there exist constants $M_2, M_3 > 0$ and $c_0 > 0$ such that with*

$$M_1 = M_\Gamma \frac{1}{L_1} \max\{M^{m_k} + M^{m_k - 1}, M^2 + M\} + 1$$

if the initial point $\Sigma_0 \in \mathcal{N}_0^d$ satisfies

$$W_2(\Sigma_0, \Sigma_*) \le \frac{1}{1 + L_2} \min\left\{ \frac{r}{L_2(mM_1 + \beta + 1)}, \frac{L_1 - L_2 + (1-\kappa)\beta L_2}{M m L_2^2} \right\}$$

then for all $k \ge 1$ we have $W_2(\Sigma_k, \Sigma_*) \le \min\left\{ \frac{r}{L_2(mM_1+\beta+1)}, \frac{L_1 - L_2 + (1-\kappa)\beta L_2}{M m L_2^2} \right\}$ and it holds that

$$\|r_{k+1}\| \le \theta_k \left(1 - \beta_k + \kappa\beta_k\right) \|r_k\| + \left(c_0 \tilde{r} m M_1^2 + M_3 + 2 M_\Gamma M_2\right) \sum_{i=0}^{m_k} \|r_{k-i}\|^2 \tag{35}$$

where $\theta_k := \frac{\|\bar{r}_k\|}{\|r_k\|} \le 1$.

Remark 8 *The special setup of the Bures–Wasserstein space allows us to simplify certain steps in the proof of Li and Bao (2023). For example, we can numerically evaluate the exponential mapping exactly, up to machine precision. This leads to improved constants with respect to* $\sum_{i=1}^{m_k} \mathcal{O}(\|r_{k-i}\|^2)$. *In our setting the hidden constants take the form*

$$\underbrace{c_0 \tilde{r} m M_1^2}_{C_1} + \underbrace{M_3 + 2M_\Gamma M_2}_{C_2}.$$

In contrast, in Li and Bao (2023) the constants read

$$\underbrace{\left((\kappa+1)(C+\sqrt{K}) + \frac{c_0\tilde{r}}{2} + \sqrt{K}\max\{1, L_2\}\right) m M_1^2}_{C_1} + \underbrace{m\left(M_3 + 2M_\Gamma M_3\right)}_{C_2}$$

with

$$M_1 = \max\{\frac{M M_\Gamma}{L_1 C} + 1, M_\Gamma \tilde{M}, \tilde{M}\}$$

where

$$\tilde{M} = \frac{1}{L_1 C}\max\{M^{m_k} + M^{m_k - 1}, M^2 + M\}.$$

Here, $C, K > 0$ *are constants arising from the retraction map and inverse exponential map. Therefore, one can observe an improvement of* C_1 *in terms of* C *and* K *as well as of* C_2 *by a factor* m, *i.e. the length of the history. We would like to point out that the overall convergence behavior is the same as for RAM and, in particular, in the Euclidean case. This is a strong indicator that we cannot expect any improvement beyond the constants in* $\mathcal{O}(\|r_i\|^2)$.

Remark 9 *One can observe that the initial value needs to be in a significantly smaller neighborhood than imposed by the smallest eigenvalue of the fixed point. This is due to the fact that recording and reusing the history might only improve the algorithm when already close to the fixed point, as the first iterates might not contain sufficient information. The bound for the initial value ensures that the algorithm does not leave a neighborhood of the fixed point allowing for an inductive argument. The numerical experiments indicate that this is a very conservative bound.*

The proof of Theorem 7 can be found in Section A.3 and follows Li and Bao (2023) very closely while improving the constants and simplifying some steps for this particular case.

Since the mapping is assumed to be contractive, the decrease in residual can be directly linked to the distance to the solution.

$$W_2(\Sigma_k, \Sigma_*) \leq W_2(G(\Sigma_k), \Sigma_*) + W_2(G(\Sigma_k), \Sigma_k) \leq \kappa W_2(\Sigma_k, \Sigma_*) + W_2(G(\Sigma_k), \Sigma_k),$$
$$W_2(\Sigma_k, \Sigma_*) \leq \frac{1}{1-\kappa}\|r_k\|$$

where, in the second inequality, we used that $W_2(\Sigma_k, G(\Sigma_k)) = W_2(\Sigma_k, \mathrm{Exp}_{\Sigma_k}(r_k)) \leq \|r_k\|_{\Sigma_k}$ due to (25).

The estimate (35) can be interpreted as follows. The method decreases the residual at least as effectively as a relaxed Picard iteration (the residual decrease in that case would be exactly $(1 - \beta_k + \kappa\beta_k)\,\|r_k\|$), up to "higher-order terms" $\mathcal{O}(\sum_i \|r_{k-i}\|^2)$. When the iterate is sufficiently close to the solution, the higher-order terms can be neglected. The coefficient $\theta_k = \frac{\|\bar{r}_k\|}{\|r_k\|}$ quantifies the gain, compared to the Picard step, and depends on the quality of the solution of the subproblem. The estimate suggests keeping the history length of the method bounded and relatively short. This intuition is indeed backed by our practical experience, where a significant acceleration can be achieved for quite modest history lengths $m \in \{1, \ldots, 5\}$. As of now, estimates of such type (i.e., including $\mathcal{O}$ terms) are the state of the art for Anderson mixing on nonlinear problems, even in Euclidean space (see Sun et al. (2021); Liu et al. (2024); Wei et al. (2023); Evans et al. (2020)).

4.2 Construction of vector transport mapping

In Algorithm 2, a vector transport mapping is employed in order to aggregate the history vectors to a common tangent space. Intuitively, the vector transport mapping can be viewed as an approximation of the parallel transport (15). The exact parallel transport naturally satisfies Assumption 4.2 since it is an isometry. The ODE for the parallel transport was derived in Malagò et al. (2018), and can be solved numerically with a suitable adaptive ODE solver. However, integrating an ordinary differential equation on the manifold can be more numerically challenging than the original fixed-point problem itself, making parallel transport unsuitable as a component of the algorithm.

For the convergence of RAM, as stated in Assumption 4.2, the mapping only has to satisfy a uniform norm bound with some constant M. Moreover, M affects only the radius of local convergence, but not the convergence estimate (1). Thus, a rough approximation of parallel transport with such bound is as good from the theoretical side as the exact mapping. From the standpoint of computational complexity, as well as practical applicability of the method, however, it is important to explore possible, more computationally tractable options.

Firstly, we consider an approximation inspired by the discrete scheme (15) using a *single* time step for the approximation. Both the transported vector field $v(x) = Vx,\ V \in \mathrm{Tan}_{\Sigma_0}\mathcal{N}_0^d$ and the optimal transport map $\tau(x) = Tx$, with T given by (17), are linear mappings. Moreover, T is invertible. Thus the vector translation $v \circ \tau^{-1}$ is also a linear mapping, with its matrix given by VT^{-1}. The formula for one time step (14) simplifies to

$$\mathcal{T}_{\Sigma_0}^{\Sigma_1} V := \Pi_{\Sigma_1}\left(VT^{-1}\right), \tag{36}$$

where $\Pi_{\Sigma_1} : \mathbb{R}^{d\times d} \to \mathrm{Tan}_{\Sigma_1}\mathcal{N}_0^d$ is the orthogonal projection onto the tangent space.

Proposition 10 *If the scalar product* (22) *is extended to the whole* $\mathbb{R}^{d\times d}$ *as* $\langle U, V\rangle_\Sigma := \mathrm{Tr}(U^T\Sigma V)$, *the orthogonal projection of* $V \in \mathbb{R}^{d\times d}$ *onto* $\mathrm{Tan}_\Sigma\mathcal{N}_0^d$

$$\Pi_\Sigma(V) = \arg\min_X \|X - V\|_\Sigma^2 \quad s.t.\ X = X^T, \tag{37}$$

can be found as the unique symmetric solution X *of the equation*

$$\Sigma X + X\Sigma = \Sigma V + V^T\Sigma. \tag{38}$$

Proof Writing out the Lagrangian and setting its derivatives to zero, one gets

$$\begin{aligned} &L(X,\lambda) = \mathrm{Tr}((X-V)^T\Sigma(X-V)) + \langle\lambda, X - X^T\rangle_F,\\ &\frac{\partial L}{\partial X} = 0 \Rightarrow 2\Sigma X - 2\Sigma V = \lambda^T - \lambda,\\ &\frac{\partial L}{\partial \lambda} = 0 \Rightarrow X = X^T. \end{aligned}$$

Averaging the first equation and its transpose, with $X = X^T$, gives

$$\Sigma X + X\Sigma = \Sigma V + V^T\Sigma.$$

The unique solution exists if no two eigenvalues of Σ add up to zero:

$$\lambda_i + \lambda_j > 0 \quad \forall(i,\ j),$$

which always holds in our case since all the eigenvalues are positive. See also Gajic and Qureshi (2008) for an overview of the numerical approaches. ■

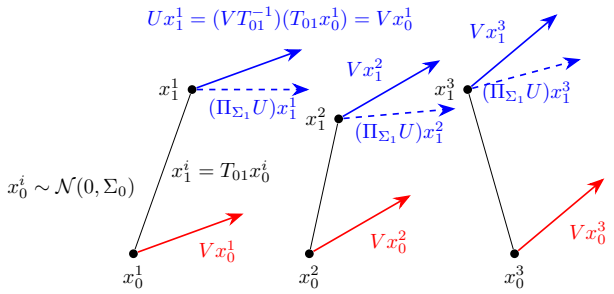


Figure 1: One-step approximation of the parallel transport. Linear OT map T pushes Σ_0 forward to Σ_1. One wishes to transport the vector field $v : x_0 \mapsto Vx_0,\ V \in \mathrm{Sym}(d)$ along this geodesic. The individual particles move in a straight line: $x_1 = Tx_0$. The vector translation of v is given by $u : x_1 \mapsto Ux_1,\ U = VT^{-1}$. $U \in \mathrm{Tan}_{\Sigma_1}\mathcal{N}_0^d$ if and only if V and T^{-1} commute, thus, orthogonal projection Π_{Σ_1} is needed.

The bound $\|\mathcal{T}_{\Sigma_0}^{\Sigma_1}V\|_{\Sigma_1} \leq \|V\|_{\Sigma_0}$ for Assumption 4.2 holds with $M = 1$ because vector translation is an isometry and because of the property of the orthogonal projection. The approximation is illustrated in Figure 1.

Secondly, in a local neighborhood of a nondegenerate covariance matrix Σ_*, the identity mapping of $\mathrm{Sym}(d)$ can be viewed as a linear operator between tangent spaces $\mathrm{Id}_{\Sigma_1}^{\Sigma_2} : \mathrm{Tan}_{\Sigma_1}\mathcal{N}_0^d \to \mathrm{Tan}_{\Sigma_2}\mathcal{N}_0^d$. Due to (28), it also satisfies the bound

$$\left\|\mathrm{Id}_{\Sigma_1}^{\Sigma_2} V\right\|_{\Sigma_2} \leq \frac{\sqrt{\lambda_1^*} + r}{\sqrt{\lambda_d^*} - r}\|V\|_{\Sigma_1} \ \forall V \in \mathrm{Tan}_{\Sigma_1}\mathcal{N}_0^d$$

which is uniform on the ball $B_r(\Sigma_*)$ with $r < \sqrt{\lambda_d^*}$. This is a perfect candidate from the computational perspective (no computation required), but the constant may affect the radius of the method's convergence. We provide the comparison of the vector transport mappings $\mathcal{T}_{\Sigma_1}^{\Sigma_2}$ and $\mathrm{Id}_{\Sigma_1}^{\Sigma_2}$ in Section 5.

Remark 11 *In this paper, we provide definitions, study the properties and derive numerical algorithms for Riemannian notions such as exponential map, vector transport, projection onto the tangent space, etc. These can be used to adapt a broad range of Riemannian numerical methods, such as Riemannian gradient descent (RGD), Riemannian conjugate gradient (RCG), Riemannian L-BFGS (RLBFGS), etc., to probabilistic tasks that can be reformulated as fixed-point or optimization problems. Although in the convergence analysis we have focused on RAM, in the numerical experiments, we also compare its performance to RGD and RCG. To the best of our knowledge, this is the first time these methods are applied in the setting considered.*

4.3 Examples of operators

In the following we describe several examples of fixed-point problems, suitable for acceleration with BWRAM. We reflect on their contractive properties and discuss the possible applications.

4.3.1 Ornstein–Uhlenbeck process

The evolution of the Ornstein–Uhlenbeck process is a suitable model problem, since it is known that the induced fixed-point operator is contractive. Recall that the operator is defined by the stochastic differential equation

$$\mathrm{d}X_t = -\Sigma_*^{-1}X_t\mathrm{d}t + \sqrt{2}\mathrm{d}W_t \tag{39}$$

as follows: if $X_0 \sim \mathcal{N}(0, \Sigma_0)$, then for some fixed time step τ, $G(\Sigma_0)$ is the law of X_τ. The process can be identified with the gradient flow of the Kullback–Leibler divergence $\mathrm{KL}(\cdot|\mathcal{N}(0, \Sigma_*))$ w.r.t. the 2-Wasserstein distance (Ambrosio et al., 2005; Wibisono, 2018).

In the case of Gaussians, there is an explicit solution (Wibisono, 2018, Example 1)

$$X_\tau = e^{-\tau\Sigma_*^{-1}} X_0 + \Sigma_*^{1/2} \left(\mathrm{Id} - e^{-2\tau\Sigma_*^{-1}}\right)^{1/2} Z,$$

where $Z \sim \mathcal{N}(0, \mathrm{Id})$ is independent of X_0. Thus, in the case of $X_0 \sim \mathcal{N}(0, \Sigma)$, the covariance after one step is given by

$$G(\Sigma) = e^{-\tau\Sigma_*^{-1}} \Sigma e^{-\tau\Sigma_*^{-1}} + \left(\mathrm{Id} - e^{-2\tau\Sigma_*^{-1}}\right)^{1/2} \Sigma_* \left(\mathrm{Id} - e^{-2\tau\Sigma_*^{-1}}\right)^{1/2}. \tag{40}$$

Since Σ_* is symmetric positive definite, the matrices Id, Σ_*, $\Sigma_*^{-1}, \Sigma_*^{1/2}$ and $e^{-2\tau\Sigma_*^{-1}}$ share the same basis of eigenvectors and, thus, commute. It is easy to simplify (40) and verify that Σ_* is a fixed point.

4.3.2 Minimization of functionals

Alternatively, given some vector field F, the operator can be defined as $G(\Sigma) = \mathrm{Exp}_\Sigma(-hF(\Sigma))$. If F is a Wasserstein gradient of some functional,

$$F(\rho) = \partial_W \mathcal{E}[\rho] := \nabla \frac{\delta}{\delta\rho} \mathcal{E}(\rho)$$

one can opt to minimize this functional with fixed-point iteration for the following operator

$$G(\rho) := \mathrm{Exp}_\rho \left(-h\partial_W \mathcal{E}[\rho]\right) \tag{41}$$

This fixed-point iteration can be seen as a generalization of the gradient descent method to spaces with Riemannian-like geometry. Minimization over the BW space finds applications in Bayesian logistic regression (Katsevich and Rigollet, 2024; Diao et al., 2023), distributionally robust optimization (Nguyen et al., 2023), etc.

4.3.3 Kullback–Leibler divergence

As a model minimization problem, we consider the Kullback–Leibler divergence:

$$\partial_W \left(\mathrm{KL}(\mathcal{N}(0, \Sigma)|\, \mathcal{N}(0, \Sigma_*))\right)(x) = \nabla_x \left(\log \rho_\Sigma(x) - \log \rho_{\Sigma_*}(x)\right) = (\Sigma_*^{-1} - \Sigma^{-1})x. \tag{42}$$

The operator can be written as

$$G_h(\Sigma) = \mathrm{Exp}_\Sigma \left(-h\partial_W \mathrm{KL}(\Sigma|\Sigma_*)\right) = (\mathrm{Id} - h(\Sigma_*^{-1} - \Sigma^{-1}))\Sigma(\mathrm{Id} - h(\Sigma_*^{-1} - \Sigma^{-1})), \tag{43}$$

where $h > 0$ is a step size parameter that can be selected so that the iterative procedure converges.

Theorem 12 *For a nondegenerate Σ_*, there exist a radius r and step size parameters τ, h such that the operators defined by* (40) *and* (43) *are contractive on $B_r(\Sigma_*)$.*

Proof According to (Ambrosio et al., 2005, Theorem 11.2.1), the functional $\mathrm{KL}(\cdot|\mathcal{N}(0, \Sigma_*))$ is $1/\lambda_1^*$-convex along generalized geodesics. Thus, the following contraction estimate for the Ornstein–Uhlenbeck process (40) can be given:

$$W_2(G(\Sigma_0), G(\Sigma_1)) \le e^{-\frac{\tau}{\lambda_1^*}} W_2(\Sigma_0, \Sigma_1). \tag{44}$$

The operator (43) can be seen as a one-step explicit time discretization of the gradient flow (39). If $\Sigma = \Sigma_*$, then $\partial_W \mathrm{KL}(\mathcal{N}(0,\Sigma)|\mathcal{N}(0,\Sigma_*)) = 0$ and Σ is a fixed point. Thus, it is natural to expect that, for a sufficiently small step size h, the discretization should capture the contractive property of the continuous process. A rigorous proof relies on establishing a property, analogous to the L-Lipschitz gradient property of a functional in $\mathbb{R}^d$, in the neighborhood of the fixed point. It is quite technical and is therefore postponed to Section A.2. ■

4.4 Complexity analysis

For fixed-point problems in $\mathbb{R}^d$, the evaluation of $G(x_k)$ typically is the most computationally expensive operation. For example, if Anderson mixing is used in the context of coupled PDE systems, each evaluation of the operator requires a numerical solution of complicated PDE(s), with parameters defined by the current iterate x_k (see Fang and Saad (2009); Aksenov et al. (2021); Jaejin and Han (2022) for more examples). The efficiency of the method is evaluated by comparing its number of operator calls with that of the Picard method. The overhead induced by the acceleration algorithm can be neglected in this case. Indeed, the additional operations consist of finding the mixing coefficients via (2) and then computing the new iterate by formula (3). The storage of historical vectors requires $\mathcal{O}(md)$ additional memory. The assembly of the matrices needed to solve (2) can be estimated as $\mathcal{O}(m^2 d)$ operations. Additional $\mathcal{O}(m^3)$ operations are required to solve the minimization problem. Since m is a small number in the range from 1 to 15, and the complexity of the operator evaluation is expected to scale worse than linearly, neglecting the overhead is justified from the computational complexity perspective, especially for large-scale problems ($d \gg m$).

If the Gaussian distributions in $\mathcal{N}_0^d$ are considered, the total number of the degrees of freedom, needed to represent the points and tangent vectors, is $\mathcal{O}(d^2)$. Thus, the storage of the matrices $X_k,\ R_k$ requires $\mathcal{O}(md^2)$ memory. In addition to the aforementioned steps, Riemannian Anderson mixing performs the vector transport. In the case when the identity map $\mathrm{Id}_{\Sigma_1}^{\Sigma_2}$ is used, there is no computation at all. For the one-step approximation $\mathcal{T}_{\Sigma_1}^{\Sigma_2}$, a computation of the OT map, and m matrix multiplications and solutions of (38) are needed, thus the total number of operations can be estimated as $\mathcal{O}(md^3)$. The computation of the matrices in (45) requires $\mathcal{O}(m^2 d^3)$, and the update using Riemannian exponential (23) costs $\mathcal{O}(d^3)$.

For example, in the averaging problems that we consider in the sequel, each operator evaluation requires computing the optimal transport map between the current distribution and n_σ other ones. This can be efficiently implemented in $\mathcal{O}(n_\sigma d^3)$ operations using the eigenvalue decomposition. In applications, n_σ can be as high as 10^3 (Chewi et al., 2020). Thus, in our examples the operator computation complexity has the same asymptotic dependence on d, as the overhead introduced by RAM. Nevertheless, we argue that it is possible to achieve overall acceleration when the operator is sufficiently "heavy", e.g., when n_σ is high.

5. Numerical experiments

The following section provides the implementation details and numerical results. The experiments were performed on a laptop with 13th Gen Intel(R) Core(TM) i7-1355U CPU with 12 cores. The execution time varied from fractions of a second in low-dimensional cases up to $\sim 100\,\mathrm{s}$ for 500 dimensions.

5.1 Implementation details

Since the explicit form of the exponential map is given in (23), we can directly use it as the retraction mapping.

To adhere to Assumption 4.3, in the minimization subproblem we impose a bound $l_{\infty,\max}$ for the weights Γ_k, and solve (using `cvxpy` package (Diamond and Boyd, 2016)) the constrained problem

$$\Gamma_k = \underset{\|\Gamma\|_\infty \le l_{\infty,\max}}{\arg\min} \|r_k - R_k\Gamma\|. \tag{45}$$

5.2 Ornstein–Uhlenbeck process

In the first set of experiments, we accelerate the Picard iteration for the operator (40), defined by the Ornstein–Uhlenbeck process (39). We vary d, $\sigma_{\max}$ and for each pair, we run the Picard iteration and BWRAM with different history lengths m, until the residual norm is smaller than $\epsilon = 10^{-6}$. The acceleration measured as $\frac{N_{\text{Picard}}}{N_{\text{BWRAM}}}$ is averaged over $n = 6$ independent covariances. The maximal mean acceleration, depending on d and $\sigma_{\max}$, is presented in Table 1a. The improved convergence of BWRAM for the Ornstein–Uhlenbeck problem, in comparison to Picard, RGD and RCG methods, is presented in Figure 2a.

Table 1: Maximal mean acceleration for BWRAM, depending on d and $\sigma_{\max}$.

$\sigma_{\max}$ / d	1.0	5.0	10.0	20.0
4	1.62	5.89	8.29	14.16
8	1.73	3.68	4.23	5.76
16	1.73	4.00	4.76	5.94
32	1.73	4.12	4.86	5.41
64	1.75	4.33	5.60	6.67
128	1.87	4.60	4.78	5.87

(a) Ornstein–Uhlenbeck dynamics

$\sigma_{\max}$ / d	5.0	10.0	20.0
4	4.09	6.33	10.77
8	3.21	4.22	5.12
16	3.15	4.02	4.89
32	3.13	4.39	4.84
64	3.10	4.61	5.01
128	3.00	4.54	5.20

(b) KL minimization

5.3 Minimization of KL divergence

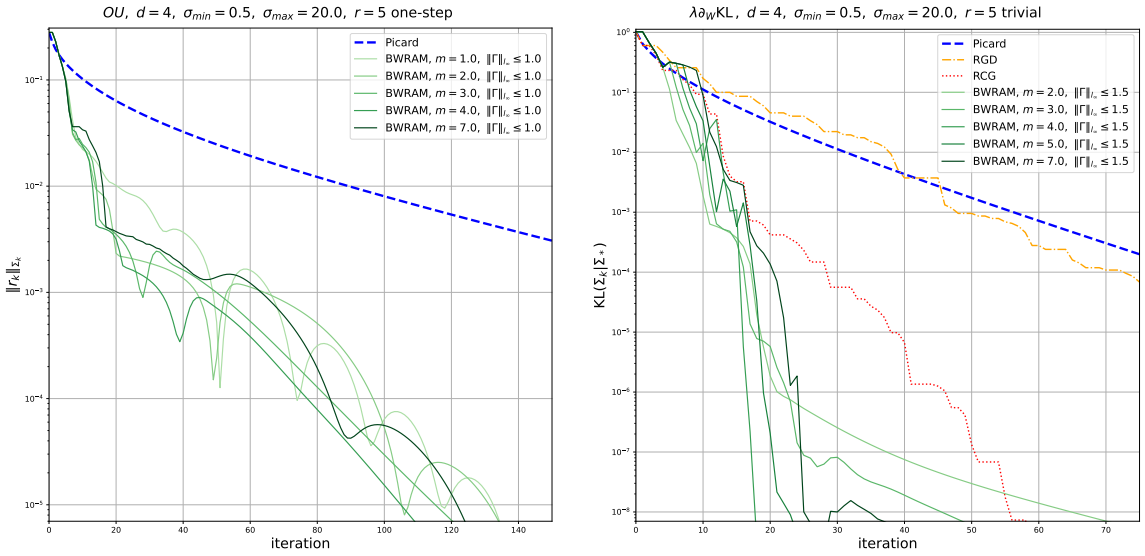
We compare the performance of BWRAM to minimization methods, in particular, RGD and RCG (Absil et al., 2009). The numerical implementation of these methods from the package `pymanopt` (Townsend et al., 2016) has been used.

The values of Σ_* were chosen in the same way as for the experiments with the Ornstein–Uhlenbeck dynamics. The step size $h = 0.3$ was chosen so that every method converges. Each method is iterated until the Kullback–Leibler divergence is reduced below $\epsilon = 10^{-10}$. The acceleration is averaged over different realizations of Σ_*. The best mean acceleration among different history lengths, depending on d and $\sigma_{\max}$, is presented in Table 1b. The improved convergence of BWRAM for the KL minimization problem, in comparison to Picard, RGD and RCG methods, is presented in Figure 2b.

5.4 Averaging probability distributions

Averaging data sources is an important problem in machine learning (see, for example, Altschuler et al. (2021) and references therein). In the case of averaging probability distributions, a common approach is to consider a Wasserstein barycenter, introduced in Agueh and Carlier (2011). Given n_σ distributions $\{\rho_i \in \mathcal{P}_2(\mathbb{R}^d)\}_{i=1}^{n_\sigma}$ and weights $\{\alpha_i > 0 \mid \alpha_1 + \cdots + \alpha_{n_\sigma} = 1\}$, the Wasserstein barycenter is a minimizer over $\mathcal{P}_2(\mathbb{R}^d)$ of

$$B(\rho) := \sum_{i=1}^{n_\sigma} \alpha_i W_2^2(\rho, \rho_i). \tag{46}$$



(a) Residual for Ornstein–Uhlenbeck

(b) Cost in the KL problem

Figure 2: Convergence of BWRAM on model problems, in comparison to Picard (blue), RGD (orange) and RCG (red) methods.

In the case of absolutely continuous ρ, the Wasserstein gradient of the cost function is given by Chewi et al. (2020)

$$\partial_W B(\rho) = \sum_{i=1}^{n_\sigma} \alpha_i (\mathrm{Id} - T_\rho^{\rho_i}), \tag{47}$$

where $T_\rho^{\rho_i}$ is the optimal transport map from ρ to ρ_i. The exponential of this gradient defines a fixed-point problem, as noted in Álvarez-Esteban et al. (2016). The authors also show that if all ρ_i are absolutely continuous and at least one of them has bounded density, then the barycenter is a fixed point of this mapping. In the case where every ρ_i is a Gaussian, the solution is unique.

Alternative approaches include the *entropy-regularized Wasserstein barycenters* and *geometric medians*, which are, respectively, the minimizers of

$$B_\gamma(\rho) := \sum_{i=1}^{n_\sigma} \alpha_i W_2^2(\rho, \rho_i) + \gamma\, \mathrm{KL}(\rho | \mathcal{N}(0, \mathrm{Id})), \tag{48}$$

$$M(\rho) := \sum_{i=1}^{n_\sigma} \alpha_i W_2(\rho, \rho_i). \tag{49}$$

The entropic penalty in the first case can be seen as a way to incorporate prior knowledge in the estimation. The geometric median is known to be more robust with respect to the perturbations of the distributions ρ_i, but poses additional difficulty to estimate. In particular, $M(\rho)$ is neither geodesically convex nor geodesically smooth. Given some smoothing parameter $\varepsilon > 0$, we work with a smoothed objective as suggested in Altschuler et al. (2021), namely

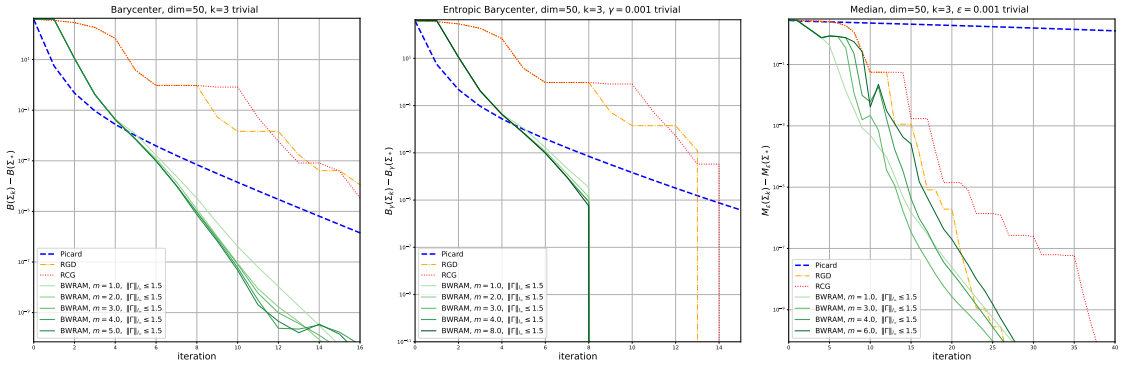
$$M_\varepsilon(\rho) = \sum_{i=1}^{n_\sigma} \alpha_i \sqrt{W_2^2(\rho, \rho_i) + \varepsilon^2}. \tag{50}$$

Constraining to the Bures–Wasserstein manifold $\mathcal{N}_0^d$, the Wasserstein gradients of these objectives are

$$\partial_W B_\gamma(\Sigma) = \sum_{i=1}^{n_\sigma} \alpha_i (\mathrm{Id} - T_\Sigma^{\Sigma_i}) + \gamma \left(\mathrm{Id} - \Sigma^{-1}\right), \tag{51}$$

$$\partial_W M_\varepsilon(\Sigma) = \sum_{i=1}^{n_\sigma} \frac{\alpha_i}{\sqrt{W_2(\Sigma, \Sigma_i)^2 + \varepsilon^2}} \left(\mathrm{Id} - T_\Sigma^{\Sigma_i}\right). \tag{52}$$

The barycenter problem has been solved with Picard iteration (Álvarez-Esteban et al., 2016) or Riemannian gradient descent (Chewi et al., 2020), which coincide if the gradient step size equals one. Stochastic gradient descent has also been considered (Chewi et al., 2020; Altschuler et al., 2021), with approximation of the sum in (47) by minibatching. For entropic barycenter and median problems, the same methods have been applied, e.g., in Altschuler et al. (2021). To the best of the authors' knowledge, this is the first attempt to apply an accelerated method to such problems.



(a) Barycenter (b) Entropic Barycenter (c) Median

Figure 3: Convergence of the cost. BWRAM (green) with different history lengths on the averaging problems, in comparison to Picard (blue), RGD (orange) and RCG (red) methods.

In Figure 3 one can see that BWRAM outperforms Picard, RGD and RCG for the averaging problems. Details on additional experiments for varying d, n_σ can be found in Section B.

5.5 Comparison of the vector transport mappings

As noted previously in Section 4.2, there are several options for the vector transport mapping. The behavior of accelerated methods (BWRAM and RCG) with different kinds of vector transport is shown in Figure 4. The computation of the exact parallel transport requires a solution of a nonlinear ODE on the manifold. This iterative procedure could be comparable in computational complexity to the solution of the fixed-point problem itself. Moreover, in some cases (e.g., the first two panels), the resulting trajectory of the method is identical to the one with a one-step approximation. These considerations rule out the usage of the exact parallel transport in the practical applications of the method.

The usage of the trivial transport map does not incur any additional computational complexity, in contrast to the one-step approximation, which can increase the per-iteration complexity. On the other hand, in the case of trivial transport, the constant M from Assumption 4.2 depends on the maximal and minimal eigenvalue ratio in the vicinity of the solution, and can be larger than one. This leads to a decrease of the theoretically guaranteed radius of the ball of converging solutions, and to an increase of the constants, hidden in the $\mathcal{O}$ terms of Theorem 7. Thus, when choosing

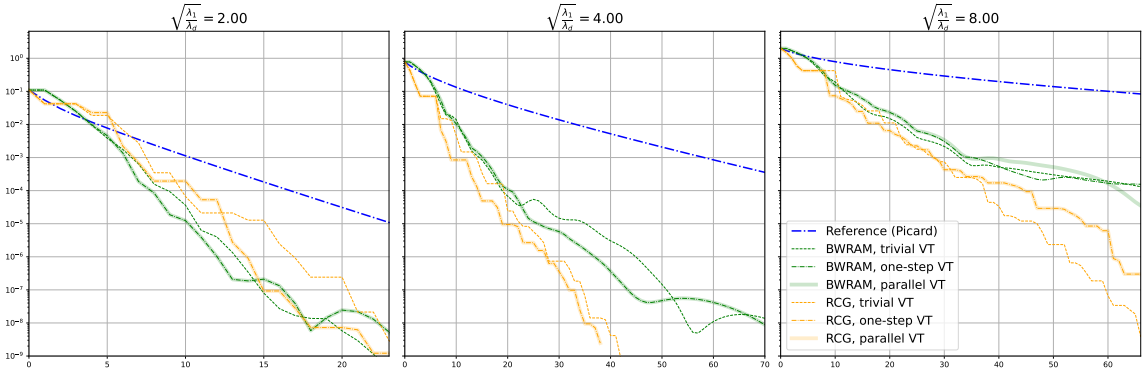


Figure 4: Performance of methods on the KL problem with varying condition number $\sqrt{\frac{\lambda_1^*}{\lambda_d^*}}$. Compared are the exact parallel transport, the one-step approximation and the trivial (identity) approximation.

between one-step and trivial transport map, there exists a trade-off between the computational efficiency of each iteration and reduced stability of the method and potentially worse acceleration. To assess the effect of this trade-off in the numerical experiments, we perform a robustness study on a subset of our test cases. The results are presented in Table 2. We consider the KL minimization

Table 2: Robustness of the performance with respect to different vector transport mappings. KL problem with $\sigma_{\max} = 20,\ \sigma_{\min} = 0.5$ and varying dimension.

		$\mathrm{Id}_{\Sigma_1}^{\Sigma_2}$		$\mathcal{T}_{\Sigma_1}^{\Sigma_2}$
d	% converged	% accelerated	% converged	% accelerated
4	60.0	53.3	98.8	64.4
8	68.0	54.0	97.3	48.4
16	63.3	53.3	94.1	53.3
32	75.8	66.4	94.7	59.7
64	84.4	74.7	94.9	65.3
128	74.0	66.0	96.4	66.5

problem with fixed maximal and minimal eigenvalues of the covariance, and vary the dimension. BWRAM is run with a set of varying parameters over several random initializations of the problem as described in Section 5.3. In this set of test runs, we compute the percentage of runs where the method converged to the prescribed tolerance, and where a speedup of at least two was achieved (in comparison with the Picard method). The outcomes show that the usage of trivial transport mapping does negatively affect the convergence. We notice, however, that this is due to divergence of the method in the early stages, which can, in our opinion, be effectively dealt with by incorporating a restarting or early stopping strategy. The probability of acceleration is, however, not changed significantly. We have also performed experiments, evaluating the maximal mean acceleration and best number of iterations, as in the previous sections, with both mentioned types of vector transport. The results were qualitatively the same, although the particular best values of the hyperparameters did not necessarily coincide. We thus argue that the use of the trivial vector transport is justified, and claim that BWRAM does improve the iteration complexity with a negligible increase of the per-iteration cost.

6. Conclusion and outlook

The numerical evidence suggests that Anderson mixing performs comparably to established Riemannian minimization methods, while being applicable to a broader class of problems, and it consistently outperforms Picard iteration. The theoretical analysis certifies the local convergence of the method and highlights the geometric structure of the Bures–Wasserstein space. It also outlines the difficulties that have to be overcome in future work on generalizing the method to arbitrary distributions.

Acknowledgments

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". MO has received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 442047500 – through the Collaborative Research Center "Sparsity and Singular Structures" (SFB 1481). The authors have no conflicts of interest to declare related to this publication.

Appendix A. Omitted proofs

This appendix contains the proofs omitted from the main text. Section A.1 establishes the local properties of $\mathcal{N}_0^d$ summarized in Theorem 5, Section A.2 proves the contractivity of the operators from Section 4.3, and Section A.3 contains the proof of Theorem 7.

A.1 Properties of $\mathcal{N}_0^d$ in a neighborhood of a distribution with a nondegenerate covariance matrix

In this section we collect technical lemmata on the local properties of the Bures–Wasserstein manifold. In particular, we relate the Bures–Wasserstein distance between the distributions and norms of the tangent vectors to Frobenius norms of the matrices in $\mathrm{Sym}(d)$ that represent them. These lemmata are used to prove that the main assumptions on the manifold, needed for the local convergence of RAM, hold in a neighborhood of a nondegenerate fixed point (cf. Theorem 18). In the following sections, they will also be utilized in order to establish the contractivity of one of the operators in question, as well as to analyze the properties of vector transport mappings.

The set $\mathcal{N}_0^d$ of Gaussian distributions with mean zero and nondegenerate covariance matrix is not complete with respect to the Wasserstein distance. But since the Wasserstein distance metrizes weak convergence and it is equivalent to pointwise convergence of the characteristic functions, one can identify the closure with the set (Takatsu, 2011)

$$\overline{\mathcal{N}_0^d} = \left\{\mu : \varphi_\mu(\xi) = e^{-\frac{1}{2}\xi^T\Sigma\xi}, \quad \Sigma^T = \Sigma \succeq 0\right\}$$

where φ_μ denotes the characteristic function of μ.

Let us denote a Bures–Wasserstein ball with center Σ_* and radius r, as

$$B_r(\Sigma_*) := \{\Sigma \in \mathcal{N}_0^d : W_2(\Sigma, \Sigma_*) \leq r\} \tag{53}$$

In the following, we show that the difference in the eigenvalues of the covariance matrices can be controlled by the Bures–Wasserstein distance between the respective distributions. Thus, the distributions in a small ball $B_r(\Sigma_*)$ around a nondegenerate distribution Σ_* have covariance matrices uniformly bounded from above and bounded away from zero. This allows us to prove several technical lemmata, useful for the analysis of the fixed-point operators and the RAM method.

Lemma 13 *Let $\Sigma_0 \succ 0,\ \Sigma_1 \succeq 0$. If the Bures–Wasserstein distance $W_2(\Sigma_0, \Sigma_1) \leq r$, then the following bound on the eigenvalues $\lambda_1^1 \geq \cdots \geq \lambda_d^1$ of Σ_1 (and analogously λ_k^0 of Σ_0) holds:*

$$\sum_k \left(\sqrt{\lambda_k^0} - \sqrt{\lambda_k^1}\right)^2 \leq r^2. \tag{54}$$

In particular,

$$\lambda_d^1 \geq \left(\sqrt{\lambda_d^0} - r\right)^2, \tag{55}$$

$$\lambda_1^1 \leq \left(\sqrt{\lambda_1^0} + r\right)^2. \tag{56}$$

Proof

$$r^2 \geq W_2^2(\Sigma_0, \Sigma_1) = \operatorname{Tr}(\Sigma_0) + \operatorname{Tr}(\Sigma_1) - 2\operatorname{Tr}(\tilde{\Sigma}) = \sum_k \left(\lambda_k^0 + \lambda_k^1 - 2\tilde{\lambda}_k\right)$$

where $\tilde{\Sigma} = \left(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\right)^{1/2}$ and $\tilde{\lambda}_k$ its eigenvalues. Since $\tilde{\Sigma}$ is symmetric, its eigenvalues and singular values coincide. Since $\tilde{\Sigma}^T\tilde{\Sigma} = \Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2} = \left(\Sigma_1^{1/2}\Sigma_0^{1/2}\right)^T \Sigma_1^{1/2}\Sigma_0^{1/2}$, the singular values of $\tilde{\Sigma}$ are the same as the singular values of $\Sigma_1^{1/2}\Sigma_0^{1/2}$. As for the singular values of $\Sigma_1^{1/2}\Sigma_0^{1/2}$,

$$\sum_k \sigma_k\left(\Sigma_1^{1/2}\Sigma_0^{1/2}\right) \leq \sum_k \sigma_k\left(\Sigma_0^{1/2}\right)\sigma_k\left(\Sigma_1^{1/2}\right) = \sum_k \sqrt{\lambda_k^0\lambda_k^1}.$$

Thus,

$$r^2 \geq \sum_k \left(\lambda_k^0 + \lambda_k^1 - 2\tilde{\lambda}_k\right) \geq \sum_k \left(\lambda_k^0 + \lambda_k^1 - 2\sqrt{\lambda_k^0\lambda_k^1}\right) = \sum_k \left(\sqrt{\lambda_k^0} - \sqrt{\lambda_k^1}\right)^2$$

and (55), (56) follow directly. ■

In the following, we need the following extended version of Proposition 3:

Proposition 14 *For $\Sigma \in \mathcal{N}_0^d$, consider an operator P_Σ*

$$P_\Sigma : (\mathrm{Sym}(d),\ \|\cdot\|_F) \to (\mathrm{Sym}(d),\ \|\cdot\|_F)$$
$$P_\Sigma[V] := \Sigma V + V\Sigma$$

P_Σ is symmetric and invertible. Denote $L_\Sigma := P_\Sigma^{-1}$.

Let $\Sigma(t) :\ [0,1] \to \mathcal{N}_0^d$ be a smooth curve, and $V(t) \in \mathrm{Tan}_{\Sigma(t)}\mathcal{N}_0^d$ its tangent vector. Then it holds

$$\dot{\Sigma}(t) = P_{\Sigma(t)}[V(t)],$$
$$V(t) = L_{\Sigma(t)}[\dot{\Sigma}(t)],$$

where $\dot{\Sigma}(t)$ stands for the time derivative of $\Sigma(t)$. The vector field $v_t(x) := V(t)x$ is the unique tangent vector field to the curve $\rho_t := \mathcal{N}(0, \Sigma(t))$ in the Wasserstein space. In particular, the continuity equation is satisfied

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

Additionally, if the eigenvalues of Σ are given by $0 < \lambda_d \leq \cdots \leq \lambda_1$, then it holds $\forall U \in \mathrm{Sym}(d)$:

$$2\lambda_d\|U\|_F^2 \leq \langle U, P_\Sigma[U]\rangle_F \leq 2\lambda_1\|U\|_F^2 \tag{57}$$

$$\frac{1}{2\lambda_1}\|U\|_F^2 \leq \langle U, L_\Sigma[U]\rangle_F \leq \frac{1}{2\lambda_d}\|U\|_F^2 \tag{58}$$

Proof Let $U,\ V \in \mathrm{Sym}(d)$. Then

$$\langle U, P_\Sigma[V]\rangle_F = \mathrm{Tr}\left(U(\Sigma V + V\Sigma)\right) = 2\,\mathrm{Tr}\left(U\Sigma V\right) = 2\,\mathrm{Tr}\left(V\Sigma U\right) = \mathrm{Tr}\left(V(\Sigma U + U\Sigma)\right) = \langle P_\Sigma[U], V\rangle_F$$

Thus, P_Σ is symmetric. Let us denote the columns of U as u_i. Then

$$\langle U, P_\Sigma[U]\rangle_F = 2\,\mathrm{Tr}(U\Sigma U) = 2\sum u_i^T \Sigma u_i \geq 2\lambda_d \sum u_i^T u_i = 2\lambda_d\,\mathrm{Tr}\left(U^2\right) = 2\lambda_d \|U\|_F^2 \tag{59}$$

Thus, P_Σ is invertible and $L_\Sigma := P_\Sigma^{-1}$ is well-defined.

Let us denote by ρ_t the time-dependent probability density induced by $\Sigma(t)$, and $v_t(x) := V(t)x$, where $V(t) := L_{\Sigma(t)}\left[\dot\Sigma(t)\right]$. We verify that the vector field satisfies the continuity equation

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

Equivalently,

$$\rho_t\left[\partial_t \log\rho_t + \langle\nabla\log\rho_t, v_t\rangle + \nabla\cdot v_t\right] = 0$$

Since ρ_t is supported on the whole $\mathbb{R}^d$, the quantity in brackets has to be zero. In our case

$$\begin{aligned}
\partial_t \log\rho_t(x) &= -\frac{1}{2}\,\mathrm{Tr}\left(\Sigma^{-1}(t)\dot\Sigma(t)\right) + \frac{1}{2}x^T\Sigma^{-1}(t)\dot\Sigma(t)\Sigma^{-1}(t)x\\
\langle\nabla\log\rho_t(x), v_t\rangle &= -x^T\Sigma^{-1}(t)V(t)x = -\frac{1}{2}x^T\left(\Sigma^{-1}(t)V(t) + V(t)\Sigma^{-1}(t)\right)x\\
\nabla\cdot v_t &= \mathrm{Tr}(V(t))
\end{aligned}$$

Summing up

$$-\frac{1}{2}\,\mathrm{Tr}\left(\Sigma^{-1}(t)(\dot\Sigma(t) - 2\Sigma(t)V(t))\right) - \frac{1}{2}x^T\Sigma^{-1}(t)\left\{\dot\Sigma(t) - (V(t)\Sigma(t) + \Sigma(t)V(t))\right\}\Sigma^{-1}(t)x = 0$$

Since both $\Sigma(t),\ V(t)$ are symmetric and $\dot\Sigma(t) = \Sigma(t)V(t) + V(t)\Sigma(t)$, we verify that the continuity equation holds. Moreover, for every t, since $V(t)x = \frac{1}{2}\nabla\langle x, V(t)x\rangle$ and $\int \|Vx\|_2^2 \mathrm{d}\rho_{\Sigma(t)} < +\infty$, the vector field is tangent in the sense of Ambrosio et al. (2005), i.e.,

$$Vx \in \mathrm{Tan}_{\mathcal{N}(0,\Sigma)}\,\mathcal{P}_2(\mathbb{R}^d) := \overline{\{\nabla\phi,\ \phi\in C_c^\infty\}}^{L^2_{\mathcal{N}(0,\Sigma)}(\mathbb{R}^d)} \tag{60}$$

As for the eigenvalue bound, the lower bound in (57) was already proven in (59). The upper bound is shown analogously because $u_i^T\Sigma u_i \leq \lambda_1 u_i^T u_i$. The bound on L_Σ follows because $L_\Sigma = P_\Sigma^{-1}$. ■

In the sequel we focus on a ball $B_r(\Sigma_*)$. We denote by $0 < \lambda_d^* \leq \cdots \leq \lambda_1^*$ the eigenvalues of Σ_*. The radius r is small enough such that $\sqrt{\lambda_d^*} - 2r > 0$.

The following lemma allows us to relate the norms of the vectors in the different tangent spaces:

Lemma 15 *If $\Sigma, \Sigma_1, \Sigma_2 \in B_r(\Sigma_*)$ with $r < \sqrt{\lambda_d^*}$, $V \in \mathrm{Tan}_{\Sigma_1}\mathcal{N}_0^d$, then the following estimates hold*

$$(\sqrt{\lambda_d^*} - r)\|V\|_F \leq \|V\|_\Sigma \leq (\sqrt{\lambda_1^*} + r)\|V\|_F \tag{61}$$

$$\|V\|_{\Sigma_2} \leq \frac{\sqrt{\lambda_1^*} + r}{\sqrt{\lambda_d^*} - r}\|V\|_{\Sigma_1} \tag{62}$$

Proof Assuming $A \in \mathbb{R}^{d\times d}, B \in \mathrm{Sym}(d)$, $\lambda_1\,\mathrm{Id} \succeq B \succeq \lambda_d\,\mathrm{Id}$

$$\begin{aligned}
\mathrm{Tr}\left(A^TBA\right) &= \sum_i a_i^T B a_i \geq \lambda_d \sum_i a_i^T a_i = \lambda_d\,\mathrm{Tr}\left(A^TA\right) = \lambda_d\|A\|_F^2\\
\mathrm{Tr}\left(A^TBA\right) &= \sum_i a_i^T B a_i \leq \lambda_1 \sum_i a_i^T a_i = \lambda_1\,\mathrm{Tr}\left(A^TA\right) = \lambda_1\|A\|_F^2
\end{aligned}$$

where a_i are columns of A. For $\Sigma \in \mathcal{N}_0^d$, $V \in \mathrm{Tan}_\Sigma \mathcal{N}_0^d$, plugging $A \leftarrow V,\ B \leftarrow \Sigma$

$$\lambda_d(\Sigma)\|V\|_F^2 \le \|V\|_\Sigma^2 \le \lambda_1(\Sigma)\|V\|_F^2 \tag{63}$$

Using the uniform bounds (55)–(56), one arrives at (61). Combining (61) twice for $\Sigma = \Sigma_{1,2}$, one gets (62). ■

Lemma 16 *Let $\Sigma_1, \Sigma_2 \in B_r(\Sigma_*)$ with $r < \frac{1}{2}\sqrt{\lambda_d^*}$. Then the Frobenius distance between the covariance matrices can be controlled by the Bures–Wasserstein distance between the distributions, i.e.*

$$\|\Sigma_1 - \Sigma_2\|_F \le 2\left(\sqrt{\lambda_1^*} + \sqrt{\lambda_d^*}\right) W_2(\Sigma_1,\ \Sigma_2) \tag{64}$$

as well as

$$2(\sqrt{\lambda_d^*} - r) W_2(\Sigma_1,\ \Sigma_2) \le \|\Sigma_1 - \Sigma_2\|_F \tag{65}$$

Proof In the dynamic OT formulation

$$W_2^2(\Sigma_1, \Sigma_2) = \inf_{\Sigma(\cdot)\in\mathcal{A}} \int_0^1 \|V(t)\|_{\Sigma(t)}^2 \,\mathrm{d}t \tag{66}$$

where $V(t) \in \mathrm{Tan}_{\Sigma(t)} \mathcal{N}_0^d$ is a tangent vector to the curve $\Sigma(t)$, and the admissible set of curves $\mathcal{A}$ is defined as

$$\mathcal{A} := \left\{ \Sigma(t) : t \in [0,1] \,\middle|\, \begin{smallmatrix} \Sigma(t)\in\mathcal{N}_0^d \\ \Sigma(0)=\Sigma_1 \\ \Sigma(1)=\Sigma_2 \end{smallmatrix} \right\} \tag{67}$$

Since $\Sigma_1, \Sigma_2 \in B_r(\Sigma_*)$, from the triangle inequality we argue that

$$W_2(\Sigma_1, \Sigma_2) \le W_2(\Sigma_1, \Sigma_*) + W_2(\Sigma_*, \Sigma_2) \le 2r$$

Considering a point Σ_t, lying on the geodesic between $\Sigma_1,\ \Sigma_2$, we can argue that

$$\begin{aligned} &W_2(\Sigma_*, \Sigma_t) \\ &\qquad \le \min\{W_2(\Sigma_*, \Sigma_1) + W_2(\Sigma_1, \Sigma_t),\ W_2(\Sigma_*, \Sigma_2) + W_2(\Sigma_2, \Sigma_t)\} \\ &\qquad\qquad \le r + \min\{W_2(\Sigma_1, \Sigma_t), W_2(\Sigma_2, \Sigma_t)\} \le r + \frac{1}{2} W_2(\Sigma_1, \Sigma_2) \le 2r, \end{aligned}$$

where the triangle inequality and the property of the geodesic (20) were used. Thus, the infimum in (66) would not change if we only consider such curves that $W_2(\Sigma(t), \Sigma_*) \le 2r\ \forall t \in [0,1]$. Let us denote the restricted admissible set of curves by

$$\mathcal{A}_r := \left\{ \Sigma(t),\ t \in [0,1]\ : \begin{smallmatrix} \Sigma(t)\in\mathcal{N}_0^d \\ W_2(\Sigma(t),\Sigma_*)\le 2r\ \forall t\in[0,1] \\ \Sigma(0)=\Sigma_1 \\ \Sigma(1)=\Sigma_2 \end{smallmatrix} \right\} \tag{68}$$

Due to Proposition 14,

$$\begin{aligned} \dot\Sigma(t) &= P_{\Sigma(t)}[V(t)] = \Sigma(t)V(t) + V(t)\Sigma(t) \\ V(t) &= L_{\Sigma(t)}[\dot\Sigma(t)] = P_{\Sigma(t)}^{-1}[\dot\Sigma(t)] \end{aligned}$$

On the admissible set of curves $\mathcal{A}_r$, using (58) and (56):

$$\langle U, L_\Sigma[U]\rangle_F \ge \frac{1}{2(\sqrt{\lambda_1^*} + 2r)^2} \|U\|_F^2 \tag{69}$$

$$\|V(t)\|_\Sigma^2 = \operatorname{Tr}\left(V(t)\Sigma(t)V(t)\right) = \frac{1}{2}\operatorname{Tr}\left(V(t)[\Sigma(t)V(t)+V(t)\Sigma(t)]\right) = \frac{1}{2}\operatorname{Tr}\left(V(t)\dot\Sigma(t)\right)$$
$$= \frac{1}{2}\operatorname{Tr}\left(L_{\Sigma(t)}[\dot\Sigma(t)]\dot\Sigma(t)\right) = \frac{1}{2}\langle\dot\Sigma(t), L_{\Sigma(t)}[\dot\Sigma(t)]\rangle_F \geq \frac{1}{4(\sqrt{\lambda_1^*}+2r)^2}\|\dot\Sigma(t)\|_F^2$$

where the final inequality holds only for the curves in $\mathcal{A}_r$.

Combining the estimates, we arrive at

$$W_2^2(\Sigma_1,\Sigma_2) = \inf_{\mathcal{A}}\int\|V(t)\|_{\Sigma(t)}^2\,\mathrm{d}t$$
$$= \inf_{\mathcal{A}_r}\int\|V(t)\|_{\Sigma(t)}^2\,\mathrm{d}t = \inf_{\mathcal{A}_r}\frac{1}{2}\int\operatorname{Tr}\left(L_{\Sigma(t)}[\dot\Sigma(t)]\dot\Sigma(t)\right)\mathrm{d}t$$
$$\overset{(*)}{\geq} \frac{1}{4(\sqrt{\lambda_1^*}+2r)^2}\inf_{\substack{\Sigma(t)\in\mathcal{N}_0^d\\ \Sigma(0)=\Sigma_1\\ \Sigma(1)=\Sigma_2}}\int\left\|\dot\Sigma(t)\right\|_F^2\mathrm{d}t = \frac{\|\Sigma_1-\Sigma_2\|_F^2}{4(\sqrt{\lambda_1^*}+2r)^2}$$

In $(*)$, we can minimize with respect to the original admissible set $\mathcal{A}$, since $\mathcal{A}_r\subset\mathcal{A}$ and thus $\inf_{\mathcal{A}}\mathcal{L}\leq\inf_{\mathcal{A}_r}\mathcal{L}$ for any functional $\mathcal{L}$. The last equality is due to the infimum being achieved at the Euclidean geodesic

$$\Sigma_e(t) := (1-t)\Sigma_1 + t\Sigma_2$$

Thus

$$\|\Sigma_1-\Sigma_2\|_F \leq 2(\sqrt{\lambda_1^*}+2r)W_2(\Sigma_1,\ \Sigma_2) \tag{70}$$

Since $2r<\sqrt{\lambda_d^*}$, we conclude

$$\|\Sigma_1-\Sigma_2\|_F \leq 2(\sqrt{\lambda_1^*}+\sqrt{\lambda_d^*})W_2(\Sigma_1,\ \Sigma_2)$$

Note that the minimal eigenvalue as a function of the matrix is an infimum over linear functions and thus is concave. This means that for the Euclidean geodesic

$$\lambda_d(\Sigma_e(t)) \geq (1-t)\lambda_d(\Sigma_1)+t\lambda_d(\Sigma_2) \geq \min\{\lambda_d(\Sigma_1),\lambda_d(\Sigma_2)\} \geq \left(\sqrt{\lambda_d^*}-r\right)^2 > 0.$$

Plugging Σ_e into (66), one gets

$$W_2^2(\Sigma_1,\Sigma_2) \leq \frac{1}{2}\int_0^1\operatorname{Tr}\left(L_{\Sigma_e(t)}[\dot\Sigma_e(t)]\dot\Sigma_e(t)\right)\mathrm{d}t$$
$$\leq \frac{1}{2}\int_0^1\frac{\|\dot\Sigma_e(t)\|_F^2}{2\lambda_d(\Sigma_e(t))}\mathrm{d}t \leq \frac{\|\Sigma_1-\Sigma_2\|_F^2}{4(\sqrt{\lambda_d^*}-r)^2}.$$

Hence $\|\Sigma_1-\Sigma_2\|_F \geq 2(\sqrt{\lambda_d^*}-r)W_2(\Sigma_1,\Sigma_2)$. ■

Combining the estimates, we establish that locally W_2 is equivalent to the Frobenius norm:

$$2(\sqrt{\lambda_d^*}-r)W_2(\Sigma_1,\ \Sigma_2) \leq \|\Sigma_1-\Sigma_2\|_F \leq 2(\sqrt{\lambda_1^*}+2r)W_2(\Sigma_1,\ \Sigma_2) \tag{71}$$

The explicit expression for the sectional curvature of the Bures–Wasserstein space is given in (Takatsu, 2010, Theorem 1.1). Its minimal and maximal values depend only on the minimal and maximal eigenvalues of the covariance matrix. We express this fact in the following corollary.

Corollary 17 *For every $\Sigma\in B_r(\Sigma_*)$, the sectional curvature K_Σ of $\mathcal{N}_0^d$ at Σ can be uniformly bounded by*

$$0 \leq K_\Sigma(e_1,e_2) \leq \frac{3(\sqrt{\lambda_1^*}+r)^4}{2(\sqrt{\lambda_d^*}-r)^6}, \tag{72}$$

where $e_1,\ e_2$ are the directions defining the plane and $r<\sqrt{\lambda_d^*}$.

We can finally formulate the theorem showing that the main assumptions for the convergence of RAM hold for $B_r(\Sigma_*)$ (analogous to Assumption 1 of Li and Bao (2023)).

Theorem 18 *Let Σ_* be a nondegenerate covariance matrix and $r < \frac{1}{2}\sqrt{\lambda_d^*}$. The ball $B_r(\Sigma_*)$ is compact in $\mathcal{N}_0^d$ with bounded sectional curvature. For every $V \in \mathrm{Tan}_\Sigma \mathcal{N}_0^d$ such that $\|V\|_\Sigma \leq \frac{1}{2}r$, it holds that $W_2(\Sigma, \mathrm{Exp}_\Sigma(V)) = \|V\|_\Sigma$ for all $\Sigma \in \mathcal{N}_0^d$.*

Proof [Proof of Theorem 18] By Lemma 13 and Lemma 16, locally around nondegenerate matrices, the Bures–Wasserstein-induced topology and the topology induced by the Frobenius norm are equivalent. As the ball is closed and bounded, it is compact in the Frobenius topology and, thus, in the Bures–Wasserstein topology as well. The sectional curvature bounds are due to Corollary 17.

Every tangent vector V induces a mapping $\tau(x) = Vx + x = (\mathrm{Id} + V)x = \nabla_x \frac{1}{2}\langle x, (\mathrm{Id} + V)x\rangle$. Using (61), we get

$$\|V\|_o \leq \|V\|_F \leq \frac{1}{\sqrt{\lambda_d^*} - r}\|V\|_\Sigma \leq \frac{r}{2(\sqrt{\lambda_d^*} - r)} < \frac{1}{2},$$

where $\|V\|_o = \sup_{x \neq 0} \frac{\|Vx\|_2}{\|x\|_2}$ is the operator norm. Since the operator norm satisfies $\|V\|_o < 1$, the operator $(\mathrm{Id} + V)$ has nonnegative eigenvalues, and the mapping τ is a gradient of a convex function. Thus, the mapping is optimal, and its $L^2_{\mathcal{N}(0,\Sigma)}$ norm is equal to the Wasserstein distance between the distributions. ■

A.2 Contraction coefficients of the operators

Recall

$$\partial_W \mathcal{E}[\rho] := \nabla \frac{\delta}{\delta \rho}\mathcal{E}(\rho).$$

For the operators of the form

$$G(\rho) = \mathrm{Exp}_\rho(-h\partial_W \mathcal{E}[\rho]) \tag{73}$$

the contraction estimate can be deduced as follows. By the definition of the exponential map in the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ (24)

$$\mathrm{Exp}_\rho(-h\partial_W \mathcal{E}[\rho]) = (\mathrm{id} - h\partial_W \mathcal{E}[\rho])_\sharp \rho,$$

if γ_{opt} is an optimal coupling between ρ_1 and ρ_2, then

$$((x, y) \mapsto (x - h\partial_W \mathcal{E}[\rho_1](x),\ y - h\partial_W \mathcal{E}[\rho_2](y)))_\sharp \gamma_{\mathrm{opt}} \in \Pi(G(\rho_1), G(\rho_2))$$

is a valid coupling between $G(\rho_1)$ and $G(\rho_2)$. Thus

$$\begin{aligned} W_2^2(G(\rho_1), G(\rho_2)) &\leq \int \|x - y - h(\partial_W \mathcal{E}[\rho_1](x) - \partial_W \mathcal{E}[\rho_2](y))\|^2 \mathrm{d}\gamma_{\mathrm{opt}} \\ &= \int \|x-y\|^2 \mathrm{d}\gamma_{\mathrm{opt}} - 2h \int \langle \partial_W \mathcal{E}[\rho_1](x) - \partial_W \mathcal{E}[\rho_2](y), x - y\rangle \mathrm{d}\gamma_{\mathrm{opt}} + h^2 \int \|\partial_W \mathcal{E}[\rho_1](x) - \partial_W \mathcal{E}[\rho_2](y)\|^2 \mathrm{d}\gamma_{\mathrm{opt}} \end{aligned} \tag{74}$$

If the functional $\mathcal{E}$ is λ-convex along generalized geodesics, then the second term can be estimated as follows (Ambrosio et al., 2005, Equation 10.1.8):

$$\int \langle \partial_W \mathcal{E}[\rho_1](x) - \partial_W \mathcal{E}[\rho_2](y), x - y\rangle \mathrm{d}\gamma_{\mathrm{opt}} \geq \frac{1}{2}\lambda W_2^2(\rho_1, \rho_2) \tag{75}$$

To estimate the final term, we can define the generalization of the Lipschitz gradient property in the following way.

Definition 19 *A functional $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ has L-Lipschitz Wasserstein gradient on some subset $\mathcal{S} \subseteq \mathcal{P}_2(\mathbb{R}^d)$, if its Wasserstein gradient*

$$\partial_W \mathcal{E}[\rho] = \nabla \frac{\delta \mathcal{E}}{\delta \rho} \in L^2_\rho \ \forall \rho \in \mathcal{S}$$

and there exists a constant $L > 0$ such that $\forall \rho_1,\ \rho_2 \in \mathcal{S}$

$$\int \|\partial_W \mathcal{E}[\rho_1](x) - \partial_W \mathcal{E}[\rho_2](y)\|^2 \mathrm{d}\gamma_{opt} \leq L^2 W_2^2(\rho_1, \rho_2) \tag{76}$$

Proposition 20 *If the functional $\mathcal{E}$ is λ-convex along generalized geodesics and has L-Lipschitz Wasserstein gradient, then the operator $\mathrm{Exp}_\rho(-h\partial_W \mathcal{E})$ is contractive for small enough h.*

Proof Plugging (75) and (76) into (74), one gets

$$W_2^2(G(\rho_1), G(\rho_2)) \leq (1 - \lambda h + h^2 L^2) W_2^2(\rho_1, \rho_2)$$

The coefficient in front of $W_2^2(\rho_1, \rho_2)$ is smaller than one if $h \in (0, \frac{\lambda}{L^2})$. In particular, the optimal step is $h = \frac{\lambda}{2L^2}$. ■

In Ambrosio et al. (2005), there are multiple examples of functionals λ-convex along generalized geodesics. In particular, for the functional $\mathcal{E} = \mathrm{KL}(\cdot|\rho_*)$, where $\rho_* \propto e^{-V}$, $\mathcal{E}$ is λ-convex when V is λ-convex. To the best of the authors' knowledge, the Lipschitz gradient property has not been thoroughly studied before. In the following, we perform this analysis for the restriction of the $\mathrm{KL}(\cdot|\rho_*)$ functional to the Bures–Wasserstein manifold. We use the shorthand notation $\mathrm{KL}(\Sigma|\Sigma_*) := \mathrm{KL}(\mathcal{N}(0, \Sigma)|\mathcal{N}(0, \Sigma_*))$.

Lemma 21 *Let Σ_* be a nondegenerate covariance matrix. The functional*

$$\mathrm{KL}(\cdot|\Sigma_*) :\ \mathcal{N}_0^d \to \mathbb{R}$$

has L-Lipschitz Wasserstein gradient on a ball $B_r(\Sigma_*)$ where

$$r < \frac{1}{32\left(1 + \sqrt{\frac{\lambda_1^*}{\lambda_d^*}}\right)} \sqrt{\lambda_d^*} \tag{77}$$

and

$$L^2 = \frac{8}{(\sqrt{\lambda_d^*} - r)^4} \left(1 + 4\frac{(\sqrt{\lambda_1^*} + 2r)^4}{(\sqrt{\lambda_d^*} - r)^4}\right)$$

Proof The idea of the proof is to work with the explicit form of (74) and split the quantity under the integral w.r.t. $\mathrm{d}\gamma_{\mathrm{opt}}$ into two parts. The first is the "linear part" of the form $\|A(x - y)\|$ for some operator A, which can be directly related to the Bures–Wasserstein distance. The second term is a "nonlinear part" $\|\Sigma_1^{-1} - \Sigma_2^{-1}\|_F$, which we treat by first considering the linearization w.r.t. $\|\Sigma_1 - \Sigma_2\|_F$, and then relating the norm difference to the Wasserstein distance using (29).

Assuming that $\Sigma_1,\ \Sigma_2 \in B_r(\Sigma_*)$ with $r < \frac{1}{2}\sqrt{\lambda_d^*}$, so that the previous lemmata can be used, we proceed with the calculation:

$$\begin{aligned}
&\int \|\partial_W \mathcal{E}[\Sigma_1](x) - \partial_W \mathcal{E}[\Sigma_2](y)\|^2 \mathrm{d}\gamma_{\mathrm{opt}} \\
&= \int \left\|(\Sigma_*^{-1} - \Sigma_1^{-1})x - (\Sigma_*^{-1} - \Sigma_2^{-1})y\right\|^2 \mathrm{d}\gamma_{\mathrm{opt}} = \int \left\|(\Sigma_*^{-1} - \Sigma_1^{-1})(x-y) + (\Sigma_2^{-1} - \Sigma_1^{-1})y\right\|^2 \mathrm{d}\gamma_{\mathrm{opt}} \\
&\leq 2\|\Sigma_1^{-1} - \Sigma_*^{-1}\|_o^2 \int \|x-y\|^2 \mathrm{d}\gamma_{\mathrm{opt}} + 2\int \left\|(\Sigma_1^{-1} - \Sigma_2^{-1})y\right\|^2 \mathrm{d}\gamma_{\mathrm{opt}} \\
&= 2\|\Sigma_1^{-1} - \Sigma_*^{-1}\|_o^2 W_2^2(\Sigma_1, \Sigma_2) + 2\left\|\Sigma_2^{1/2}(\Sigma_1^{-1} - \Sigma_2^{-1})\right\|_F^2 \\
&\leq 2\|\Sigma_1^{-1} - \Sigma_*^{-1}\|_o^2 W_2^2(\Sigma_1, \Sigma_2) + 2\|\Sigma_2\|_o \|\Sigma_1^{-1} - \Sigma_2^{-1}\|_F^2 \qquad (78)
\end{aligned}$$

where $\|\cdot\|_o$ and $\|\cdot\|_F$ stand for operator and Frobenius norm, respectively. The eigenvalues for the inverse of matrices can be bounded as

$$\frac{1}{\lambda_1^*}\,\mathrm{Id} \preceq \Sigma_*^{-1} \preceq \frac{1}{\lambda_d^*}\,\mathrm{Id} \qquad (79)$$

$$\frac{1}{(\sqrt{\lambda_1^*} + r)^2}\,\mathrm{Id} \preceq \Sigma_{1,2}^{-1} \preceq \frac{1}{(\sqrt{\lambda_d^*} - r)^2}\,\mathrm{Id} \qquad (80)$$

The operator norm in front of the first term can thus be bounded as

$$\|\Sigma_1^{-1} - \Sigma_*^{-1}\|_o^2 \leq (\|\Sigma_1^{-1}\|_o + \|\Sigma_*^{-1}\|_o)^2 \leq \left(\frac{(\sqrt{\lambda_d^*} - r)^2 + \lambda_d^*}{(\sqrt{\lambda_d^*} - r)^2 \lambda_d^*}\right)^2 \leq \frac{4}{(\sqrt{\lambda_d^*} - r)^4}$$

Considering the quantity $\|\Sigma_1^{-1} - \Sigma_2^{-1}\|_F$, we claim that if Σ_1 and Σ_2 are sufficiently close, the following Neumann series expansion can be used

$$\begin{aligned}
\Sigma_2^{-1} = [\Sigma_1 - (\Sigma_1 - \Sigma_2)]^{-1} &= \Sigma_1^{-1}\left[\mathrm{Id} - \underbrace{(\Sigma_1 - \Sigma_2)\Sigma_1^{-1}}_{:=\Delta\Sigma}\right]^{-1} \\
&= \Sigma_1^{-1}\left[\mathrm{Id} + \Delta\Sigma + \Delta\Sigma^2 + \dots\right] = \Sigma_1^{-1}\left[\mathrm{Id} + \Delta\Sigma + R\right]
\end{aligned}$$

Before we proceed, let us find the exact condition to bound the remainder R of the series. Since $\Sigma_1,\ \Sigma_2$ are both symmetric,

$$\begin{aligned}
\|\Delta\Sigma\|_F^2 &= \|(\Sigma_1 - \Sigma_2)\Sigma_1^{-1}\|_F^2 \\
&= \mathrm{Tr}\left(\Sigma_1^{-1}(\Sigma_1 - \Sigma_2)^2\Sigma_1^{-1}\right) = \mathrm{Tr}\left((\Sigma_1^{-1}(\Sigma_1 - \Sigma_2))^T(\Sigma_1^{-1}(\Sigma_1 - \Sigma_2))\right) \\
&= \|\Sigma_1^{-1}(\Sigma_1 - \Sigma_2)\|_F^2
\end{aligned}$$

where symmetry of the matrices and the cyclic property of the trace was used. Using (54) and (70)

$$\begin{aligned}
\|\Delta\Sigma\|_F &= \|\Sigma_1^{-1}(\Sigma_1 - \Sigma_2)\|_F \\
&\leq \|\Sigma_1^{-1}\|_o \|\Sigma_1 - \Sigma_2\|_F \leq \frac{2(\sqrt{\lambda_1^*} + 2r)}{(\sqrt{\lambda_d^*} - r)^2} W_2(\Sigma_1, \Sigma_2) \\
&\leq \frac{8(\sqrt{\lambda_1^*} + 2r)}{\lambda_d^*} W_2(\Sigma_1, \Sigma_2) \leq \frac{8(\sqrt{\lambda_1^*} + \sqrt{\lambda_d^*})}{\lambda_d^*} W_2(\Sigma_1, \Sigma_2) \qquad (81)
\end{aligned}$$

Thus, if

$$W_2(\Sigma_1,\Sigma_2) \leq \frac{1}{16\left(1+\sqrt{\frac{\lambda_1^*}{\lambda_d^*}}\right)}\sqrt{\lambda_d^*} \implies \|\Delta\Sigma\|_F \leq \frac{1}{2}$$

If $\|\Delta\Sigma\|_F \leq \frac{1}{2}$, the remainder of the series R can be estimated as

$$\begin{aligned}\|R\|_F = \left\|\Delta\Sigma^2+\Delta\Sigma^3+\dots\right\|_F &\leq \|\Delta\Sigma\|_F^2+\|\Delta\Sigma\|_F^3+\dots \\ &= \|\Delta\Sigma\|_F^2\left(1+\|\Delta\Sigma\|_F+\|\Delta\Sigma\|_F^2+\dots\right) = \frac{\|\Delta\Sigma\|_F^2}{1-\|\Delta\Sigma\|_F} \leq 2\|\Delta\Sigma\|_F^2\end{aligned}$$

Since $W_2(\Sigma_1,\Sigma_2) \leq W_2(\Sigma_1,\Sigma_*)+W_2(\Sigma_*,\Sigma_2) \leq 2r$ for $\Sigma_1,\Sigma_2 \in B_r(\Sigma_*)$, we constrain the radius of the ball r to be

$$r \leq \frac{1}{32\left(1+\sqrt{\frac{\lambda_1^*}{\lambda_d^*}}\right)}\sqrt{\lambda_d^*} \leq \frac{1}{64}\sqrt{\lambda_d^*}$$

Since $\frac{1}{64} < \frac{1}{2}$, the usage of (29) was justified. Using this relation, we can get

$$\begin{aligned}\|\Sigma_1^{-1}-\Sigma_2^{-1}\|_F^2 = \|\Sigma_1^{-1}(\Sigma_1-\Sigma_2)\Sigma_1^{-1}+\Sigma_1^{-1}R\|_F^2 \\ \leq \|\Sigma_1^{-1}\|_o^2\|\Delta\Sigma+R\|_F^2 \leq 2\|\Sigma_1^{-1}\|_o^2(\|\Delta\Sigma\|_F^2+\|R\|_F^2) \\ \leq 2\|\Sigma_1^{-1}\|_o^2(\|\Delta\Sigma\|_F^2+4\|\Delta\Sigma\|_F^4) \leq 4\|\Sigma_1^{-1}\|_o^2\|\Delta\Sigma\|_F^2 \quad (82)\end{aligned}$$

Using (80) and (81), one gets

$$\|\Sigma_1^{-1}-\Sigma_2^{-1}\|_F^2 \leq 4\|\Sigma_1^{-1}\|_o^2\|\Delta\Sigma\|_F^2 \leq \frac{4}{(\sqrt{\lambda_d^*}-r)^4}\frac{4(\sqrt{\lambda_1^*}+2r)^2}{(\sqrt{\lambda_d^*}-r)^4}W_2^2(\Sigma_1,\Sigma_2)$$

We can finish the bound in (78)

$$\begin{aligned}\int \|\partial_W\mathcal{E}[\Sigma_1](x)-\partial_W\mathcal{E}[\Sigma_2](y)\|^2 \mathrm{d}\gamma_{\mathrm{opt}} \\ \leq \left(2\|\Sigma_1^{-1}-\Sigma_*^{-1}\|_o^2+2\|\Sigma_2\|_o\frac{4}{(\sqrt{\lambda_d^*}-r)^4}\frac{4(\sqrt{\lambda_1^*}+2r)^2}{(\sqrt{\lambda_d^*}-r)^4}\right)W_2^2(\Sigma_1,\Sigma_2) \\ \leq \frac{8}{(\sqrt{\lambda_d^*}-r)^4}\left(1+4\frac{(\sqrt{\lambda_1^*}+2r)^4}{(\sqrt{\lambda_d^*}-r)^4}\right)W_2^2(\Sigma_1,\Sigma_2) \quad (83)\end{aligned}$$

■

Remark 22 *The sufficient condition on the radius, provided in the lemma, as well as the estimates on the step size h and the contractivity coefficient κ, that can be derived using Proposition 20, are in no way optimal. In our numerical experiments, we empirically observe stable linear convergence for initial solutions orders of magnitude further away from the fixed point than guaranteed by the estimate above. We believe that these estimates can be improved, but a systematic treatment of this problem is left to future work.*

A.3 Proof of Theorem 7

Throughout this section, we write x_k for the iterates Σ_k and x_* for the fixed point Σ_*, and r denotes the radius from Assumption 4.1.

Lemma 23 (cf. Lemma 8 in Li and Bao (2023)) *Let $x_1, \dots, x_k \in B_{r/2}(\Sigma_*)$. Then there is M_1 such that*

$$\|X_k\Gamma_k\| + \|r_k\| \le M_1 \sum_{j=0}^{k} \|r_{k-j}\|.$$

Proof We observe that

$$\|\operatorname{Exp}_{x_\ell}^{-1}(x_{\ell+1})\| = W_2(x_\ell, x_{\ell+1})$$

by optimality of Exp^{-1}. Then by the Lipschitz assumption in Assumption 4.1, we have

$$\begin{aligned}
\|\Delta x_{k-j}^{(k)}\| &= \|\mathcal{T}_{x_{k-1}}^{x_k} \cdots \mathcal{T}_{x_{k-j}}^{x_{k-j+1}} \operatorname{Exp}_{x_{k-j}}^{-1}(x_{k-j+1})\| \\
&\le M^j \|\operatorname{Exp}_{x_{k-j}}^{-1}(x_{k-j+1})\| \\
&= M^j W_2(x_{k-j}, x_{k-j+1}) \\
&\le \frac{M^j}{L_1} \|r_{k-j+1} - \mathcal{P}_{x_{k-j}}^{x_{k-j+1}} r_{k-j}\|.
\end{aligned}$$

It holds

$$\begin{aligned}
\|\sum_{j=1}^{i} \Delta x_{k-j}^{(k)}\| &\le \sum_{j=1}^{i} \|\Delta x_{k-j}^{(k)}\| \le \sum_{j=1}^{i} \frac{M^j}{L_1} (\|r_{k-j+1}\| + \|r_{k-j}\|) \\
&\le \frac{M}{L_1}\|r_k\| + \sum_{j=2}^{i} \frac{M^{j-1}(M+1)}{L_1} \|r_{k-j+1}\| + \frac{M^i}{L_1}\|r_{k-i}\|.
\end{aligned}$$

Observe $2\frac{\max\{M^m, M\}}{L_1} \ge \frac{1}{L_1}\max\{M^m + M^{m-1}, M^2 + M\}$. By Assumption 4.3 we have $\|\Gamma_k\| \le M_\Gamma$. Hence,

$$\begin{aligned}
\|X_k\Gamma_k\| + \|r_k\| &\le \|\Gamma_k\| \sum_{j=1}^{k} \|\Delta x_{k-j}^{(k)}\| + \|r_k\| \\
&\le 2M_\Gamma \frac{\max\{M^m, M\}}{L_1} (\|r_k\| + \sum_{j=2}^{k} \|r_{k-j+1}\| + \|r_0\|) + \|r_k\| \\
&\le (2M_\Gamma \frac{\max\{M^m, M\}}{L_1} + 1) \sum_{j=0}^{k} \|r_{k-j}\|.
\end{aligned}$$

■

Lemma 24 (cf. Propositions 2 and 3 in Li and Bao (2023)) *Let $x_1, \dots, x_k \in B_\varrho(\Sigma_*)$ with $\varrho = \frac{r}{L_2(mM_1+\beta+1)}$. Let $z_k^i = \sum_{j=1}^{i} \Delta x_{k-j}^{(k)}$ and $y_k^i = \operatorname{Exp}_{x_k}(-z_k^i)$. Then there exist $M_2, M_3 > 0$ such that*

$$\|\mathcal{P}_{y_k^i}^{x_k} F(y_k^i) - \mathcal{T}_{x_{k-1}}^{x_k} \mathcal{T}_{x_{k-2}}^{x_{k-1}} \cdots \mathcal{T}_{x_{k-i}}^{x_{k-i+1}} F(x_{k-i})\| \le M_2 \sum_{j=0}^{i} \|r_{k-j}\|^2.$$

Furthermore, for $w_k^i = \sum_{j=0}^{i} \gamma_j^k \Delta x_{k-j}^{(k)}$ and $v_k^i = \operatorname{Exp}_{x_k}(-w_k^i)$ it holds

$$\|\mathcal{P}_{v_k^i}^{x_k} F(v_k^i) - \mathcal{P}_{v_k^{i-1}}^{x_k} F(v_k^{i-1}) - \gamma_i^k (\mathcal{P}_{y_k^i}^{x_k} F(y_k^i) - \mathcal{P}_{y_k^{i-1}}^{x_k} F(y_k^{i-1}))\| \le M_3 \sum_{j=0}^{i} \|r_{k-j}\|^2.$$

Now let us prove Theorem 7.

Proof First we assume there are some $x_1, \dots, x_k \in B_{\tilde{r}}(\Sigma_*)$ with $\tilde{r} < \frac{r}{2L_2(mM_1+\beta+1)}$. Define

$$\begin{cases} \bar{x}_k = \operatorname{Exp}_{x_k}(-X_k\Gamma_k) \\ x_{k+1} = \operatorname{Exp}_{x_k}(-X_k\Gamma_k + \beta_k\bar{r}_k) \\ \tilde{x}_{k+1} = \operatorname{Exp}_{x_k}(-X_k\Gamma_k + \bar{r}_k) \end{cases}$$

with $\bar{r}_k = r_k - R_k\Gamma_k$. Then $\|\bar{r}_k\| \le \|r_k\|$.

We observe

$$\begin{aligned} \|F(x_{k+1})\| \le& L_2 W_2(x_{k+1}, x_*) \\ \le& L_2(W_2(x_{k+1}, x_k) + W_2(x_k, x_*)) \\ \le& L_2(W_2(x_{k+1}, x_k) + \tilde{r}) \\ \le& L_2(\|X_k\Gamma_k + \beta\bar{r}_k\| + \tilde{r}) \\ \le& L_2(M_1(\sum_{i=0}^{\min\{k,m\}} \|r_{k-i}\| + \beta\|r_k\|) + \tilde{r}) \\ \le& L_2(M_1 m\tilde{r} + \beta\tilde{r} + \tilde{r}) \\ \le& L_2(M_1 m + \beta + 1)\tilde{r} \le \frac{1}{2}r \end{aligned}$$

Similarly, $W_2(\bar{x}_k, x_*), W_2(\tilde{x}_{k+1}, x_*) \le L_2(M_1 m + \beta + 1)\tilde{r}$.

Hence, with Lemma 23 we get

$$\begin{aligned} \|r_{k+1}\| =& W_2(G(x_{k+1}), x_{k+1}) \\ \le& W_2(G(x_{k+1}), G(\bar{x}_k)) + W_2(G(\bar{x}_k), x_{k+1}). \end{aligned}$$

By assumption,

$$W_2(G(x_{k+1}), G(\bar{x}_k)) \le \kappa W_2(\operatorname{Exp}_{x_k}(X_k\Gamma_k), \operatorname{Exp}_{x_k}(X_k\Gamma_k + \beta_k\bar{r}_k)) \le \kappa\|\beta_k\bar{r}_k\|.$$

Now

$$W_2(x_{k+1}, G(\bar{x}_k)) \le W_2(G(\bar{x}_k), \tilde{x}_{k+1}) + W_2(\tilde{x}_{k+1}, x_{k+1}).$$

By assumption

$$W_2(\tilde{x}_{k+1}, x_{k+1}) \le \|X_k\Gamma_k - \beta_k\bar{r}_k - X_k\Gamma_k + \bar{r}_k\| = (1 - \beta_k)\|\bar{r}_k\|.$$

Also

$$W_2(G(\bar{x}_k), \tilde{x}_{k+1}) \le W_2(G(\bar{x}_k), \operatorname{Exp}_{\bar{x}_k}(\mathcal{T}_{x_k}^{\bar{x}_k}\bar{r}_k)) + W_2(\operatorname{Exp}_{\bar{x}_k}(\mathcal{T}_{x_k}^{\bar{x}_k}\bar{r}_k), \tilde{x}_{k+1}).$$

By Lemma 5 in Li and Bao (2023), using the footnote 3 of Sun et al. (2019) regarding generalizing beyond complete manifolds by restricting the tangent vectors to be inside the injectivity radius, there is a constant c_0 depending on the sectional curvature such that

$$W_2(\operatorname{Exp}_{\bar{x}_k}(\mathcal{T}_{x_k}^{\bar{x}_k}\bar{r}_k), \tilde{x}_{k+1}) \le c_0 \underbrace{\min\{\|X_k\Gamma_k\|, \|\bar{r}_k\|\}}_{\le\tilde{r}}(\|X_k\Gamma_k\| + \|\bar{r}_k\|)^2.$$

With Lemma 23 we get

$$(\|X_k\Gamma_k\| + \|\bar{r}_k\|)^2 \le mM_1^2 \sum_{j=0}^{k} \|r_{k-j}\|^2.$$

Furthermore,

$$W_2(G(\bar{x}_k), \operatorname{Exp}_{\bar{x}_k}(\mathcal{T}_{x_k}^{\bar{x}_k}\bar{r}_k)) = W_2(\operatorname{Exp}_{\bar{x}_k}(-F(\bar{x}_k)), \operatorname{Exp}_{\bar{x}_k}(\mathcal{T}_{x_k}^{\bar{x}_k}\bar{r}_k)) \leq \| - F(\bar{x}_k) - \mathcal{T}_{x_k}^{\bar{x}_k}\bar{r}_k\|.$$

Defining $w_k^i = \sum_{j=0}^i \gamma_j^k \Delta x_{k-j}^{(k)}$, $z_k^i = \sum_{j=1}^i \Delta x_{k-j}^{(k)}$, $v_k^i = \operatorname{Exp}_{x_k}(-w_k^i)$ and $y_k^i = \operatorname{Exp}_{x_k}(-z_k^i)$ we get by Lemma 24

$$\begin{aligned}
\| - F(\bar{x}_k) - \mathcal{T}_{x_k}^{\bar{x}_k}\bar{r}_k\| &= \|\mathcal{T}_{\bar{x}_k}^{x_k}F(\bar{x}_k) + \bar{r}_k\| \\
&= \|\mathcal{T}_{\bar{x}_k}^{x_k}F(\bar{x}_k) - F(x_k) - \sum_{i=1}^k \gamma_i^k \Delta r_{k-i}\| \\
&\leq \underbrace{\|\mathcal{P}_{v_k^i}^{x_k}F(v_k^i) - \mathcal{P}_{v_k^{i-1}}^{x_k}F(v_k^{i-1}) - \gamma_i^k(\mathcal{P}_{y_k^i}^{x_k}F(y_k^i) - \mathcal{P}_{y_k^{i-1}}^{x_k}F(y_k^{i-1}))\|}_{\leq M_3 \sum_{j=0}^i \|r_{k-j}\|^2} \\
&\quad + \underbrace{\sum_{i=1}^k |\gamma_i^k| \| - \Delta r_{k-i} + \mathcal{P}_{y_k^i}^{x_k}F(y_k^i) - \mathcal{P}_{y_k^{i-1}}^{x_k}F(y_k^{i-1})\|}_{\substack{\leq M_\Gamma(\|\mathcal{T}_{x_{k-1}}^{x_k}\cdots\mathcal{T}_{x_{k-i+1}}^{x_{k-i+2}}r_{k-i+1} + \mathcal{P}_{y_k^{i-1}}^{x_k}F(y_k^{i-1})\| \\ + \|\mathcal{T}_{x_{k-1}}^{x_k}\cdots\mathcal{T}_{x_{k-i}}^{x_{k-i+1}}r_{k-i} + \mathcal{P}_{y_k^i}^{x_k}F(y_k^i)\|)}} \\
&\leq (M_3 + 2M_\Gamma M_2)\sum_{i=0}^k \|r_{k-i}\|^2.
\end{aligned}$$

Altogether, we get

$$\|r_{k+1}\| \leq (1 - \beta_k + \kappa\beta_k)\|\bar{r}_k\| + (c_0\tilde{r}mM_1^2 + M_3 + 2M_\Gamma M_2)\sum_{i=0}^k \|r_{k-i}\|^2.$$

Now we want to show that the history produced by the algorithm satisfies these assumptions. To that end we use induction. Define $\hat{r} = \min\{\frac{r}{L_2(mM_1+\beta+1)}, \frac{L_1 - L_2 + (1-\kappa)\beta L_2}{MmL_2^2}\}$. If $W_2(x_0, x_*) < \frac{\hat{r}}{1+L_2}$ we get

$$\|F(x_0)\| \leq L_2 W_2(x_0, x_*) \leq \frac{L_2}{1+L_2}\hat{r}$$

and

$$\begin{aligned}
W_2(x_1, x_*) &\leq W_2(x_1, x_0) + W_2(x_0, x_*) \\
&\leq \underbrace{\|\operatorname{Exp}_{x_0}^{-1}(x_1)\|}_{=\|r_0\|} + \frac{1}{1+L_2}\hat{r} \\
&\leq \hat{r}.
\end{aligned}$$

Assume now that the claim holds for all $j \leq k$. Then by above considerations we have

$$\begin{aligned}
\|r_{k+1}\| &\leq (1 - \beta_k + \kappa\beta_k)\|r_k\| + M\sum_{j=0}^k \|r_{k-j}\|^2 \\
&\leq (1 - \beta_k + \kappa\beta_k)L_2\hat{r} + MmL_2^2\hat{r}^2 \\
&\leq L_1\hat{r}.
\end{aligned}$$

Then we have

$$W_2(x_{k+1}, x_*) \leq \frac{1}{L_1}\|r_{k+1}\| \leq \hat{r}$$

and the claim follows by induction. ∎

Appendix B. Additional numerical results

For all the mentioned problems, a comprehensive set of numerical experiments has been conducted. We study the performance of the method depending on both the problem settings and the hyperparameters of the numerical solver. The selection of the former will be described for each problem individually. As for the method hyperparameters, for the BWRAM method we vary the relaxation parameter β_k, the maximal number of history vectors m, and the constraint parameter $l_{\infty,\max}$. The number of history vectors varies in the range from 1 to 15. In the current work, these are constant during the iterations. In principle, adaptive strategies for restarting the iteration (Novák et al., 2023; Wei et al., 2023) or for selecting the relaxation parameter (Irons and Tuck, 1969) could be employed, but these strategies are outside the scope of the current work.

B.1 Ornstein–Uhlenbeck process

For the Ornstein–Uhlenbeck process and KL minimization, we consider multiple invariant distributions with covariance matrices being i.i.d. realizations of the following random variable:

$$\Sigma_* = Q^T D Q, \quad D = \operatorname{diag}\left(\sigma_{\min} + \frac{i-1}{d-1}(\sigma_{\max} - \sigma_{\min})\right)_{i=1}^{d}, \quad Q \text{ a Haar-distributed random orthogonal matrix.}$$

We fix $\sigma_{\min}$ and vary d and $\sigma_{\max}$. For each pair of d, $\sigma_{\max}$, we run the Picard iteration and BWRAM with different history lengths m. The acceleration measured as $\frac{N_{\text{Picard}}}{N_{\text{BWRAM}}}$ is averaged over several independent realizations of Q.

For the Ornstein–Uhlenbeck problem, the relaxation parameter and the constraint parameter are set to $\beta = 1.0$ and $l_{\infty,\max} = 1.0$. In Table 1a, one can see the maximal mean acceleration (averaged over 6 independent runs) of BWRAM, compared to Picard.

B.2 KL minimization

For the KL minimization problem, where the operator takes the form

$$G(\Sigma) = \operatorname{Exp}_\Sigma(-h\partial_W \operatorname{KL}(\Sigma|\Sigma_*)),$$

the scaling parameter is set to $h = 0.3$, and the values of d, $\sigma_{\min}$, $\sigma_{\max}$ are the same as in the Ornstein–Uhlenbeck problem. As for the hyperparameters of BWRAM, the constraint parameter takes values $l_{\infty,\max} \in \{1.5, 10.0\}$. In the case of fixed-point problems with an operator defined by the Wasserstein gradient of some functional $\mathcal{E}$ as $\operatorname{Exp}(-h\partial_W \mathcal{E})$, the optimal step size h may have different values for the basic Picard method and for RAM. From the structure of the update formula in Algorithm 2, one concludes that using relaxation $\beta \neq 1$ is equivalent to using RAM with $\tilde{\beta} = 1$ and $\tilde{h} = \beta h$. Thus, one can compensate for the mismatch of RAM and Picard step size by choosing the relaxation factor β_k smaller or larger than one. Empirically it has been observed that AM can converge even in cases where Picard would not (Willert et al., 2014). Thus, one can leverage over-relaxation, i.e., setting β larger than 1. We test the idea by varying the relaxation factor in the KL problem in the set $\beta \in \{0.9, 1.0, 1.2\}$.

For the KL problem, we try to find heuristics for the choice of the method's hyperparameters. The average number of iterations, depending on the choice of l_∞ norm bound, relaxation β and the number m of history vectors, is presented in Figure 5. In our experiments, the choice of the l_∞ constraint has proven the most important. With the optimal value of the constraint, the method performs well with a large enough history ($m \geq 3$), with the optimal m frequently being in the range 3–10.

Even though the performance is mostly affected by the choice of one parameter, the dependence on its manual tuning hinders practical applications. Approaches with adaptive regularization have been considered in Wei et al. (2021). Adaptive choice of the method's step size h can also be employed to

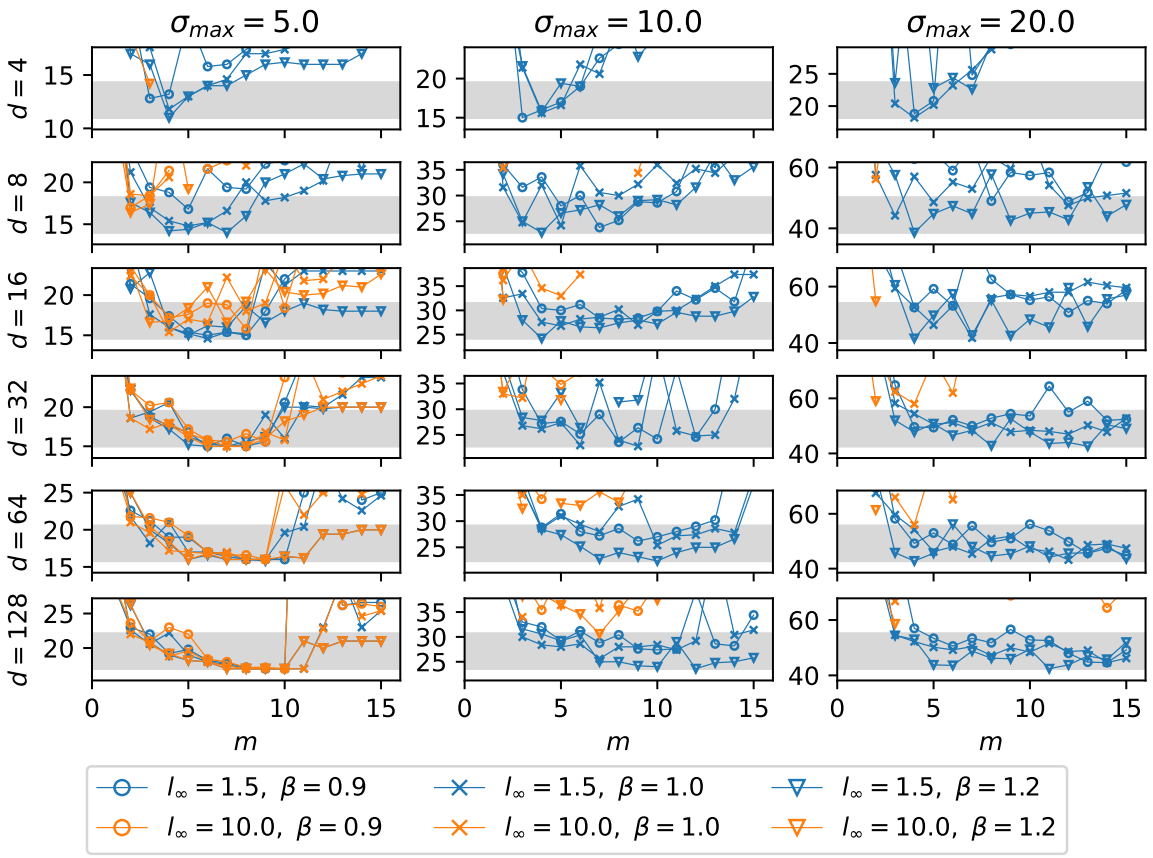


Figure 5: Hyperparameter search for the KL minimization problem. Each subplot represents a combination of the problem's parameters. Each curve presents the number of BWRAM steps to convergence, averaged over $n = 6$ random realizations of the problem, depending on the number m of history vectors, with fixed values of relaxation β and l_∞ norm constraint in the least-squares problem. Only solutions with number of steps within 60% of the minimal are depicted. The most sensitive parameter is the constraint l_∞, with smaller value 1.5 (blue) performing significantly better than the larger one (orange). Across the experiments, we see that larger value of β improves convergence speed. As for m, the optimal value is typically between 5 and 10. For a broad range of parameters, the number of steps is within 30% of the optimal (grey area).

improve the method's robustness. We speculate that standard line-search and trust-region methods, as well as simplified techniques from quasi-Newton methods (Agafonov et al., 2025) can be extended to fixed-point problems, but the detailed investigation of this question is left to future work.

We also compare our method to established methods of Riemannian optimization. In particular, we consider RGD with backtracking line search and RCG, provided by the `pymanopt` (Townsend et al., 2016) package. The results of the experiments are presented in Table 3. Here, the number of iterations is first averaged over $n = 6$ runs for random Σ_*, and then the minimal one is chosen among different values of the hyperparameters. The best result for each problem is marked in bold.

B.3 Averaging problems

In the case of the averaging problems (Barycenter, Entropic Barycenter, Median), the parameters of the invariant distributions are the dimension d and the number of distributions n_σ, taking values

Table 3: The mean number of steps for the KL minimization problem for each method.

d	$\sigma_{\max}$	BWRAM	Picard	RCG	RGD
4	5.0	**11.0**	45.0	25.0	53.0
	10.0	**15.0**	95.0	42.0	113.0
	20.0	**18.2**	196.0	56.0	162.0
8	5.0	**14.0**	45.0	33.0	39.0
	10.0	**22.8**	96.0	53.0	124.0
	20.0	**38.6**	197.0	68.0	243.0
16	5.0	**14.6**	46.0	29.0	64.0
	10.0	**24.2**	97.0	39.0	109.0
	20.0	**41.6**	200.0	69.0	230.0
32	5.0	**15.0**	47.0	31.0	51.0
	10.0	**22.8**	100.0	42.0	110.0
	20.0	**42.6**	206.0	66.0	265.0
64	5.0	**15.8**	49.0	31.0	58.0
	10.0	**22.4**	103.0	57.0	117.0
	20.0	**42.8**	212.0	68.0	233.0
128	5.0	**17.0**	51.0	30.0	51.0
	10.0	**23.6**	107.0	54.0	124.0
	20.0	**42.4**	220.0	74.0	273.0

from the sets $d \in \{3, 5, 10, 20, 50\}$ and $n_\sigma \in \{3, 5, 10, 20\}$, respectively. The number of iterations for each method is averaged over 5 random realizations of the distributions Σ_i, which are drawn i.i.d. from the Wishart distribution $\mathcal{W}_d(\mathrm{Id}, d)$. The entropic regularization parameter in the Entropic Barycenter problem is set to $\gamma = 0.001$ and the smoothing parameter in the Median problem is set to $\varepsilon = 0.001$. In all three experiments, the iteration proceeds until the cost converges to the minimum up to the tolerance of $\epsilon = 10^{-10}$. As for the hyperparameters of the method, for the Barycenter and Entropic Barycenter problems, they are the same as for the KL problem. For the Median problem, a broader range of relaxations and regularizations was considered, namely $l_{\infty,\max} \in \{0.1,\ 5.0,\ 15.0\}$ and $\beta_k \in \{1.0,\ 5.0,\ 10.0\}$. The mean number of steps taken by each method with the best set of hyperparameters is given in Table 4. We can see that BWRAM always provides acceleration compared to plain Picard iteration, in fact more than an order of magnitude in certain cases. Riemannian methods are also outperformed in most cases, the exception being the Median problem in dimension $d = 50$ with $n_\sigma \in \{3, 5\}$.

In these cases, the mean number of iterations for RGD is lower. We note, however, that the performance of BWRAM in that case is still comparable, and the difference of these means is within the variance (w.r.t. random realizations of the problem).

Table 4: Mean number of iterations for the three averaging problems: Barycenter, Entropic Barycenter and Median. The best result for each pair of $(d,\ n_\sigma)$ is given in bold.

		Barycenter				Entropic Barycenter				Median			
		BWRAM	Picard	RCG	RGD	BWRAM	Picard	RCG	RGD	BWRAM	Picard	RCG	RGD
d	n_σ												
3	3	**5.4**	6.2	31.6	14.4	**4.4**	6.6	15.8	8.8	**12.8**	27.0	31.0	30.3
	5	**7.0**	9.8	29.0	16.6	**5.8**	10.6	12.2	9.0	**13.8**	26.5	24.5	21.5
	10	**6.8**	8.6	28.6	15.8	**4.8**	10.4	16.4	9.8	**13.4**	23.3	22.4	15.8
	20	**6.6**	9.0	26.2	15.2	**5.0**	9.2	13.4	10.0	**13.3**	22.9	26.1	16.3
5	3	**7.4**	9.8	32.4	18.0	**6.4**	8.8	14.2	9.0	**23.2**	72.1	32.5	33.3
	5	**7.2**	9.6	34.8	18.2	**5.6**	8.2	15.8	9.8	**13.5**	46.1	24.5	17.2
	10	**7.4**	10.6	38.4	18.8	**6.0**	9.0	17.0	10.6	**11.2**	41.8	21.3	16.6
	20	**7.2**	10.0	36.0	19.0	**5.0**	9.4	18.4	10.8	**10.0**	39.3	22.1	14.8
10	3	**8.6**	13.6	26.4	19.2	**7.0**	14.2	9.0	8.6	**11.9**	120.9	30.8	25.2
	5	**9.0**	13.8	33.4	20.4	**6.2**	13.2	14.0	11.2	**10.3**	94.1	26.7	18.7
	10	**8.8**	13.8	40.6	22.6	**6.6**	11.0	15.6	12.6	**10.4**	89.7	22.7	15.6
	20	**8.0**	11.2	39.4	22.0	**5.4**	11.4	16.2	11.0	**9.8**	82.8	16.5	13.5
20	3	**10.6**	20.6	34.6	23.8	**7.8**	23.2	11.2	11.4	**11.8**	245.7	28.5	23.9
	5	**11.2**	19.0	36.6	21.6	**7.0**	23.8	12.2	11.0	**11.3**	206.0	30.5	19.7
	10	**10.0**	17.0	44.6	22.8	**7.0**	19.4	15.2	12.2	**10.8**	192.8	27.5	18.5
	20	**8.8**	13.4	44.8	21.8	**6.8**	15.8	18.0	12.4	**10.6**	177.5	14.8	14.5
50	3	**13.4**	26.4	37.2	30.8	**8.6**	23.4	16.0	14.0	28.4	707.5	32.2	**28.0**
	5	**12.8**	23.8	50.2	29.2	**8.6**	28.8	21.8	14.6	26.5	601.3	35.1	**22.8**
	10	**11.2**	18.6	44.4	27.4	**7.0**	22.4	21.0	14.4	**25.5**	506.9	44.3	25.8
	20	**9.2**	14.2	44.0	26.4	**6.4**	15.8	19.8	13.8	**20.5**	461.2	40.5	23.5

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Artem Agafonov, Vladislav Ryspayev, Samuel Horváth, Alexander Gasnikov, Martin Takáč, and Slavomir Hanzely. Simple stepsize for quasi-newton methods with global convergence guarantees. *arXiv preprint arXiv:2508.19712*, 2025.

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

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

Jason Altschuler, Sinho Chewi, Patrik R Gerber, and Austin Stromme. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*, 34:22132–22145, 2021.

Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2): 744–762, 2016.

Luigi Ambrosio and Nicola Gigli. Construction of the parallel transport in the Wasserstein space. 2008.

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

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.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones mathematicae*, 37(2):165–191, 2019.

Shi Chen, Qin Li, Oliver Tse, and Stephen J Wright. Accelerating optimization over the space of probability measures. *Journal of Machine Learning Research*, 26(31):1–40, 2025.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 2016. URL `https://stanford.edu/~boyd/papers/pdf/cvxpy_paper.pdf`. To appear.

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward gaussian variational inference via jko in the bures-wasserstein space. In *International Conference on Machine Learning*, pages 7960–7991. PMLR, 2023.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023.

Claire Evans, Sara Pollock, Leo G Rebholz, and Mengying Xiao. A proof that anderson acceleration improves the convergence rate in linearly converging fixed-point methods (but not in those converging quadratically). *SIAM Journal on Numerical Analysis*, 58(1):788–810, 2020.

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

Zoran Gajic and Muhammad Tahir Javed Qureshi. *Lyapunov matrix equation in system stability and control*. Courier Corporation, 2008.

Nicola Gigli. On the geometry of the space of probability measures in $\mathbb{R}^n$ endowed with the quadratic optimal transport distance. 2008.

Andi Han, Bamdev Mishra, Pratik Kumar Jawanpuria, and Junbin Gao. On riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. *Advances in Neural Information Processing Systems*, 34:8940–8953, 2021.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97. URL `https://doi.org/10.1093/biomet/57.1.97`.

Bruce M Irons and Robert C Tuck. A version of the Aitken accelerator for computer iteration. *International Journal for Numerical Methods in Engineering*, 1(3):275–277, 1969.

Lee Jaejin and Gyu Joo Han. Convergence analysis of fixed-point iteration with anderson acceleration on a simplified neutronics/thermal-hydraulics system. *Nuclear Engineering and Technology*, 54(2):532–545, 2022. ISSN 1738-5733. doi: https://doi.org/10.1016/j.net.2021.08.005. URL `https://www.sciencedirect.com/science/article/pii/S1738573321004915`.

Anya Katsevich and Philippe Rigollet. On the approximation accuracy of gaussian variational inference. *The Annals of Statistics*, 52(4):1384–1409, 2024.

Carson Kent, Jiajin Li, Jose Blanchet, and Peter W Glynn. Modified Frank Wolfe in probability space. *Advances in Neural Information Processing Systems*, 34:14448–14462, 2021.

Zanyu Li and Chenglong Bao. Riemannian Anderson mixing methods for minimizing C^2-functions on Riemannian manifolds. *arXiv preprint arXiv:2309.04091*, 2023.

Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, Jun Zhu, and Lawrence Carin. Accelerated first-order methods on the Wasserstein space for Bayesian inference. *stat*, 1050:4, 2018.

Hailiang Liu, Jia-Hao He, and Xuping Tian. Anderson acceleration of gradient methods with energy for optimization problems. *Communications on Applied Mathematics and Computation*, 6(2): 1299–1318, 2024.

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

Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein riemannian geometry of Gaussian densities. *Information Geometry*, 1:137–179, 2018.

Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Bridging bayesian and minimax mean square error estimation via wasserstein distributionally robust optimization. *Mathematics of Operations Research*, 48(1):1–37, 2023.

Matyáš Novák, Jiří Vackář, Robert Cimrman, and Ondřej Šipr. Adaptive Anderson mixing for electronic structure calculations. *Computer Physics Communications*, 292:108865, 2023.

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

Ke Sun, Yafei Wang, Yi Liu, Bo Pan, Shangling Jui, Bei Jiang, Linglong Kong, et al. Damped anderson mixing for deep reinforcement learning: Acceleration, convergence, and stabilization. *Advances in Neural Information Processing Systems*, 34:3732–3743, 2021.

Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on riemannian manifolds. In *Neural Information Processing Systems*, 2019. URL `https://api.semanticscholar.org/CorpusID:189998682`.

Asuka Takatsu. On wasserstein geometry of Gaussian measures. *Probabilistic approach to geometry*, 57:463–472, 2010.

Asuka Takatsu. Wasserstein geometry of Gaussian measures. 2011.

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

James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016. URL `http://jmlr.org/papers/v17/16-177.html`.

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

Melanie Weber and Suvrit Sra. Riemannian optimization via Frank-Wolfe methods. *Mathematical Programming*, 199(1):525–556, 2023.

Fuchao Wei, Chenglong Bao, and Yang Liu. Stochastic anderson mixing for nonconvex stochastic optimization. *Advances in Neural Information Processing Systems*, 34:22995–23008, 2021.

Fuchao Wei, Chenglong Bao, and Yang Liu. A class of short-term recurrence Anderson mixing methods and their applications. In *International Conference on Learning Representations*, 2022.

Fuchao Wei, Chenglong Bao, Yang Liu, and Guangwen Yang. Convergence analysis for restarted Anderson mixing and beyond. *arXiv preprint arXiv:2307.02062*, 2023.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Jeffrey Willert, William T Taitano, and Dana Knoll. Leveraging anderson acceleration for improved convergence of iterative solutions to transport systems. *Journal of Computational Physics*, 273: 278–286, 2014.