2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Moving Average Estimation by Geometric Optimization

Alex Nguyen-Le and Nikolai Matni

***Abstract*— We propose a new geometric-optimization framework for maximum likelihood estimation of moving-average models. Instead of optimizing directly over the moving average parameters, we formulate the estimation problem over the reflection coefficients and show how to perform gradient descent over a reflection-coefficient manifold. This choice leads to simpler expressions in the objective function and in the constraints, which can yield more convenient expressions for theoretical analysis. Finally, we numerically implement and compare the proposed estimation schemes in the reflection coefficients to those based on moving-average parmeterizations. We show that our novel formulation works in practice and yields equivalent solutions to currently employed formulations.**

I. INTRODUCTION

Noise signals encountered in practice are often correlated across time in a manner that is well described by moving-average (hereafter MA(q)) models,

$$x_t = b_0 w_t + b_1 w_{t-1} + \ldots + b_q w_{t-q}, \tag{1}$$

where w_t is a scalar zero mean Gaussian white-noise signal, and q is the order of the moving average model. Unfortunately this simplicity in form is not met with straightforward parameter identification, as a direct maximum-likelihood (ML) formulation yields a nonconvex optimization problem. There are other choices of objective, such as spectral fitting [1], but our we focus on likelihood-based formulations due to their broader compatibility with model-selection principles. Parameter estimation typically proceeds by sweeping over possible orders q and then selecting a model by examining the corresponding likelihoods [2].

Directly attacking this problem by writing a likelihood over the model coefficients $\beta = [b_0, \ldots, b_q]$ and observed data sequence, $\xi = [x_0, \ldots, x_{T-1}]$ leads to a nonconvex optimization problem, with both a nonconvex objective function and a nonconvex feasible set. We refer to this problem as the maximum likelihood moving average (ML-MA) problem. State-of-the-art techniques for tackling the ML-MA problem [3] typically embed the MA model (1) into a state-space model to gain access to the Kalman Filter recursions, but these reparameterizations do not address the nonconvexity inherent to the original formulation; rather, they only speed up the gradient descent search step. Over the past two decades, there has been increasing interest in reparameterizations of this problem by its autocovariances $\gamma \in \mathbb{R}^q$, defined as:

$$\gamma_i := \sum_{j=0}^{q-i} b_j b_{j+i},\ i = 0, \ldots, q. \tag{2}$$

The set of autocovariances compatible with the above definition is semidefinite-representable, and hence convex. Furthermore, under suitable assumptions, the autocovariances $\{\gamma_i\}$ can be mapped back to the moving average parameters β [4], [5]. Several closely related problems, such as structured covariance estimation [6], have revealed that a likelihood objective function over the autocovariances γ is locally convex [7] in a region containing the global maximum with high probability, under suitable technical conditions. However, reconciling these technical conditions together into formulations amenable for analysis is challenging and unresolved; even studying the stationary points of the objective function is formidable [8].

We consider an adjacent problem to the autocovariance estimation and reparameterize the ML covariance estimation problem into one over its reflection coefficients, which are bijectively related to the autocovariances. This formulation leads to a novel manifold optimization problem, and, in many ways, grants simpler expressions to both the objective function and the constraints.

A. Main Results

1) We find new formulations for the ML-MA problem. These formulations are geometric, operating directly on the reflection coefficients, so we also connect MA and structured covariance estimation to reflection coefficient estimation.
2) We show how to formulate ML-MA and ML structured covariance problems so that they identify equivalent models, and implement all schemes in a directly comparable way. Further, we demonstrate these equivalences numerically to illustrate that practical optimization can be carried out using the geometric formulation we propose.

B. Organization

We begin by expressing ML-MA estimation as an optimization problem, which makes its connection to structured covariance estimation clear; this is the main subject of Section II. We show that structured covariance estimation is a relaxation to MA(q) estimation, and while we can make the problems completely equivalent with additional constraints, we leave this for Section IV. Enforcing strict equivalence too early complicates the analysis of Section III, where the bulk of the main analytic results are. In Section III, we

A. Nguyen-Le* and N. Matni are with the Department of Electrical and Systems Engineering, University of Pennsylvania, PA, USA. {atn,nmatni}@seas.upenn.edu. Corresponds to atn.

show how to pose the relaxed MA(q) problem as a manifold optimization problem, as well as derive a retraction, allowing for numerical optimization. In Section IV, we show how to use results from spectral factorization theory to constrain both the structured covariance and geometric formulations to account for their relaxation of the ML-MA problem. Finally, we demonstrate some numerical experiments in Section V, and end with conclusions in Section VI.

II. Moving Average Estimation

To motivate reparameterizing the ML-MA problem into a structured covariance estimation problem, we first analyze the feasible set of covariance matrices that arise from MA models, clarifying the connection of autocovariance parameterizations to the standard system identification formulation. This formulation is identical to the exact-likelihood formulation of [9]. Let $w := [w_0, \ldots, w_{t+q}]$ denote a (latent) zero mean Gaussian white-noise sequence and $\xi := \begin{bmatrix} x_0 & \ldots & x_{T-1} \end{bmatrix}$ denote the observed data sequence. Given a dataset ξ we begin by forming a likelihood optimization problem over the model parameters β. Since Gaussian random variables are fully specified by their mean and covariance, we first consider the feasible set.

Lemma 1 (MA $\Rightarrow$ Toeplitz Covariance): The observed data ξ is distributed as a multivariate Gaussian random variable with a Toeplitz covariance matrix of bandwidth q. A suitable symmetric matrix basis for the covariance matrix of ξ is:

$$C_k = \begin{bmatrix} 0 & \cdots & 1 & & \\ \vdots & & & \ddots & \\ 1 & & & & 1 \\ & \ddots & & & \vdots \\ & & 1 & \cdots & 0 \end{bmatrix}, \quad [C_k]_{ij} = \begin{cases} 1 & |j-i| = k \\ 0 & \text{otherwise} \end{cases},$$

where every $C_k \in \mathbb{R}^{T \times T}$.

Proof: Since,

$$\begin{bmatrix} x_0 \\ \vdots \\ x_{T-1} \end{bmatrix} = \underbrace{\begin{bmatrix} b_q & \cdots & b_0 & & \\ & \ddots & & \ddots & \\ & & b_q & \cdots & b_0 \end{bmatrix}}_{B} \begin{bmatrix} w_{-q} \\ \vdots \\ w_{T-1} \end{bmatrix}$$

we have that ξ is a zero-mean multivariate Gaussian and:

$$[\text{cov}(\xi)]_{i,j} = [B\text{cov}(w)B^{\mathsf{T}}]_{i,j} = \sum_{k=0}^{q-|j-i|} b_k b_{k+|j-i|} = \gamma_{|j-i|}$$

which describes a Toeplitz matrix. The symmetric matrix basis we have chosen accommodates individual contributions of the autocovariances γ_k, as in (2), to the bands of the covariance matrix. ∎

We note that expressing the autocovariances γ in terms of the MA parameters β readily yields the ML-MA problem described in the introduction. The ML-MA estimation problem is given as,

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & \tfrac{1}{2}\log\det(\Sigma(\beta)) + \tfrac{1}{2}\text{tr}\big(\xi\xi^*\Sigma(\beta)^{-1}\big) \\ \text{subject to:} \quad & \Sigma(\beta) = \sum_{i=0}^{q}\sum_{j=0}^{q-i} b_j b_{j+i} C_i \\ & b_0 z^0 + \ldots + b_q z^q \neq 0 \text{ for all } |z| > 1 \end{aligned} \tag{3}$$

where the objective function is the likelihood function of a Gaussian random variable with covariance $\Sigma(\beta)$ given observation ξ (up to an additive constant). The last constraint is often referred to as an invertibility or stability requirement, and leads to an identifiable model [9]. It requires that all roots of the polynomial in z be smaller than 1 in magnitude, and resolves the identifiability issue that different values of MA-parameters β can have the same autocovariances γ, and are therefore equally valid solutions [10]. The ML-MA estimation problem (3) is a common starting point for moving average estimation in time-series analysis/system identification [9], [10].

The objective function of (3) is often rewritten in terms of an LDL^* factorization of $\Sigma(\beta)$, or in terms of Kalman Filter recursions [9], with the constraint that β defines a stable polynomial implicitly enforced [9]. This constraint complicates analysis because the set of stable polynomials coefficients is nonconvex, and hence one might instead consider directly estimating the autocovariances γ.

This direct formulation over autocovariances γ is based on semidefinite programming, and leads to the following structured covariance estimation problem, which we show to be a relaxation of the ML-MA problem (3). While the problem is still nonconvex, the nonconvexity is now only present in the objective function, as the feasible set is semidefinite representable. For two symmetric matrices A, B of compatible size, we write $A \succ B$ to denote that $A - B$ is positive definite.

Lemma 2 (Structured Covariance Estimation): The ML structured covariance estimation problem,

$$\begin{aligned} \underset{\gamma}{\text{minimize}} \quad & \tfrac{1}{2}\log\det(\Sigma(\gamma)) + \tfrac{1}{2}\text{tr}\big(\xi\xi^*\Sigma(\gamma)^{-1}\big) \\ \text{subject to:} \quad & \Sigma(\gamma) = \gamma_0 C_0 + \ldots + \gamma_q C_q \succ 0 \end{aligned} \tag{4}$$

is a relaxation to the ML-MA problem (3).

Proof: We construct an explicit example of a positive definite Toeplitz covariance matrix with no associated MA parameters. Assume for a contradiction that there exists b_0, b_1 such that:

$$\Sigma_2 = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} b_1 & b_0 & \\ & b_1 & b_0 \end{bmatrix} \begin{bmatrix} b_1 & b_0 & \\ & b_1 & b_0 \end{bmatrix}^*$$

Σ_2's eigenvalues are $\{1, 7\}$, so it is feasible. By construction of the autocovariances, we must also have that:

$$\Sigma_3 = \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & 3 \\ 0 & 3 & 4 \end{bmatrix} = \begin{bmatrix} b_1 & b_0 & & \\ & b_1 & b_0 & \\ & & b_1 & b_0 \end{bmatrix} \begin{bmatrix} b_1 & b_0 & & \\ & b_1 & b_0 & \\ & & b_1 & b_0 \end{bmatrix}^*$$

The right hand side of this equation implies that Σ_3 is positive definite. However, Σ_3's eigenvalues are $\{-0.24, 4, 8.24\}$, which contradicts its positive-definiteness

requirement. It follows that no b_0 or b_1 can exist satisfying our original specification. Our example for the two-observation case can be modified as needed for more observations. ∎

Succinctly put, the structured covariance estimation problem (4) is a relaxed MA(q) problem because we do not enforce the existence of moving average parameters, that is:

$$\exists \beta : \Sigma(\gamma) = \Sigma(\beta), \quad b_0 z^0 + \ldots + b_q z^q \neq 0 \text{ for all } |z| > 1$$

The relaxed problem (4) will be our starting point for our analysis and estimation because it is simpler. We defer further constraining the covariance matrix to enforce equivalence with the original MA(q) setting, as described in [5] or [11], for later.

III. A Geometric Formulation for Structured Covariance Estimation

This section defines and analyzes a manifold optimization-based approach to structured covariance estimation. We provide a short roadmap for this section in which we introduce and leverage tools from the manifold optimization literature [12]. To pose the structured covariance estimation problem (4), we first need to show that the feasible set of problem (4) is a manifold: we do so by establishing a diffeomorphism between the feasible set and a Euclidean space using the Levinson-Trench algorithm. Once we have a manifold, we need a way to move around on it, say for implementing gradient descent. A simple tool for accomplishing this is a retraction map, which in our case, can be implemented via projection.

A. Levinson-Trench Algorithm

The celebrated Levinson-Trench algorithm [13], [14] directly computes the Cholesky factors for the inverse of a Toeplitz matrix, and we provide a very explicit construction of this function because we rely on its properties to define a manifold. Consider a positive definite Toeplitz matrix, Σ_T, and its associated inverse, $P = \Sigma_T^{-1}$.

$$\Sigma_T = \gamma_0 C_0 + \ldots + \gamma_{T-1} C_{T-1}$$

Let $P = UU^*$ be the upper-triangular Cholesky decomposition of P:

$$U = \begin{bmatrix} u_{00} & u_{01} & \cdots & u_{0,T-1} \\ & u_{11} & \cdots & u_{1,T-1} \\ & & \ddots & \vdots \\ & & & u_{T-1,T-1} \end{bmatrix}, \tilde{v}_i = \begin{bmatrix} u_{0i} \\ \vdots \\ u_{ii} \end{bmatrix}, \tilde{v}_0 = \begin{bmatrix} u_{00} \end{bmatrix},$$

where $\tilde{v}_i$ is the non-zero part of of column i, and v_i is the corresponding full column of U with the appropriate number of zeros. The coefficient u_0 also obeys the special relationship $u_0 = \gamma_0^{-\frac{1}{2}}$, and we will call it a *seed-value* because it is used to start the Levinson recursions, which encodes the relationship between successive columns of U. The Levinson recursions are:

$$\tilde{v}_{i+1} = \frac{1}{\sqrt{1-\rho_i^2}} \begin{bmatrix} 0 & 0 & 0 \\ 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \tilde{v}_i - \frac{\rho_i}{\sqrt{1-\rho_i^2}} \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \\ 0 & 0 & 0 \end{bmatrix} \tilde{v}_i. \quad (5)$$

The quantities, ρ_i are known as the reflection coefficients, belong to the open interval $(-1, 1)$ [13], and are given by the Trench equation:

$$\rho_i = \langle [\gamma_1, \ldots, \gamma_{i+1}], \tilde{v}_i \rangle u_{ii}^{1/2}, \quad i \in \{1, \ldots, T\}. \quad (6)$$

Our goal is to define a map from a seed value u_0 and the reflection coefficients ρ_i to the autocovariances γ. Further, we define the set of autocovariances Γ_T and seed value and reflection coefficients ϱ_i as:

$$\Gamma_T := \left\{ \gamma \in \mathbb{R}^T : \sum_{i=0}^{T-1} \gamma_i C_i \succ 0 \right\}, \; \varrho_T := \mathbb{R}_{>0} \times (-1, 1)^{T-1}.$$

We now define the function $f_{\Gamma_T \leftarrow \varrho_T} : \varrho_T \to \Gamma_T$ that takes as inputs a seed value u_0, and reflection coefficients $(\rho_1, \ldots, \rho_{T-1})$ and outputs autocovariances γ, as follows:

1) Compute the Levinson parameters $\tilde{v}_i$ (and zero-padded v_i) using the Levinson recursions (5).
2) Form $U = [v_0, \ldots, v_{T-1}]$.
3) Form $\Sigma_T^{-1} := UU^*$, take the inverse of the result to obtain a positive definite Toeplitz matrix Σ_T, and return the first column of the resulting matrix.

Similarly, we define the function $f_{\varrho_T \leftarrow \Gamma_T} : \Gamma_T \to \varrho_T$, which takes as input autocovariances γ and outputs a seed value u_0 and reflection coefficients $(\rho_1, \ldots, \rho_{T-1})$, as follows:

1) Compute $\Sigma_T = \sum_{i=0}^{T-1} \gamma_i C_i$.
2) Form the Cholesky decomposition, $LL^* = \Sigma_T$, and compute the inverse, $U = L^{-*}$
3) Using the columns of U, compute the reflection coefficients ρ_i using the Trench equations (6). Also return $\tilde{v}_0$ as the seed value.

In the sequel, we overload the terminology *reflection coefficients* to denote the vector containing both the seed value u_0 and the reflection coefficients ρ_i.

Theorem 1 (Reflection Coefficient Bijections): There is a bijection between the positive definite Toeplitz matrices described by γ and the reflection coefficients $(\gamma_0, \rho) := (\gamma_0, \rho_1, \ldots, \rho_T)$ if every $|\rho_i| < 1$.

Proof: First, we show that $f_{\varrho_T \leftarrow \Gamma_T} \circ f_{\Gamma_T \leftarrow \varrho_T}(\rho) = \rho$. The uniqueness of U in the Cholesky factorization guarantees that when we first obtain the upper triangular matrix in $f_{\Gamma_T \leftarrow \varrho_T}$ in step 2 of $f_{\Gamma_T \leftarrow \varrho_T}$, which we call U_1, and when we later obtain an upper triangular matrix in $f_{\varrho_T \leftarrow \Gamma_T}$ in step 2 of $f_{\varrho_T \leftarrow \Gamma_T}(\gamma)$, which we call U_2, we must obtain the same matrix $U_1 = U_2$. It follows that the original reflection coefficients will come out of the Trench equations so that $f_{\varrho_T \leftarrow \Gamma_T} \circ f_{\Gamma_T \leftarrow \varrho_T}(\rho) = \rho$. The other direction is the same, where we rely on the uniqueness of U to show that $f_{\Gamma_T \leftarrow \varrho_T} \circ f_{\varrho_T \leftarrow \Gamma_T}(\gamma) = \gamma$. ∎

Most surprisingly, this parameterization converts the semidefinite constraints to interval constraints,

$$\gamma_0 C_0 + \ldots + \gamma_{T-1} C_T \succ 0 \quad \Leftrightarrow \quad -1 < \rho_i < 1$$

which are considerably simpler. Further, if we ever obtain reflection coefficients ρ from $f_{\varrho_T \leftarrow \Gamma_T}(\gamma)$ that lie outside of the interval $(-1, 1)$, then the associated autocovariances do not parameterize a positive definite Toeplitz matrix. There are additional requirements on the reflection coefficients ρ, as they should describe *finite bandwidth* Toeplitz matrices. To enforce this constraint, we employ tools from manifold optimization theory.

B. Reflection Coefficients Manifold

We work with the following definition of an embedded submanifold of a Euclidean space.

Definition 1 (Thm. 3.12, [12]): Let $\mathcal{E}$ be a linear space of dimension T. A subset $\mathcal{M}$ is an embedded submanifold of $\mathcal{E}$ of dimension d if for all $x \in \mathcal{M}$, there exists an open neighborhood $\mathcal{U}$ of x, an open set $\mathcal{V} \subseteq \mathbb{R}^T$, and a diffeomorphism $F : \mathcal{U} \to \mathcal{V}$ such that $F(\mathcal{M} \cap \mathcal{U}) = \mathcal{V} \cap \mathcal{L}$, where $\mathcal{L}$ is the subspace $\{y \in \mathbb{R}^T : y_d = \ldots = y_{T-1} = 0\}$.

We now show that the reflection coefficients that describe a finite bandwidth Toeplitz matrix describe an embedded submanifold. We have slightly adapted the definition for our situation, as the first element in the vector γ has index 0. Leveraging the bijections developed in the previous section $f_{\varrho_T \leftarrow \Gamma_T}$ and $f_{\Gamma_T \leftarrow \varrho_T}$ to construct the diffeomorphism F, let

$$\Gamma_q := \Big\{ \gamma \in \mathbb{R}^T : \sum_{i=0}^{T-1} \gamma_i C_i \succ 0,\ \gamma_{q+1} = \ldots = \gamma_{T-1} = 0 \Big\}$$

denote the feasible set of the structured covariance estimation problem (4), i.e., the set of Toeplitz covariance matrices of bandwidth q. There are two equivalent definitions for the feasible set of reflection coefficients. Observe that we may use our diffeomorphism as part of a *defining function* description,

$$h(\rho) = \begin{bmatrix} [f_{\Gamma_T \leftarrow \varrho_T}(\rho)]_{q+1} \\ \vdots \\ [f_{\Gamma_T \leftarrow \varrho_T}(\rho)]_{T-1} \end{bmatrix} = 0,$$

with $\mathcal{M}_\varrho = \{\rho \in \mathbb{R}^T : h(\rho) = 0\}$ which is the typical way to define a manifold. The brackets are used to index specific outputs of $f_{\Gamma_T \leftarrow \varrho_T}$. We may also define $\mathcal{M}_\varrho$ as the image of Γ_q under $f_{\varrho_T \leftarrow \Gamma_T}$; the second definition is useful for establishing that $\mathcal{M}_\varrho$ is a manifold.

Theorem 2: Let $\mathcal{M}_\varrho = f_{\varrho_T \leftarrow \Gamma_T}\{\Gamma_q\}$ be the image of the set Γ_q of finite bandwidth positive definite Toeplitz matrices under $f_{\varrho_T \leftarrow \Gamma_T}$, that is:

$$\mathcal{M}_\varrho := \Big\{ (t, \rho) \in \mathbb{R}^T \Big| \exists \gamma \in \Gamma_q : (t, \rho) = f_{\varrho_T \leftarrow \Gamma_T}(\gamma) \Big\}.$$

Then $\mathcal{M}_\varrho$ is an embedded submanifold of dimension q.

Proof: Our definition of $f_{\Gamma_T \leftarrow \varrho_T}(\cdot)$ constitutes a diffeomorphism between the set of reflection coefficients and the autocovariances because it is bijective and can be constructed by sums, products, and compositions of smooth functions. Next, let $\mathcal{E} = \mathbb{R}^T$, and consider a point $x \in \mathcal{M}_\varrho$ with corresponding open neighborhood $\mathcal{U}$. Let $\mathcal{V} = f_{\Gamma_T \leftarrow \varrho_T}\{\mathcal{U}\}$ be the image of $\mathcal{U}$ under $f_{\Gamma \leftarrow \varrho}$. By definition of $\mathcal{M}_\varrho$, we have that $f_{\Gamma_T \leftarrow \varrho_T}\{\mathcal{M}_\varrho \cap \mathcal{U}\} = \mathcal{V} \cap L$, where $L = \{y \in \mathbb{R}^T : y_d = \ldots = y_{T-1} = 0\}$ is a subspace of $\mathcal{E}$. ∎

We conclude this section by stating the manifold optimization formulation for structured covariance estimation.

Corollary 1 (Geometric Covariance Estimation): The structured covariance estimation problem (4) is equivalent to:

$$\begin{aligned} \underset{\rho,\, \gamma_0}{\text{minimize}} \quad & T \log(\gamma_0) - \sum_{k=1}^{T-1} (T-k) \log(1 - \rho_i^2) \\ & + \sum_{k=0}^{T-1} \langle \upsilon_k, \xi \rangle^2 \\ \text{subject to:} \quad & (t_0, \rho) \in \mathcal{M}_\varrho \\ & \tilde{\upsilon}_{k+1} = \text{lev}(\rho_k, \tilde{\upsilon}_k), \ \ \tilde{\upsilon}_0 = \gamma_0^{-1/2}. \end{aligned} \tag{7}$$

Here, we use $\text{lev}(\rho_k, \tilde{\upsilon}_k)$ as shorthand for the right-hand side of the recursion (5).

Manipulating the optimization problem (4) to obtain problem (7) is mechanical but otherwise straightforward, and hence we defer the details to the Appendix. The objective function is now in a canonical form for the *curved* exponential family [15], which suggests that tasks such as certifying global optimality or finding sufficient statistics can be accomplished geometrically. However, we remark that this manifold formulation cannot resolve issues with spurious minima on its own because the stationary points of the *nonconvex* problem over Γ_q map to one another [16]—we return to this point later when discussing future work. For now, we return to finding a mechanism enabling us to move within the manifold to enable gradient descent. This is accomplished through the use of a suitably constructed *retraction*.

C. Retractions onto Reflection Coefficient Manifolds

To enable gradient steps on the reflection coefficient manifold, we define a retraction that associates curves on $\mathcal{M}_\rho$ to lines on its *tangent bundle*. This simplifies function evaluation and differentiation, as developed in [12]. We begin with necessary definitions.

Definition 2 (Tangent Spaces and Bundles): Let $\mathcal{M}$ be an embedded submanifold, and $x \in \mathcal{M}$. The tangent space at x, $\mathcal{T}_x\mathcal{M}$, is the subspace, $\ker\{Dh(x)\}$, where $Dh(x)$ is the Jacobian of the defining function h at x. The collection of tangent spaces, $\big\{(x, v) : x \in \mathcal{M}, v \in \ker\{Dh(x)\}\big\}$ is called the tangent bundle, which we denote by $\mathcal{TM}$.

Definition 3: Let $R : \mathcal{TM} \to \mathcal{M}$ be a differentiable map. R is said to be a local retraction if given $(x, v) \in \mathcal{TM}$ for v sufficiently small, we have:

$$R(x, 0) = x, \qquad DR(x, 0) \cdot v = v$$

In the interest of clarity, we occasionally use $X \cdot v$ in the sequel, for X and v matrices of compatible dimension, to denote matrix multiplication. We equip each tangent space

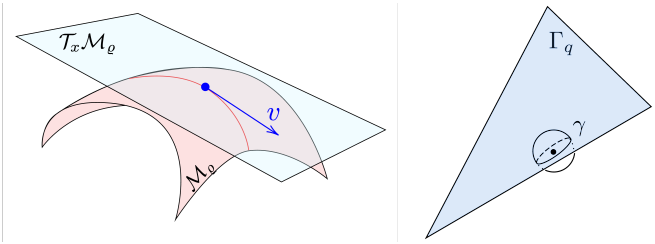

Fig. 1. The left subfigure illustrates a tangent space at x, and a retraction which maps the point $v \in \mathcal{T}_x\mathcal{M}_\varrho$ to a point in $\mathcal{M}_\varrho$. If we parameterize a curve in the tangent space by $c(t) = tv$, then we obtain the red curve on $\mathcal{M}_\varrho$. The right subfigure illustrates the intuition behind Lemma 3, where the projection to the set Γ_q is given by a subspace projection in small neighborhoods of $\gamma \in \Gamma_q$.

of $\mathcal{M}_\rho$ with the standard Euclidean Metric, giving us a Riemannian submanifold of the linear space, $\mathbb{R}^T$. This leads to gradients that align with their usual ones and (eventually) grants us compatibility with automatic differentiation when paired with a suitable retraction. We refer the reader to [12] for additional context and details.

Next we define a local retraction, which we illustrate in Figure 1 (left). We start with a technical lemma before providing an explicit construction.

Lemma 3 (Local Γ_q Projection): There exists open sets $\mathcal{U} \in \mathbb{R}^T$ where:

$$\text{Proj}_{\Gamma_q}\big(f_{\Gamma_q \leftarrow \varrho}\{\mathcal{U}\}\big) = \begin{bmatrix} I_q & \\ & \mathbf{0}_{T-q} \end{bmatrix} f_{\Gamma_q \leftarrow \varrho}\{\mathcal{U}\}$$

where $\text{Proj}_{\Gamma_q}(\cdot)$ is Euclidean projection onto a set.

This lemma encodes the idea that for sufficiently small perturbations to points in Γ_q, projection to the set Γ_q can be accomplished by subspace projection. We illustrate this in Figure 1 (right); the proof of this lemma can be found in the Appendix.

Theorem 3 (Retraction to $\mathcal{M}_\rho$): Let $R : T\mathcal{M} \to \mathcal{M}$, where:

$$R(x, v) = f_{\varrho_T \leftarrow \Gamma_T} \circ \text{Proj}_{\Gamma_q} \circ f_{\Gamma_T \leftarrow \varrho_T}(x + v)$$

R is a valid local retraction onto the manifold $\mathcal{M}_\rho$ for v sufficiently small, where the projection is given by Lemma 3.

Proof: We need to check that R satisfies the properties of a retraction map. The bijectivity of the maps, $f_{\varrho \leftarrow \Gamma}$ and $f_{\Gamma \leftarrow \varrho}$ implies that $R(x, 0) = x$. Using the chain rule, we compute:

$$\begin{aligned} DR(x,0) = {} & Df_{\varrho_T \leftarrow \Gamma_T}(\text{Proj}_{\Gamma_q} \circ f_{\Gamma_T \leftarrow \varrho_T}(x)) \\ & \cdot D\text{Proj}_{\Gamma_q}(f_{\Gamma_T \leftarrow \varrho_T}(x)) \\ & \cdot Df_{\Gamma_T \leftarrow \varrho_T}(x) \cdot v \end{aligned}$$

By construction, (x, v) is in the tangent bundle to $\mathcal{M}_\rho$, so $v \in \ker\{Dh(x)\}$. This implies that $Df_{\Gamma_T \leftarrow \varrho_T}(x) \cdot v$ satisfies,

$$\begin{bmatrix} \tilde{v} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} [Df_{\Gamma_T \leftarrow \varrho_T}(x)]_{0:,q} \\ [Df_{\Gamma_T \leftarrow \varrho_T}(x)]_{q+1:T} \end{bmatrix} v, \tag{8}$$

where $\tilde{v}$ is the non-zero part of the product $Df_{\Gamma_T \leftarrow \varrho_T}(x) \cdot v$. If we restrict ourselves to a set where v is sufficiently small, then the projection is given by a subspace projection as in Lemma 3, and hence the projection operator reduces to the identify function, because:

$$D\text{Proj}_{\Gamma_q}(x) = \begin{bmatrix} I_q & \\ & \mathbf{0}_{T-q} \end{bmatrix} \implies \begin{bmatrix} \tilde{v} \\ 0 \end{bmatrix} = \begin{bmatrix} I_q & \\ & \mathbf{0}_{T-q} \end{bmatrix} \begin{bmatrix} \tilde{v} \\ 0 \end{bmatrix}.$$

Finally, we have that

$$\begin{aligned} Df_{\varrho \leftarrow \Gamma}(\text{Proj}_{\Gamma_q} \circ f_{\Gamma \leftarrow \varrho}(x)) &= Df_{\varrho \leftarrow \Gamma}(f_{\Gamma \leftarrow \varrho}(x)) \\ &= \big(Df_{\Gamma \leftarrow \varrho}(x)\big)^{-1}. \end{aligned}$$

The last line follows from the inverse function theorem [12], which we have access to because $f_{\varrho_T \leftarrow \Gamma_T}$ is a diffeomorphism. This all combined leads to,

$$\begin{aligned} & Df_{\varrho_T \leftarrow \Gamma_T}(\text{Proj}_{\Gamma_q} \circ f_{\Gamma_T \leftarrow \varrho_T}(x)) \begin{bmatrix} \tilde{v} \\ 0 \end{bmatrix} \\ & \qquad = \big(Df_{\Gamma_T \leftarrow \varrho_T}(x)\big)^{-1} Df_{\Gamma_T \leftarrow \varrho_T}(x) v = v \end{aligned}$$

which establishes that R is a retraction. ∎

We conclude our theoretical analysis in the next section by addressing that our study thus far has been on the relaxation to the ML-MA problem (3).

IV. Spectral Factorizability

We return to the considering the relaxation we employed before performing our geometric analysis, and show how to further constrain the relaxed problem (4) to obtain a form equivalent to the original problem (3). Schemes for removing the unfactorizable autocovariances introduced by the structured covariance relaxation (4) are categorized in [17], which prescribes algorithms for accomplishing *spectral factorization* which maps autocovariances to back to MA coefficients. These algorithms follow from the Riesz-Fejér theorem, which also establishes that factorization existence is equivalent to positivity of a particular polynomial in γ. This can be expressed in a form useful for optimization by clever use of the positive real lemma [4].

Proposition 1: Define

$$A = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$C = \begin{bmatrix} \gamma_q & \gamma_{q-1} & \cdots & \gamma_1 \end{bmatrix}, \quad D = \frac{1}{2}[\gamma_0].$$

If there exists $P \succeq 0$ satisfying

$$\begin{bmatrix} P - A^*PA & C^* - A^*PB \\ C - B^*PA & D + D^* + B^*PB \end{bmatrix} \succeq 0, \tag{9}$$

then there exists moving-average parameters β satisfying

$$\exists \beta : \Sigma(\gamma) = \Sigma(\beta), \quad b_0 z^0 + \ldots + b_q z^q \neq 0 \text{ for all } |z| > 1.$$

Further, all possible moving average sequences can be constructed by an appropriate choice of γ.

The following corollary is then immediate by applying Proposition 1 to the ML-MA problem (3).

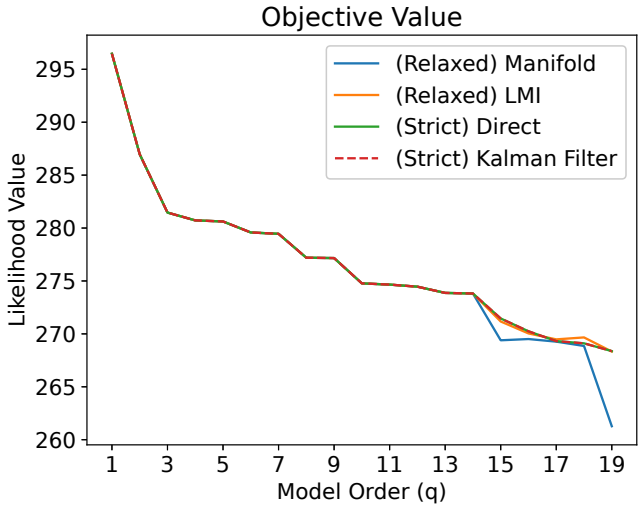

Fig. 2. Likelihood values from 100 time points sampled from MA(3) process with parameters, $\beta = [4, 3, 2, 1]$, identified using 4 different schemes.

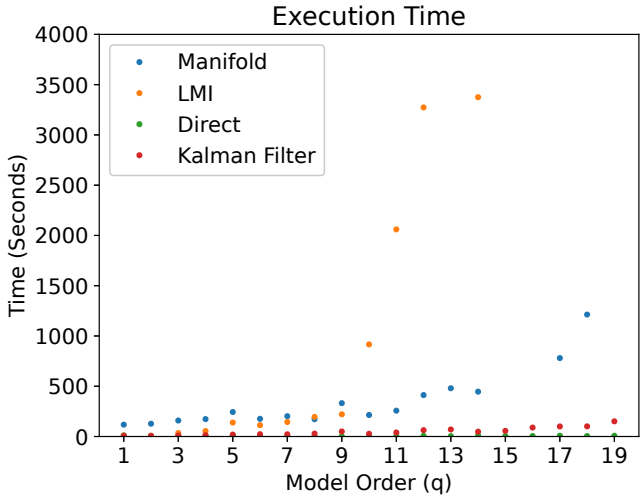

Fig. 3. Run-times from the same set of experiment. We mask out a few values of running time in the Manifold or LMI formulations when these instances terminate due to a maximum iterations safety term rather than the usual small gradient termination criteria.

Corollary 2: The optimization problem,

$$\begin{aligned} \underset{\gamma,\, P \succeq 0}{\text{minimize}} \quad & \tfrac{1}{2}\log\det(\Sigma(\gamma)) + \tfrac{1}{2}\mathrm{tr}\big(\xi\xi^*\Sigma(\gamma)^{-1}\big) \\ \text{subject to:} \quad & \Sigma(\gamma) = \gamma_0 C_0 + \gamma_1 C_1 + \ldots + \gamma_q C_q \\ & \begin{bmatrix} P - A^*PA & C^* - A^*PB \\ C - B^*PA & D + D^* + B^*PB \end{bmatrix} \succeq 0, \end{aligned} \tag{10}$$

with matrices A, B, C, and D defined as in Proposition 1 is equivalent to the ML-MA estimation problem 3.

Following (10), if we treat P as a certificate for the feasibility of γ, then [5] has noted that the set of feasible autocovariances in optimization problem 10 forms a convex cone, which we call $\Gamma_q^F \subseteq \Gamma_q$. From this point, addressing factorizability in the geometric setting is straight-forward using the diffeomorphisms $f_{\varrho_T \leftarrow \Gamma_T}$ and $f_{\Gamma_T \leftarrow \varrho_T}$, which are defined on all positive definite Toeplitz matrices, including the factorizable ones.

Corollary 3: Replace Γ_q with Γ_q^F in our definition of $\mathcal{M}_\varrho$ for a defintion of $\mathcal{M}_\varrho^F$. The set $\mathcal{M}_\varrho^F$ is a manifold.

Using this manifold instead of Γ_q in (7) leads to a formulation that is equivalent to (3); we do not need to worry about non-factorizability with this adjustment.

V. Numerical Validation

We write four different implementations of the estimation problem: two strict and two relaxed versions of the MA estimation problem. The strict versions require factorizability, while the relaxed versions (4) and (7) do not. Imposing factorizability on the manfiold formulation (7) is far more complicated than just a symbol substitution because we need to check for factorizability in addition to positive definiteness. This involves verifying that the positive real lemma holds, which is typically accomplished by solving a semidefinite feasibility problem. It is completely impractical to repeatedly call a semidefinite program in a subroutine that implements gradient descent, so our numerical experiments solve the relaxed problem instead. We note that the relaxation is often factorizable, so this complication may not be necessary in practice. The formulations we compare are:

1) The ML-MA formulation (3), which we call direct parameter search.
2) The Kalman Filter formulation, which performs a the direct parameter search within a state-space model embedding, see [18].
3) The relaxed structured covariance optimization formulation (4).
4) The relaxed manifold formulation (7).

All are implemented using back-tracking gradient descent. In the manifold setting, this scheme simply involves composing the objective function with the retraction and picking a suitable starting point on the manifold. Using a Riemmanian submanifold leads to gradients that line up with their usual ones, so we implement all optimization problems in JAX [19] for its automatic differentiation capabilities. Our goal is to show that our formulations can achieve the same results as the standard optimization schemes, so we refer the reader to [12] for extra details about how to implement backtracking gradient descent on manifolds. A basic description is given in the Appendix.

In terms of likelihood attained, all methods work equally as well as the benchmark Kalman Filter based formulation when identifying low model orders, up to $q = 14$ as illustrated in Fig. 2. When the model order increases, eventually, the geometric formulation occasionally find points that achieve lower likelihood than the strict formulations. Unfortunately, this is merely an artifact of the relaxation because the parameters identified in these situations lead to non-factorizable autocovariance parameters.

The long run-times, on the other hand, suggest that the formulations we have analyzed and implemented are better

used for theoretical analyses rather than for practical identification of MA(q) systems. Our contributions are more theoretical than computational, and hence the purpose of our experiments is to validate that the various formulations we have described are equivalent, rather than to demonstrate speed of computation. And indeed, we see that such equivalences largely hold in our numerical implementation, as shown in Fig. 2. Simply put, the well-studied Kalman-Filter/direct objective implementations are best suited for numerical optimization as illustrated in Fig. 3, but the autocovariance/geometric formulations may be more useful for theoretical analysis of MA(q) estimation due to the simplicity of the resulting optimization problem.

VI. Discussion and Future Prospects

We demonstrated that the moving average problem can be encoded as a manifold optimization problem over the reflection coefficients, and that the ML-MA, structured covariance, and manifold formulations all yield identical solutions in theory and in (limited) practice. We can therefore consider any of these problems for theoretical analysis, depending on how convenient they are to the problem at hand. The reflection coefficient parameterization may be particularly convenient for analysis since it is in the simple and canonical form of a curved-exponential family member. For example, with the formulation of (7), some straight-forward manipulations of the data yield non-trivial sufficient statistics, but we save investigating this for future work.

We have not resolved the most pressing question of global optimality. A reasonable conjecture is that all of the formulations we propose here are actually finding global optima whenever the parameters estimated are factorizable. Unfortunately, certifying this is beyond what we present here; translating local optimality to global optimality is more complicated when manifold constraints are present in the formulation. We close by mentioning that the analyses of geometric statistics [20] can provide these types of optimality guarantees in the exponential families, which broadly work by generalizing techniques from convex optimization to formulations on statistical manifolds. This line too, we save for future work.

Acknowledgements

This work was supported in part by NSF Awards SLES-2331880, ECCS-2045834, ECCS-2231349, and AFOSR Award FA9550-24-1-0102. We thank the anonymous reviewers for their excellent feedback and careful reading.

References

[1] P. Stoica, T. McKelvey, and J. Mari, "Ma estimation in polynomial time," *IEEE Transactions on Signal Processing*, vol. 48, no. 7, pp. 1999–2012, 2000. [Online]. Available: https://doi.org/10.1109/78.847786

[2] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974. [Online]. Available: https://doi.org/10.1109/TAC.1974.1100705

[3] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: the forecast package for r," *Journal of statistical software*, vol. 27, pp. 1–22, 2008. [Online]. Available: https://doi.org/10.18637/jss.v027.i03

[4] S.-P. Wu, S. Boyd, and L. Vandenberghe, "Fir filter design via semidefinite programming and spectral factorization," in *Proceedings of 35th IEEE Conference on Decision and Control*, vol. 1. IEEE, 1996, pp. 271–276. [Online]. Available: https://doi.org/10.1109/CDC.1996.574313

[5] B. Alkire and L. Vandenberghe, "Convex optimization problems involving finite autocorrelation sequences," *Mathematical Programming*, vol. 93, no. 3, pp. 331–359, 2002. [Online]. Available: https://doi.org/10.1007/s10107-002-0334-x

[6] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 963–974, 1982. [Online]. Available: https://doi.org/10.1109/PROC.1982.12427

[7] P. Zwiernik, C. Uhler, and D. Richards, "Maximum likelihood estimation for linear gaussian covariance models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 79, no. 4, pp. 1269–1292, 2017. [Online]. Available: https://doi.org/10.1111/rssb.12217

[8] B. Sturmfels, S. Timme, and P. Zwiernik, "Estimating linear covariance models with numerical nonlinear algebra," *Algebraic Statistics*, vol. 11, no. 1, pp. 31–52, 2020. [Online]. Available: https://doi.org/10.2140/astat.2020.11.31

[9] J. D. Hamilton, *Time series analysis*. Princeton University Press, 1994.

[10] L. Ljung, *System identification*. Prentice Hall, 1999.

[11] B. Dumitrescu, I. Tabus, and P. Stoica, "On the parameterization of positive real sequences and ma parameter estimation," *IEEE Transactions on Signal Processing*, vol. 49, no. 11, pp. 2630–2639, 2001.

[12] N. Boumal, *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[13] W. F. Trench, "An algorithm for the inversion of finite toeplitz matrices," *Journal of the Society for Industrial and Applied Mathematics*, vol. 12, no. 3, pp. 515–522, 1964. [Online]. Available: https://doi.org/10.1137/0112045

[14] N. Levinson, "The wiener (root mean square) error criterion in filter design and prediction," *Journal of Mathematics and Physics*, vol. 25, no. 1-4, pp. 261–278, 1946. [Online]. Available: https://doi.org/10.1002/sapm1946251261

[15] B. Efron, "The geometry of exponential families," *The Annals of Statistics*, pp. 362–376, 1978.

[16] E. Levin, J. Kileel, and N. Boumal, "The effect of smooth parametrizations on nonconvex optimization landscapes," *Mathematical Programming*, pp. 1–49, 2024. [Online]. Available: https://doi.org/10.1007/s10107-024-02058-3

[17] B. Dumitrescu, *Positive trigonometric polynomials and signal processing applications*. Springer, 2007, vol. 103.

[18] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation*. Prentice Hall, 2000.

[19] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs," 2018. [Online]. Available: http://github.com/google/jax

[20] S.-i. Amari, *Information geometry and its applications*. Springer, 2016, vol. 194.

Appendix

A. Omitted Proofs

Lemma 3 (Local Γ_q Projection): There exists open sets $\mathcal{U} \subseteq \mathbb{R}^T$ where:

$$\text{Proj}_{\Gamma_q}\left(f_{\Gamma_q \leftarrow \varrho}\{\mathcal{U}\}\right) = \begin{bmatrix} I_q & \\ & \mathbf{0}_{T-q} \end{bmatrix} f_{\Gamma_q \leftarrow \varrho}\{\mathcal{U}\}$$

Proof: Let $\gamma \in \Gamma_q$, which implies the existence of $m > 0$ where,

$$G = \gamma_0 C_0 + \ldots + \gamma_q C_q \succ mI$$

We construct an open neighborhood of γ in $\mathbb{R}^T$ where the projection behaves like a subspace projection $\mathcal{V}$, as well as

an open set $\mathcal{U}$ whose image under $f_{\Gamma_q \leftarrow \varrho}(\mathcal{U})$ is entirely contained in $\mathcal{V}$. First, let,

$$\mathcal{V}_1 = \left\{ \pi \in \mathbb{R}^T : -\frac{m}{2} I \prec \sum_{k=0}^{q} \pi_k C_k \right\}$$

and:

$$\mathcal{V}_2 = \left\{ \pi \in \mathbb{R}^T : -\frac{m}{2} I \prec \sum_{k=q+1}^{T-1} \pi_k C_k \prec \frac{m}{2} I \right\}$$

and define the open neighborhood of γ:

$$\mathcal{V} = \big(\mathcal{V}_1 \cap \mathcal{V}_2\big) + \gamma$$

Notice that $\mathcal{V}$ is nonempty because at any γ, we can choose any $\pi \in \mathbb{R}^T$, and simply scale π down until it satisfies the matrix inequalities in $\mathcal{V}_1$ and $\mathcal{V}_2$. Define the matrices P, P_1, P_2 as:

$$P_1 = \sum_{k=0}^{q} \pi_k C_k, \quad P_2 = \sum_{k=q+1}^{T} \pi_k C_k, \quad P = P_1 + P_2$$

and let $T = P + G$. Using Weyl's inequality repeatedly, we see that the minimum eigenvalue of T is bounded below by:

$$\begin{aligned} \lambda_{\min}(T) &\geqslant \lambda_{\min}(G) + \lambda_{\min}(P) \\ &\geqslant \lambda_{\min}(G) + \lambda_{\min}(P_1) + \lambda_{\min}(P_2) \\ &\geqslant \lambda_{\min}(G) - m > 0 \end{aligned}$$

So elements in $\mathcal{V}$ parameterize positive definite Toeplitz matrices in Γ_T. Let $\mathcal{U} = f_{\varrho_T \leftarrow \Gamma_T}\{\mathcal{V}\}$; this is an open set in $\mathbb{R}^T$ whose image is $\mathcal{V}$. Finally, we verify that subtracting off the tail terms in elements of $\mathcal{V}$ still leads to a positive definite matrix. Notice for $\tilde{\gamma} \in \mathcal{V}$, this is given by:

$$\lambda_{\min}\left(\sum_{k=0}^{q} \tilde{\gamma}_k C_k\right) \geqslant \lambda_{\min}\left(\sum_{k=0}^{T} \tilde{\gamma}_k C_k\right) - \frac{m}{2} > 0$$

due to the constraint on $\mathcal{V}_2$, implying positive definiteness. Note that the projection operation is the solution to the optmization problem:

$$\underset{\gamma \in \Gamma_q}{\text{minimize}} \quad \|\tilde{\gamma} - \gamma\|_2$$

and its solution can be expressed in matrix form as:

$$\text{Proj}_{\Gamma_q}(v) = \begin{bmatrix} I_q & \\ & \mathbf{0}_{T-q} \end{bmatrix} v.$$

This achieves the smallest possible perturbation to $\tilde{\gamma}$ because all other feasible choices must also alter components in the first q entries, thus incurring higher cost. ∎

Corollary 1 (Geometric Covariance Estimation): The structured covariance estimation problem (4) is equivalent to:

$$\begin{aligned} \min_{\rho} \quad & T\log(\gamma_0) + \sum_{k=1}^{T-1}(T-k)\log(1-\rho_i^2) + \sum_{k=0}^{T-1}\langle \upsilon_k, \xi\rangle^2 \\ \text{s.t.} \quad & (t_0, \rho) \in \mathcal{M}_\varrho \end{aligned}$$

Proof: We omit a factor of $1/2$ in the objective function because this does not affect identified optima. The inner-product term follows directly from block partitioning $\xi^* U U^* \xi$. The $\log\det(\cdot)$ term follows from first noting that:

$$-\log\det(\Sigma(\gamma)) = \log\det(UU^*) = \sum_{i=0}^{T} \log(u_{ii})$$

because U is triangular. Next, we observe that the Levinson-recursions for just the diagonal elements is,

$$u_{i+1,i+1} = \frac{u_{i,i}}{\sqrt{1-\rho_i^2}} \implies u_{i,i} = \frac{u_{0,0}}{\sqrt{1-\rho_0^2}\cdots\sqrt{1-\rho_i^2}}.$$

The rest follows from expanding the sum of logs, and appropriately to omitting factors of 2, which leads to the defined objective function. ∎

B. Backtracking Riemannian Gradient Descent

We provide a simple scheme for accomplishing Riemannian gradient descent (RGD) on manifolds, which is analogous to projected gradient descent. While the retraction we derive is only a local map about $\big((\gamma_0, \rho), 0\big)$ on the tangent bundle, checking if we have exited the region where it's valid is simple because the retraction only returns reflection coefficients belonging to $(-1, 1)$ when $\tilde{\gamma} = \text{Proj}_{\Gamma_q} \circ f_{\Gamma_q \leftarrow \varrho}(x + v)$ also belongs to Γ_q due to the properties of the Levinson-Trench recursions. When this fails, then the subspace projection has returned a point outside of Γ_q and our analysis in Lemma 3 no longer holds. The simple adjustment is to do two backtracking steps; one to first ensure that we are inside a region where the retraction is valid, then another to ensure we end up with a decrease in our objective value. For the sake of simplicity, let $\mathcal{L}(\rho; \xi)$ be the likelihood of the optimization problem in Corollary 1. We define:

$$\mathcal{RL}((\gamma_0, \rho); \xi) =: \mathcal{L}(\cdot; \xi) \circ R((\gamma_0, \rho), 0)$$

that is, the likelihood function composed with the retraction.

Algorithm 1 Backtracking RGD

Input: $(\gamma_0, \rho)^{(k)}$
Hyperparameters: $\alpha \in (0, 0.5), \mu \in (0, 1)$
Output: $(\gamma_0, \rho)^{(k+1)}$

Initialisation : $(\gamma, \rho) = (1, 0, \ldots, 0), \quad t = 1$
$d \leftarrow \nabla_{(\gamma_0, \rho)}\big\{\mathcal{RL}((\gamma_0, \rho)^{(k)}; \xi)\big\}$
Backtrack until retraction is valid
while $\|R((\gamma_0, \rho)^{(k)} - td)\|_\infty > 1$ **do**
  $t \leftarrow \mu t$
end while
Backtrack until sufficient objective decrease
while $\mathcal{RL}((\gamma_0, \rho)^{(k)} - td; \xi) > \mathcal{RL}((\gamma_0, \rho)^{(k)}; \xi) - \alpha t \|d\|^2$ **do**
  $t \leftarrow \mu t$
end while
return $(\gamma_0, \rho)^{(k+1)}$