The derivatives of the SEDS optimization cost function and constraints with respect to the learning parameters

S. Mohammad Khansari-Zadeh and Aude Billard

I. Introduction

This technical report provides supplementary information for the optimization problems defined for Stable Estimator of Dynamical Systems (SEDS) [1]. The source code of SEDS can be downloaded from:

http://lasa.epfl.ch/sourcecode/

Reading of this report is *not necessary* for researchers who only want to use SEDS learning algorithm. The report is aimed at helping those persons who want to develop SEDS, or to write their own optimization program. Reading and understanding of [1] is a prerequisite for this document. All the formulations reported here are developed for SEDS models; however, they can also be used for general Gaussian Mixture Model (GMM) formulations. In the case of the latter, they should be slightly modified to consider the general form of GMM. Hopefully, the report should be clear enough to help readers in that.

To facilitate reading of the paper, a list of main variables and mathematical notations is provided in Table I. Furthermore, to have a clean summary of the final results, all the derivatives are summarized in Tables II-VI.

The remainder of this document is structured as follows. Section II gives a recap of the SEDS formulations taken from [1]. Sections III and V provide analytical formulations to compute the derivatives of MSE and Likelihood cost functions with respect to the optimization parameters, respectively. In addition, Sections IV and VI present two alternative optimization problems that automatically satisfy 4 out of 5 constraints of the original optimization problem through a change of variable. Finally, Section VII defines a proper mathematical representation of the optimization constraints, and provides the analytical derivatives of these constraints with respect to the optimization parameters.

II. SEDS Formulation

Let us consider a robot motion that is defined as an autonomous Dynamical System (DS). We formulate this DS as a mixture of Gaussian functions:

$$\hat{\dot{\xi}} = \hat{f}(\xi) = \sum_{k=1}^{K} h^k(\xi)(A^k\xi + b^k) \tag{1}$$

where

$$\begin{cases} A^k = \Sigma^k_{\dot{\xi}\xi}(\Sigma^k_{\xi})^{-1} \\ b^k = \mu^k_{\dot{\xi}} - A^k\mu^k_{\xi} \\ h^k(\xi) = \frac{\mathcal{P}(k)\mathcal{P}(\xi|k)}{\sum_{i=1}^{K}\mathcal{P}(i)\mathcal{P}(\xi|i)} \end{cases} \tag{2}$$

S.M. Khansari-Zadeh and A. Billard are with LASA Laboratory, School of Engineering, Ecole Polytechnique Federale de Lausanne, Switzerland {mohammad.khansari,Aude.Billard}@epfl.ch

TABLE I
Nomenclature

Variable	Type (size)	Description
d	Scalar	Dimension of DS
ξ	Vector (d)	Input variable, e.g. position
ξ^*	Vector (d)	Target point
$\dot{\xi}$	Vector (d)	Output variable, e.g. velocity
π	Scalar	Prior of the Gaussian function
μ	Vector (d)	Center of the Gaussian function
Σ	Matrix ($2d \times 2d$)	Covariance matrix of the Gaussian fun.
f	Function ($d \mapsto d$)	Unknown original DS
J	Scalar	Optimization cost function
$\boldsymbol{\theta}$	Structure	Optimization parameters
L	Matrix ($2d \times 2d$)	Lower triangle matrix
A	Matrix ($d \times d$)	Matrix of the linear DS
b	Vector (d)	Intersection point of the linear DS
$\boldsymbol{I}$	Matrix	Identity matrix
$\mathbf{0}$	Vector	Zero vector
K	Scalar	Number of Gaussian functions
N	Scalar	Number of demonstrations

Notation	Description
$\hat{(.)}$	Estimated value of a variable
$(.)^k$	Of the k-th Gaussian function
$(.)^T$	Transpose of a Vector/matrix
$(.)^{t,n}$	The t-th datapoint of the n-th demonstration
$(.)_i$	The i-th component of a vector
$(.)_{ij}$	The (i,j)-th component of a matrix
$(\text{vec})_{\xi}$	Sub-vector of vec with indices $1:d$
$(\text{vec})_{\dot{\xi}}$	Sub-vector of vec with indices $d{+}1:2d$
$(\text{mat})_{\xi}$	Sub-matrix of mat with indices $(1:d, 1:d)$
$(\text{mat})_{\dot{\xi}\xi}$	Sub-matrix of mat with indices $(d{+}1:2d, 1:d)$
$(.)_{1:c,1:c}$	A slice of a matrix with indices $(1:c, 1:c)$
$\mathbf{0}^{\{i\}}$	A zero vector with the exception that its i-th component is 1
$\mathbf{0}^{\{ij\}}$	A matrix of zeros with the exception that its (i,j)-th component is one.
$\mathbf{0}^{\{\overline{ij}\}}$	A matrix of zeros with the exception that its (i,j) and (j,i)-th components are one.
$adj(.)$	Adjugate of a matrix
$tr(.)$	Trace of a matrix
$ln(.)$	The natural logarithm
$Chol(.)$	Cholesky decomposition of a matrix

$$\mathcal{P}(\xi|k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma^k_{\xi}|}} e^{-\frac{1}{2}(\xi-\mu^k_{\xi})^T(\Sigma^k_{\xi})^{-1}(\xi-\mu^k_{\xi})} \tag{3}$$

The unknown parameters of $\hat{f}(\xi)$ that should be learned based on demonstrations are the priors $\pi^k = \mathcal{P}(k)$, means μ^k and covariance matrices Σ^k of the $k = 1..K$ Gaussian functions. Given a set of N demonstrations $\{\xi^{t,n}, \dot{\xi}^{t,n}\}_{t=0,n=1}^{T^n,N}$ of the motion, these parameters can be estimated by solving an optimization problem under the constraint of ensuring the model's global asymptotic stability. We consider two different optimization cost functions: 1) log-likelihood, and 2) Mean Square Error (MSE), which we explain next.

TABLE II
DERIVATIVES OF THE MSE COST FUNCTION TAKEN FROM SECTION III.

$\boldsymbol{\theta} = \{\pi^1..\pi^K; \mu_\xi^1..\mu_\xi^K; \Sigma_\xi^1..\Sigma_\xi^K; \Sigma_{\dot\xi\xi}^1..\Sigma_{\dot\xi\xi}^K\}$

Cost function: $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})^T(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})$

Indices range: $k \in 1..K, \quad i \in 1..d$

$\frac{\partial J}{\partial \pi^k} = \frac{1}{\pi^k N}\sum_{n=1}^{N}\sum_{t=0}^{T^n} h^k(\xi^{t,n})(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})^T(A^k\xi^{t,n} - \hat{\dot\xi}^{t,n})$

$\frac{\partial J}{\partial \mu_{\xi,i}^k} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n} h^k(\xi^{t,n})\Big((\xi^{t,n} - \mu_\xi^k)^T(\Sigma_\xi^k)^{-1}\mathbf{0}^{\{i\}}\Big)(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})^T(A^k\xi^{t,n} - \hat{\dot\xi}^{t,n})$

$\frac{\partial J}{\partial \Sigma_{\xi,ij}^k} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n} h^k(\xi^{t,n})(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})^T\Big(0.5(\xi^{t,n} - \mu_\xi^k)^T(\Sigma_\xi^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}(\Sigma_\xi^k)^{-1}(\xi^{t,n} - \mu_\xi^k)(A^k\xi^{t,n} - \hat{\dot\xi}^{t,n})$
$- 0.5tr\big((\Sigma_\xi^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}\big)(A^k\xi^{t,n} - \hat{\dot\xi}^{t,n}) - A^k\mathbf{0}^{\{\overline{ij}\}}(\Sigma_\xi^k)^{-1}\xi^{t,n}\Big) \qquad j \in 1..i$

$\frac{\partial J}{\partial \Sigma_{\dot\xi\xi,ij}^k} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n} h^k(\xi^{t,n})(\xi^{t,n} - \mu_\xi^k)^T\mathbf{0}^{\{ij\}}(\Sigma_\xi^k)^{-1}\xi^{t,n} \qquad j \in 1..d$

TABLE III
DERIVATIVES OF THE ALTERNATIVE MSE COST FUNCTION TAKEN FROM SECTION IV.

$\boldsymbol{\theta} = \{\tilde\pi^1..\tilde\pi^K; \mu_\xi^1..\mu_\xi^K; L_\xi^1..L_\xi^K; A^1..A^K\}$

Cost function: $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})^T(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})$

Indices range: $k \in 1..K, \quad i \in 1..d$

Change of variables: $\tilde\pi^k = ln(\pi^k), \quad L_\xi^k = Chol(\Sigma_\xi^k)$

$\frac{\partial J}{\partial \tilde\pi^k} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n} h^k(\xi^{t,n})(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})^T(A^k\xi^{t,n} - \hat{\dot\xi}^{t,n})$

$\frac{\partial J}{\partial \mu_{\xi,i}^k} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n} h^k(\xi^{t,n})\Big((\xi^{t,n} - \mu_\xi^k)^T(\Sigma_\xi^k)^{-1}\mathbf{0}^{\{i\}}\Big)(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})^T(A^k\xi^{t,n} - \hat{\dot\xi}^{t,n})$

$\frac{\partial J}{\partial L_{\xi,ij}^k} = \frac{1}{2N}\sum_{n=1}^{N}\sum_{t=0}^{T^n} h^k(\xi^{t,n})\Big((\xi^{t,n} - \mu_\xi^k)^T(\Sigma_\xi^k)^{-1}\Phi(\Sigma_\xi^k)^{-1}(\xi^{t,n} - \mu_\xi^k) - tr\Big((\Sigma_\xi^k)^{-1}\Phi\Big)\Big)(\hat{\dot\xi}^{t,n} - \dot\xi^{t,n})^T(A^k\xi^{t,n} - \hat{\dot\xi}^{t,n})$
where $\Phi = \mathbf{0}^{\{ij\}}(L^k)^T + L^k(\mathbf{0}^{\{ij\}})^T \qquad j \in 1..i$

$\frac{\partial J}{\partial A_{ij}^k} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n} h^k(\xi^{t,n})(\xi^{t,n} - \mu_\xi^k)^T\mathbf{0}^{\{ij\}}\xi^{t,n} \qquad j \in 1..d$

Reconstruction of GMM from the optimization parameters: $\pi^k = e^{\tilde\pi^k}/(\sum_{i=1}^{K} e^{\tilde\pi^i}), \quad \Sigma_\xi^k = L_\xi^k(L_\xi^k)^T, \quad \Sigma_{\dot\xi\xi}^k = A^k\Sigma_\xi^k$

TABLE IV
DERIVATIVES OF THE LIKELIHOOD COST FUNCTION TAKEN FROM SECTION V.

$\boldsymbol{\theta} = \{\pi^1..\pi^K; \mu_\xi^1..\mu_\xi^K; \Sigma^1..\Sigma^K\}$

Cost function: $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\log\mathcal{P}(\xi^{t,n}, \dot\xi^{t,n}|\boldsymbol{\theta})$

Indices range: $k \in 1..K$

$\frac{\partial J}{\partial \pi^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\big(\frac{\mathcal{P}(\xi^{t,n},\dot\xi^{t,n}|k)}{\mathcal{P}^{t,n}} - 1\big)$

$\frac{\partial J}{\partial \mu_{\xi,i}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\mathcal{P}(k)\mathcal{P}(\xi^{t,n},\dot\xi^{t,n}|k)}{\mathcal{P}^{t,n}}(\mathbf{0}^{\{i\}})^T\big[\boldsymbol{I} \quad (A^k)^T\big](\Sigma^k)^{-1}\big([\xi^{t,n};\dot\xi^{t,n}] - \mu^k\big) \qquad \forall i \in 1..d$

$\frac{\partial J}{\partial \Sigma_{ij}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\mathcal{P}(k)\mathcal{P}(\xi^{t,n},\dot\xi^{t,n}|k)}{\mathcal{P}^{t,n}}\Big(0.5(\xi^{t,n} - \mu^k)^T(\Sigma^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}(\Sigma^k)^{-1}(\xi^{t,n} - \mu^k) - 0.5tr\big((\Sigma^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}\big)$
$+ (\xi^{t,n} - \mu^k)^T(\Sigma^k)^{-1}\mathcal{S}^k\Big) \qquad \forall i \in 1..2d, \quad j \in 1..i$

where $\mathcal{S}^k = \begin{bmatrix} \mathbf{0} \\ \Big(-A^k\big[\mathbf{0}^{\{\overline{ij}\}}\big]_\xi + \big[\mathbf{0}^{\{\overline{ij}\}}\big]_{\dot\xi\xi}\Big)(\Sigma_\xi^k)^{-1}\mu_\xi^k \end{bmatrix}$

TABLE V
DERIVATIVES OF THE LIKELIHOOD COST FUNCTION TAKEN FROM SECTION VI.

$\boldsymbol{\theta} = \{\tilde{\pi}^1..\tilde{\pi}^K; \mu_\xi^1..\mu_\xi^K; L^1..L^K\}$

Cost function: $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n} \log \mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n}|\boldsymbol{\theta})$

Indices range: $k \in 1..K$

Change of variables: $\tilde{\pi}^k = ln(\pi^k), \quad L^k = Chol(\Sigma^k)$

$\frac{\partial J}{\partial \tilde{\pi}^k} = -\frac{e^{\tilde{\pi}^k}}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\left(\frac{\mathcal{P}(\xi^{t,n},\dot{\xi}^{t,n}|k)}{\mathcal{P}^{t,n}} - 1\right)$

$\frac{\partial J}{\partial \mu_{\xi,i}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\mathcal{P}(k)\mathcal{P}(\xi^{t,n},\dot{\xi}^{t,n}|k)}{\mathcal{P}^{t,n}}(\mathbf{0}^{\{i\}})^T\left[\boldsymbol{I} \quad (A^k)^T\right](\Sigma^k)^{-1}\left([\xi^{t,n};\dot{\xi}^{t,n}] - \mu^k\right) \qquad i \in 1..d$

$\frac{\partial J}{\partial L_{ij}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\mathcal{P}(k)\mathcal{P}(\xi^{t,n},\dot{\xi}^{t,n}|k)}{\mathcal{P}^{t,n}}\Big(0.5(\xi^{t,n}-\mu^k)^T(\Sigma^k)^{-1}\Phi(\Sigma^k)^{-1}(\xi^{t,n}-\mu^k) - 0.5tr\big((\Sigma^k)^{-1}\Phi\big) + (\xi^{t,n}-\mu^k)^T(\Sigma^k)^{-1}\tilde{\mathcal{S}}^k\Big) \qquad i \in 1..2d, \quad j \in 1..i$

where $\Phi = \mathbf{0}^{\{ij\}}(L^k)^T + L^k(\mathbf{0}^{\{ij\}})^T, \quad \tilde{\mathcal{S}}^k = \begin{bmatrix} \mathbf{0} \\ \big(-A^k\Phi_\xi + \Phi_{\dot{\xi}\xi}\big)(\Sigma_\xi^k)^{-1}\mu_\xi^k \end{bmatrix}$

Reconstruction of GMM from the optimization parameters: $\pi^k = e^{\tilde{\pi}^k}/(\sum_{i=1}^{K} e^{\tilde{\pi}^i}), \quad \Sigma^k = L^k(L^k)^T$

TABLE VI
CONSTRAINTS FORMULATION AND THEIR DERIVATIVES FOR THE ALTERNATIVE LIKELIHOOD AND MSE COST FUNCTIONS TAKEN FROM SECTION VII.

Indices range: $k \in 1..K, \quad c \in 1..d$

Constraint: $A^k + (A^k)^T < 0$

The equivalence of the constraint used in the code: $\mathcal{C}_{(k-1)d+c}: \quad (-1)^{c+1}|B_{1:c,1:c}| < 0$

$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial \tilde{\pi}^k} = 0$ (valid for both the MSE and Likelihood cost functions)

$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial \mu_i^k} = 0 \qquad i \in 1..d$ (valid for both the MSE and Likelihood cost functions)

The derivatives specific to the MSE cost function:

$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial L_{ij}^k} = 0 \qquad i \in 1..d, \quad j \in 1..d$

$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial A_{ij}^k} = (-1)^{c+1}tr\Big(adj\big(B_{1:c,1:c}\big)\big[\mathbf{0}^{\{\overline{ij}\}}\big]_{1:c,1:c}\Big) \qquad i \in 1..d, \quad j \in 1..d$

The derivative specific to the Likelihood cost function:

$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial L_{ij}^k} = (-1)^{c+1}tr\Big(adj\big(B_{1:c,1:c}\big)\mathcal{X}_{1:c,1:c}\Big) \qquad i \in 1..2d, \quad j \in 1..i$

where $\Phi = \mathbf{0}^{\{ij\}}(L^k)^T + L^k(\mathbf{0}^{\{ij\}})^T, \quad \Psi = (-A^k\Phi_\xi + \Phi_{\dot{\xi}\xi}(\Sigma_\xi)^{-1}, \quad \mathcal{X} = \Psi + (\Psi)^T$

III. Mean Square Error Optimization

Mean Square Error (MSE) is a means to quantify the accuracy of estimations based on demonstrations, and it is defined as:

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}(\hat{\dot{\xi}}^{t,n} - \dot{\xi}^{t,n})^T(\hat{\dot{\xi}}^{t,n} - \dot{\xi}^{t,n}) \tag{4}$$

subject to

$$\begin{cases} \textbf{(a)}\ b^k = -A^k\xi^* \\ \textbf{(b)}\ A^k + (A^k)^T < 0 \\ \textbf{(c)}\ \Sigma_\xi^k > 0 \\ \textbf{(d)}\ 0 < \pi^k \leq 1 \\ \textbf{(e)}\ \sum_{k=1}^{K}\pi^k = 1 \end{cases} \quad \forall k \in 1..K \tag{5}$$

where $\hat{\dot{\xi}}^{t,n} = \hat{f}(\xi^{t,n})$ are computed from Eq. (1). The optimization parameters for this objective function are: $\boldsymbol{\theta} = \{\pi^1..\pi^K; \mu_\xi^1..\mu_\xi^K; \Sigma_\xi^1..\Sigma_\xi^K; \Sigma_{\dot{\xi}\xi}^1..\Sigma_{\dot{\xi}\xi}^K\}$. Solving the above optimization requires a user to provide the derivative of the cost function w.r.t. the optimization parameters. These derivatives are provided next.

A. Derivatives w.r.t. Priors π^k

$$\frac{\partial J}{\partial \pi^k} = \frac{1}{2N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}}\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \pi^k} \qquad \forall k \in 1..K \tag{6}$$

The partial derivatives $\frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}}$ and $\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \pi^k}$ can be computed from Eqs. (7) and (8), respectively:

$$\frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}(\hat{\dot{\xi}}^{t,n} - \dot{\xi}^{t,n})^T \tag{7}$$

$$\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \pi^k} = \frac{h^k(\xi^{t,n})}{\pi^k}(A^k\xi^{t,n} - \hat{\dot{\xi}}^{t,n}) \tag{8}$$

Substituting Eqs. (7) and (8) into Eq. (6) yields:

$$\frac{\partial J}{\partial \pi^k} = \frac{1}{\pi^k N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}h^k(\xi^{t,n})(\hat{\dot{\xi}}^{t,n} - \dot{\xi}^{t,n})^T(A^k\xi^{t,n} - \hat{\dot{\xi}}^{t,n}) \tag{9}$$

B. Derivatives w.r.t. Means μ_ξ^k

Since μ_ξ^k is a d-dimensional vector, we need to compute the derivative w.r.t. each component of μ_ξ^k separately:

$$\frac{\partial J}{\partial \mu_{\xi,i}^k} = \frac{1}{2N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}}\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \mu_{\xi,i}^k} \quad \forall i \in 1..d,\ k = 1..K \tag{10}$$

The partial derivative $\frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}}$ is given by Eq. (7), and $\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \mu_{\xi,i}^k}$ is:

$$\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \mu_{\xi,i}^k} = h^k(\xi^{t,n})\left((\xi^{t,n} - \mu_\xi^k)^T(\Sigma_\xi^k)^{-1}\mathbf{0}^{\{i\}}\right)(A^k\xi^{t,n} - \hat{\dot{\xi}}^{t,n}) \tag{11}$$

where $\mathbf{0}^{\{i\}}$ has the dimension of d.

C. Derivatives w.r.t. Means $\mu_{\dot{\xi}}^k$

By substituting directly the constraint Eq. (5)-(a) into Eq. (1), the partial derivative $\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \mu_{\dot{\xi},i}^k}$ is always zero because $\hat{f}(\xi)$ no longer depends on $\mu_{\dot{\xi}}^k$. Therefore, $\mu_{\dot{\xi},i}^k$ can be dropped from the list of the optimization parameters. In fact, at each iteration $\mu_{\dot{\xi}}^k$ is exploited to satisfy this constraint, and its value can be directly computed from Eq. (5)-(a).

D. Derivatives w.r.t. Σ_ξ^k

Since Σ_ξ^k is a $d \times d$ matrix, we will compute the derivative w.r.t. its each component separately. Since Σ_ξ^k is a symmetric matrix, we compute the derivatives only for the components on the lower triangle matrix.

$$\frac{\partial J}{\partial \Sigma_{\xi,ij}^k} = \frac{1}{2N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}}\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \Sigma_{\xi,ij}^k} \quad \begin{cases} \forall i \in 1..d \\ \forall j \in 1..i \\ \forall k \in 1..K \end{cases} \tag{12}$$

The partial derivative $\partial \hat{\dot{\xi}}^{t,n}/\partial \Sigma_{\xi,ij}^k$ is:

$$\begin{aligned} \frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \Sigma_{\xi,ij}^k} = &-h^k(\xi^{t,n})A^k\mathbf{0}^{\{\overline{ij}\}}(\Sigma_\xi^k)^{-1}\xi^{t,n} + \\ &\frac{h^k(\xi^{t,n})}{2}\Big((\xi^{t,n} - \mu_\xi^k)^T(\Sigma_\xi^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}(\Sigma_\xi^k)^{-1}(\xi^{t,n} - \mu_\xi^k) \\ &- tr\big((\Sigma_\xi^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}\big)\Big)(A^k\xi^{t,n} - \hat{\dot{\xi}}^{t,n}) \end{aligned} \tag{13}$$

where $\mathbf{0}^{\{\overline{ij}\}}$ has the dimension of $d \times d$.

E. Derivatives w.r.t. $\Sigma_{\dot{\xi}\xi}^k$

The partial derivatives of the cost function w.r.t. the components of $\Sigma_{\dot{\xi}\xi}^k$ are

$$\frac{\partial J}{\partial \Sigma_{\dot{\xi}\xi,ij}^k} = \frac{1}{2N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}}\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \Sigma_{\dot{\xi}\xi,ij}^k} \quad \begin{cases} \forall i \in 1..d \\ \forall j \in 1..d \\ \forall k \in 1..K \end{cases} \tag{14}$$

The partial derivative $\partial \hat{\dot{\xi}}^{t,n}/\partial \Sigma_{\dot{\xi}\xi,ij}^k$ is:

$$\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \Sigma_{\dot{\xi}\xi,ij}^k} = h^k(\xi^{t,n})\mathbf{0}^{\{ij\}}(\Sigma_\xi^k)^{-1}\xi^{t,n} \tag{15}$$

where $\mathbf{0}^{\{ij\}}$ has the dimension of $d \times d$.

IV. Alternative MSE Optimization

Though the MSE optimization provided in Section III is sufficient to estimate a stable DS, its performance can be significantly increased through a change of optimization parameters. Let us define:

$$\begin{cases} \tilde{\pi}^k = ln(\pi^k) \\ L_\xi^k = Chol(\Sigma_\xi^k) \end{cases} \tag{16}$$

where L_ξ^k is a $d \times d$ lower triangle matrix. Since Σ_ξ^k are positive definite matrix, their Cholesky decomposition L_ξ^k always exist. Furthermore, as it was pointed out before, by substituting Eq. (5)-(a) into Eq. (1), we can define the evolution of motion with:

$$\hat{\dot{\xi}} = \hat{f}(\xi) = \sum_{k=1}^{K} h^k(\xi) A^k (\xi - \xi^*) \tag{17}$$

Considering Eqs. (16) and (17) and defining the optimization parameters to be $\boldsymbol{\theta} = \{\tilde{\pi}^1..\tilde{\pi}^K; \mu_\xi^1..\mu_\xi^K; L_\xi^1..L_\xi^K; A^1..A^K\}$, the alternative MSE optimization can be expressed as:

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{2N} \sum_{n=1}^{N} \sum_{t=0}^{T^n} (\hat{\dot{\xi}}^{t,n} - \dot{\xi}^{t,n})^T (\hat{\dot{\xi}}^{t,n} - \dot{\xi}^{t,n}) \tag{18}$$

subject to

$$A^k + (A^k)^T < 0 \qquad \forall k \in 1..K \tag{19}$$

where $\hat{\dot{\xi}}^{t,n} = \hat{f}(\xi^{t,n})$ are computed from Eq. (17). Once the optimization finished, the parameters of GMM can be reconstructed as follows:

$$\begin{cases} \pi^k = e^{\tilde{\pi}^k} / (\sum_{i=1}^{K} e^{\tilde{\pi}^i}) \\ \Sigma_\xi^k = L_\xi^k (L_\xi^k)^T \\ \Sigma_{\dot{\xi}\xi}^k = A^k \Sigma_\xi^k \end{cases} \tag{20}$$

In fact the proposed change of parameters allows us to automatically satisfy the last three optimization constraints of Eq. (5). The first constraint of Eq. (5) is also removed since it is directly considered in Eq. (17). The derivatives of the new optimization problem are provided in the following subsections.

A. Derivatives w.r.t. Priors π^k

$$\frac{\partial J}{\partial \tilde{\pi}^k} = \frac{1}{2N} \sum_{n=1}^{N} \sum_{t=0}^{T^n} \frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}} \frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial \pi^k} \frac{\partial \pi^k}{\partial \tilde{\pi}^k} \qquad \forall k \in 1..K \tag{21}$$

The partial derivatives $\partial J / \partial \hat{\dot{\xi}}^{t,n}$ and $\partial \hat{\dot{\xi}}^{t,n} / \partial \pi^k$ are given by Eqs. (7) and (8), and the derivative $\partial \pi^k / \partial \tilde{\pi}^k$ is simply:

$$\frac{\partial \pi^k}{\partial \tilde{\pi}^k} = e^{\tilde{\pi}^k} \tag{22}$$

B. Derivatives w.r.t. Means μ_ξ^k

These derivative can be similarly computed from Eq. (10).

C. Derivatives w.r.t. L^k

L_ξ^k is a $d \times d$ lower triangle matrix. The partial derivatives of the cost function w.r.t. its parameters are:

$$\frac{\partial J}{\partial L_{\xi,ij}^k} = \frac{1}{2N} \sum_{n=1}^{N} \sum_{t=0}^{T^n} \frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}} \frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial L_{\xi,ij}^k} \qquad \begin{cases} \forall i \in 1..d \\ \forall j \in 1..i \\ \forall k \in 1..K \end{cases} \tag{23}$$

The partial derivative $\partial \hat{\dot{\xi}}^{t,n} / \partial L_{\xi,ij}^k$ is:

$$\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial L_{\xi,ij}^k} = \frac{h^k(\xi^{t,n})}{2} \Bigg((\xi^{t,n} - \mu_\xi^k)^T (\Sigma_\xi^k)^{-1} \Phi (\Sigma_\xi^k)^{-1} (\xi^{t,n} - \mu_\xi^k) - tr\Big((\Sigma_\xi^k)^{-1} \Phi \Big) \Bigg) (A^k \xi^{t,n} - \hat{\dot{\xi}}^{t,n}) \tag{24}$$

where $\Phi = \mathbf{0}^{\{ij\}} (L_\xi^k)^T + L_\xi^k (\mathbf{0}^{\{ij\}})^T$, and has the dimension of $d \times d$..

D. Derivatives w.r.t. A^k

The partial derivatives of the cost function w.r.t. the components of A^k are

$$\frac{\partial J}{\partial A_{ij}^k} = \frac{1}{2N} \sum_{n=1}^{N} \sum_{t=0}^{T^n} \frac{\partial J}{\partial \hat{\dot{\xi}}^{t,n}} \frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial A_{ij}^k} \qquad \begin{cases} \forall i \in 1..d \\ \forall j \in 1..d \\ \forall k \in 1..K \end{cases} \tag{25}$$

The partial derivative $\partial \hat{\dot{\xi}}^{t,n} / \partial A_{ij}^k$ is:

$$\frac{\partial \hat{\dot{\xi}}^{t,n}}{\partial A_{ij}^k} = h^k(\xi^{t,n}) \mathbf{0}^{\{ij\}} \xi^{t,n} \tag{26}$$

where $\mathbf{0}^{\{ij\}}$ has the dimension of $d \times d$.

V. Likelihood Optimization

The likelihood optimization is defined as:

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{T^n} \log \mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n} | \boldsymbol{\theta}) \tag{27}$$

subject to the same constrains as given by Eq. (5). In Eq. (27), $\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n} | \boldsymbol{\theta})$ is computed from:

$$\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \mathcal{P}(k) \mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n} | k) \qquad \begin{cases} \forall n \in 1..N \\ t \in 0..T^n \end{cases} \tag{28}$$

where $\mathcal{P}(k) = \pi^k / (\sum_{i=1}^{K} \pi^i)$ is the prior and $\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n} | k)$ is the conditional probability density function given by:

$$\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n} | k) = \mathcal{N}(\xi^{t,n}, \dot{\xi}^{t,n}; \mu^k, \Sigma^k) = \frac{1}{\sqrt{(2\pi)^{2d} |\Sigma^k|}} e^{-\frac{1}{2} ([\xi^{t,n}, \dot{\xi}^{t,n}] - \mu^k)^T (\Sigma^k)^{-1} ([\xi^{t,n}, \dot{\xi}^{t,n}] - \mu^k)} \tag{29}$$

The optimization parameters for this objective function are: $\boldsymbol{\theta} = \{\pi^1..\pi^K; \mu_\xi^1..\mu_\xi^K; \Sigma^1..\Sigma^K\}$. Next we compute these derivatives with respect to $\boldsymbol{\theta}$.

A. Derivatives w.r.t. Priors π^k

$$\frac{\partial J}{\partial \pi^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\partial J}{\partial \mathcal{P}^{t,n}}\frac{\partial \mathcal{P}^{t,n}}{\partial \pi^k} \qquad \forall k \in 1..K \qquad (30)$$

where for simplicity we shorten the notation $\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n}; \boldsymbol{\theta})$ to $\mathcal{P}^{t,n}$. The partial derivatives $\frac{\partial J}{\partial \mathcal{P}^{t,n}}$ and $\frac{\partial \mathcal{P}^{t,n}}{\partial \pi^k}$ can be computed from Eqs. (31) and (32), respectively:

$$\frac{\partial J}{\partial \mathcal{P}^{t,n}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{1}{\mathcal{P}^{t,n}} \qquad (31)$$

$$\frac{\partial \mathcal{P}^{t,n}}{\partial \pi^k} = \mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n}|k) - \mathcal{P}^{t,n} \qquad (32)$$

Substituting Eqs. (31) and (32) into Eq. (30) yields:

$$\frac{\partial J}{\partial \pi^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\Big(\frac{\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n}|k)}{\mathcal{P}^{t,n}} - 1\Big) \qquad (33)$$

B. Derivatives w.r.t. Means μ_ξ^k

Special attention should be considered in computing derivatives with respect to μ_ξ^k. As it is already discussed in Section III, there is a direct relation between μ_ξ^k and $\mu_{\dot{\xi}}^k$ through the constraint Eq. (5)-(a). By substituting the corresponding value of $\mu_{\dot{\xi}}^k$ into the cost function given by Eq. (27), the optimization no longer depends on $\mu_{\dot{\xi}}^k$. Hence, we can drop $\mu_{\dot{\xi}}^k$ from the optimization parameters and the constraint Eq. (5)-(a) is always satisfied. However, this substitution should be considered when computing the derivatives with respect to μ_ξ^k:

$$\frac{\partial J}{\partial \mu_{\xi,i}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\partial J}{\partial \mathcal{P}^{t,n}}\Big(\frac{\partial \mathcal{P}^{t,n}}{\partial \mu_{\xi,i}^k} + \sum_{j=1}^{d}\frac{\partial \mathcal{P}^{t,n}}{\partial \mu_{\dot{\xi},j}^k}\frac{\partial \mu_{\dot{\xi},j}^k}{\partial \mu_{\xi,i}^k}\Big) \qquad (34)$$

The partial derivative $\frac{\partial J}{\partial \mathcal{P}^{t,n}}$ is given by Eq. (31), and $\partial \mathcal{P}^{t,n}/\partial \mu_{\xi,i}^k$ is:

$$\frac{\partial \mathcal{P}^{t,n}}{\partial \mu_{\xi,i}^k} = (\mathbf{0}^{\{i\}})^T(\Sigma^k)^{-1}\big([\xi^{t,n}; \dot{\xi}^{t,n}] - \mu^k\big)\mathcal{P}(k)\,\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n}|k) \qquad \forall i \in 1..d \qquad (35)$$

where $\mathbf{0}^{\{i\}}$ is a vector of dimension $2d$.

The partial derivative $\partial \mathcal{P}^{t,n}/\partial \mu_{\dot{\xi},j}^k$ can be computed similarly to Eq. (34); however by replacing $\mathbf{0}^{\{i\}}$ with $\mathbf{0}^{\{i+d\}}$.

The derivative $\frac{\partial \mu_{\dot{\xi},j}^k}{\partial \mu_{\xi,i}^k}$ can be computed by differentiating Eq. (5)-(a) with respect to $\mu_{\xi,i}^k$:

$$\frac{\partial \mu_{\dot{\xi},j}^k}{\partial \mu_{\xi,i}^k} = A_{ji}^k \qquad \forall i \in 1..d,\ j \in 1..d \qquad (36)$$

Thanks to matrix multiplication, we can significantly simplify the multiplications by substituting Eqs. (35), (36), and (31) into Eq. (34), and compute $\frac{\partial J}{\partial \mu_\xi^k}$:

$$\frac{\partial J}{\partial \mu_\xi^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\mathcal{P}(k)\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n}|k)}{\mathcal{P}^{t,n}}\big[\boldsymbol{I} \quad (A^k)^T\big](\Sigma^k)^{-1}\big([\xi^{t,n}; \dot{\xi}^{t,n}] - \mu^k\big) \qquad \forall i \in 1..d \qquad (37)$$

where $\boldsymbol{I}$ has the dimension of $d \times d$. Note that $\frac{\partial J}{\partial \mu_\xi^k}$ is now a *vector* of dimension d, and each $\frac{\partial J}{\partial \mu_{\xi,i}^k}$ is in fact one element of this vector.

C. Derivatives w.r.t. Means $\mu_{\dot{\xi}}^k$

By substituting directly the constraint Eq. (5)-(a) into Eq. (1), the partial derivative $\frac{\partial \mathcal{P}^{t,n}}{\partial \mu_{\dot{\xi},i}^k}$ is always zero because $\hat{f}(\xi)$ no longer depends on $\mu_{\dot{\xi}}^k$. Therefore, $\mu_{\dot{\xi}}^k$ can be dropped from the list of the optimization parameters. For more information see Section V-B.

D. Derivatives w.r.t. Σ^k

Similar to Section V-B, we need to consider the effect of substitution of $\mu_{\dot{\xi}}^k$ when computing the derivatives of Σ^k. All Σ^k are $2d \times 2d$ symmetric matrices, hence we compute the derivatives only for the components on the lower triangle matrix.

$$\frac{\partial J}{\partial \Sigma_{ij}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\partial J}{\partial \mathcal{P}^{t,n}}\Big(\frac{\partial \mathcal{P}^{t,n}}{\partial \Sigma_{ij}^k} + \frac{\partial \mathcal{P}^{t,n}}{\partial \Sigma_{ij}^k}\Big|_{\mu_{\dot{\xi}}^k}\Big) \qquad \begin{cases}\forall i \in 1..2d\\ \forall j \in 1..i\\ \forall k \in 1..K\end{cases} \qquad (38)$$

where $\frac{\partial \mathcal{P}^{t,n}}{\partial \Sigma_{ij}^k}\Big|_{\mu_{\dot{\xi}}^k}$ corresponds to the portion of derivatives due to the effect of $\mu_{\dot{\xi}}^k$, and can be computed from:

$$\frac{\partial \mathcal{P}^{t,n}}{\partial \Sigma_{ij}^k}\Big|_{\mu_{\dot{\xi}}^k} = \sum_{l=1}^{d}\sum_{m=1}^{d}\frac{\partial \mathcal{P}^{t,n}}{\partial \mu_{\dot{\xi},l}^k}\frac{\partial \mu_{\dot{\xi},l}^k}{\partial A_{lm}^k}\frac{\partial A_{lm}^k}{\partial \Sigma_{ij}^k} \qquad (39)$$

The partial derivative $\partial \mathcal{P}^{t,n}/\partial \Sigma_{ij}^k$ is:

$$\frac{\partial \mathcal{P}^{t,n}}{\partial \Sigma_{ij}^k} = 0.5\Big((\xi^{t,n} - \mu^k)^T(\Sigma^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}(\Sigma^k)^{-1}(\xi^{t,n} - \mu^k) - tr\big((\Sigma^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}\big)\Big)\mathcal{P}(k)\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n}|k) \qquad (40)$$

where $\mathbf{0}^{\{\overline{ij}\}}$ has the dimension of $2d \times 2d$.

The partial derivative $\frac{\partial \mathcal{P}^{t,n}}{\partial \Sigma_{ij}^k}\Big|_{\mu_{\dot{\xi}}^k}$ could be significantly simplified if it is computed in the matrix form (because we can drop the both summations on l and m):

$$\frac{\partial \mathcal{P}^{t,n}}{\partial \Sigma_{ij}^k}\Big|_{\mu_{\dot{\xi}}^k} = \mathcal{P}(k)\mathcal{P}(\xi^{t,n}, \dot{\xi}^{t,n}|k)(\xi^{t,n} - \mu^k)^T(\Sigma^k)^{-1}\mathcal{S}^k \qquad (41)$$

where $\mathcal{S}^k$ is a vector of dimension $2d$ and is equal to:

$$\mathcal{S}^k = \begin{bmatrix}\mathbf{0}\\ \Big(-A^k\big[\mathbf{0}^{\{\overline{ij}\}}\big]_\xi + \big[\mathbf{0}^{\{\overline{ij}\}}\big]_{\dot{\xi}\xi}\Big)(\Sigma_\xi^k)^{-1}\mu_\xi^k\end{bmatrix} \qquad (42)$$

In Eq. (42), $\mathbf{0}$ is a zero column vector of dimension d, and $\mathbf{0}_{\xi}^{\{\overline{ij}\}}$ and $\mathbf{0}_{\dot{\xi}\xi}^{\{\overline{ij}\}}$ are partitions of $\mathbf{0}^{\{\overline{ij}\}}$. Finally, by substituting Eqs. (40), (41), and (31) into Eq. (38) we have:

$$\frac{\partial J}{\partial \Sigma_{ij}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\mathcal{P}(k)\mathcal{P}(\xi^{t,n},\dot{\xi}^{t,n}|k)}{\mathcal{P}^{t,n}} \Big(0.5(\xi^{t,n}-\mu^k)^T(\Sigma^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}(\Sigma^k)^{-1}(\xi^{t,n}-\mu^k) - 0.5tr\big((\Sigma^k)^{-1}\mathbf{0}^{\{\overline{ij}\}}\big) + (\xi^{t,n}-\mu^k)^T(\Sigma^k)^{-1}\mathcal{S}^k\Big) \quad (43)$$

VI. Alternative Likelihood Optimization

Similarly to Section IV, we can define an alternative likelihood optimization so that 4 out of 5 optimization constraints can be automatically satisfied through a change of variable:

$$\begin{cases}\tilde{\pi}^k = ln(\pi^k)\\ L^k = Chol(\Sigma^k)\end{cases} \quad (44)$$

where L^k are $2d \times 2d$ lower triangle matrices. Since Σ^k are positive definite matrices, their Cholesky decomposition always exist. The alternative likelihood optimization can be expressed as:

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\log \mathcal{P}(\xi^{t,n},\dot{\xi}^{t,n}|\boldsymbol{\theta}) \quad (45)$$

subject to

$$A^k + (A^k)^T < 0 \qquad \forall k \in 1..K \quad (46)$$

where $\boldsymbol{\theta} = \{\tilde{\pi}^1..\tilde{\pi}^K; \mu_{\xi}^1..\mu_{\xi}^K; L^1..L^K\}$. Once the optimization finished, the parameters of GMM can be reconstructed as follows:

$$\begin{cases}\pi^k = e^{\tilde{\pi}^k}/(\sum_{i=1}^{K} e^{\tilde{\pi}^i})\\ \Sigma^k = L^k(L^k)^T\end{cases} \quad (47)$$

In fact the proposed change of parameters allows us to automatically satisfy the last three optimization constraints of Eq. (5). The first constraint of Eq. (5) is also removed since it is directly considered in Eq. (17). The derivatives of the new optimization problem are provided in the following subsections.

A. Derivatives w.r.t. Priors π^k

$$\frac{\partial J}{\partial \tilde{\pi}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\partial J}{\partial \mathcal{P}^{t,n}}\frac{\partial \mathcal{P}^{t,n}}{\partial \pi^k}\frac{\partial \pi^k}{\partial \tilde{\pi}^k} \qquad \forall k \in 1..K \quad (48)$$

The partial derivatives $\partial J/\partial \mathcal{P}^{t,n}$, $\partial \mathcal{P}^{t,n}/\partial \pi^k$ and $\partial \pi^k/\partial \tilde{\pi}^k$ are given by Eqs. (31), (32), and (22), respectively.

B. Derivatives w.r.t. Means μ_{ξ}^k

These derivative can be similarly computed from Eq. (34).

C. Derivatives w.r.t. L^k

L^k is a $2d \times 2d$ lower triangle matrix. The partial derivatives of the cost function with respect to the optimization parameters are:

$$\frac{\partial J}{\partial L_{ij}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\partial J}{\partial \mathcal{P}^{t,n}}\frac{\partial \mathcal{P}^{t,n}}{\partial L_{ij}^k} \qquad \begin{cases}\forall i \in 1..2d\\ \forall j \in 1..i\\ \forall k \in 1..K\end{cases} \quad (49)$$

The partial derivative $\partial \mathcal{P}^{t,n}/\partial L_{ij}^k$ is:

$$\frac{\partial J}{\partial L_{ij}^k} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T^n}\frac{\mathcal{P}(k)\mathcal{P}(\xi^{t,n},\dot{\xi}^{t,n}|k)}{\mathcal{P}^{t,n}} \Big(0.5(\xi^{t,n}-\mu^k)^T(\Sigma^k)^{-1}\Phi(\Sigma^k)^{-1}(\xi^{t,n}-\mu^k) - 0.5tr\big((\Sigma^k)^{-1}\Phi\big) + (\xi^{t,n}-\mu^k)^T(\Sigma^k)^{-1}\tilde{\mathcal{S}}^k\Big) \quad (50)$$

where $\Phi = \mathbf{0}^{\{ij\}}(L^k)^T + L^k(\mathbf{0}^{\{ij\}})^T$, and has the dimension of $2d \times 2d$. The $2d$ dimension vector $\tilde{\mathcal{S}}^k$ is:

$$\tilde{\mathcal{S}}^k = \begin{bmatrix}\mathbf{0}\\ (-A^k\Phi_{\xi} + \Phi_{\dot{\xi}\xi})(\Sigma_{\xi}^k)^{-1}\mu_{\xi}^k\end{bmatrix} \quad (51)$$

where $\mathbf{0}$ is a zero column vector of dimension d.

VII. Optimization constraints and their derivative

In this section we provide formulations for the optimization problems defined in Sections IV and VI, where the only constraint is the negative definiteness of matrices A^k. To ensure this constraint, we first need to define a method to mathematically determine whether a matrix is negative definite. There are several ways to ensure whether a symmetric matrix B is negative definite, among which the two most famous ones are 1) all eigenvalues of B are strictly negative, 2) using Sylvesters criterion. In our work, we use Sylvesters criterion because it provides us with an analytical formulation to verify negative definiteness (compared to computing eigenvalues which is an iterative procedure).

Sylvester's criterion states that a Hermitian matrix B is negative-definite if and only if the determinant of all i-th order leading principal minors[1] are negative if i is odd and positive if i is even [2]. Each $d \times d$ symmetric matrix has d principal minors. By defining $B^k = A^k + (A^k)^T$, the optimization constraint given by Eq. (46) is equal to:

$$\mathcal{C}_{(k-1)d+c}: \quad (-1)^{c+1}|B_{1:c,1:c}| < 0 \qquad \begin{cases}\forall c \in 1..d\\ \forall k \in 1..K\end{cases} \quad (52)$$

where we use $\mathcal{C}_{(k-1)d+c}$ to refer to the $((k-1)d+c)$-th constraint. Thus for a GMM model composed of K Gaussian functions, there are $K \times d$ constraints that should be satisfied

[1]The i-th principal minor of a $d \times d$ symmetric matrix B is a quadratic upper-left part of B, which consists of matrix elements in rows and columns from 1 to d.

during the optimization. The derivative of these constraints with respect to π^k and μ^k are always zero, irrespective of which cost function is used:

$$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial \tilde{\pi}^k} = 0 \qquad \begin{cases} \forall c \in 1..d \\ \forall k \in 1..K \end{cases} \tag{53}$$

$$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial \mu_i^k} = 0 \qquad \begin{cases} \forall c \in 1..d \\ \forall i \in 1..2d \\ \forall k \in 1..K \end{cases} \tag{54}$$

For the MSE optimization defined by Eq. (18) we have:

$$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial L_{ij}^k} = 0 \qquad \begin{cases} \forall c \in 1..d \\ \forall i \in 1..d \\ \forall j \in 1..i \\ \forall k \in 1..K \end{cases} \tag{55}$$

$$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial A_{ij}^k} = (-1)^{c+1} tr\Big(adj\big(B_{1:c,1:c}\big)\big[\mathbf{0}^{\{\overline{ij}\}}\big]_{1:c,1:c}\Big) \begin{cases} \forall c \in 1..d \\ \forall i \in 1..d \\ \forall j \in 1..d \\ \forall k \in 1..K \end{cases} \tag{56}$$

where $\mathbf{0}^{\{\overline{ij}\}}$ has the dimension of $d \times d$. For the likelihood optimization defined by Eq. (45) we have:

$$\frac{\partial \mathcal{C}_{(k-1)d+c}}{\partial L_{ij}^k} = (-1)^{c+1} tr\Big(adj\big(B_{1:c,1:c}\big)\mathcal{X}_{1:c,1:c}\Big) \begin{cases} \forall c \in 1..d \\ \forall i \in 1..2d \\ \forall j \in 1..i \\ \forall k \in 1..K \end{cases} \tag{57}$$

where $\mathcal{X}$ is a $d \times d$ symmetric matrix defined by:

$$\Phi = \mathbf{0}^{\{ij\}}(L^k)^T + L^k(\mathbf{0}^{\{ij\}})^T \tag{58}$$

$$\Psi = \big(-A^k\Phi_\xi + \Phi_{\dot{\xi}\xi}\big)(\Sigma_\xi)^{-1} \tag{59}$$

$$\mathcal{X} = \Psi + (\Psi)^T \tag{60}$$

where Φ is a $2d \times 2d$ matrix.

REFERENCES

[1] S.-M. Khansari-Zadeh and A. Billard, "Imitation learning of globally stable non-linear point-to-point robot motions using nonlinear programming," in *Proceeding of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 2676–2683.

[2] G. T. Gilbert, "Positive definite matrices and Sylvester's criterion," *The American Mathematical Monthly (Mathematical Association of America)*, vol. 98(1), pp. 44–46, 1991.