Spatial Gaussian Process Regression for Probabilistic Optical Flow Estimation

Andy Wang

May 2025

Abstract

Classical computer vision methods have long been used to estimate optical flow fields for image sequences. These methods produce deterministic point estimates, failing to capture uncertainty inherent in optical flow estimation. I discuss spatial Gaussian process regression (GPR) to specify probability distributions for latent optical flow fields conditioned on observing these estimates. I expose the theory of GPR, describe procedures to fit these distributions to observed data, and apply spatial GPR to estimate optical flows of a real image sequence.

1 Introduction

Optical flow fields are an invaluable interdisciplinary tool, used throughout computer vision, robotics, human-computer interaction, medical imaging, and scientific research to perform object detection, motion segmentation, object tracking, action recognition, video stabilization, video compression, image registration, etc. A motion field is a 3-dimensional vector field that describes the velocity of objects moving in space: an optical flow field is a projection of such a motion field onto a 2-dimensional plane, viewed alternatively as a 2D vector field expressing the motion occurring between two frames in a sequence of images at each pixel. Optical flow estimation is the inverse problem of estimating this field given only a pair of images.

As a natural consequence of optical flows' widespread use, accurate optical flow estimation has been an active subject of research for decades. A simple correlation-based method found, for each window in the first image, that window in the second image most similar to it, then used the difference between these window locations as a flow vector (Micheli, 2025). An early gradient-based method assumed the image as a function of a parametric curve describing an object's motion is constant, then used the minimizer of its first derivative as a flow (Micheli, 2025). Two seminal methods were (Horn & Schunck, 1981) and (Lucas & Kanade, 1981). The former introduced global smoothness by regularizating flow gradients to prevent drastic variation in flow estimates of nearby pixels: the latter introduced local smoothness by computing flows for all pixels in a neighborhood via least-squares estimation. Modern deep learning approaches include FlowNet (Dosovitskiy et al., 2015) and Recurrent All-Pairs Field Transforms (RAFT) (Teed & Deng, 2020), which employ a convolutional neural network and recurrent neural network, respectively.

These methods are all deterministic in nature, producing point (exact) optical flow estimates without any measure of confidence or uncertainty in their predictions. Uncertainty is inherent in optical flow estimation, due to image noise, brightness changes, lack of texture, object occlusion, the aperture problem, and multiple incompatible motions in localized regions. Furthermore, quantification of confidence is paramount in applications where erroneous estimates might have catastropic consequences, such as autonomous navigation, computer-integrated surgery, and real-time surveillance. As such, significant efforts have been made to make optical flow estimation probabilistic in nature, by specifying distributions of optical flows: some of thse efforts have employed simple maximum-a-posteriori estimation (Simoncelli et al., 1991), Markov random fields (Roy & Govindu, 2000), Markov chain Monte Carlo (Sun et al., 2017), and mean field approximation (Wannenwetsch et al., 2017).

I propose a nonparametric Bayesian approach that uses existing deterministic methods to inform probabilistic modeling. I assume a joint Gaussian prior over the latent true optical flow field, making the flows random vectors following a 2-dimensional Gaussian process determined by a chosen mean and covariance kernel. I use the Horn-Schunck and Lucas-Kanade methods to generate observed flow fields, then compute the posterior distribution of the latent field conditioned on each observed field. The kernel parameters, and subsequently the latent mean and covariance, are optimized by maximizing the likelihood of the observations via gradient descent. After optimization, the posterior mean and covariance provide each pixel's optical flow estimate and associated uncertainty, respectively. I discuss how this approach may be extended to the interpolation of optical flows, via a posterior distribution conditioned on an incomplete set of observations. I apply this method to optical flow computation of the Yosemite sequence (Barron et al., 1994), a notable optical flow benchmark dataset consisting of 15 synthetic images, generated to resemble those captured by a drone flying over Yosemite National Park. Finally, I discuss my results and potential directions for improvement.

2 Methodology

2.1 Spatial Gaussian Processes

A 2-dimensional *spatial Gaussian process* is a collection of random variables:

$$\{\mathbf{f}(\mathbf{x})\}_{\mathbf{x}\in\mathcal{D}} = \left\{ \begin{bmatrix} u(\mathbf{x}) \\ v(\mathbf{x}) \end{bmatrix} \right\}_{\mathbf{x}\in\mathcal{D}} \sim \mathcal{GP}(\mathbf{m}(\mathbf{x}), \mathbf{k}(\mathbf{x}, \mathbf{x}'))$$

such that every finite subset follows the joint Gaussian distribution:

$$\begin{bmatrix} \mathbf{f}(\mathbf{x}_1) \\ \vdots \\ \mathbf{f}(\mathbf{x}_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}(\mathbf{x}_1) \\ \ddots \\ \mathbf{m}(\mathbf{x}_n) \end{bmatrix}, \begin{bmatrix} \mathbf{k}(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \mathbf{k}(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \mathbf{k}(\mathbf{x}_n, \mathbf{x}_1) & \cdots & \mathbf{k}(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \right)$$

determined by mean kernel $\mathbf{m}$ and covariance kernel $\mathbf{k}$, the choice of which encodes the prior over the process (Álvarez et al., 2012). Common $\mathbf{m}$ include zero, constant, and affine kernels:

$$\mathbf{m}(\mathbf{x}) = \mathbf{0} \qquad \mathbf{m}(\mathbf{x}) = \mathbf{c} \qquad \mathbf{m}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

Common $\mathbf{k}$ include Gaussian (L^2) and Laplace (L^1) kernels:

$$\mathbf{k}_{L^2}(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\lambda^2} \right) \mathbf{\Sigma} \qquad \mathbf{k}_{L^1}(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{\lambda} \right) \mathbf{\Sigma}$$

I chose a constant mean and Gaussian covariance, encoding a constant drift and loss of correlation as flows grow farther apart. Note that while kernel formulations are user-chosen, kernel parameters are optimized via likelihood maximization. An equivalent joint distribution formulation neatly separating horizontal, vertical, and cross covariances is:

$$\mathbf{F} := \begin{bmatrix} \mathbf{U}(\mathbf{X}) \\ \mathbf{V}(\mathbf{X}) \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu} := \begin{bmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{bmatrix}, \mathbf{K} := \begin{bmatrix} \mathbf{K}_{uu} & \mathbf{K}_{uv} \\ \mathbf{K}_{vu} & \mathbf{K}_{vv} \end{bmatrix} \right)$$

$$\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N,\ \mathbf{U}(\mathbf{X})_i = u(\mathbf{x}_i),\ \mathbf{V}(\mathbf{X})_i = v(\mathbf{x}_i),\ (\boldsymbol{\mu}_u)_i = \mathbf{m}(\mathbf{x}_i)_1,\ (\boldsymbol{\mu}_v)_i = \mathbf{m}(\mathbf{x}_i)_2$$

$$\tilde{\mathbf{k}} = \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) \implies (\mathbf{K}_{uu})_{ij} = \tilde{\mathbf{k}}_{11},\ (\mathbf{K}_{uv})_{ij} = \tilde{\mathbf{k}}_{12},\ (\mathbf{K}_{vu})_{ij} = \tilde{\mathbf{k}}_{21},\ (\mathbf{K}_{vu})_{ij} = \tilde{\mathbf{k}}_{22}.$$

In this work, $\mathbf{F} \in \mathbb{R}^{2N}$, where N is the number of pixels, is the true flow field to be estimated, with concatenated horizontal and vertical components.

2.2 Classical Observations

We work with two sets of optical flow observations, computed via the Horn-Schunck and Lucas-Kanade methods. Each is structured as a vector:

$$\tilde{\mathbf{F}} = \begin{bmatrix} \tilde{u}(\mathbf{x}_1) & \cdots & \tilde{u}(\mathbf{x}_N) & \tilde{v}(\mathbf{x}_1) & \cdots & \tilde{v}(\mathbf{x}_N) \end{bmatrix}^T \in \mathbb{R}^{2N}$$

where N is the number of image pixels, $\{\mathbf{x}_i\}_{i=1}^N$ are pixel indices, and $\{\tilde{u}(\mathbf{x}_i)\}_{i=1}^N$ and $\{\tilde{v}(\mathbf{x}_i)\}_{i=1}^N$ are horizontal and vertical flow components, respectively. For mathematical details, see (Horn & Schunck, 1981), (Lucas & Kanade, 1981).

2.3 Gaussian Process Regression

We assume each Horn-Schunck/Lucas-Kanade observed flow is its corresponding true flow plus some independent additive Gaussian noise:

$$\tilde{\mathbf{f}}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) + \boldsymbol{\eta}(\mathbf{x}), \quad \boldsymbol{\eta}(\mathbf{x}) \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\boldsymbol{\eta}}^2 \mathbf{I}\right)$$

We want to determine marginal distribution $\tilde{\mathbf{F}}|\mathbf{X}$ via:

$$p(\tilde{\mathbf{F}}|\mathbf{X}) = \int p(\tilde{\mathbf{F}}|\mathbf{F}, \mathbf{X}) p(\mathbf{F}|\mathbf{X})\, d\mathbf{F}$$

The convolution of two Gaussians is Gaussian (Rasmussen & Williams, 2006).

$$\tilde{\mathbf{F}}|\mathbf{F}, \mathbf{X} \equiv \tilde{\mathbf{F}}|\mathbf{F} \sim \mathcal{N}\left(\mathbf{F}, \sigma_{\boldsymbol{\eta}}^2 \mathbf{I}\right) \qquad \mathbf{F}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{K}\right)$$

We use these to compute the mean and variance of $\tilde{\mathbf{F}}|\mathbf{X}$:

$$\mathbb{E}[\tilde{\mathbf{F}}|\mathbf{X}] = \mathbb{E}[\mathbb{E}[\tilde{\mathbf{F}}|\mathbf{F}]|\mathbf{X}] = \mathbb{E}[\mathbf{F}|\mathbf{X}] = \boldsymbol{\mu}$$

$$\mathrm{Var}(\tilde{\mathbf{F}}|\mathbf{X}) = \mathrm{Var}(\mathbb{E}[\tilde{\mathbf{F}}|\mathbf{F}]|\mathbf{X}) + \mathbb{E}[\mathrm{Var}(\tilde{\mathbf{F}}|\mathbf{F})|\mathbf{X}] = \mathrm{Var}(\mathbf{F}|\mathbf{X}) + \mathbb{E}[\sigma_{\boldsymbol{\eta}}^2 \mathbf{I}|\mathbf{X}] = \mathbf{K} + \sigma_{\boldsymbol{\eta}}^2 \mathbf{I}$$

Thus $\tilde{\mathbf{F}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{K} + \sigma_{\boldsymbol{\eta}}^2 \mathbf{I}\right)$. See Appendix A for a more through proof. Also:

$$\mathrm{Cov}(\mathbf{F}, \tilde{\mathbf{F}}) = \mathrm{Cov}(\mathbf{F}, \mathbf{F}) + \mathrm{Cov}(\mathbf{F}, \sigma_{\boldsymbol{\eta}}^2 \mathbf{I}) = \mathbf{K} + \mathbf{0} = \mathbf{K}$$

All this allows us to formulate the joint distribution of $(\mathbf{F}, \tilde{\mathbf{F}})$:

$$\begin{bmatrix} \mathbf{F} \\ \tilde{\mathbf{F}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{K} \\ \mathbf{K} & \mathbf{K} + \sigma_{\boldsymbol{\eta}}^2 \mathbf{I} \end{bmatrix} \right)$$

Finally, we derive the posterior distribution $\mathbf{F}|\tilde{\mathbf{F}}$ (Rasmussen & Williams, 2006):

$$\mathbf{F}|\tilde{\mathbf{F}} \sim \mathcal{N}\left(\boldsymbol{\mu} + \mathbf{K}\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2 \mathbf{I}\right)^{-1} \left(\tilde{\mathbf{F}} - \boldsymbol{\mu}\right), \mathbf{K} - \mathbf{K}\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2 \mathbf{I}\right)^{-1} \mathbf{K} \right)$$

See Appendix B for a full derivation. This posterior gives us our updated optical flow estimates and uncertainties. It is computed after kernel parameters (e.g. $\sigma, \lambda, \sigma_{\boldsymbol{\eta}}$) are optimized via likelihood maximization (Rasmussen & Williams, 2006) using gradient descent. See Appendix C for mathematical details. Gaussian process regression is also used to make predictions. We form the joint distribution of incomplete observations $\mathbf{F}_{\mathbf{X}}$, and latent flows $\mathbf{F}_*$ (different pixel locations) using an analogous mathematical formulation:

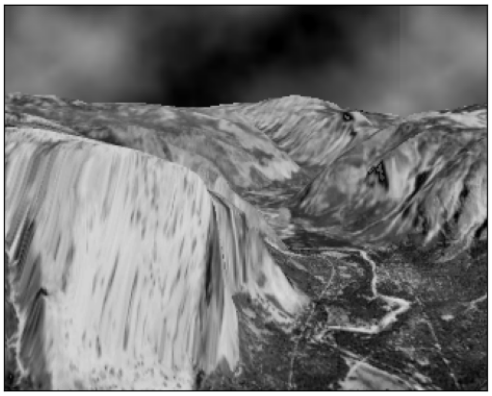
$$\begin{bmatrix} \mathbf{F}_* \\ \mathbf{F}_X \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_* \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{**} & \mathbf{K}_{*X} \\ \mathbf{K}_{X*} & \mathbf{K}_{XX} \end{bmatrix} \right)$$

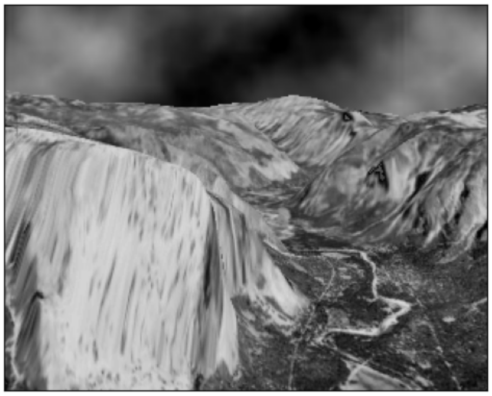
Predictions for $\mathbf{F}_*$ are then given by posterior distribution $\mathbf{F}_*|\mathbf{F}_{\mathbf{X}}$:

$$\mathbf{F}_*|\mathbf{F}_X \sim \mathcal{N}\left(\boldsymbol{\mu}_* + \mathbf{K}_{*X}\left(\mathbf{K}_{XX} + \sigma_{\boldsymbol{\eta}}^2 \mathbf{I}\right)^{-1}(\mathbf{F}_X - \boldsymbol{\mu}_X), \mathbf{K}_{**} - \mathbf{K}_{*X}\left(\mathbf{K}_{XX} + \sigma_{\boldsymbol{\eta}}^2 \mathbf{I}\right)^{-1}\mathbf{K}_{X*} \right)$$

3 Implementation

The images below are the first two images of the Yosemite sequence. Motion is minute, perhaps most visible at the images' lower left corners:

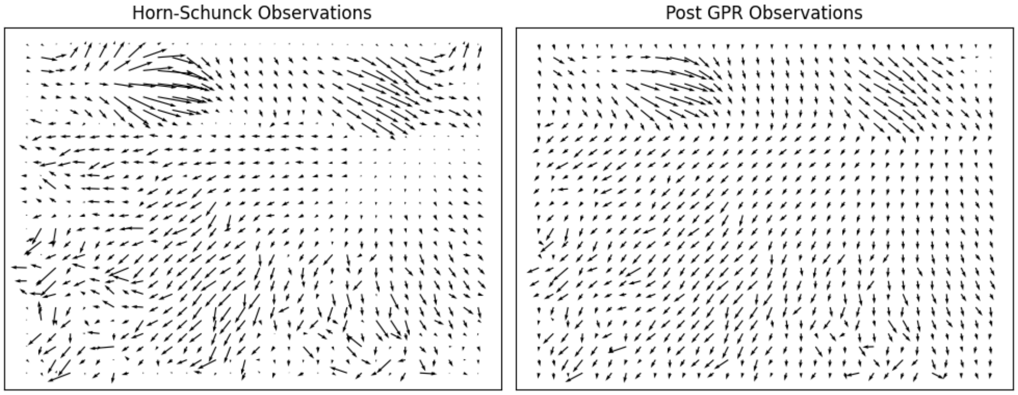




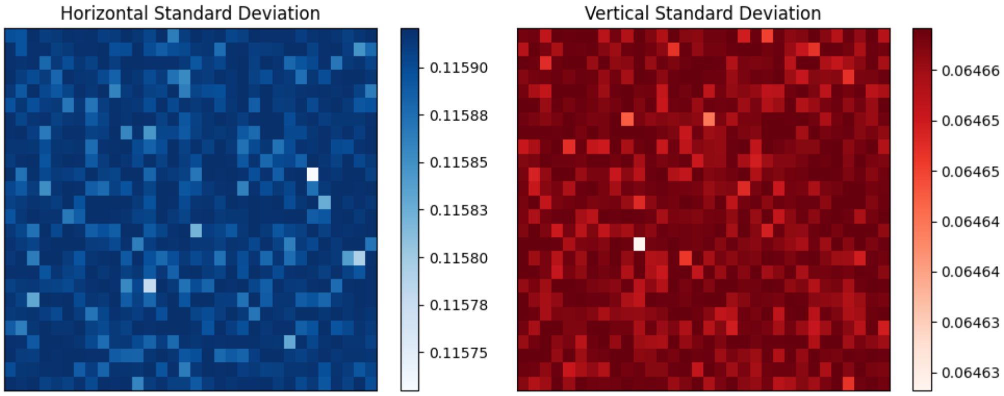
The images are 252 × 316 RGBA images, which I grayscaled and contrast enhanced via histogram equalization. Spatial and temporal derivatives were computed using Sobel filters and simple differencing, respectively. Horn-Schunck computation was performed using 100 iterations with smoothness parameter $\lambda = 300$: Lucas-Kanade computation was performed using 35×35 windows. GPR was performed separately for the two flow fields: in each case, a Gaussian process with constant mean and Gaussian covariance kernels was fit to a downsampled 25×31 field, due to memory constraints, using the Adam optimizer (Kingma & Ba, 2014) to perform gradient descent with learning rate $\eta = 0.1$. All code was written in Python.

4 Results

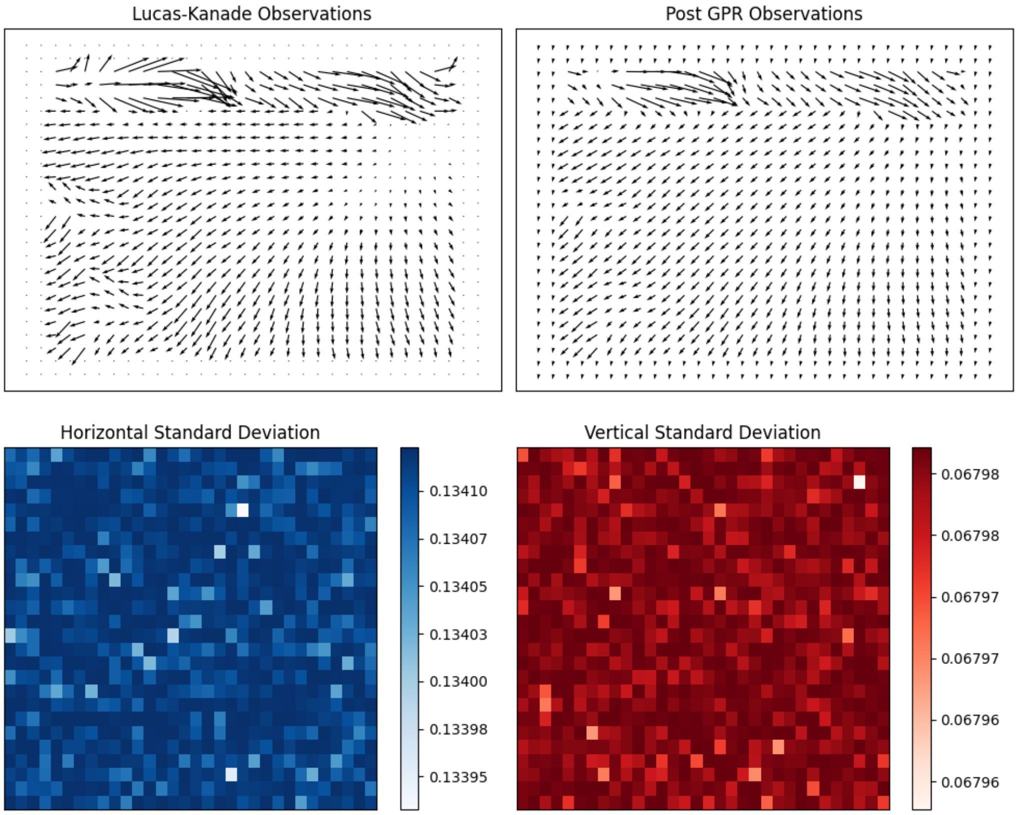
We illustrate the results of Horn-Schunck-based GPR. Downsampled for visual clarity, the figure below depicts each pixel's prior estimate (left) and posterior mean (right). Notice that GPR flows have uniformly smoothed the classical flows, denoising the image and incorporating correlation of neighboring pixels. Consider how the loss of information at the images' lower left edge, yielding motion in all directions, has been corrected: the clouds, however, with motion distinct from that of the cliffs, have been smoothed erroneously.



Each pixel's posterior covariance is a 2 × 2 matrix: its diagonal entries quantify GPR's uncertainty (variance) in that individual prediction, decomposed into components. The figure below displays these entries, again downsampled:



Below are the analogous figures for the results of Lucas-Kanade-based GPR:



5 Discussion

My novel optical flow estimation methodology combining Horn-Schunck and Lucas-Kanade optical flow computation with spatial Gaussian process regression proved effective in enforcing spatial smoothness between optical flows and quantifying prediction uncertainties. This methodology was built upon a few key assumptions, however, that must be addressed. The first is an accurate set of prior optical flow estimates. While this was achieved for the Yosemite sequence, a well-known benchmark dataset, this may not always be possible for more complex real-world image sequences. Classical methods work best in highly textured regions (e.g. object corners and edges), but may perform suboptimally in regions of low texture or contrast (Micheli, 2025). Additionally, we must also have identically distributed additive Gaussian noise for each pixel: in practice, image noise may not be Gaussian, and may depend on pixel locations (e.g. more noise at object boundaries).

This project showed that spatial Gaussian process regression has potential for enhancement and confidence quantification of optical flow estimation. There are a myriad of directions in which to extend this work, and I list the most interesting of these below.

In legitimizing this work as a research endeavor:

- Validate the potential of the GPR approach by experimenting with complex real-world datasets (e.g. KITTI, MPI Sintel, FlyingChairs).
- Evaluate pre-GPR and post-GPR flows against ground truth vectors when possible to quantify GPR's efficacy (e.g. using endpoint error).

In interpreting results:

- Identify which hyperparameters suit which image sequences, through theoretical exploration or empirical comparison. For example, when is Gaussian or Laplace covariance preferred over the other?
- Analyze posterior covariance off-diagonal elements, encoding correlation between individual pixels' horizontal and vertical motions.

In improving results:

- Obtain better observations, likely via deep learning (e.g. FlowNet, RAFT).
- Perform semantic/instance segmentation, then perform GPR on each segment independently, to capture different motions in localized regions.
- Establish more complex priors (e.g. non-uniform/non-Gaussian noise).
- Exploit efficient representations of sparse flows for reduced computation time, allowing minimal image downsampling in running GPR.

In extending existing theory:

- Determine if a 3-dimensional Gaussian process can model sequences of optical flow fields, capturing spatial and temporal flow correlations.

Optical flow fields are a vastly valuable tool across industries and applications. This work has demonstrated the potential for probabilistic optical flow computation through Gaussian processes, to offer a small contribution to the continual improvement of this great aid for innovation and insight.

References

Álvarez, M. A., Rosasco, L., & Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning, 4*(3), 195-266.

Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1994). Performance of Optical Flow Techniques. *International Journal of Computer Vision, 12*(1), 43-77.

Dosovitskiy A., Fischer P., Ilg E., Häusser P., Hazirbaş C., & Golkov V. (2015). FlowNet: Learning Optical Flow with Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2758-2766.

Horn, B. K. P., & Schunck, B. G. (1981). Determining Optical Flow. *Artificial Intelligence, 17*(1-3), 185-203.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *3rd International Conference for Learning Representations.*

Lucas, B. D., & Kanade, T. (1981). An Iterative Image Registration Technique with an Application to Stereo Vision. *7th International Joint Conference on Artificial Intelligence, 2*, 674-679.

Magnus, J. R., & Neudecker, H. (1999). Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley.

Micheli, M. (2025). Mathematical Image Analysis [Unpublished lecture notes]. Johns Hopkins University.

Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian Process for Machine Learning. MIT Press.

Roy, S., & Govindu, V. (2000). MRF Solutions for Probabilistic Optical Flow Formulations. *Proceedings 15th International Conference on Pattern Recognition, 3*, 1041-1047.

Simoncelli, E. P., Adelson, E. H., & Heeger, D. J. (1991). Probability Distributions of Optical Flow. *CVPR, 91*, 310-315.

Sun, J., Quevedo, F. J., & Bollt, E. (2018). Bayesian optical flow with uncertainty quantification. *Inverse Problems 34*(10), 105008.

Teed, Z., & Deng, J. (2020). RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. *Computer Vision–ECCV 2020: 16th European Conference, 2*(16), 402-419.

Wannenwetsch, A. S., Keuper, M., & Roth, S. (2017). ProbFlow: Joint Optical Flow and Uncertainty Estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 1173-1182.

Zhang, F. (2005). The Schur complement and its applications. Springer Science & Business Media.

Appendix A

Here I prove $\tilde{\mathbf{F}}|\mathbf{F} \sim \mathcal{N}\left(\mathbf{F}, \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)$ and $\mathbf{F}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{K}\right)$ implies $\tilde{\mathbf{F}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{K}\right)$:

$$\begin{aligned}
p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) &= \int p\left(\tilde{\mathbf{F}}|\mathbf{F}, \mathbf{X}\right) p\left(\mathbf{F}|\mathbf{X}\right)\, d\mathbf{F} \\
&\propto \int \exp\left(-\frac{1}{2}\left(\tilde{\mathbf{F}} - \mathbf{F}\right)^T \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\mathbf{I}\left(\tilde{\mathbf{F}} - \mathbf{F}\right)\right) \exp\left(-\frac{1}{2}\left(\mathbf{F} - \boldsymbol{\mu}\right)^T \mathbf{K}^{-1}\left(\mathbf{F} - \boldsymbol{\mu}\right)\right) d\mathbf{F} \\
&= \int \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_{\boldsymbol{\eta}}^2}\mathbf{F}^T\mathbf{F} - \frac{2}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}}^T\mathbf{F} + \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}}^T\tilde{\mathbf{F}} + \mathbf{F}^T\mathbf{K}^{-1}\mathbf{F} - 2\boldsymbol{\mu}^T\mathbf{K}^{-1}\mathbf{F} + \boldsymbol{\mu}^T\mathbf{K}^{-1}\boldsymbol{\mu}\right]\right) d\mathbf{F}
\end{aligned}$$

We are integrating over $\mathbf{F}$ and briefly group all terms not involving $\mathbf{F}$ into a constant:

$$p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) \propto \int \exp\left(-\frac{1}{2}\left[\mathbf{F}^T\left(\frac{1}{\sigma_{\boldsymbol{\eta}}^2}\mathbf{I} + \mathbf{K}^{-1}\right)\mathbf{F} - 2\left(\frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}}^T + \boldsymbol{\mu}^T\mathbf{K}^{-1}\right)\mathbf{F} + \text{const.}\right]\right) d\mathbf{F}$$

Let $\mathbf{A} := \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\mathbf{I} + \mathbf{K}^{-1}$, $\mathbf{b}^T := \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}}^T - \boldsymbol{\mu}^T\mathbf{K}^{-1}$. We complete the square:

$$\mathbf{F}^T\mathbf{A}\mathbf{F} - 2\mathbf{b}^T\mathbf{F} = \left(\mathbf{F} - \mathbf{A}^{-1}\mathbf{b}\right)^T \mathbf{A}\left(\mathbf{F} - \mathbf{A}^{-1}\mathbf{b}\right) - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$$

This yields:

$$p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) \propto \exp\left(-\frac{1}{2}\left(\mathbf{F} - \mathbf{A}^{-1}\mathbf{b}\right)^T \mathbf{A}\left(\mathbf{F} - \mathbf{A}^{-1}\mathbf{b}\right)\right) \exp\left(\frac{1}{2}\left[\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} - \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}}^T\tilde{\mathbf{F}} - \boldsymbol{\mu}^T\mathbf{K}^{-1}\boldsymbol{\mu}\right]\right) d\mathbf{F}$$

The first term integrates to the inverse of the Gaussian normalizing factor. Meanwhile, the second term does not depend on $\mathbf{F}$. Reincorporating the initial normalizing factor, we obtain:

$$p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) = \frac{1}{\left(2\pi\sigma_{\boldsymbol{\eta}}\right)^{2N}\sqrt{\det\left(\mathbf{K}\right)}} \sqrt{\left(2\pi\right)^{2N}\det\left(\mathbf{A}^{-1}\right)} \exp\left(\frac{1}{2}\left[\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} - \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}}^T\tilde{\mathbf{F}} - \boldsymbol{\mu}^T\mathbf{K}^{-1}\boldsymbol{\mu}\right]\right)$$

We use the intermediate result:

$$\mathbf{A} = \mathbf{K}^{-1} + \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\mathbf{I} = \mathbf{K}^{-1}\left(\mathbf{I} + \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\mathbf{K}\right) = \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\mathbf{K}^{-1}\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)$$

to compute $\det\left(\mathbf{A}^{-1}\right)$:

$$\det\left(\mathbf{A}\right)^{-\frac{1}{2}} = \left(\frac{1}{\sigma_{\boldsymbol{\eta}}^{2n}}\right)^{-\frac{1}{2}} \det\left(\mathbf{K}^{-1}\right)^{-\frac{1}{2}} \det\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)^{-\frac{1}{2}} = \sigma_{\boldsymbol{\eta}}^n \det\left(\mathbf{K}\right)^{\frac{1}{2}} \det\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)^{-\frac{1}{2}}$$

Substituting this into the above expression yields:

$$p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) = \frac{1}{\sqrt{2\pi\det\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)}} \exp\left(\frac{1}{2}\left[\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} - \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}}^T\tilde{\mathbf{F}} - \boldsymbol{\mu}^T\mathbf{K}^{-1}\boldsymbol{\mu}\right]\right)$$

We express $\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$ in terms of $\tilde{\mathbf{F}}, \boldsymbol{\mu}, \mathbf{K}$. Let $\mathbf{K}_\sigma \coloneqq \mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}$:

$$\begin{aligned}
\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} &= \left(\frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}} - \mathbf{K}^{-1}\boldsymbol{\mu}\right)^T \left(\mathbf{I} + \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\mathbf{K}\right)^{-1} \mathbf{K}\left(\frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}} + \mathbf{K}^{-1}\boldsymbol{\mu}\right) \\
&= \sigma_{\boldsymbol{\eta}}^2 \left(\frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}} + \mathbf{K}^{-1}\boldsymbol{\mu}\right)^T \mathbf{K}_\sigma^{-1}\mathbf{K}\left(\frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}} + \mathbf{K}^{-1}\boldsymbol{\mu}\right) \\
&= \sigma_{\boldsymbol{\eta}}^2 \left[\frac{1}{\sigma_{\boldsymbol{\eta}}^4}\tilde{\mathbf{F}}^T\mathbf{K}_\sigma^{-1}\mathbf{K}\tilde{\mathbf{F}} + \frac{2}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}}^T\mathbf{K}_\sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\mathbf{K}^{-1}\mathbf{K}_\sigma^{-1}\boldsymbol{\mu}\right] \\
&= \frac{1}{\sigma_{\boldsymbol{\eta}}^2}\tilde{\mathbf{F}}^T\tilde{\mathbf{F}} - \tilde{\mathbf{F}}^T\mathbf{K}_\sigma^{-1}\tilde{\mathbf{F}} + 2\tilde{\mathbf{F}}^T\mathbf{K}_\sigma^{-1}\boldsymbol{\mu} + \sigma_{\boldsymbol{\eta}}^2\boldsymbol{\mu}^T\mathbf{K}^{-1}\mathbf{K}_\sigma^{-1}\boldsymbol{\mu}
\end{aligned}$$

Using this, the previous expression becomes:

$$\begin{aligned}
p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) &\propto \exp\left(\frac{1}{2}\left[-\tilde{\mathbf{F}}^T\mathbf{K}_\sigma^{-1}\tilde{\mathbf{F}} + 2\tilde{\mathbf{F}}\mathbf{K}_\sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\left(\sigma_\eta^2\mathbf{K}^{-1}\mathbf{K}_\sigma^{-1} - \mathbf{K}^{-1}\right)\boldsymbol{\mu}\right]\right) \\
&= \exp\left(\frac{1}{2}\left[-\tilde{\mathbf{F}}^T\mathbf{K}_\sigma^{-1}\tilde{\mathbf{F}} + 2\tilde{\mathbf{F}}^T\mathbf{K}_\sigma^{-1}\boldsymbol{\mu} + \text{const.}\right]\right)
\end{aligned}$$

We again complete the square:

$$-\tilde{\mathbf{F}}^T\mathbf{K}_\sigma^{-1}\tilde{\mathbf{F}} + 2\tilde{\mathbf{F}}\mathbf{K}_\sigma^{-1}\boldsymbol{\mu} = -\left(\tilde{\mathbf{F}} - \boldsymbol{\mu}\right)^T\mathbf{K}_\sigma^{-1}\left(\tilde{\mathbf{F}} - \boldsymbol{\mu}\right) + \boldsymbol{\mu}^T\mathbf{K}_\sigma^{-1}\boldsymbol{\mu}$$

This yields:

$$p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) \propto \exp\left(-\frac{1}{2}\left(\tilde{\mathbf{F}} - \boldsymbol{\mu}\right)^T\mathbf{K}_\sigma^{-1}\left(\tilde{\mathbf{F}} - \boldsymbol{\mu}\right)\right)\exp\left(\frac{1}{2}\boldsymbol{\mu}^T\left[\mathbf{K}_\sigma^{-1} + \sigma_\eta^2\mathbf{K}^{-1}\mathbf{K}_\sigma^{-1} - \mathbf{K}^{-1}\right]\boldsymbol{\mu}\right)$$

The middle term in the second exponent is:

$$\mathbf{K}_\sigma^{-1} + \sigma_\eta^2\mathbf{K}^{-1}\mathbf{K}_\sigma^{-1} - \mathbf{K}^{-1} = \left(\mathbf{I} + \sigma_{\boldsymbol{\eta}}^2\mathbf{K}^{-1}\right)\mathbf{K}_\sigma^{-1} - \mathbf{K}^{-1} = \mathbf{K}^{-1}\mathbf{K}_\sigma\mathbf{K}_\sigma^{-1} - \mathbf{K}^{-1} = \mathbf{O}$$

Thus we achieve the final expression for the density:

$$p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) = \frac{1}{\sqrt{2\pi\det\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)}}\exp\left(-\frac{1}{2}\left(\tilde{\mathbf{F}} - \boldsymbol{\mu}\right)^T\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)^{-1}\left(\tilde{\mathbf{F}} - \boldsymbol{\mu}\right)\right)$$

and subsequently the target distribution:

$$\tilde{\mathbf{F}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)$$

Appendix B

Here I prove that if $\mathbf{X}_1, \mathbf{X}_2$ are multivariate normal, i.e.:

$$\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right)$$

then the conditional distribution of $\mathbf{X}_1|\mathbf{X}_2$ is:

$$\mathbf{X}_1|\mathbf{X}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{X}_2 - \boldsymbol{\mu}_2\right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

Conditioning on $\mathbf{X}_2$ means $\mathbf{X}_2$ is fixed. Thus:

$$p_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2) = \frac{p_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2)}{p_{\mathbf{X}_2}(\mathbf{x}_2)} \propto p_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2)$$

Let $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{-1}$. This density is proportional to:

$$\exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{x}_1 - \boldsymbol{\mu}_1\\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{bmatrix}^T \begin{bmatrix}\tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12}\\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 - \boldsymbol{\mu}_1\\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{bmatrix}\right)$$

Notice that $\tilde{\boldsymbol{\Sigma}}_{ij}$ equals $\left(\boldsymbol{\Sigma}^{-1}\right)_{ij}$ and not $\left(\boldsymbol{\Sigma}_{ij}\right)^{-1}$. Let Q be the exponent. We expand:

$$Q \propto \left(\mathbf{x}_1 - \boldsymbol{\mu}_1\right)^T \tilde{\boldsymbol{\Sigma}}_{11}\left(\mathbf{x}_1 - \boldsymbol{\mu}_1\right) + 2\left(\mathbf{x}_1 - \boldsymbol{\mu}_1\right)^T \tilde{\boldsymbol{\Sigma}}_{12}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right) + \left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)^T \tilde{\boldsymbol{\Sigma}}_{22}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)$$

We expand further and ignore all constant terms (those not involving $\mathbf{x}_1$):

$$Q = \mathbf{x}_1^T\tilde{\boldsymbol{\Sigma}}_{11}\mathbf{x}_1 - 2\mathbf{x}_1^T\tilde{\boldsymbol{\Sigma}}_{11}\boldsymbol{\mu}_1 + 2\mathbf{x}_1^T\tilde{\boldsymbol{\Sigma}}_{12}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right) + \text{const.}$$

We complete the square by finding $\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*$:

$$\left(\mathbf{x}_1 - \boldsymbol{\mu}_*\right)^T \boldsymbol{\Sigma}_*^{-1}\left(\mathbf{x}_1 - \boldsymbol{\mu}_*\right) = \mathbf{x}_1^T\boldsymbol{\Sigma}_*^{-1}\mathbf{x}_1 - 2\mathbf{x}_1\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\mu}_* = Q$$

The quadratic terms must equate:

$$\mathbf{x}_1^T\boldsymbol{\Sigma}_*^{-1}\mathbf{x}_1 = \mathbf{x}_1^T\tilde{\boldsymbol{\Sigma}}_{11}\mathbf{x}_1 \implies \boldsymbol{\Sigma}_*^{-1} = \tilde{\boldsymbol{\Sigma}}_{11} \implies \boldsymbol{\Sigma}_* = \tilde{\boldsymbol{\Sigma}}_{11}^{-1}$$

The linear terms must also equate:

$$-2\mathbf{x}_1\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\mu}_* = -2\mathbf{x}_1^T\tilde{\boldsymbol{\Sigma}}_{11}\boldsymbol{\mu}_1 + 2\mathbf{x}_1^T\tilde{\boldsymbol{\Sigma}}_{12}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)$$

Substituting $\boldsymbol{\Sigma}_*^{-1} = \tilde{\boldsymbol{\Sigma}}_{11}$ and simplifying yields:

$$\tilde{\boldsymbol{\Sigma}}_{11}\boldsymbol{\mu}_* = \tilde{\boldsymbol{\Sigma}}_{11}\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\Sigma}}_{12}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right) \implies \boldsymbol{\mu}_* = \boldsymbol{\mu}_1 - \tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)$$

Thus, we have:

$$p_{\mathbf{X}_1|\mathbf{X}_2} \propto \exp\left(-\frac{1}{2}\left(\mathbf{x}_1 - \boldsymbol{\mu}_*\right)^T \boldsymbol{\Sigma}_* \left(\mathbf{x}_1 - \boldsymbol{\mu}_*\right)\right) \implies \mathbf{X}_1|\mathbf{X}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*\right)$$

We have absorbed all constants into the final density's normalizing factor.

All that remains is to determine $\tilde{\boldsymbol{\Sigma}}_{11}, \tilde{\boldsymbol{\Sigma}}_{12}$. Zhang (2005) gives:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{BD}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{S} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix}$$

where Schur complement $\mathbf{S} \coloneqq \mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}$. We invert:

$$\begin{aligned} \mathbf{M}^{-1} &= \begin{bmatrix} \mathbf{I} & -\mathbf{BD}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{S}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{S}^{-1} & -\mathbf{S}^{-1}\mathbf{BD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CS}^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CS}^{-1}\mathbf{BD}^{-1} \end{bmatrix} \end{aligned}$$

For $\mathbf{M} = \boldsymbol{\Sigma}$, Schur complement $\boldsymbol{\Sigma}_S \coloneqq \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$, and:

$$\tilde{\boldsymbol{\Sigma}}_{11} = \left(\boldsymbol{\Sigma}^{-1}\right)_{11} = \boldsymbol{\Sigma}_S^{-1} \qquad \tilde{\boldsymbol{\Sigma}}_{12} = \left(\boldsymbol{\Sigma}^{-1}\right)_{12} = -\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$$

From these we derive the intermediate result:

$$\tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12} = \boldsymbol{\Sigma}_S\left(-\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\right) = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$$

These results yield expressions for $\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*$ in terms of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$:

$$\boldsymbol{\mu}_* = \boldsymbol{\mu}_1 - \tilde{\boldsymbol{\Sigma}}_{11}^{-1}\tilde{\boldsymbol{\Sigma}}_{12}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)$$

$$\boldsymbol{\Sigma}_* = \tilde{\boldsymbol{\Sigma}}_{11}^{-1} = \boldsymbol{\Sigma}_S = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

We have fully derived conditional distribution $\mathbf{X}_1|\mathbf{X}_2$. Substituting the appropriate mean and covariance parameters from the formulations of Gaussian process regression into this general expression yields the posterior distributions:

$$\mathbf{F}|\tilde{\mathbf{F}} \sim \mathcal{N}\left(\boldsymbol{\mu} + \mathbf{K}\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)^{-1}\left(\tilde{\mathbf{F}} - \boldsymbol{\mu}\right), \mathbf{K} - \mathbf{K}\left(\mathbf{K} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)^{-1}\mathbf{K}\right)$$

$$\mathbf{F}_*|\mathbf{F}_X \sim \mathcal{N}\left(\boldsymbol{\mu}_* + \mathbf{K}_{*X}\left(\mathbf{K}_{XX} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)^{-1}\left(\mathbf{F}_X - \boldsymbol{\mu}_X\right), \mathbf{K}_{**} - \mathbf{K}_{*X}\left(\mathbf{K}_{XX} + \sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)^{-1}\mathbf{K}_{X*}\right)$$

Appendix C

Here I show how to maximize the marginal likelihood of $\tilde{\mathbf{F}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{K}+\sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)$:

$$p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) = \frac{1}{\sqrt{(2\pi)^{2N}\det\left(\mathbf{K}+\sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)}}\exp\left(-\frac{1}{2}\left(\tilde{\mathbf{F}}-\boldsymbol{\mu}\right)^T\left(\mathbf{K}+\sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)^{-1}\left(\tilde{\mathbf{F}}-\boldsymbol{\mu}\right)\right)$$

Maximizing the likelihood is equivalent to minimizing the negative log-likelihood:

$$-\log p\left(\tilde{\mathbf{F}}|\mathbf{X}\right) = \frac{1}{2}\left(\tilde{\mathbf{F}}-\boldsymbol{\mu}\right)^T\left(\mathbf{K}+\sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)^{-1}\left(\tilde{\mathbf{F}}-\boldsymbol{\mu}\right)+\frac{1}{2}\log\det\left(\mathbf{K}+\sigma_{\boldsymbol{\eta}}^2\mathbf{I}\right)+N\log 2\pi$$

Let $\mathcal{L} = -\log p\left(\tilde{\mathbf{F}}|\mathbf{X}\right)$, $\mathbf{y} = \tilde{\mathbf{F}}-\boldsymbol{\mu}$, $\mathbf{K}_y = \mathbf{K}+\sigma_{\boldsymbol{\eta}}^2\mathbf{I}$, and φ be a parameter of $\mathbf{m}$:

$$\frac{\partial\mathcal{L}}{\partial\varphi} = \frac{1}{2}\left[\left(\frac{\partial\mathbf{y}^T}{\partial\varphi}\right)\mathbf{K}_y^{-1}\mathbf{y}+\mathbf{y}^T\mathbf{K}_y\left(\frac{\partial\mathbf{y}}{\partial\varphi}\right)\right] = \frac{1}{2}\left(2\mathbf{y}^T\mathbf{K}_y\left(\frac{\partial\mathbf{y}}{\partial\varphi}\right)\right) = -\mathbf{y}^T\mathbf{K}_y\left(\frac{\partial\boldsymbol{\mu}}{\partial\varphi}\right)$$

Let θ be a parameter of $\mathbf{k}$. Jacobi's formula (Magnus & Neudecker, 1999) gives:

$$\frac{\partial\mathcal{L}}{\partial\theta} = \mathbf{y}^T\left(\frac{\partial\mathbf{K}_y^{-1}}{\partial\theta}\right)\mathbf{y}+\frac{\partial|\mathbf{K}_y|}{\partial\theta} = -\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\left(\frac{\partial\mathbf{K}_y}{\partial\theta}\right)\mathbf{K}^{-1}\mathbf{y}+\frac{1}{2}\mathrm{tr}\left(\mathbf{K}_y^{-1}\left(\frac{\partial\mathbf{K}_y}{\partial\theta}\right)\right)$$

Partials $\frac{\partial\mathbf{m}}{\partial\varphi}$ and $\frac{\partial\mathbf{K}_y}{\partial\theta}$ depend on the specific kernels. My choices were:

$$\mathbf{m}(\mathbf{x}) = \mathbf{c} \qquad \mathbf{k}(\mathbf{x},\mathbf{x}') = \sigma^2\exp\left(-z\right)\boldsymbol{\Sigma} \qquad z = \frac{\|\mathbf{x}-\mathbf{x}\|^2}{2\lambda^2} \qquad \boldsymbol{\Sigma} = \begin{bmatrix}\sigma_{uu} & \sigma_{uv}\\ \sigma_{vu} & \sigma_{vv}\end{bmatrix}$$

I set $\boldsymbol{\Sigma} = \mathbf{I}$, but let's consider $\frac{\partial\mathbf{K}}{\partial\sigma_{uu}}$ (the others are analogous). We have:

$$\frac{\partial\mathbf{m}}{\partial\mathbf{c}} = \mathbf{1} \qquad \frac{\partial\mathbf{k}}{\partial\lambda} = \sigma^2\exp\left(-z\right)\Sigma\left(\frac{\|\mathbf{x}-\mathbf{x}\|^2}{\lambda^3}\right) = \frac{\mathbf{k}\|\mathbf{x}-\mathbf{x}\|^2}{\lambda^3}$$

$$\frac{\partial\mathbf{k}}{\partial\sigma} = 2\sigma\exp\left(-z\right)\boldsymbol{\Sigma} = \frac{2\mathbf{k}}{\sigma} \qquad \frac{\partial\mathbf{k}}{\partial\sigma_{uu}} = \sigma^2\exp\left(-z\right)\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$$

These extend naturally to the partials below. We also include $\frac{\partial\mathbf{K}}{\partial\sigma_{\boldsymbol{\eta}}}$:

$$\frac{\partial\boldsymbol{\mu}}{\partial\mathbf{c}} = \mathbf{1} \qquad \frac{\partial\mathbf{K}}{\partial\lambda} = K\odot\frac{\|\mathbf{x}-\mathbf{x}\|^2}{\lambda^3} \qquad \frac{\partial\mathbf{K}}{\partial\sigma} = \frac{2\mathbf{K}}{\sigma}$$

$$\frac{\partial\mathbf{K}}{\partial\sigma_{uu}} = \sigma^2\exp\left(-z\right)\begin{bmatrix}\mathbf{1}_{n\times n} & \mathbf{O}_{n\times n}\\ \mathbf{O}_{n\times n} & \mathbf{O}_{n\times n}\end{bmatrix} \qquad \frac{\partial\mathbf{K}}{\partial\sigma_{\boldsymbol{\eta}}} = 2\sigma_\eta\mathbf{I}$$

These derivatives compose the gradient vector at each descent step.