0.1. **Comments on notation.**

- $\triangleright$ $\mathbf{A}$ is a matrix, $\mathbf{a}$ is a column vector, and a is a scalar.
- $\triangleright$ $\mathrm{Tr}\left(\mathbf{A}\right)$ is the matrix trace and $|\mathbf{A}|$ the determinant.
- $\triangleright$ a_i is the i-th element of vector $\mathbf{a}$. $a_{i,j}$ is the element in row i and column j of matrix $\mathbf{A}$, and $\mathbf{a}_i^T$ is the i-th row. For more complex matrix notations or where lowercase may be confused with another function or object, an element may be written as $\left(\mathbf{A}\right)_{i,j}$ and the i-th row or j-th column as $\left(\mathbf{A}\right)_{i,:}$ or $\left(\mathbf{A}\right)_{:,j}$ respectively.
- $\triangleright$ $\mathbf{A}_{1:K}$ is the matrix containing rows $\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_K^T$.
- $\triangleright$ $\left(\mathbf{A}^{\circ n}\right)_{i,j} = a_{i,j}^n$ and $\left(\mathbf{A} \circ \mathbf{B}\right)_{i,j} = a_{i,j} b_{i,j}$ (i.e. element-wise exponentation and multiplication), where $\circ$ is the Hadamard operator.
- $\triangleright$ $\mathbf{a}^{\Delta}$ is a matrix containing the differences between each pair of elements in vector $\mathbf{a}$; i.e. $\left(\mathbf{a}^{\Delta}\right)_{i,j} = a_i - a_j$.
- $\triangleright$ $\|\mathbf{a}\| = \sqrt{\mathbf{a}^T\mathbf{a}} = \sqrt{a_1^2 + \cdots + a_k^2}$ is the L^2 (Euclidean) norm of the k-vector $\mathbf{a}$.
- $\triangleright$ $\mathbf{1}$ is a column vector of ones; this is used for summation of matrices where the size of the vector is implicit given the dimensions of the matrix. For example, if $\mathbf{A}$ is an $M \times N$ matrix, then $\mathbf{1}^T\mathbf{A}\mathbf{1}$ is the sum of all elements in $\mathbf{A}$.
- $\triangleright$ $\mathrm{diag}\left(\mathbf{a}\right)$ is the (square) matrix with the vector $\mathbf{a}$ on its diagonal and zeros elsewhere, and $\mathrm{diag}^{-1}\left(\mathbf{A}\right)$ is the vector of the diagonal elements of the matrix $\mathbf{A}$.

0.2. **Useful identities.**

- $\triangleright$ $$\frac{\partial \mathbf{A}}{\partial x} = \begin{bmatrix} \frac{\partial a_{1,1}}{\partial x} & \frac{\partial a_{1,2}}{\partial x} & \cdots & \frac{\partial a_{1,n}}{\partial x} \\ \frac{\partial a_{2,1}}{\partial x} & \frac{\partial a_{2,2}}{\partial x} & \cdots & \frac{\partial a_{2,n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a_{m,1}}{\partial x} & \frac{\partial a_{m,2}}{\partial x} & \cdots & \frac{\partial a_{m,n}}{\partial x} \end{bmatrix}$$
- $\triangleright$ $$\frac{\partial \ln|\mathbf{A}|}{\partial x} = \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right)$$
- $\triangleright$ $\mathrm{d}\mathrm{Tr}\left(\mathbf{A}\right) = \mathrm{Tr}\left(\mathrm{d}\mathbf{A}\right)$
- $\triangleright$ $\mathrm{Tr}\left(\mathbf{A} + \mathbf{B}\right) = \mathrm{Tr}\left(\mathbf{A}\right) + \mathrm{Tr}\left(\mathbf{B}\right)$ for any matrices $\mathbf{A}$ and $\mathbf{B}$.
- $\triangleright$ $\mathrm{Tr}\left(c\mathbf{A}\right) = c\mathrm{Tr}\left(\mathbf{A}\right)$ where c is a scalar.
- $\triangleright$ $\mathrm{Tr}\left(\mathbf{A}\mathbf{C}\right) = \mathbf{1}^T\left(\mathbf{C}^T \circ \mathbf{A}\right)\mathbf{1}$
- $\triangleright$ $\mathbf{a}^{\Delta} = \mathbf{a}\mathbf{1}^T - \mathbf{1}\mathbf{a}^T$ where $\mathbf{a}$ is a column vector.

0.3. **Gradients of the negative log likelihood.** The negative log likelihood has the form

$$\begin{aligned}
-\ln p\left(\mathbf{X},\bar{\alpha},\bar{\beta}|\mathbf{Y}\right) = \mathcal{L} = & \frac{D}{2}\ln|\mathbf{K}_{\mathrm{Y}}| + \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^{2}\mathbf{Y}^{T}\right) - N\ln|\mathbf{W}| \\
& + \frac{d}{2}\ln|\mathbf{K}_{\mathrm{X}}| + \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}\mathbf{X}_{2:N}^{T}\right) + \frac{1}{2}\mathbf{x}_{1}^{T}\mathbf{x}_{1} \\
& + \sum_{j}\ln\beta_{j} + \frac{1}{2\kappa^{2}}\mathrm{Tr}\left(\mathbf{W}^{2}\right) + \sum_{j}\ln\alpha_{j}
\end{aligned}$$

Given that N is the number of time steps in the pose sequence, and D and d are the number of dimensions in the measurement space and latent space, respectively: $\mathbf{Y}$ is the $N \times D$ matrix of measurement coordinates (e.g. joint angles), $\mathbf{W}$ is the $D \times D$ diagonal matrix of scaling factors for each dimension in the measurement space, and $\mathbf{X}$ is the $N \times d$ matrix of latent coordinates. $\mathbf{K}_{\mathrm{Y}}$ and $\mathbf{K}_{\mathrm{X}}$ are $N \times N$ and $(N-1) \times (N-1)$ covariance matrices. $\mathbf{K}_{\mathrm{Y}}$ is defined for all pairs of rows $(\mathbf{x}, \mathbf{x}')$ in $\mathbf{X}$, whereas $\mathbf{K}_{\mathrm{X}}$ is only defined for pairs in $\mathbf{X}_{1:N-1}$; their elements are defined by the functions

$$\begin{aligned}
k_{\mathrm{Y}}\left(\mathbf{x},\mathbf{x}'\right) &= \beta_{1}\exp\left(-\frac{\beta_{2}}{2}\left\|\mathbf{x}-\mathbf{x}'\right\|^{2}\right) + \beta_{3}^{-1}\delta_{\mathbf{x},\mathbf{x}'} \\
k_{\mathrm{X}}\left(\mathbf{x},\mathbf{x}'\right) &= \alpha_{1}\exp\left(-\frac{\alpha_{2}}{2}\left\|\mathbf{x}-\mathbf{x}'\right\|^{2}\right) + \alpha_{3}\mathbf{x}^{T}\mathbf{x}' + \alpha_{4}^{-1}\delta_{\mathbf{x},\mathbf{x}'}
\end{aligned}$$

Where δ is the Kronecker delta function and $\bar{\alpha} = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ and $\bar{\beta} = \{\beta_1, \beta_2, \beta_3\}$ are hyperparameters. In matrix form,

$$\begin{aligned}
\mathbf{K}_{\mathrm{Y}} &= \mathbf{K}_{\mathrm{Y}}^{\mathrm{rbf}} + \beta_{3}^{-1}\mathbf{I} \\
\mathbf{K}_{\mathrm{X}} &= \mathbf{K}_{\mathrm{X}}^{\mathrm{rbf}} + \alpha_{3}\mathbf{X}_{1:N-1}\mathbf{X}_{1:N-1}^{T} + \alpha_{4}^{-1}\mathbf{I}
\end{aligned}$$

Where $\mathbf{K}_{\mathrm{Y}}^{\mathrm{rbf}}$ and $\mathbf{K}_{\mathrm{X}}^{\mathrm{rbf}}$ are the radial basis covariance matrices defined by

$$k_{\mathrm{Y}}^{\mathrm{rbf}}\left(\mathbf{x},\mathbf{x}'\right) = \beta_{1}\exp\left(-\frac{\beta_{2}}{2}\left\|\mathbf{x}-\mathbf{x}'\right\|^{2}\right)$$

$$k_{\mathrm{X}}^{\mathrm{rbf}}\left(\mathbf{x},\mathbf{x}'\right) = \alpha_{1}\exp\left(-\frac{\alpha_{2}}{2}\left\|\mathbf{x}-\mathbf{x}'\right\|^{2}\right)$$

Now we will derive the gradient of the negative log posterior with respect to the individual variables in the MAP optimization.

0.3.1. *With respect to the latent mapping hyperparameters $\bar{\beta}$.* The terms in the negative log posterior that depend on $\bar{\beta}$ are:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \beta_i} &= \frac{\partial}{\partial \beta_i}\left(\frac{D}{2}\ln\left|\mathbf{K}_{\mathrm{Y}}\right| + \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right) + \sum_j \ln \beta_j\right) \\
&= \frac{D}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i}\right) + \frac{1}{2}\frac{\partial\left(\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\right)}{\partial \beta_i} + \beta_i^{-1}
\end{aligned}$$

Although it is true that $\frac{\partial \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{A}\right)}{\partial \mathbf{K}_{\mathrm{Y}}} = -\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{A}\mathbf{K}_{\mathrm{Y}}^{-1}$ for any $\mathbf{K}_{\mathrm{Y}}$ and any $\mathbf{A}$ that is not a function of $\mathbf{K}_{\mathrm{Y}}$ (which is true for $\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T$), and though $\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i}$ can be derived from the matrix form of the covariance function, the chain rule cannot be directly applied to scalar derivatives of matrices (if we try, the solution is a matrix, while the posterior is a scalar). However, the differential can be manipulated:

$$\begin{aligned}
\mathrm{d}\left(\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\right) &= \mathrm{Tr}\left(\mathrm{d}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\right) \\
&= \mathrm{Tr}\left(\mathrm{d}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\right)\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right) \\
&= \mathrm{Tr}\left(-\mathbf{K}_{\mathrm{Y}}^{-1}\mathrm{d}\left(\mathbf{K}_{\mathrm{Y}}\right)\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)
\end{aligned}$$

Thus:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \beta_i} &= \frac{D}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i}\right) + \frac{1}{2}\frac{\mathrm{Tr}\left(-\mathbf{K}_{\mathrm{Y}}^{-1}\partial\left(\mathbf{K}_{\mathrm{Y}}\right)\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)}{\partial \beta_i} + \beta_i^{-1} \\
&= \frac{D}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i}\right) + \frac{1}{2}\mathrm{Tr}\left(-\mathbf{K}_{\mathrm{Y}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i}\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right) + \beta_i^{-1} \\
&= \frac{D}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i}\right) - \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\mathbf{K}_{\mathrm{Y}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i}\right) + \beta_i^{-1} \\
&= \frac{1}{2}\mathrm{Tr}\left(\left(D\mathbf{I} - \mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\mathbf{K}_{\mathrm{Y}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i}\right) + \beta_i^{-1} \\
&= \frac{1}{2}\mathbf{1}^T\left(\left(\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i}\right)^T \circ \left(\left(D\mathbf{I} - \mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\mathbf{K}_{\mathrm{Y}}^{-1}\right)\right)\mathbf{1} + \beta_i^{-1} \\
&= \frac{1}{2}\mathbf{1}^T\left(\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_i} \circ \left(\left(D\mathbf{I} - \mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\mathbf{K}_{\mathrm{Y}}^{-1}\right)\right)\mathbf{1} + \beta_i^{-1}
\end{aligned}$$

The trace of a matrix multiplication has been replaced here with the sum over the elements of a Hadamard product; this notation is less clear but reflects an efficient implementation, as it does not imply the computation of any non-diagonal elements in a matrix product. The derivative of the covariance function k_{Y} with respect to the hyperparameters is

$$\frac{\partial k_{\mathrm{Y}}\left(\mathbf{x},\mathbf{x}'\right)}{\partial \beta_1} = \exp\left(-\frac{\beta_2}{2}\left\|\mathbf{x}-\mathbf{x}'\right\|^2\right)$$

$$\frac{\partial k_{\mathrm{Y}}\left(\mathbf{x},\mathbf{x}'\right)}{\partial \beta_2} = -\frac{\beta_1}{2}\left\|\mathbf{x}-\mathbf{x}'\right\|^2 \exp\left(-\frac{\beta_2}{2}\left\|\mathbf{x}-\mathbf{x}'\right\|^2\right)$$

$$\frac{\partial k_{\mathrm{Y}}\left(\mathbf{x},\mathbf{x}'\right)}{\partial \beta_3} = -\beta_3^{-2}\delta_{\mathbf{x},\mathbf{x}'}$$

In matrix form,

$$\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_1} = \mathbf{K}_{\mathrm{Y}}^{\mathrm{rbf}}/\beta_1$$

$$\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_2} = -\frac{1}{2}\mathbf{L}_{\mathrm{X}}^2 \circ \mathbf{K}_{\mathrm{Y}}^{\mathrm{rbf}}$$

$$\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial \beta_3} = -\beta_3^{-2}\mathbf{I}$$

where $\mathbf{L}_{\mathrm{X}}^2$ is the matrix of pairwise squared L^2 norms between the rows of $\mathbf{X}$; that is, $\left(\mathbf{L}_{\mathrm{X}}^2\right)_{i,j} = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2$.

Thus:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_1} &= \frac{1}{2}\mathbf{1}^T\left(\left(\mathbf{K}_{\mathrm{Y}}^{\mathrm{rbf}}/\beta_1\right)\circ\mathbf{U}\right)\mathbf{1}+\beta_1^{-1}\\ &= \frac{1}{2\beta_1}\mathbf{1}^T\left(\mathbf{K}_{\mathrm{Y}}^{\mathrm{rbf}}\circ\mathbf{U}\right)\mathbf{1}+\beta_1^{-1}\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_2} &= \frac{1}{2}\mathbf{1}^T\left(\left(-\frac{1}{2}\mathbf{L}_{\mathrm{X}}^2\circ\mathbf{K}_{\mathrm{Y}}^{\mathrm{rbf}}\right)\circ\mathbf{U}\right)\mathbf{1}+\beta_2^{-1}\\ &= -\frac{1}{4}\mathbf{1}^T\left(\mathbf{L}_{\mathrm{X}}^2\circ\mathbf{K}_{\mathrm{Y}}^{\mathrm{rbf}}\circ\mathbf{U}\right)\mathbf{1}+\beta_2^{-1}\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta_3} &= \frac{1}{2}\mathbf{1}^T\left(\left(-\beta_3^{-2}\mathbf{I}\right)\circ\mathbf{U}\right)\mathbf{1}+\beta_3^{-1}\\ &= -\frac{1}{2\beta_3^2}\mathrm{Tr}\left(\mathbf{U}\right)+\beta_3^{-1}\end{aligned}$$

where $\mathbf{U}=\left(D\mathbf{I}-\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\mathbf{K}_{\mathrm{Y}}^{-1}$.

0.3.2. *With respect to the dynamic mapping hyperparameters $\bar{\alpha}$.* The terms that depend on $\bar{\alpha}$ have a similar form to those that depend on $\bar{\beta}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha_i} &= \frac{\partial}{\partial \alpha_i}\left(\frac{d}{2}\ln|\mathbf{K}_{\mathrm{X}}| + \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}\mathbf{X}_{2:N}^T\right) + \sum_j \ln \alpha_j\right) \\
&= \frac{d}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}}^{-1}\frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial \alpha_i}\right) + \frac{1}{2}\frac{\partial\left(\mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}\mathbf{X}_{2:N}^T\right)\right)}{\partial \alpha_i} + \alpha_i^{-1} \\
&= \frac{1}{2}\mathbf{1}^T\left(\frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial \alpha_i} \circ \left(\left(d\mathbf{I} - \mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}\mathbf{X}_{2:N}^T\right)\mathbf{K}_{\mathrm{X}}^{-1}\right)\right)\mathbf{1} + \alpha_i^{-1}
\end{aligned}$$

The derivatives of the covariance matrix $\mathbf{K}_{\mathrm{X}}$ with respect to the hyperparameters are

$$\begin{aligned}
\frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial \alpha_1} &= \mathbf{K}_{\mathrm{X}}^{\mathrm{rbf}}/\alpha_1 & \frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial \alpha_2} &= -\tfrac{1}{2}\mathbf{L}_{\mathrm{X_{1:N-1}}}^2 \circ \mathbf{K}_{\mathrm{X}}^{\mathrm{rbf}} \\
\frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial \alpha_3} &= \mathbf{X}_{1:N-1}\mathbf{X}_{1:N-1}^T & \frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial \alpha_4} &= -\alpha_4^{-2}\mathbf{I}
\end{aligned}$$

So:

$$\frac{\partial \mathcal{L}}{\partial \alpha_1} = \frac{1}{2\alpha_1}\mathbf{1}^T\left(\mathbf{K}_{\mathrm{X}}^{\mathrm{rbf}} \circ \mathbf{V}\right)\mathbf{1} + \alpha_1^{-1}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_2} = -\frac{1}{4}\mathbf{1}^T\left(\mathbf{L}_{\mathrm{X_{1:N-1}}}^2 \circ \mathbf{K}_{\mathrm{X}}^{\mathrm{rbf}} \circ \mathbf{V}\right)\mathbf{1} + \alpha_2^{-1}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_3} = \frac{1}{2}\mathbf{1}^T\left(\left(\mathbf{X}_{1:N-1}\mathbf{X}_{1:N-1}^T\right) \circ \mathbf{V}\right)\mathbf{1} + \alpha_3^{-1}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_4} = -\frac{1}{2\alpha_4^2}\mathrm{Tr}\left(\mathbf{V}\right) + \alpha_4^{-1}$$

where $\mathbf{V} = \left(d\mathbf{I} - \mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}\mathbf{X}_{2:N}^T\right)\mathbf{K}_{\mathrm{X}}^{-1}$.

0.3.3. *With respect to the latent coordinates $x_{n,m}$.* Given a single latent variable $x_{n,m}$ belonging to latent pose $\mathbf{x}_n$, for any $n = 1, \ldots, N$ and $m = 1, \ldots, d$, the corresponding entry in the gradient is

$$\begin{aligned}
\left(\frac{\partial \mathcal{L}}{\partial \mathbf{X}}\right)_{n,m} &= \frac{\partial}{\partial x_{n,m}} \left(\frac{D}{2} \ln |\mathbf{K}_{\mathrm{Y}}| + \frac{1}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1} \mathbf{Y} \mathbf{W}^2 \mathbf{Y}^T\right) \right. \\
&\quad \left. + \frac{d}{2} \ln |\mathbf{K}_{\mathrm{X}}| + \frac{1}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}}^{-1} \mathbf{X}_{2:N} \mathbf{X}_{2:N}^T\right) + \frac{1}{2} \mathbf{x}_1^T \mathbf{x}_1 \right) \\
&= \frac{D}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1} \frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial x_{n,m}}\right) - \frac{1}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1} \mathbf{Y} \mathbf{W}^2 \mathbf{Y}^T \mathbf{K}_{\mathrm{Y}}^{-1} \frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial x_{n,m}}\right) \\
&\quad + \frac{d}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}}^{-1} \frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial x_{n,m}}\right) + \frac{1}{2} \mathrm{Tr}\left(-\mathbf{K}_{\mathrm{X}}^{-1} \frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial x_{n,m}} \mathbf{K}_{\mathrm{X}}^{-1} \mathbf{X}_{2:N} \mathbf{X}_{2:N}^T\right) \\
&\quad + \frac{1}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}}^{-1} \frac{\partial \left(\mathbf{X}_{2:N} \mathbf{X}_{2:N}^T\right)}{\partial x_{n,m}}\right) + x_{1,m} \delta_{\mathbf{x}_n, \mathbf{x}_1} \\
&= \frac{1}{2} \mathbf{1}^T \left(\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial x_{n,m}} \circ \mathbf{U}\right) \mathbf{1} + \frac{1}{2} \mathbf{1}^T \left(\frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial x_{n,m}} \circ \mathbf{V}\right) \mathbf{1} \\
&\quad + \frac{1}{2} \mathbf{1}^T \left(\frac{\partial \left(\mathbf{X}_{2:N} \mathbf{X}_{2:N}^T\right)}{\partial x_{n,m}} \circ \mathbf{K}_{\mathrm{X}}^{-1}\right) \mathbf{1} + x_{1,m} \delta_{\mathbf{x}_n, \mathbf{x}_1}
\end{aligned}$$

Where, as before, $\mathbf{U} = \left(D\mathbf{I} - \mathbf{K}_{\mathrm{Y}}^{-1} \mathbf{Y} \mathbf{W}^2 \mathbf{Y}^T\right) \mathbf{K}_{\mathrm{Y}}^{-1}$ and $\mathbf{V} = \left(d\mathbf{I} - \mathbf{K}_{\mathrm{X}}^{-1} \mathbf{X}_{2:N} \mathbf{X}_{2:N}^T\right) \mathbf{K}_{\mathrm{X}}^{-1}$. This can be separated into components:

$$\left(\frac{\partial \mathcal{L}}{\partial \mathbf{X}}\right)_{n,m} = \left(\frac{\partial \mathcal{L}_{\mathrm{Y}}}{\partial \mathbf{X}}\right)_{n,m} + \left(\frac{\partial \mathcal{L}_{\mathrm{X}}}{\partial \mathbf{X}}\right)_{n,m}$$

$$\left(\frac{\partial \mathcal{L}_{\mathrm{Y}}}{\partial \mathbf{X}}\right)_{n,m} = \frac{1}{2} \mathbf{1}^T \left(\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial x_{n,m}} \circ \mathbf{U}\right) \mathbf{1}$$

$$\begin{aligned}
\left(\frac{\partial \mathcal{L}_{\mathrm{X}}}{\partial \mathbf{X}}\right)_{n,m} &= \frac{1}{2} \mathbf{1}^T \left(\frac{\partial \mathbf{K}_{\mathrm{X}}}{\partial x_{n,m}} \circ \mathbf{V}\right) \mathbf{1} + \frac{1}{2} \mathbf{1}^T \left(\frac{\partial \left(\mathbf{X}_{2:N} \mathbf{X}_{2:N}^T\right)}{\partial x_{n,m}} \circ \mathbf{K}_{\mathrm{X}}^{-1}\right) \mathbf{1} \\
&\quad + x_{1,m} \delta_{\mathbf{x}_n, \mathbf{x}_1}
\end{aligned}$$

Note that since $\mathbf{Y}$ and $\mathbf{W}$ do not depend on $x_{n,m}$, the $\mathcal{L}_{\mathrm{Y}}$ component has a similar form to that of the derivatives by $\bar{\beta}$. However, the $\mathcal{L}_{\mathrm{X}}$ component is more complex, with extra terms due to the dependence of $\mathbf{X}_{2:N}$ and $\mathbf{x}_1$ on $x_{n,m}$.. The derivative of the term $\mathbf{x}_1^T \mathbf{x}_1 / 2$ is zero except in the derivatives by elements of $\mathbf{x}_1$, in which case $\frac{\partial}{\partial x_{1,m}} \left(\frac{1}{2} \mathbf{x}_1^T \mathbf{x}_1\right) = x_{1,m}$.

Consider the derivative of k_{Y} for arbitrary rows $\mathbf{x}_i^T$, $\mathbf{x}_j^T$ in $\mathbf{X}$:

$$\begin{aligned}\frac{\partial k_{\mathrm{Y}}\left(\mathbf{x}_i, \mathbf{x}_j\right)}{\partial x_{n,m}} &= \frac{\partial}{\partial x_{n,m}}\left(\beta_1 \exp\left(-\frac{\beta_2}{2}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)\right) \\ &= -\beta_1\beta_2\left(x_{i,m} - x_{j,m}\right)\exp\left(-\frac{\beta_2}{2}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)\left(\delta_{\mathbf{x}_i,\mathbf{x}_n} + \delta_{\mathbf{x}_j,\mathbf{x}_n}\right)\end{aligned}$$

Here, delta functions have been used to indicate that the derivative will only be non-zero if $x_{n,m}$ is either in $\mathbf{x}_i$ or $\mathbf{x}_j$; i.e. the matrix $\frac{\partial \mathbf{K}_{\mathrm{Y}}}{\partial x_{n,m}}$ will have one non-zero row and one non-zero column. (If $x_{n,m}$ is in both $\mathbf{x}_i$ and $\mathbf{x}_j$, the derivative is zero even though both delta functions are not.) This row and column have identical elements (i.e. the matrix is symmetric), which the above equation assumes by multiplying both delta functions by the same term; this can be shown by:

$$\begin{aligned}\frac{\partial}{\partial x_{n,m}}\left\{\beta_1 \exp\left(-\frac{\beta_2}{2}\left\|\mathbf{x}_n - \mathbf{x}_j\right\|^{2}\right)\right\} &= -\beta_1\beta_2\left(x_{n,m} - x_{j,m}\right)\exp\left(-\frac{\beta_2}{2}\left\|\mathbf{x}_n - \mathbf{x}_j\right\|^{2}\right) \\ &= -\beta_1\beta_2\left(-1\right)\left(x_{j,m} - x_{n,m}\right)\exp\left(-\frac{\beta_2}{2}\left\|\mathbf{x}_j - \mathbf{x}_n\right\|^{2}\right) \\ &= \frac{\partial}{\partial x_{n,m}}\left\{\beta_1 \exp\left(-\frac{\beta_2}{2}\left\|\mathbf{x}_j - \mathbf{x}_n\right\|^{2}\right)\right\}\end{aligned}$$

As the rows and columns are identical, they may be memorized as a single vector. All such vectors for a given latent dimension m may be represented as the rows in a matrix

$$\mathbf{Q}^{(m)} = -\beta_2\left(\mathbf{X}\right)^{\Delta}_{:,m} \circ \mathbf{K}^{\mathrm{rbf}}_{\mathrm{Y}}$$

where $\left(\mathbf{X}\right)^{\Delta}_{:,m}$ contains all the pairwise differences $\left(x_{i,m} - x_{j,m}\right)$ between elements of $\left(\mathbf{X}\right)_{:,m}$. Then column m of the derivative of $\mathcal{L}_{\mathrm{Y}}$ by $\mathbf{X}$ is:

$$\begin{aligned}\left(\frac{\partial \mathcal{L}_{\mathrm{Y}}}{\partial \mathbf{X}}\right)_{:,m} &= \left(\mathbf{Q}^{(m)T} \circ \mathbf{U}\right)\mathbf{1} - \frac{1}{2}\mathrm{diag}^{-1}\left(\mathbf{Q}^{(m)T} \circ \mathbf{U}\right) \\ &= \left(\mathbf{Q}^{(m)T} \circ \mathbf{U}\right)\mathbf{1} - \frac{1}{2}\left(\mathrm{diag}^{-1}\left(\mathbf{Q}^{(m)}\right) \circ \mathrm{diag}^{-1}\left(\mathbf{U}\right)\right)\end{aligned}$$

For the derivative with respect to k_{X},

$$\begin{aligned}\frac{\partial k_{\mathrm{X}}\left(\mathbf{x}_i,\mathbf{x}_j\right)}{\partial x_{n,m}} &= \frac{\partial}{\partial x_{n,m}}\left(\alpha_1\exp\left(-\frac{\alpha_2}{2}\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2\right)+\alpha_3\mathbf{x}_i^T\mathbf{x}_j\right)\\ &= -\alpha_1\alpha_2\left(x_{i,m}-x_{j,m}\right)\exp\left(-\frac{\alpha_2}{2}\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2\right)\left(\delta_{\mathbf{x}_i,\mathbf{x}_n}+\delta_{\mathbf{x}_j,\mathbf{x}_n}\right)\\ &\quad+\alpha_3\left(x_{j,m}\delta_{\mathbf{x}_i,\mathbf{x}_n}+x_{i,m}\delta_{\mathbf{x}_j,\mathbf{x}_n}\right)\end{aligned}$$

And:

$$\frac{\partial\mathbf{K}_{\mathrm{X}}^{(m)}}{\partial\left(\mathbf{X}_{1:N-1}\right)_{:,m}} = -\alpha_2\left(\mathbf{X}_{1:N-1}\right)^{\Delta}_{:,m}\circ\mathbf{K}_{\mathrm{X}}^{\mathrm{rbf}}+\alpha_3\left(\mathbf{1}\left(\mathbf{X}_{1:N-1}\right)^T_{:,m}+\operatorname{diag}\left(\left(\mathbf{X}_{1:N-1}\right)_{:,m}\right)\right)$$

For $n>1$, the term $x_{n,m}$ appears only in row $n-1$ and column $n-1$ of $\mathbf{X}_{2:N}\mathbf{X}_{2:N}^T$, and so these are the only non-zero row and column in the derivative. Since $\mathbf{X}_{2:N}\mathbf{X}_{2:N}^T$ is symmetric, so is its derivative by $x_{n,m}$; the vector specifying both is equal to $\left(\mathbf{X}\right)_{:,m}$, except that the diagonal element is the derivative of a quadratic term and is multiplied by 2. Since $\mathbf{K}_{\mathrm{X}}^{-1}$ is also symmetric, its Hadamard product with the derivative is symmetric, and the sum of all its elements is

$$\begin{aligned}\frac{1}{2}\mathbf{1}^T\left(\frac{\partial\left(\mathbf{X}_{2:N}\mathbf{X}_{2:N}^T\right)}{\partial x_{n,m}}\circ\mathbf{K}_{\mathrm{X}}^{-1}\right)\mathbf{1} &= \frac{1}{2}\left(2\left(\mathbf{K}_{\mathrm{X}}^{-1}\right)_{n,:}\cdot\left(\mathbf{X}_{2:N}\right)_{:,m}\right)\\ &= \left(\mathbf{K}_{\mathrm{X}}^{-1}\right)_{n,:}\cdot\left(\mathbf{X}_{2:N}\right)_{:,m}\end{aligned}$$

This is just one term from $\mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}$; for any $x_{n,m}$ where $n>1$, the term found in its derivative is $\left(\mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}\right)_{n-1,m}$. Prepending a single row of zeros to this matrix such that $\left(\mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}\right)^*_{n,m}=\left(\mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}\right)_{n-1,m}$, and padding $x_{1,m}$ as the first row following by $N-1$ rows of zeros in $\mathbf{X}^*_{1,m}$:

$$\begin{aligned}\left(\frac{\partial\mathcal{L}_{\mathrm{X}}}{\partial\mathbf{X}}\right)_{:,m} &= 2\left(\frac{\partial\mathbf{K}_{\mathrm{X}}^{(m)}}{\partial\left(\mathbf{X}_{1:N-1}\right)_{:,m}}\circ\mathbf{V}\right)\mathbf{1}\\ &\quad-\operatorname{diag}^{-1}\left(\frac{\partial\mathbf{K}_{\mathrm{X}}^{(m)}}{\partial\left(\mathbf{X}_{1:N-1}\right)_{:,m}}\circ\mathbf{V}\right)\\ &\quad+\left(\mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}\right)^*+\mathbf{X}^*_{1,m}\end{aligned}$$

0.3.4. *With respect to the weights w_k.*

$$\frac{\partial\mathcal{L}}{\partial w_k}=\frac{\partial}{\partial w_k}\left(\frac{1}{2}\operatorname{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)-N\ln\left|\mathbf{W}\right|+\frac{1}{2\kappa^2}\operatorname{Tr}\left(\mathbf{W}^2\right)\right)$$

As $\mathbf{W}$ is a diagonal matrix, its determinant is the product of the weights w_k, and $\operatorname{Tr}\left(\mathbf{W}^2\right)$ is their sum of squares. Thus for the last two terms in the above derivative:

$$\begin{aligned}\frac{\partial}{\partial w_k}\left(-N\ln|\mathbf{W}|+\frac{1}{2\kappa^2}\mathrm{Tr}\left(\mathbf{W}^2\right)\right)&=\frac{\partial}{\partial w_k}\left(-N\ln\prod_{\ell=1}^{D}w_\ell+\frac{1}{2\kappa^2}\sum_{\ell=1}^{D}w_\ell^2\right)\\&=\frac{\partial}{\partial w_k}\left(-N\left(\sum_{\ell=1}^{D}\ln w_\ell\right)+\frac{1}{2\kappa^2}\sum_{\ell=1}^{D}w_\ell^2\right)\\&=\frac{\partial}{\partial w_k}\left(-N\ln w_k+\frac{1}{2\kappa^2}w_k^2\right)\\&=-Nw_k^{-1}+\frac{1}{\kappa^2}w_k\end{aligned}$$

For the remaining term,

$$\begin{aligned}\mathrm{dTr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)&=\mathrm{Tr}\left(\mathrm{d}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)\right)\\&=\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathrm{d}\left(\mathbf{W}^2\right)\mathbf{Y}^T\right)\end{aligned}$$

$$\frac{\partial\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T\right)}{\partial w_k}=\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\frac{\partial\left(\mathbf{W}^2\right)}{\partial w_k}\mathbf{Y}^T\right)$$

So, noting that $\frac{\partial(\mathbf{W}^2)}{\partial w_k}$ has a single non-zero diagonal element $(\mathbf{W})_{k,k}$:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial w_k}&=\frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}\frac{\partial\left(\mathbf{W}^2\right)}{\partial w_k}\mathbf{Y}^T\right)-Nw_k^{-1}+\frac{1}{\kappa^2}w_k\\&=\left(\mathbf{Y}\right)_{:,k}^T\mathbf{K}_{\mathrm{Y}}^{-1}\left(\mathbf{Y}\right)_{:,k}w_k-Nw_k^{-1}+\frac{1}{\kappa^2}w_k\end{aligned}$$

The optimization of this derivative has a closed form solution:

$$\begin{aligned}0&=\left(\mathbf{Y}\right)_{:,k}^T\mathbf{K}_{\mathrm{Y}}^{-1}\left(\mathbf{Y}\right)_{:,k}w_k-Nw_k^{-1}+\frac{1}{\kappa^2}w_k\\w_k^2&=N\left(\left(\mathbf{Y}\right)_{:,k}^T\mathbf{K}_{\mathrm{Y}}^{-1}\left(\mathbf{Y}\right)_{:,k}+\kappa^{-2}\right)^{-1}\end{aligned}$$

0.4. **Conditional GPDM.** The probability of some new pose $(\mathbf{X}^*, \mathbf{Y}^*)$ can be evaluated given a trained GPDM prior, $\Gamma \equiv (\mathbf{X}, \mathbf{Y}, \bar{\alpha}, \bar{\beta}, \mathbf{W})$.

(Include conditional GPDM stuff here?)

The joint probability of the new pose given the learned model is:

$$\begin{aligned}
-\ln p(\mathbf{X}^*, \mathbf{Y}^* | \Gamma) &= -\ln p(\mathbf{Y}^* | \mathbf{X}^*, \Gamma) - \ln p(\mathbf{X}^* | \Gamma) \\
&= \frac{D}{2} \ln |\mathbf{K}_{\mathrm{Y}^*}| + \frac{MD}{2} \ln(2\pi) + \frac{1}{2} \mathrm{Tr} \left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T \right) - M \ln |\mathbf{W}| \\
&\quad + \frac{(M-1)d}{2} \ln(2\pi) + \frac{d}{2} \ln |\mathbf{K}_{\mathrm{X}^*}| + \frac{1}{2} \mathrm{Tr} \left(\mathbf{K}_{\mathrm{X}^*}^{-1} \mathbf{Z}_{\mathrm{X}} \mathbf{Z}_{\mathrm{X}}^T \right) - \frac{1}{2} {\mathbf{x}_1^*}^T \mathbf{x}_1^*
\end{aligned}$$

where $\mathbf{Z}_{\mathrm{Y}} = \mathbf{Y}^* - \mathbf{A}^T \mathbf{K}_{\mathrm{Y}}^{-1} \mathbf{Y}$, $\mathbf{K}_{\mathrm{Y}^*} = \mathbf{B} - \mathbf{A}^T \mathbf{K}_{\mathrm{Y}}^{-1} \mathbf{A}$, $(\mathbf{A})_{ij} = k_{\mathrm{Y}}(\mathbf{x}_i, \mathbf{x}_j^*)$, $(\mathbf{B})_{ij} = k_{\mathrm{Y}}(\mathbf{x}_i^*, \mathbf{x}_j^*)$, $\mathbf{Z}_{\mathrm{X}} = \mathbf{X}_{2:N}^* - \mathbf{C}^T \mathbf{K}_{\mathrm{X}}^{-1} \mathbf{X}_{2:N}$, $\mathbf{K}_{\mathrm{X}^*} = \mathbf{D} - \mathbf{C}^T \mathbf{K}_{\mathrm{X}}^{-1} \mathbf{C}$, $(\mathbf{C})_{ij} = k_{\mathrm{X}}(\mathbf{x}_i, \mathbf{x}_j^*)$, $(\mathbf{D})_{ij} = k_{\mathrm{X}}(\mathbf{x}_i^*, \mathbf{x}_j^*)$. Note that $\mathbf{Z}_{\mathrm{Y}}$ is the difference between the given observations and the observations reconstructed from $\mathbf{X}^*$ given the learned model, and likewise $\mathbf{Z}_{\mathrm{X}}$ is the difference between the given latent poses and the dynamic prediction from the learned model.

0.4.1. *Gradient with respect to the latent coordinates $x_{n,m}$ in the conditional case.* Instead of calculating the conditional probability for some given $(\mathbf{X}^*, \mathbf{Y}^*)$, in some cases it may be desired to optimize $(\mathbf{X}^*, \mathbf{Y}^*)$ given a trained model and some incomplete pose observations (e.g. 2D joint location estimates). Then the posterior probability of the current estimate of $(\mathbf{X}^*, \mathbf{Y}^*)$ is proportional to the product of its conditional probability under the trained model (i.e. the model prior for the estimate) with the likelihood of the incomplete observations given the estimate. In this case, it is only necessary to optimize over $\mathbf{X}^*$ because it determines the most probable $\mathbf{Y}^*$ through the learned mapping. Regardless of the form of the likelihood, which varies depending on the nature of the incomplete observations, we can find the gradient of the model prior with respect to $\mathbf{X}^*$. Considering only the terms that depend on $\mathbf{X}^*$,

$$\begin{aligned}
\mathcal{P}^* = -\ln p(\mathbf{X}^*, \mathbf{Y}^* | \Gamma) &= \frac{D}{2} \ln |\mathbf{K}_{\mathrm{Y}^*}| + \frac{1}{2} \mathrm{Tr} \left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T \right) \\
&\quad + \frac{d}{2} \ln |\mathbf{K}_{\mathrm{X}^*}| + \frac{1}{2} \mathrm{Tr} \left(\mathbf{K}_{\mathrm{X}^*}^{-1} \mathbf{Z}_{\mathrm{X}} \mathbf{Z}_{\mathrm{X}}^T \right) - \frac{1}{2} {\mathbf{x}_1^*}^T \mathbf{x}_1^*
\end{aligned}$$

Thus:

$$\begin{aligned}
\left(\frac{\partial \mathcal{P}^*}{\partial \mathbf{X}^*}\right)_{n,m} &= \frac{\partial}{\partial x^*_{n,m}} \left(\frac{D}{2} \ln |\mathbf{K}_{\mathrm{Y}^*}| + \frac{1}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T\right)\right. \\
&\quad \left. + \frac{d}{2} \ln |\mathbf{K}_{\mathrm{X}^*}| + \frac{1}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}^*}^{-1} \mathbf{Z}_{\mathrm{X}} \mathbf{Z}_{\mathrm{X}}^T\right) - \frac{1}{2} \mathbf{x}_1^{*^T} \mathbf{x}_1^*\right) \\
&= \frac{D}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \frac{\partial \mathbf{K}_{\mathrm{Y}^*}}{\partial x^*_{n,m}}\right) + \frac{1}{2} \frac{\partial}{\partial x^*_{n,m}} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T\right) \\
&\quad + \frac{d}{2} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}^*}^{-1} \frac{\partial \mathbf{K}_{\mathrm{X}^*}}{\partial x^*_{n,m}}\right) + \frac{1}{2} \frac{\partial}{\partial x^*_{n,m}} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}^*}^{-1} \mathbf{Z}_{\mathrm{X}} \mathbf{Z}_{\mathrm{X}}^T\right) \\
&\quad + x^*_{1,m} \delta_{\mathbf{x}^*_n, \mathbf{x}^*_1}
\end{aligned}$$

Expanding the differentials for the remaining undifferentiated terms:

$$\begin{aligned}
\mathrm{dTr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T\right) &= \mathrm{Tr}\left(\mathrm{d}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1}\right) \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T\right) + \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathrm{d}\left(\mathbf{Z}_{\mathrm{Y}}\right) \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T\right) + \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathrm{d}\left(\mathbf{Z}_{\mathrm{Y}}^T\right)\right) \\
&= -\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathrm{d}\left(\mathbf{K}_{\mathrm{Y}^*}\right) \mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T\right) + \mathrm{Tr}\left(\mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T \mathbf{K}_{\mathrm{Y}^*}^{-1} \mathrm{d}\mathbf{Z}_{\mathrm{Y}}\right) + \mathrm{Tr}\left(\mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T \mathbf{K}_{\mathrm{Y}^*}^{-1} \mathrm{d}\mathbf{Z}_{\mathrm{Y}}\right) \\
&= 2\mathrm{Tr}\left(\mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T \mathbf{K}_{\mathrm{Y}^*}^{-1} \mathrm{d}\mathbf{Z}_{\mathrm{Y}}\right) - \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T \mathbf{K}_{\mathrm{Y}^*}^{-1} \mathrm{d}\mathbf{K}_{\mathrm{Y}^*}\right)
\end{aligned}$$

Similarly:

$$\mathrm{dTr}\left(\mathbf{K}_{\mathrm{X}^*}^{-1} \mathbf{Z}_{\mathrm{X}} \mathbf{Z}_{\mathrm{X}}^T\right) = 2\mathrm{Tr}\left(\mathbf{Z}_{\mathrm{X}}^T \mathbf{K}_{\mathrm{X}^*}^{-1} \mathrm{d}\mathbf{Z}_{\mathrm{X}}\right) - \mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}^*}^{-1} \mathbf{Z}_{\mathrm{X}} \mathbf{Z}_{\mathrm{X}}^T \mathbf{K}_{\mathrm{X}^*}^{-1} \mathrm{d}\mathbf{K}_{\mathrm{X}^*}\right)$$

So:

$$\frac{\partial}{\partial x^*_{n,m}} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T\right) = 2\mathrm{Tr}\left(\mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T \mathbf{K}_{\mathrm{Y}^*}^{-1} \frac{\partial \mathbf{Z}_{\mathrm{Y}}}{\partial x^*_{n,m}}\right) - \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1} \mathbf{Z}_{\mathrm{Y}} \mathbf{W}^2 \mathbf{Z}_{\mathrm{Y}}^T \mathbf{K}_{\mathrm{Y}^*}^{-1} \frac{\partial \mathbf{K}_{\mathrm{Y}^*}}{\partial x^*_{n,m}}\right)$$

$$\frac{\partial}{\partial x^*_{n,m}} \mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}^*}^{-1} \mathbf{Z}_{\mathrm{X}} \mathbf{Z}_{\mathrm{X}}^T\right) = 2\mathrm{Tr}\left(\mathbf{Z}_{\mathrm{X}}^T \mathbf{K}_{\mathrm{X}^*}^{-1} \frac{\partial \mathbf{Z}_{\mathrm{X}}}{\partial x^*_{n,m}}\right) - \mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}^*}^{-1} \mathbf{Z}_{\mathrm{X}} \mathbf{Z}_{\mathrm{X}}^T \mathbf{K}_{\mathrm{X}^*}^{-1} \frac{\partial \mathbf{K}_{\mathrm{X}^*}}{\partial x^*_{n,m}}\right)$$

Taking the derivatives of the composite terms:

$$\begin{aligned}
\frac{\partial \mathbf{Z}_{\mathrm{Y}}}{\partial x^*_{n,m}} &= \frac{\partial}{\partial x^*_{n,m}} \left(\mathbf{Y}^* - \mathbf{A}^T \mathbf{K}_{\mathrm{Y}}^{-1} \mathbf{Y}\right) \\
&= \frac{\partial \mathbf{Y}^*}{\partial x^*_{n,m}} - \left(\frac{\partial \mathbf{A}}{\partial x^*_{n,m}}\right)^T \mathbf{K}_{\mathrm{Y}}^{-1} \mathbf{Y}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \mathbf{K}_{\mathrm{Y}^*}}{\partial x^*_{n,m}} &= \frac{\partial}{\partial x^*_{n,m}}\left(\mathbf{B} - \mathbf{A}^T\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{A}\right) \\
&= \frac{\partial \mathbf{B}}{\partial x^*_{n,m}} - \mathbf{A}^T\mathbf{K}_{\mathrm{Y}}^{-1}\frac{\partial \mathbf{A}}{\partial x^*_{n,m}} - \left(\frac{\partial \mathbf{A}}{\partial x^*_{n,m}}\right)^T \mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{A} \\
&= \frac{\partial \mathbf{B}}{\partial x^*_{n,m}} - \mathbf{S} - \mathbf{S}^T
\end{aligned}
$$

where $\mathbf{S} = \mathbf{A}^T\mathbf{K}_{\mathrm{Y}}^{-1}\frac{\partial \mathbf{A}}{\partial x^*_{n,m}}$. Similarly:

$$
\frac{\partial \mathbf{Z}_{\mathrm{X}}}{\partial x^*_{n,m}} = \frac{\partial \mathbf{X}^*_{2:N}}{\partial x^*_{n,m}} - \left(\frac{\partial \mathbf{C}}{\partial x^*_{n,m}}\right)^T \mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}
$$

$$
\frac{\partial \mathbf{K}_{\mathrm{X}^*}}{\partial x^*_{n,m}} = \frac{\partial \mathbf{D}}{\partial x^*_{n,m}} - \mathbf{R} - \mathbf{R}^T
$$

where $\mathbf{R} = \mathbf{C}^T\mathbf{K}_{\mathrm{X}}^{-1}\mathbf{X}_{2:N}$.

Assume that during optimization, $\mathbf{Y}^*$ is reconstructed from $\mathbf{X}^*$ at each step given the learned mapping. Since the mapping is probabilistic, we take the most probable reconstruction (i.e. the mean, given that the mapping is Gaussian). For a given pose $\mathbf{x}^*$:

$$
\mu_{\mathrm{Y}}\left(\mathbf{x}^*\right) = \mathbf{Y}^T\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{k}_{\mathrm{Y}}\left(\mathbf{x}^*\right)
$$

where $\mathbf{k}_{\mathrm{Y}}\left(\mathbf{x}^*\right)$ is the vector with $\left(\mathbf{k}_{\mathrm{Y}}\left(\mathbf{x}^*\right)\right)_i = k_{\mathrm{Y}}\left(\mathbf{x}^*, \mathbf{x}_i\right)$ for the i-th training pose; note that this is one row in $\mathbf{A}$. Then the most probable reconstruction is:

$$
\mathbf{Y}^* = \mathbf{A}^T\mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y}
$$

Thus:

$$
\begin{aligned}
\frac{\partial \mathbf{Z}_{\mathrm{Y}}}{\partial x^*_{n,m}} &= \frac{\partial \mathbf{Y}^*}{\partial x^*_{n,m}} - \left(\frac{\partial \mathbf{A}}{\partial x^*_{n,m}}\right)^T \mathbf{K}_{\mathrm{Y}}^{-1}\mathbf{Y} \\
&= 0
\end{aligned}
$$

$$
\frac{\partial}{\partial x^*_{n,m}}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1}\mathbf{Z}_{\mathrm{Y}}\mathbf{W}^2\mathbf{Z}_{\mathrm{Y}}^T\right) = 0 - \mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1}\mathbf{Z}_{\mathrm{Y}}\mathbf{W}^2\mathbf{Z}_{\mathrm{Y}}^T\mathbf{K}_{\mathrm{Y}^*}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}^*}}{\partial x^*_{n,m}}\right)
$$

That is, the optimization does not depend on $\mathbf{Z}_{\mathrm{Y}}$, because $\mathbf{Y}^*$ is determined by $\mathbf{X}^*$. So:

$$\begin{aligned}\left(\frac{\partial \mathcal{P}^*}{\partial \mathbf{X}^*}\right)_{n,m} &= \frac{D}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}^*}}{\partial x^*_{n,m}}\right) - \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{Y}^*}^{-1}\mathbf{Z}_{\mathrm{Y}}\mathbf{W}^2\mathbf{Z}_{\mathrm{Y}}^T\mathbf{K}_{\mathrm{Y}^*}^{-1}\frac{\partial \mathbf{K}_{\mathrm{Y}^*}}{\partial x^*_{n,m}}\right)\\ &+ \frac{d}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}^*}^{-1}\frac{\partial \mathbf{K}_{\mathrm{X}^*}}{\partial x^*_{n,m}}\right) + \mathrm{Tr}\left(\mathbf{Z}_{\mathrm{X}}^T\mathbf{K}_{\mathrm{X}^*}^{-1}\frac{\partial \mathbf{Z}_{\mathrm{X}}}{\partial x^*_{n,m}}\right)\\ &- \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\mathrm{X}^*}^{-1}\mathbf{Z}_{\mathrm{X}}\mathbf{Z}_{\mathrm{X}}^T\mathbf{K}_{\mathrm{X}^*}^{-1}\frac{\partial \mathbf{K}_{\mathrm{X}^*}}{\partial x^*_{n,m}}\right) + x^*_{1,m}\delta_{\mathbf{x}^*_n,\mathbf{x}^*_1}\end{aligned}$$

The only terms that remain to be determined are $\frac{\partial \mathbf{A}}{\partial x^*_{n,m}}$, $\frac{\partial \mathbf{B}}{\partial x^*_{n,m}}$, $\frac{\partial \mathbf{C}}{\partial x^*_{n,m}}$, $\frac{\partial \mathbf{D}}{\partial x^*_{n,m}}$, and $\frac{\partial \mathbf{X}^*_{2:N}}{\partial x^*_{n,m}}$.

0.5. **Tracking from image observations.** Learned GPDM: $\Gamma \equiv (\mathbf{X}, \mathbf{Y}, \bar{\alpha}, \bar{\beta})$. State estimate: $\mathbf{\Phi} \equiv (\mathbf{X}^*, \mathbf{Y}^*)$.

$$p(\mathbf{\Phi}|\mathbf{I}, \Gamma) \propto p(\mathbf{I}|\mathbf{\Phi})p(\mathbf{\Phi}|\Gamma)$$

$$p(\mathbf{X}^*, \mathbf{Y}^*|\mathbf{I}, \Gamma) \propto p(\mathbf{I}|\mathbf{X}^*, \mathbf{Y}^*)p(\mathbf{X}^*, \mathbf{Y}^*|\Gamma)$$

The prior probability $p(\phi_{1:T}|\Gamma)$ is the conditional probability $p(\mathbf{Y}^*, \mathbf{X}^*|\Gamma)$ from the GPDM.

Measurement errors may be correlated over time, but as usual assume the image measurements conditioned on states are independent:

$$p(\mathbf{I}|\mathbf{\Phi}) = \prod_{t=1}^{T} p(\mathbf{i}_t|\phi_t)$$

$$p(\mathbf{I}|\mathbf{X}^*, \mathbf{Y}^*) = \prod_{t=1}^{T} p(\mathbf{i}_t|\mathbf{x}^*_t, \mathbf{y}^*_t)$$

Assume zero-mean Gaussian measurement noise

$$-\ln p(\mathbf{i}_t|\phi_t) = \frac{1}{2\sigma_e^2}\sum_{j=1}^{J}\left\|\hat{\mathbf{m}}_t^j - P(\mathbf{p}^j(\phi_t))\right\|^2$$

$$-\ln p(\mathbf{i}_t|\mathbf{x}^*_t, \mathbf{y}^*_t) = \frac{1}{2\sigma_e^2}\sum_{j=1}^{J}\left\|\hat{\mathbf{m}}_t^j - P(\mathbf{p}^j(\mathbf{x}^*_t, \mathbf{y}^*_t))\right\|^2$$

Thus:

$$-\ln p(\mathbf{I}|\mathbf{\Phi}) = \frac{1}{2\sigma_e^2}\sum_{t=1}^{T}\sum_{j=1}^{J}\left\|\hat{\mathbf{m}}_t^j - P(\mathbf{p}^j(\phi_t))\right\|^2$$

where $P(\mathbf{p}^j(\phi_t))$ is the perspective projection of the j-th body point in pose ϕ_t, and $\hat{\mathbf{m}}_t^j$ is the corresponding image measurement from the tracker. The measurement variance σ_e is set to 10 px, "based on empirical results".

0.5.1. *Perspective projection of state estimates.* Camera transform:

0.5.2. *Optimization.* Minimize the following with respect to $\phi_{1:T}$:

$$\begin{aligned}\mathcal{E} &= -\sum_{t=1}^{T} \ln p(\mathbf{i}_t|\phi_t) - \ln p(\mathbf{\Phi}|\Gamma) \\ &= -\sum_{t=1}^{T} \ln p(\mathbf{i}_t|\mathbf{X}^*, \mathbf{Y}^*) - \ln p(\mathbf{X}^*, \mathbf{Y}^*|\Gamma)\end{aligned}$$