

Composition of Movement Primitives

Andrea Pierré

June 19, 2025

Contents

1 ProMPs	1
1.1 Recap	1
1.2 Coupling between joints	2
1.3 Hierarchical Bayesian Model	2
1.4 Via-Points Modulation	3
1.4.1 Do we actually get the desired mean by applying the conditioning update?	3
1.4.2 Multi via-points	3
2 Gaussian mixture modeling (GMM)/Gaussian mixture regression (GMR) recap	4
2.1 Gaussian Mixture Modeling (GMM)	4
2.2 Gaussian Mixture Regression (GMR)	5
3 Composition of MPs	7
3.1 Stitching	7
3.2 Mixture of ProMPs	8
3.2.1 Hierarchical Bayesian Model	8
3.2.2 Via-Points Modulation	8
3.3 Piecewise Gaussian Process	8
A Hierarchical Bayesian Model proof	8
B Via-Points conditioning proof	9
C Gaussian mixture regression approximated by a single normal distribution	10

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

$$\Phi_t = \begin{bmatrix} \phi_1 & \dot{\phi}_1 \\ \vdots & \vdots \\ \phi_n & \dot{\phi}_n \end{bmatrix} \quad (1)$$

$$\mathbf{y}_t = \begin{bmatrix} q_t \\ \dot{q}_t \end{bmatrix} = \Phi_t^\top \mathbf{w} + \epsilon_y \quad (2)$$

$$p(\tau|\mathbf{w}) = \prod_t \mathcal{N}(\mathbf{y}_t | \Phi_t^\top \mathbf{w}, \Sigma_y) \quad (3)$$

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

1.2 Coupling between joints

$$p(\mathbf{y}_t|\mathbf{w}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{y}_{1,t} \\ \vdots \\ \mathbf{y}_{d,t} \end{bmatrix} \middle| \begin{bmatrix} \Phi_t^\top & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \Phi_t^\top \end{bmatrix} \mathbf{w}, \Sigma_y\right) = \mathcal{N}(\mathbf{y}_t | \Psi_t \mathbf{w}, \Sigma_y) \quad (5)$$

with:

Table 2: Notation	
$\mathbf{w} = [\mathbf{w}_1^\top, \dots, \mathbf{w}_n^\top]^\top$	combined weight vector $[n \times n]$
Ψ_t	block-diagonal basis matrix containing the basis functions and their derivatives for each dimension
$\mathbf{y}_{i,t} = [q_{i,t}, \dot{q}_{i,t}]^\top$	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.

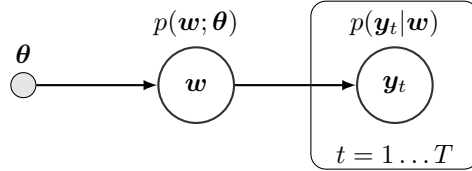


Figure 1: Hierarchical Bayesian Model used in ProMPs.

$$p(\mathbf{y}_t; \theta) = \int \mathcal{N}(\mathbf{y}_t | \Psi_t^\top \mathbf{w}, \Sigma_y) \cdot p(\mathbf{w}; \theta) d\mathbf{w} \quad (6)$$

$$= \int \mathcal{N}(\mathbf{y}_t | \Psi_t^\top \mathbf{w}, \Sigma_y) \cdot \mathcal{N}(\mathbf{w} | \mu_w, \Sigma_w) d\mathbf{w} \quad (7)$$

$$= \mathcal{N}(\mathbf{y}_t | \Psi_t^\top \mu_w, \Psi_t^\top \Sigma_w \Psi_t + \Sigma_y) \quad (8)$$

See Appendix A for the proof.

Table 3: Notation	
$\theta = \{\mu_w, \Sigma_w\}$	parameters
$p(\mathbf{w}; \theta) = \mathcal{N}(\mathbf{w} \mu_w, \Sigma_w)$	prior over the weight vector \mathbf{w} , with parameters θ , assumed to be Gaussian

Table 4: Notation

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

1.4 Via-Points Modulation

Using Bayes rule:

$$p(\mathbf{w}|\mathbf{x}_t^*) = \frac{p(\mathbf{x}_t^*|\mathbf{w}) \cdot p(\mathbf{w})}{p(\mathbf{x}_t^*)} \quad (9)$$

$$p(\mathbf{w}|\mathbf{x}_t^*) \propto \mathcal{N}(\mathbf{y}_t^*|\boldsymbol{\Psi}_t^\top \mathbf{w}, \boldsymbol{\Sigma}_t^*) \cdot \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w) \quad (10)$$

$$\boldsymbol{\mu}_w^{[new]} = \boldsymbol{\mu}_w + \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t \left(\boldsymbol{\Sigma}_y^* + \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t \right)^{-1} (\mathbf{y}_t^* - \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w) \quad (11)$$

$$\boldsymbol{\Sigma}_w^{[new]} = \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t \left(\boldsymbol{\Sigma}_y^* + \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t \right)^{-1} \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \quad (12)$$

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.

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

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

$$\boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w|\mathbf{x}_t^* = \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w + \cancel{\boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t} \left(\cancel{\boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t} \right)^{-1} (\mathbf{y}_t^* - \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w) \quad (14)$$

$$= \cancel{\boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w} + \mathbf{y}_t^* - \cancel{\boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w} \quad (15)$$

$$= \mathbf{y}_t^* \quad (16)$$

□

1.4.2 Multi via-points

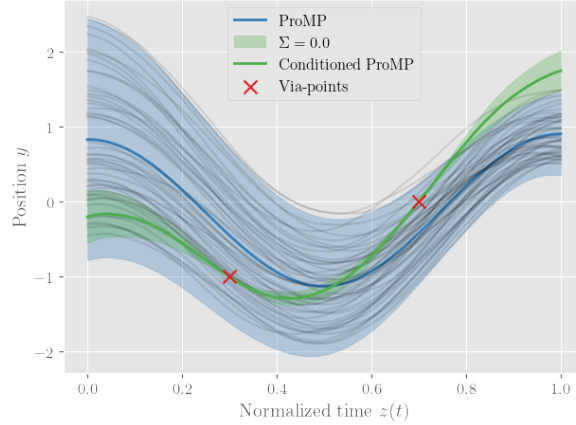


Figure 2: Example of ProMP with two via-points.

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

$$\mu_{w|\mathbf{x}_{t_1}^*} = \mu_w + \Sigma_w \Psi_{t_1} \left(\Sigma_{t_1}^* + \Psi_{t_1}^\top \Sigma_w \Psi_{t_1} \right)^{-1} (\mathbf{y}_{t_1}^* - \Psi_{t_1}^\top \mu_w) \quad (17)$$

$$\Sigma_{w|\mathbf{x}_{t_1}^*} = \Sigma_w - \Sigma_w \Psi_{t_1} \left(\Sigma_{t_1}^* + \Psi_{t_1}^\top \Sigma_w \Psi_{t_1} \right)^{-1} \Psi_{t_1}^\top \Sigma_w \quad (18)$$

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

$$\mu_{w|\mathbf{x}_{t_1}^*, \mathbf{x}_{t_2}^*} = \mu_{w|\mathbf{x}_{t_1}^*} + \Sigma_{w|\mathbf{x}_{t_1}^*} \Psi_{t_2} \left(\Sigma_{t_2}^* + \Psi_{t_2}^\top \Sigma_{w|\mathbf{x}_{t_1}^*} \Psi_{t_2} \right)^{-1} (\mathbf{y}_{t_2}^* - \Psi_{t_2}^\top \mu_{w|\mathbf{x}_{t_1}^*}) \quad (19)$$

$$\Sigma_{w|\mathbf{x}_{t_1}^*, \mathbf{x}_{t_2}^*} = \Sigma_{w|\mathbf{x}_{t_1}^*} - \Sigma_{w|\mathbf{x}_{t_1}^*} \Psi_{t_2} \left(\Sigma_{t_2}^* + \Psi_{t_2}^\top \Sigma_{w|\mathbf{x}_{t_1}^*} \Psi_{t_2} \right)^{-1} \Psi_{t_2}^\top \Sigma_{w|\mathbf{x}_{t_1}^*} \quad (20)$$

3. For the k^{th} via-point conditioning update with the observed via-point $\mathbf{x}_{t_k}^* = [\mathbf{y}_{t_k}^*, \Sigma_{t_k}^*]$, the prior is the posterior after conditioning on the previous $k-1$ via-points, *i.e.*, $\mathbf{w} \sim \mathcal{N}(\mu_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*}, \Sigma_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*})$, the likelihood is $\mathbf{y}_{t_k}^* \sim \mathcal{N}(\Psi_{t_k}^\top \mathbf{w}, \Sigma_{t_k}^*)$, with Ψ_{t_k} the observation matrix at time t_k , and the posterior update becomes:

$$\begin{aligned} \mu_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_k}^*} &= \mu_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*} \\ &+ \Sigma_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*} \Psi_{t_k} \left(\Sigma_{t_k}^* + \Psi_{t_k}^\top \Sigma_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*} \Psi_{t_k} \right)^{-1} (\mathbf{y}_{t_k}^* - \Psi_{t_k}^\top \mu_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*}) \end{aligned} \quad (21)$$

$$\begin{aligned} \Sigma_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_k}^*} &= \Sigma_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*} \\ &- \Sigma_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*} \Psi_{t_k} \left(\Sigma_{t_k}^* + \Psi_{t_k}^\top \Sigma_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*} \Psi_{t_k} \right)^{-1} \Psi_{t_k}^\top \Sigma_{w|\mathbf{x}_{t_1}^*, \dots, \mathbf{x}_{t_{k-1}}^*} \end{aligned} \quad (22)$$

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

$$\mathbf{y}^* = \begin{bmatrix} \mathbf{y}_{t_1}^* \\ \vdots \\ \mathbf{y}_{t_k}^* \end{bmatrix}, \quad \Psi = \begin{bmatrix} \Psi_{t_1} \\ \vdots \\ \Psi_{t_k} \end{bmatrix}, \quad \Sigma^* = \text{diag}(\Sigma_{t_1}^*, \dots, \Sigma_{t_k}^*) \quad (23)$$

$$\mu_{w|\{\mathbf{x}_{t_k}^*\}_{k=1}^K} = \mu_w + \Sigma_w \Psi \left(\Sigma^* + \Psi^\top \Sigma_w \Psi \right)^{-1} (\mathbf{y}^* - \Psi^\top \mu_w) \quad (24)$$

$$\Sigma_{w|\{\mathbf{x}_{t_k}^*\}_{k=1}^K} = \Sigma_w - \Sigma_w \Psi \left(\Sigma^* + \Psi^\top \Sigma_w \Psi \right)^{-1} \Psi^\top \Sigma_w \quad (25)$$

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

2.1 Gaussian Mixture Modeling (GMM)

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

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

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

Table 5: Notation

π_k	mixture weights
$\boldsymbol{\theta} := \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k : k = 1, \dots, K\}$	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:

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{n=1}^N r_{nk} \mathbf{x}_n}{\sum_{n=1}^N r_{nk}} \quad (29)$$

Update of the GMM covariances:

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

Update of the GMM mixture weights:

$$\pi_k^{\text{new}} = \frac{N_k}{N}, \quad k = 1, \dots, K \quad (31)$$

The parameters are updated using Algorithm 1.

2.2 Gaussian Mixture Regression (GMR)

At each iteration step t , the datapoint \mathbf{x}_t can be decomposed as two subvectors \mathbf{x}_t^I and \mathbf{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 $p(\mathbf{x}_t^I, \mathbf{x}_t^O)$ with GMM through EM (Algorithm 1) with:

$$p(\mathbf{x}_t^I, \mathbf{x}_t^O) = \sum_{k=1}^K \pi_k \mathcal{N}_k(\mathbf{x}_t^I, \mathbf{x}_t^O | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (32)$$

$$= \sum_{k=1}^K \pi_k p(\mathbf{x}_t^O | \mathbf{x}_t^I) \cdot p(\mathbf{x}_t^I) \quad (33)$$

$$= \sum_{k=1}^K \pi_k \mathcal{N}_k(\hat{\boldsymbol{\mu}}_k^O(\boldsymbol{\mu}_k^I), \hat{\boldsymbol{\Sigma}}_k^O) \cdot \mathcal{N}_k(\boldsymbol{\mu}_k^I, \boldsymbol{\Sigma}_k^I) \quad (34)$$

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

$$(36)$$

$$\hat{\boldsymbol{\mu}}_k^O(\mathbf{x}_t^I) = \boldsymbol{\mu}_k^O + \boldsymbol{\Sigma}_k^{OI} \boldsymbol{\Sigma}_k^{I-1} (\mathbf{x}_t^I - \boldsymbol{\mu}_k^I) \quad (37)$$

$$\hat{\boldsymbol{\Sigma}}_k^O = \boldsymbol{\Sigma}_k^O - \boldsymbol{\Sigma}_k^{OI} \boldsymbol{\Sigma}_k^{I-1} \cdot \boldsymbol{\Sigma}_k^{IO} \quad (38)$$

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

$$p(\mathbf{x}_t^I) = \int p(\mathbf{x}_t^I, \mathbf{x}_t^O) d\mathbf{x}_t^O = \sum_{k=1}^K \pi_k \cdot \mathcal{N}_k(\boldsymbol{\mu}_k^I, \boldsymbol{\Sigma}_k^I) \quad (39)$$

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 $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$

Output: Final model parameters $\{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\}, \{\pi_k\}$

repeat

 // E-step

for $n \in \{1, \dots, N\}$ **do**

for $k \in \{1, \dots, K\}$ **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 $k \in \{1, \dots, K\}$ **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	
$\mathbf{x}_t \in \mathbb{R}^D$	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 $p(\mathbf{x}_t^O | \mathbf{x}_t^I)$ for each input/timestep we have:

$$p(\mathbf{x}_t^O | \mathbf{x}_t^I) = \frac{p(\mathbf{x}_t^O, \mathbf{x}_t^I)}{p(\mathbf{x}_t^I)} \quad (40)$$

$$= \frac{\sum_{k=1}^K \pi_k \mathcal{N}(\hat{\boldsymbol{\mu}}_k^O(\boldsymbol{\mu}_k^I), \hat{\boldsymbol{\Sigma}}_k^O) \cdot \mathcal{N}(\boldsymbol{\mu}_k^I, \boldsymbol{\Sigma}_k^I)}{\sum_{k=1}^K \pi_k \cdot \mathcal{N}(\boldsymbol{\mu}_k^I, \boldsymbol{\Sigma}_k^I)} \quad (41)$$

$$= \sum_{k=1}^K r_{nk} \cdot \mathcal{N}(\hat{\boldsymbol{\mu}}_k^O(\boldsymbol{\mu}_k^I), \hat{\boldsymbol{\Sigma}}_k^O) \quad (42)$$

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

$$m(x) = \mathbb{E}(x_j | t_j) = \sum_{i=1}^K r_{nk} \cdot m_i(t_j) \quad (43)$$

$$var(x) = \sum_{j=1}^K r_{nk} \cdot cov_i \quad (44)$$

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.

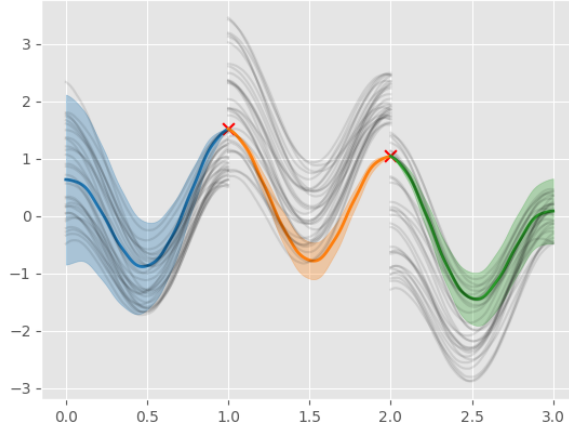


Figure 3: Stitching three ProMPs.

3.2 Mixture of ProMPs

3.2.1 Hierarchical Bayesian Model

Learn the joint distribution $p(t, \mathbf{y}_t; \boldsymbol{\theta})$ as a mixture of ProMPs, and retrieve the conditional distribution $p(\mathbf{y}_t|t; \boldsymbol{\theta})$ for each timestep t .

$$p(t, \mathbf{y}_t; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{y}_t | \Psi_t^\top \boldsymbol{\mu}_w, \Psi_t^\top \Sigma_w \Psi_t + \Sigma_y) \quad (45)$$

ToDo

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:

$$p(\mathbf{x}_a, \mathbf{x}_b) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right) \quad (46)$$

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

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_a, \Sigma_{aa}) \quad (47)$$

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

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{w} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\mathbf{y}_t} \\ \boldsymbol{\mu}_{\mathbf{w}} \end{bmatrix}, \begin{bmatrix} \text{Cov}[\mathbf{y}_t, \mathbf{y}_t] & \text{Cov}[\mathbf{y}_t, \mathbf{w}] \\ \text{Cov}[\mathbf{w}, \mathbf{y}_t] & \text{Cov}[\mathbf{w}, \mathbf{w}] \end{bmatrix}\right) \quad (48)$$

$$\boldsymbol{\mu}_{y_t} = \mathbb{E}[\mathbf{y}_t] \quad (49)$$

$$= \mathbb{E}[\boldsymbol{\Psi}_t^\top \mathbf{w} + \boldsymbol{\epsilon}_y] \quad (50)$$

$$= \boldsymbol{\Psi}_t^\top \mathbb{E}[\mathbf{w}] + \mathbb{E}[\boldsymbol{\epsilon}_y] \quad (51)$$

$$= \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w + 0 \quad (52)$$

$$= \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w \quad (53)$$

$$\text{Cov}[\mathbf{y}_t, \mathbf{y}_t] = \text{Cov}[\boldsymbol{\Psi}_t^\top \mathbf{w} + \boldsymbol{\epsilon}_y] \quad (54)$$

$$= \text{Cov}[\boldsymbol{\Psi}_t^\top \mathbf{w}] + \text{Cov}[\boldsymbol{\epsilon}_y] \quad (55)$$

$$= \boldsymbol{\Psi}_t^\top \text{Cov}[\mathbf{w}] \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y \quad (56)$$

$$= \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y \quad (57)$$

$$\begin{bmatrix} \mathbf{y}_t \\ \mathbf{w} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w \\ \boldsymbol{\mu}_w \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y & \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \\ \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t & \boldsymbol{\Sigma}_w \end{bmatrix} \right) \quad (58)$$

$$p(\mathbf{y}_t; \boldsymbol{\theta}) = \mathcal{N} \left(\mathbf{y}_t | \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w, \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_y \right) \quad (59)$$

□

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. (46), and from (Bishop and Bishop, 2024), the parameters of a conditional multivariate Gaussian $p(\mathbf{x}_a | \mathbf{x}_b) = \mathcal{N}(\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$ are the following:

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} (\mathbf{x}_b - \boldsymbol{\mu}_b) \quad (60)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} \boldsymbol{\Sigma}_{ba} \quad (61)$$

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

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{x}_t^* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w \end{bmatrix}, \begin{bmatrix} \text{Cov}[\mathbf{w}, \mathbf{w}] & \text{Cov}[\mathbf{w}, \mathbf{x}_t^*] \\ \text{Cov}[\mathbf{x}_t^*, \mathbf{w}] & \text{Cov}[\mathbf{x}_t^*, \mathbf{x}_t^*] \end{bmatrix} \right) \quad (62)$$

$\text{Cov}[\mathbf{x}_t^*, \mathbf{x}_t^*]$ follows from Eq. (57).

$$\text{Cov}[\mathbf{w}, \mathbf{x}_t^*] = \text{Cov}[\mathbf{w}, \boldsymbol{\Psi}_t^\top \mathbf{w} + \boldsymbol{\epsilon}_y] \quad (63)$$

$$= \text{Cov}[\mathbf{w}, \boldsymbol{\Psi}_t^\top \mathbf{w}] \quad (\text{Cov}[\mathbf{w}, \boldsymbol{\epsilon}_y] = 0 \text{ since } \boldsymbol{\epsilon}_y \text{ is independent of } \mathbf{w}) \quad (64)$$

$$= \mathbb{E}[(\mathbf{w} - \boldsymbol{\mu}_w)(\boldsymbol{\Psi}_t^\top \mathbf{w} - \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w)^\top] \quad (65)$$

$$= \mathbb{E}[(\mathbf{w} - \boldsymbol{\mu}_w)(\mathbf{w} - \boldsymbol{\mu}_w)^\top \boldsymbol{\Psi}_t] \quad (66)$$

$$= \text{Cov}[\mathbf{w}, \mathbf{w}] \cdot \boldsymbol{\Psi}_t \quad (67)$$

$$= \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t \quad (68)$$

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{x}_t^* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\Psi}_t^\top \boldsymbol{\mu}_w \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_w & \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t \\ \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w & \boldsymbol{\Psi}_t^\top \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_t + \boldsymbol{\Sigma}_t^* \end{bmatrix} \right) \quad (69)$$

Using Eq. (60) we get:

$$\mu_{w|x_t^*} = \mu_w + \Sigma_w \Psi_t \left(\Sigma_t^* + \Psi_t^\top \Sigma_w \Psi_t \right)^{-1} (y_t^* - \Psi_t^\top \mu_w) \quad (70)$$

Using Eq. (61) we get:

$$\Sigma_{w|x_t^*} = \Sigma_w - \Sigma_w \Psi_t \left(\Sigma_t^* + \Psi_t^\top \Sigma_w \Psi_t \right)^{-1} \Psi_t^\top \Sigma_w \quad (71)$$

□

C Gaussian mixture regression approximated by a single normal distribution

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

ToDo

□