Composition of Movement Primitives

Andrea Pierré

June 20, 2025

Contents

1 ProMPs

1.1 Recap

From (Paraschos et al., 2013, 2018):

Table 1: Notation

q_t	joint angle over time
$\dot{q}_t$	joint velocity over time
$\boldsymbol{\tau} = \{q_t\}_{t=0\dots T}$	trajectory
$\boldsymbol{w}$	weight vector of a single trajectory $[n \times 1]$
ϕ_t	basis function
n	number of basis functions
$\boldsymbol{\Phi}_t = [\phi_t, \dot{\phi}_t]$	$n \times 2$ dimensional time-dependent basis matrix
$z(t)$	monotonically increasing phase variable
$\boldsymbol{\epsilon}_y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_y)$	zero-mean i.i.d. Gaussian noise

$$\boldsymbol{\Phi}_t = \begin{bmatrix} \phi_1 & \dot{\phi}_1 \\ \vdots & \vdots \\ \phi_n & \dot{\phi}_n \end{bmatrix} \tag{1}$$

$$\boldsymbol{y}_t = \begin{bmatrix} q_t \\ \dot{q}_t \end{bmatrix} = \boldsymbol{\Phi}_t^\top \boldsymbol{w} + \boldsymbol{\epsilon}_y \tag{2}$$

$$p(\boldsymbol{\tau}|\boldsymbol{w}) = \prod_t \mathcal{N}\Big(\boldsymbol{y}_t|\boldsymbol{\Phi}_t^\top \boldsymbol{w}, \boldsymbol{\Sigma}_y\Big) \tag{3}$$

$$p(\boldsymbol{\tau};\boldsymbol{\theta}) = \int p(\boldsymbol{\tau}|\boldsymbol{w}) \cdot p(\boldsymbol{w};\boldsymbol{\theta}) d\boldsymbol{w} \tag{4}$$

1.2 Coupling between joints

$$p(\boldsymbol{y}_t|\boldsymbol{w}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{y}_{1,t} \\ \vdots \\ \boldsymbol{y}_{d,t} \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\Phi}_t^\top & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{\Phi}_t^\top \end{bmatrix} \boldsymbol{w}, \boldsymbol{\Sigma}_y \right) = \mathcal{N}\Big(\boldsymbol{y}_t|\boldsymbol{\Psi}_t \boldsymbol{w}, \boldsymbol{\Sigma}_y\Big) \tag{5}$$

with:

Table 2: Notation

$\boldsymbol{w} = [\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_n^\top]^\top$	combined weight vector $[n \times n]$
$\boldsymbol{\Psi}_t$	block-diagonal basis matrix containing the basis functions and their derivatives for each dimension
$\boldsymbol{y}_{i,t} = [q_{i,t}, \dot{q}_{i,t}]^\top$	joint angle and velocity for the i^{th} joint

1.3 Hierarchical Bayesian Model

The Hierarchical Bayesian Model used in ProMPs is illustrated in Fig. 1.

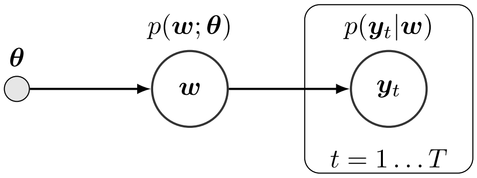
$\boldsymbol{\theta}$
$p(\boldsymbol{w};\boldsymbol{\theta})$
$\boldsymbol{w}$
$p(\boldsymbol{y}_t|\boldsymbol{w})$
$\boldsymbol{y}_t$
$t = 1 \ldots T$

Figure 1: Hierarchical Bayesian Model used in ProMPs.

$$\begin{aligned} p(\boldsymbol{y}_t;\boldsymbol{\theta}) &= \int \mathcal{N}\Big(\boldsymbol{y}_t|\boldsymbol{\Psi}_t^\top \boldsymbol{w}, \boldsymbol{\Sigma}_y\Big) \cdot p(\boldsymbol{w};\boldsymbol{\theta})\ d\boldsymbol{w} & (6) \\ &= \int \mathcal{N}\Big(\boldsymbol{y}_t|\boldsymbol{\Psi}_t^\top \boldsymbol{w}, \boldsymbol{\Sigma}_y\Big) \cdot \mathcal{N}\Big(\boldsymbol{w}|\boldsymbol{\mu}_{\boldsymbol{w}}, \boldsymbol{\Sigma}_{\boldsymbol{w}}\Big)\ d\boldsymbol{w} & (7) \\ &= \mathcal{N}\Big(\boldsymbol{y}_t|\boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}}, \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y\Big) & (8) \end{aligned}$$

See Appendix A for the proof.

Table 3: Notation

$\boldsymbol{\theta} = \{\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w\}$	parameters
$p(\boldsymbol{w};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$	prior over the weight vector $\boldsymbol{w}$, with parameters $\boldsymbol{\theta}$, assumed to be Gaussian

Table 4: Notation

$\boldsymbol{x}_t^\star = [\boldsymbol{y}_t^\star, \boldsymbol{\Sigma}_t^\star]$	desired observation
$\boldsymbol{y}_t^\star$	desired position and velocity vector at time t
$\boldsymbol{\Sigma}_t^\star$	accuracy of the desired observation

1.4 Via-Points Modulation

Using Bayes rule:

$$p(\boldsymbol{w}|\boldsymbol{x}_t^\star) = \frac{p(\boldsymbol{x}_t^\star|\boldsymbol{w}) \cdot p(\boldsymbol{w})}{p(\boldsymbol{x}_t^\star)} \tag{9}$$

$$p(\boldsymbol{w}|\boldsymbol{x}_t^\star) \propto \mathcal{N}\left(\boldsymbol{y}_t^\star|\boldsymbol{\Psi}_t^\top \boldsymbol{w}, \boldsymbol{\Sigma}_t^\star\right) \cdot \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w) \tag{10}$$

$$\boldsymbol{\mu}_{\boldsymbol{w}}^{[new]} = \boldsymbol{\mu}_{\boldsymbol{w}} + \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t \left(\boldsymbol{\Sigma}_y^\star + \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t\right)^{-1} (\boldsymbol{y}_t^\star - \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}}) \tag{11}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w}}^{[new]} = \boldsymbol{\Sigma}_{\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t \left(\boldsymbol{\Sigma}_y^\star + \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t\right)^{-1} \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \tag{12}$$

See Appendix B for the proof.

1.4.1 Do we actually get the desired mean by applying the conditioning update?

Proof that the posterior mean equals the observed mean.

$$\mathbb{E}[\boldsymbol{y}_t|\boldsymbol{x}_t^\star] = \boldsymbol{\mu}_{\boldsymbol{y}_t|\boldsymbol{x}_t^\star} = \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}_t^\star} = \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}} + \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t \left(\boldsymbol{\Sigma}_t^\star + \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t\right)^{-1} (\boldsymbol{y}_t^\star - \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}}) \tag{13}$$

We set the observed covariance $\boldsymbol{\Sigma}_t^\star$ to 0 so as to have perfect accuracy around our observed position.

$$\begin{aligned} \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}_t^\star} &= \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}} + \cancel{\boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t} \left(\cancel{\boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t}\right)^{-1} (\boldsymbol{y}_t^\star - \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}}) && (14) \\ &= \cancel{\boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}}} + \boldsymbol{y}_t^\star - \cancel{\boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}}} && (15) \\ &= \boldsymbol{y}_t^\star && (16) \end{aligned}$$

□

1.4.2 Multi via-points

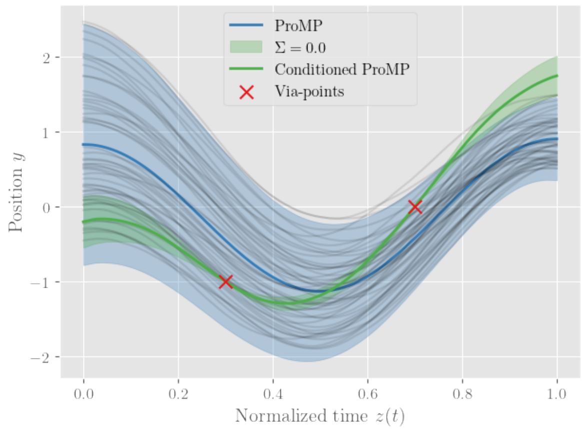

Figure 2: Example of ProMP with two via-points.

1. For the first via-point conditioning update with the observed via-point $\boldsymbol{x}^\star_{t_1} = [\boldsymbol{y}^\star_{t_1}, \boldsymbol{\Sigma}^\star_{t_1}]$, we can directly apply Eq. (11) and (12), with $\boldsymbol{\Psi}_{t_1}$ the observation matrix at time t_1:

$$\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}} = \boldsymbol{\mu}_{\boldsymbol{w}} + \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_{t_1} \left(\boldsymbol{\Sigma}^\star_{t_1} + \boldsymbol{\Psi}^\top_{t_1} \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_{t_1} \right)^{-1} (\boldsymbol{y}^\star_{t_1} - \boldsymbol{\Psi}^\top_{t_1} \boldsymbol{\mu}_{\boldsymbol{w}}) \tag{17}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}} = \boldsymbol{\Sigma}_{\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_{t_1} \left(\boldsymbol{\Sigma}^\star_{t_1} + \boldsymbol{\Psi}^\top_{t_1} \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_{t_1} \right)^{-1} \boldsymbol{\Psi}^\top_{t_1} \boldsymbol{\Sigma}_{\boldsymbol{w}} \tag{18}$$

2. For the second via-point conditioning update with the observed via-point $\boldsymbol{x}^\star_{t_2} = [\boldsymbol{y}^\star_{t_2}, \boldsymbol{\Sigma}^\star_{t_2}]$, the prior is the posterior from the first via-point, *i.e.*, $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}}, \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}})$, the likelihood is $\boldsymbol{y}^\star_{t_2} \sim \mathcal{N}(\boldsymbol{\Psi}^\top_{t_2} \boldsymbol{w}, \boldsymbol{\Sigma}^\star_{t_2})$, with $\boldsymbol{\Psi}_{t_2}$ the observation matrix at time t_2, and the posterior update becomes:

$$\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}, \boldsymbol{x}^\star_{t_2}} = \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}} + \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}} \boldsymbol{\Psi}_{t_2} \left(\boldsymbol{\Sigma}^\star_{t_2} + \boldsymbol{\Psi}^\top_{t_2} \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}} \boldsymbol{\Psi}_{t_2} \right)^{-1} (\boldsymbol{y}^\star_{t_2} - \boldsymbol{\Psi}^\top_{t_2} \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}}) \tag{19}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}, \boldsymbol{x}^\star_{t_2}} = \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}} - \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}} \boldsymbol{\Psi}_{t_2} \left(\boldsymbol{\Sigma}^\star_{t_2} + \boldsymbol{\Psi}^\top_{t_2} \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}} \boldsymbol{\Psi}_{t_2} \right)^{-1} \boldsymbol{\Psi}^\top_{t_2} \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1}} \tag{20}$$

3. For the k^{th} via-point conditioning update with the observed via-point $\boldsymbol{x}^\star_{t_k} = [\boldsymbol{y}^\star_{t_k}, \boldsymbol{\Sigma}^\star_{t_k}]$, the prior is the posterior after conditioning on the previous $k-1$ via-points, *i.e.*, $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}}, \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}})$, the likelihood is $\boldsymbol{y}^\star_{t_k} \sim \mathcal{N}(\boldsymbol{\Psi}^\top_{t_k} \boldsymbol{w}, \boldsymbol{\Sigma}^\star_{t_k})$, with $\boldsymbol{\Psi}_{t_k}$ the observation matrix at time t_k, and the posterior update becomes:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_k}} &= \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}} \\ &\quad + \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}} \boldsymbol{\Psi}_{t_k} \left(\boldsymbol{\Sigma}^\star_{t_k} + \boldsymbol{\Psi}^\top_{t_k} \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}} \boldsymbol{\Psi}_{t_k} \right)^{-1} (\boldsymbol{y}^\star_{t_k} - \boldsymbol{\Psi}^\top_{t_k} \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}}) \end{aligned} \tag{21}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_k}} &= \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}} \\ &\quad - \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}} \boldsymbol{\Psi}_{t_k} \left(\boldsymbol{\Sigma}^\star_{t_k} + \boldsymbol{\Psi}^\top_{t_k} \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}} \boldsymbol{\Psi}_{t_k} \right)^{-1} \boldsymbol{\Psi}^\top_{t_k} \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}^\star_{t_1},\ldots,\boldsymbol{x}^\star_{t_{k-1}}} \end{aligned} \tag{22}$$

Alternative Batch Formulation Instead of iterative updates, we could condition on all via-points simultaneously by stacking the observations:

$$\boldsymbol{y}^\star = \begin{bmatrix} \boldsymbol{y}^\star_{t_1} \\ \vdots \\ \boldsymbol{y}^\star_{t_k} \end{bmatrix}, \quad \boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_{t_1} \\ \vdots \\ \boldsymbol{\Psi}_{t_k} \end{bmatrix}, \quad \boldsymbol{\Sigma}^\star = \operatorname{diag}(\boldsymbol{\Sigma}^\star_{t_1}, \ldots, \boldsymbol{\Sigma}^\star_{t_k}) \tag{23}$$

$$\boldsymbol{\mu}_{\boldsymbol{w}|\{\boldsymbol{x}^\star_{t_k}\}_{k=1}^K} = \boldsymbol{\mu}_{\boldsymbol{w}} + \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi} \left(\boldsymbol{\Sigma}^\star + \boldsymbol{\Psi}^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi} \right)^{-1} (\boldsymbol{y}^\star - \boldsymbol{\Psi}^\top \boldsymbol{\mu}_{\boldsymbol{w}}) \tag{24}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w}|\{\boldsymbol{x}^\star_{t_k}\}_{k=1}^K} = \boldsymbol{\Sigma}_{\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi} \left(\boldsymbol{\Sigma}^\star + \boldsymbol{\Psi}^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi} \right)^{-1} \boldsymbol{\Psi}^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \tag{25}$$

2 Gaussian mixture modeling (GMM)/Gaussian mixture regression (GMR) recap

2.1 Gaussian Mixture Modeling (GMM)

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{26}$$

$$0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1 \tag{27}$$

$$r_{nk} := \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{28}$$

Table 5: Notation

π_k	mixture weights
$\boldsymbol{\theta} := \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k : k = 1, \ldots, K\}$	collection of all parameters of the model
r_{nk}	responsibility of the k^{th} mixture component for the n^{th} data point
N	number of data points
$N_k := \sum_{n=1}^{N} r_{nk}$	total responsibility of the k^{th} mixture component for the entire dataset

Update of the GMM means:

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^{N} r_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}} \tag{29}$$

Update of the GMM covariances:

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^{\top} \tag{30}$$

Update of the GMM mixture weights:

$$\pi_k^{new} = \frac{N_k}{N}, \quad k = 1, \ldots, K \tag{31}$$

The parameters are updated using Algorithm 1.

2.2 Gaussian Mixture Regression (GMR)

At each iteration step t, the datapoint $\boldsymbol{x}_t$ can be decomposed as two subvectors $\boldsymbol{x}_t^I$ and $\boldsymbol{x}_t^O$ spanning for the input and output dimensions. For trajectory encoding in task space, I corresponds to the time input dimension (*e.g.*, value of a decay term), and O corresponds to the output dimensions describing a path (*e.g.*, end-effector position in task space).

During the training phase we learn the joint probability $p(\boldsymbol{x}_t^I, \boldsymbol{x}_t^O)$ with GMM through EM (Algorithm 1) with:

$$\begin{aligned} p(\boldsymbol{x}_t^I, \boldsymbol{x}_t^O) &= \sum_{k=1}^{K} \pi_k \mathcal{N}_k(\boldsymbol{x}_t^I, \boldsymbol{x}_t^O | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) && (32)\\ &= \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_t^O | \boldsymbol{x}_t^I) \cdot p(\boldsymbol{x}_t^I) && (33)\\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}_k(\boldsymbol{x}_t^O | \hat{\boldsymbol{\mu}}_k^O(\boldsymbol{x}_t^I), \hat{\boldsymbol{\Sigma}}_k^O) \cdot \mathcal{N}_k(\boldsymbol{\mu}_k^I, \boldsymbol{\Sigma}_k^I) && (34) \end{aligned}$$

$$\boldsymbol{x}_t = \begin{bmatrix} \boldsymbol{x}_t^I \\ \boldsymbol{x}_t^O \end{bmatrix}, \quad \boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^I \\ \boldsymbol{\mu}_k^O \end{bmatrix}, \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^I & \boldsymbol{\Sigma}_k^{IO} \\ \boldsymbol{\Sigma}_k^{OI} & \boldsymbol{\Sigma}_k^O \end{bmatrix} \tag{35}$$

$$\begin{aligned} \hat{\boldsymbol{\mu}}_k^O(\boldsymbol{x}_t^I) &= \boldsymbol{\mu}_k^O + \boldsymbol{\Sigma}_k^{OI}(\boldsymbol{\Sigma}_k^I)^{-1}(\boldsymbol{x}_t^I - \boldsymbol{\mu}_k^I) && (36)\\ \hat{\boldsymbol{\Sigma}}_k^O &= \boldsymbol{\Sigma}_k^O - \boldsymbol{\Sigma}_k^{OI}(\boldsymbol{\Sigma}_k^I)^{-1} \cdot \boldsymbol{\Sigma}_k^{IO} && (37) \end{aligned}$$

The marginal probability $p(\boldsymbol{x}_t^I)$ is:

$$p(\boldsymbol{x}_t^I) = \int p(\boldsymbol{x}_t^I, \boldsymbol{x}_t^O) d\boldsymbol{x}_t^O = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}_k(\boldsymbol{x}_t^I | \boldsymbol{\mu}_k^I, \boldsymbol{\Sigma}_k^I) \tag{38}$$

Algorithm 1: EXPECTATION MAXIMIZATION (EM) algorithm for a Gaussian mixture model

```
Input  : Initial model parameters $\{\boldsymbol{\mu}_k\}$, $\{\boldsymbol{\Sigma}_k\}$, $\{\pi_k\}$
Input  : Data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
Output: Final model parameters $\{\boldsymbol{\mu}_k\}$, $\{\boldsymbol{\Sigma}_k\}$, $\{\pi_k\}$
repeat
    // E-step
    for $n \in \{1, \ldots, N\}$ do
        for $k \in \{1, \ldots, K\}$ do
            $r_{nk} \leftarrow \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$
        end
    end
    // M-step
    for $k \in \{1, \ldots, K\}$ do
        $N_k = \sum_{n=1}^{N} r_{nk}$
        $\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n$
        $\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top$
        $\pi_k = \frac{N_k}{N}$
    end
    // Log likelihood
    $\mathcal{L} \leftarrow \sum_{n=1}^{N} \ln \left[ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$
until convergence
return $\{\boldsymbol{\mu}_k\}$, $\{\boldsymbol{\Sigma}_k\}$, $\{\pi_k\}$
```

Table 6: Notation

$\boldsymbol{x}_t \in \mathbb{R}^D$	datapoint at timestep t
$\boldsymbol{\mu}_k$	center of the k^{th} Gaussian
$\boldsymbol{\Sigma}_k$	covariance of the k^{th} Gaussian

To retrieve the multimodal conditional distribution $p(\boldsymbol{x}_t^O|\boldsymbol{x}_t^I)$ for each input/timestep we have:

$$p(\boldsymbol{x}_t^O|\boldsymbol{x}_t^I) = \frac{p(\boldsymbol{x}_t^O, \boldsymbol{x}_t^I)}{p(\boldsymbol{x}_t^I)} \tag{39}$$

$$= \frac{\sum_{k=1}^K \pi_k \mathcal{N}_k(\hat{\boldsymbol{\mu}}_k^O(\boldsymbol{x}_t^I), \hat{\boldsymbol{\Sigma}}_k^O) \cdot \mathcal{N}_k(\boldsymbol{x}_t^I|\boldsymbol{\mu}_k^I, \boldsymbol{\Sigma}_k^I)}{\sum_{k=1}^K \pi_k \cdot \mathcal{N}_k(\boldsymbol{x}_t^I|\boldsymbol{\mu}_k^I, \boldsymbol{\Sigma}_k^I)} \tag{40}$$

$$= \sum_{k=1}^K r_{nk} \cdot \mathcal{N}_k(\hat{\boldsymbol{\mu}}_k^O(\boldsymbol{x}_t^I), \hat{\boldsymbol{\Sigma}}_k^O) \tag{41}$$

When a unimodal output distribution is required, the law of total mean and variance can be used to approximate the distribution with the Gaussian (see Appendix C):

ToDo

$$m(x) = \mathbb{E}(x_j|t_j) = \sum_{i=1}^K r_{nk} \cdot m_i(t_j) \tag{42}$$

$$var(x) = \sum_{j=1}^K r_{nk} \cdot cov_i \tag{43}$$

3 Composition of MPs

3.1 Stitching

The main issue with stitching is the smoothness of the mean and covariance between ProMPs, see Fig. 3.

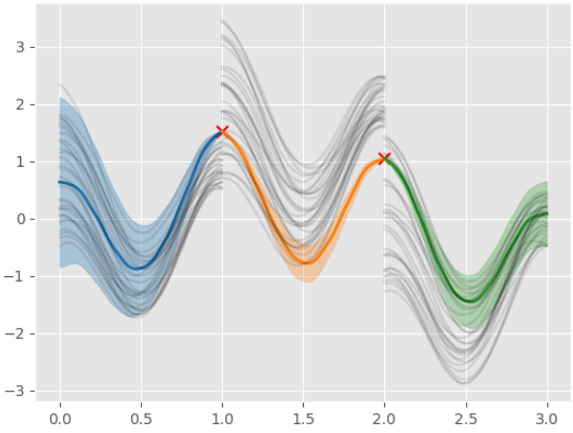

Figure 3: Stitching three ProMPs.

Table 7: Notation

$\boldsymbol{t} = [t_1, \ldots, t_m]$	timestep vector
$\boldsymbol{\mu}_{kt}$	mean of t_i for the k^{th} Gaussian cluster
$h_k(t_i)$	responsibility of the k^{th} mixture component at timestep t_i

3.2 Mixture of ProMPs

3.2.1 Hierarchical Bayesian Model

Learn the joint distribution $p(t_i, y_i)$ as a mixture of ProMPs, and retrieve the conditional distribution $p(y_i|t_i)$ for each timestep t_i.

$$p(t_i, y_i) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}_k\big(y_i|t_i; \hat{\boldsymbol{\mu}}_k(t_i), \hat{\boldsymbol{\Sigma}}_k\big) \cdot \mathcal{N}_k\big(t_i|\mu_{kt}, \boldsymbol{\Sigma}_{kt}\big) \tag{44}$$

$$\boldsymbol{x}_t = \begin{bmatrix} \boldsymbol{t} \\ \boldsymbol{y}_t \end{bmatrix}, \quad \boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{kt} \\ \boldsymbol{\mu}_{ky} \end{bmatrix} = \begin{bmatrix} \cdot \\ \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}} \end{bmatrix}, \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_{ktt} & \boldsymbol{\Sigma}_{kty} \\ \boldsymbol{\Sigma}_{kyt} & \boldsymbol{\Sigma}_{kyy} \end{bmatrix} = \begin{bmatrix} \cdot & \cdot \\ \cdot & \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y \end{bmatrix} \tag{45}$$

$$\hat{\boldsymbol{\mu}}_k(t_i) = \boldsymbol{\mu}_{ky} + \boldsymbol{\Sigma}_{kyt} \boldsymbol{\Sigma}_{ktt}^{-1} (t_i - \boldsymbol{\mu}_{kt}) \tag{46}$$

$$\hat{\boldsymbol{\Sigma}}_k = \boldsymbol{\Sigma}_{kyy} - \boldsymbol{\Sigma}_{kyt} \boldsymbol{\Sigma}_{ktt}^{-1} \boldsymbol{\Sigma}_{kty} \tag{47}$$

Marginal probability $p(t_i)$:

$$p(t_i) = \int p(t_i, y_i) dy = \sum_{k=1}^{K} \pi_k \mathcal{N}_k(t_i|\boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_{kt}) \tag{48}$$

Conditional probability $p\big(y_i|t_i; \hat{\boldsymbol{\mu}}_k(t_i), \hat{\boldsymbol{\Sigma}}_k\big)$:

$$p\big(y_i|t_i; \hat{\boldsymbol{\mu}}_k(t_i), \hat{\boldsymbol{\Sigma}}_k\big) = \frac{p(t_i, y_i)}{p(t_i)} \tag{49}$$

$$= \frac{\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}_k\big(y_i|t_i; \hat{\boldsymbol{\mu}}_k(t_i), \hat{\boldsymbol{\Sigma}}_k\big) \cdot \mathcal{N}_k\big(t_i|\mu_{kt}, \boldsymbol{\Sigma}_{kt}\big)}{\sum_{k=1}^{K} \pi_k \mathcal{N}_k\big(t_i|\boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_{kt}\big)} \tag{50}$$

$$= \sum_{k=1}^{K} h_k(t_i) \mathcal{N}_k\big(y_i|t_i; \hat{\boldsymbol{\mu}}_k(t_i), \hat{\boldsymbol{\Sigma}}_k\big) \tag{51}$$

with:

$$h_k(t_i) = \frac{\pi_k \mathcal{N}_k\big(t_i|\mu_{kt}, \boldsymbol{\Sigma}_{kt}\big)}{\sum_{k=1}^{K} \pi_k \mathcal{N}_k\big(t_i|\boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_{kt}\big)} \tag{52}$$

3.2.2 Via-Points Modulation

ToDo

3.3 Piecewise Gaussian Process

References

A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann, "Probabilistic Movement Primitives," in *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc., 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/e53a0a2978c28872a4505bdb51db06dc-Abstract.html

A. Paraschos, C. Daniel, J. Peters, and G. Neumann, “Using probabilistic movement primitives in robotics,” *Autonomous Robots*, vol. 42, no. 3, pp. 529–551, Mar. 2018. [Online]. Available: https://doi.org/10.1007/s10514-017-9648-7

M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning.* Cambridge University Press, 2020.

C. M. Bishop and H. Bishop, *Deep Learning: Foundations and Concepts.* Springer International Publishing, 2024. [Online]. Available: https://doi.org/10.1007/978-3-031-45468-4

S. Calinon, “A tutorial on task-parameterized movement learning and retrieval,” *Intelligent Service Robotics*, vol. 9, no. 1, pp. 1–29, Jan. 2016. [Online]. Available: https://doi.org/10.1007/s11370-015-0187-9

A Hierarchical Bayesian Model proof

Proof of Eq. (8). From (Deisenroth et al., 2020), we have the joint distribution:

$$p(\mathbf{x}_a, \mathbf{x}_b) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \right) \tag{53}$$

and the marginal distribution $p(\mathbf{x}_a)$ of a joint Gaussian distribution $p(\mathbf{x}_a, \mathbf{x}_b)$:

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \tag{54}$$

Since $\boldsymbol{y}_t$ and $\boldsymbol{w}$ are jointly Gaussian, we have:

$$\begin{bmatrix} \boldsymbol{y}_t \\ \boldsymbol{w} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{y}_t} \\ \boldsymbol{\mu}_{\boldsymbol{w}} \end{bmatrix}, \begin{bmatrix} \mathrm{Cov}[\boldsymbol{y}_t, \boldsymbol{y}_t] & \mathrm{Cov}[\boldsymbol{y}_t, \boldsymbol{w}] \\ \mathrm{Cov}[\boldsymbol{w}, \boldsymbol{y}_t] & \mathrm{Cov}[\boldsymbol{w}, \boldsymbol{w}] \end{bmatrix} \right) \tag{55}$$

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{y}_t} &= \mathbb{E}[\boldsymbol{y}_t] && (56) \\ &= \mathbb{E}[\boldsymbol{\Psi}_t^\top \boldsymbol{w} + \boldsymbol{\epsilon}_y] && (57) \\ &= \boldsymbol{\Psi}_t^\top \mathbb{E}[\boldsymbol{w}] + \mathbb{E}[\boldsymbol{\epsilon}_y] && (58) \\ &= \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}} + 0 && (59) \\ &= \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}} && (60) \end{aligned}$$

$$\begin{aligned} \mathrm{Cov}[\boldsymbol{y}_t, \boldsymbol{y}_t] &= \mathrm{Cov}[\boldsymbol{\Psi}_t^\top \boldsymbol{w} + \boldsymbol{\epsilon}_y] && (61) \\ &= \mathrm{Cov}[\boldsymbol{\Psi}_t^\top \boldsymbol{w}] + \mathrm{Cov}[\boldsymbol{\epsilon}_y] && (62) \\ &= \boldsymbol{\Psi}_t^\top \mathrm{Cov}[\boldsymbol{w}] \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y && (63) \\ &= \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y && (64) \end{aligned}$$

$$\begin{bmatrix} \boldsymbol{y}_t \\ \boldsymbol{w} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}} \\ \boldsymbol{\mu}_{\boldsymbol{w}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y & \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \\ \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t & \boldsymbol{\Sigma}_{\boldsymbol{w}} \end{bmatrix} \right) \tag{65}$$

$$p(\boldsymbol{y}_t; \boldsymbol{\theta}) = \mathcal{N}\Big(\boldsymbol{y}_t | \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_{\boldsymbol{w}}, \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y \Big) \tag{66}$$

□

B Via-Points conditioning proof

Proof of Eq. (11) *and Eq.* (12). With the joint distribution $p(\mathbf{x}_a, \mathbf{x}_b)$ in Eq. (53), and from (Bishop and Bishop, 2024), the parameters of a conditional multivariate Gaussian $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}\big(\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}\big)$ are the following:

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \tag{67}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \tag{68}$$

We want the posterior $p(\boldsymbol{w}|\boldsymbol{x}_t^\star)$, knowing the likelihood $\boldsymbol{x}_t^\star|\boldsymbol{w} \sim \mathcal{N}\Big(\boldsymbol{y}_t^\star|\boldsymbol{\Psi}_t^\top\boldsymbol{w}, \boldsymbol{\Sigma}_t^\star\Big)$, and the prior $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$.

$$\begin{bmatrix}\boldsymbol{w}\\ \boldsymbol{x}_t^\star\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{\boldsymbol{w}}\\ \boldsymbol{\Psi}_t^\top\boldsymbol{\mu}_{\boldsymbol{w}}\end{bmatrix}, \begin{bmatrix}\mathrm{Cov}[\boldsymbol{w}, \boldsymbol{w}] & \mathrm{Cov}[\boldsymbol{w}, \boldsymbol{x}_t^\star]\\ \mathrm{Cov}[\boldsymbol{x}_t^\star, \boldsymbol{w}] & \mathrm{Cov}[\boldsymbol{x}_t^\star, \boldsymbol{x}_t^\star]\end{bmatrix}\right) \tag{69}$$

$\mathrm{Cov}[\boldsymbol{x}_t^\star, \boldsymbol{x}_t^\star]$ follows from Eq. (64).

$$\begin{aligned}
\mathrm{Cov}[\boldsymbol{w}, \boldsymbol{x}_t^\star] &= \mathrm{Cov}[\boldsymbol{w}, \boldsymbol{\Psi}_t^\top\boldsymbol{w} + \boldsymbol{\epsilon}_y] && (70)\\
&= \mathrm{Cov}[\boldsymbol{w}, \boldsymbol{\Psi}_t^\top\boldsymbol{w}] \qquad (\mathrm{Cov}[\boldsymbol{w}, \boldsymbol{\epsilon}_{\boldsymbol{y}}] = 0 \text{ since } \boldsymbol{\epsilon}_{\boldsymbol{y}} \text{ is independent of } \boldsymbol{w}) && (71)\\
&= \mathbb{E}[(\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}})(\boldsymbol{\Psi}_t^\top\boldsymbol{w} - \boldsymbol{\Psi}_t^\top\boldsymbol{\mu}_{\boldsymbol{w}})^\top] && (72)\\
&= \mathbb{E}[(\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}})(\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}})^\top\boldsymbol{\Psi}_t] && (73)\\
&= \mathrm{Cov}[\boldsymbol{w}, \boldsymbol{w}] \cdot \boldsymbol{\Psi}_t && (74)\\
&= \boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{\Psi}_t && (75)
\end{aligned}$$

$$\begin{bmatrix}\boldsymbol{w}\\ \boldsymbol{x}_t^\star\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{\boldsymbol{w}}\\ \boldsymbol{\Psi}_t^\top\boldsymbol{\mu}_{\boldsymbol{w}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{w}} & \boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{\Psi}_t\\ \boldsymbol{\Psi}_t^\top\boldsymbol{\Sigma}_{\boldsymbol{w}} & \boldsymbol{\Psi}_t^\top\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_t^\star\end{bmatrix}\right) \tag{76}$$

Using Eq. (67) we get:

$$\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{x}_t^\star} = \boldsymbol{\mu}_{\boldsymbol{w}} + \boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{\Psi}_t\Big(\boldsymbol{\Sigma}_t^\star + \boldsymbol{\Psi}_t^\top\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{\Psi}_t\Big)^{-1}(\boldsymbol{y}_t^\star - \boldsymbol{\Psi}_t^\top\boldsymbol{\mu}_{\boldsymbol{w}}) \tag{77}$$

Using Eq. (68) we get:

$$\boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{x}_t^\star} = \boldsymbol{\Sigma}_{\boldsymbol{w}} - \boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{\Psi}_t\Big(\boldsymbol{\Sigma}_t^\star + \boldsymbol{\Psi}_t^\top\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{\Psi}_t\Big)^{-1}\boldsymbol{\Psi}_t^\top\boldsymbol{\Sigma}_{\boldsymbol{w}} \tag{78}$$

□

C Gaussian mixture regression approximated by a single normal distribution

Proof of Eq. (11) *and Eq.* (12). From (Calinon, 2016), we have:

ToDo

□