Spatial Gaussian Process Regression for Bayesian Optical Flow Estimation

Andy Wang & Luca Orquiza

December 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. We propose a probabilistic variation of the Lucas-Kanade method, which returns initial observed flows and error covariance estimates. We discuss spatial Gaussian process regression (GPR) to specify prior probability distributions on latent optical flow fields conditioned on these observations. We describe our method, expose GPR's statistical formulation and parameter fitting via empirical Bayes, and apply GPR to estimate optical flows for a benchmark 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. Notably, (Lucas & Kanade, 1981) introduced local smoothness by assuming equal motion for all pixels in image windows, and computing flows via least-squares estimation. (Horn & Schunck, 1981) estimated optical flow by minimizing a global energy functional, with a smoothness regularization term, via calculus of variations. 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: (Simoncelli et al., 1991) model brightness constraint errors with Gaussian noise and derive a posterior flow distribution, (Roy & Govindu, 2000) formulate optical flow as a Markov Random Field labeling problem and solve it via graph cuts on angle and magnitude parameters, and (Wannenwestch et al., 2017) perform variational inference on an energy-based model using mean-field approximation to predict optical flow and uncertainty (entropy of the variational distribution).

We propose a novel method for obtaining maximum-likelihood based optical flow estimates, a probabilistic variation of the Lucas-Kanade method that incorporates additive Gaussian noise. We then discuss Gaussian process regression, a nonparameteric Bayesian method, which uses our ML estimates as its likelihood. We assume a Gaussian process prior over our latent "flow-generating function", meaning our flows flow a multivariate Gaussian distribution determined by a chosen mean and covariance kernel. The kernel hyperparameters, and thus the latent mean and covariance, are optimized by maximizing the marginal 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. We apply this method on the Yosemite image sequence (Barron et al., 1994), a well-known optical flow benchmark dataset consisting of 15 synthetic images, "captured" by a drone flying over Yosemite National Park. Finally, we discuss our results and potential directions for improvement.

2 Methodology

2.1 Wang-Orquiza Method

Consider latent vector $\boldsymbol{x} \in \mathbb{R}^n$, and observed vector $\boldsymbol{y} \in \mathbb{R}^n$. Assume $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\eta}$, for some fixed $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and additive Gaussian noise $\boldsymbol{\eta} \in \mathbb{R}^n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H})$. We may compute a maximum-likelihood estimate for $\boldsymbol{x}$ based on $\boldsymbol{y}$, where the likelihood we maximize is that of $\boldsymbol{\eta}$. Namely:

$$\hat{\boldsymbol{x}}_{\text{MLE}}(\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} f_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{x}) = \arg\max_{\boldsymbol{x}} f_{\boldsymbol{\eta}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}} \boldsymbol{A}^T \boldsymbol{H}^{-1} \boldsymbol{y}$$

Here we used error covariance matrix:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}} = \text{Cov}[\hat{\boldsymbol{x}}_{\text{MLE}}(\boldsymbol{y}) - \boldsymbol{x}] = (\boldsymbol{A}^T \boldsymbol{H}^{-1} \boldsymbol{A})^{-1}$$

We use this result to propose a reformulation of the Lucas-Kanade equation, where we include additive Gaussian noise:

$$\boldsymbol{I_t} = -\boldsymbol{\nabla I}^T \boldsymbol{f} + \boldsymbol{\eta} \qquad \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H})$$

Here $\boldsymbol{\nabla I} \in \mathbb{R}^{2 \times n}$ is our horizontal and vertical image derivatives, and $\boldsymbol{I}_t \in \mathbb{R}^n$ is our temporal derivatives, and $\boldsymbol{f} \in \mathbb{R}^2$ is our flow vector, for all pixels in a window. Our maximum-likelihood method then gives us our estimated optical flow vector, and its error covariance matrix, for each pixel:

$$\tilde{\boldsymbol{f}} = -\Sigma_{\tilde{\boldsymbol{f}}} \boldsymbol{\nabla I H}^{-1} \boldsymbol{I}_t \in \mathbb{R}^2 \qquad \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}} = (\boldsymbol{\nabla I H}^{-1} \boldsymbol{\nabla I}^T)^{-1} \in \mathbb{R}^{2 \times 2}$$

We see if an object at pixel $\boldsymbol{x}$ has low texture (meaning small image brightness change from that pixel to its neighbors), it will have small gradient, $\boldsymbol{\nabla} I$, which in turn implies a large error covariance $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}(\boldsymbol{x})$ at $\boldsymbol{x}$.

(We've excluded a detailed exposition of optical flow computation for brevity, since it doesn't pertain to statistics: see (Lucas & Kanade, 1981) for details. All that's important is that our primary innovation with this method is the production of an error covariance $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}$, which quantifies prediction confidence).

2.2 Spatial Gaussian Processes

A 2-dimensional *spatial Gaussian process* is a distribution of functions with 2-dimensional outputs:

$$\boldsymbol{f}(\cdot) = \begin{bmatrix} u(\cdot) \\ v(\cdot) \end{bmatrix} \sim \mathcal{GP}(\boldsymbol{m}(\cdot), \boldsymbol{k}(\cdot, \cdot))$$

Such a function evaluated at any finite subset of points follows a joint Gaussian distribution:

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

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

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

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

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

We 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 hyper parameters are optimized via empirical Bayes. An equivalent joint distribution formulation neatly separating horizontal, vertical, and cross covariances is:

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

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

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

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

2.3 Gaussian Process Regression

We assume each Wang-Orquiza observation is its corresponding true flow plus some independent additive Gaussian noise:

$$\tilde{\boldsymbol{f}}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{\eta}(\boldsymbol{x}), \quad \boldsymbol{\eta}(\boldsymbol{x}) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}(\boldsymbol{x})\right)$$

We want to determine marginal distribution $\tilde{\boldsymbol{F}} \in \mathbb{R}^{2N}$ via:

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

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

$$\tilde{\boldsymbol{F}}|\boldsymbol{F} \sim \mathcal{N}\left(\boldsymbol{F}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right) \qquad \boldsymbol{F} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{K}\right)$$

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

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

$$\mathrm{Var}(\tilde{\boldsymbol{F}}) = \mathrm{Var}(\mathbb{E}[\tilde{\boldsymbol{F}}|\boldsymbol{F}]) + \mathbb{E}[\mathrm{Var}(\tilde{\boldsymbol{F}}|\boldsymbol{F})] = \mathrm{Var}(\boldsymbol{F}) + \mathbb{E}[\boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}] = \boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}$$

Thus $\tilde{\boldsymbol{F}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)$. See Appendix A for a more through proof. Also:

$$\mathrm{Cov}(\boldsymbol{F}, \tilde{\boldsymbol{F}}) = \mathrm{Cov}(\boldsymbol{F}, \boldsymbol{F}) + \mathrm{Cov}(\boldsymbol{F}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}) = \boldsymbol{K} + \mathbf{0} = \boldsymbol{K}$$

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

$$\begin{bmatrix} \boldsymbol{F} \\ \tilde{\boldsymbol{F}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K} \\ \boldsymbol{K} & \boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}} \end{bmatrix} \right)$$

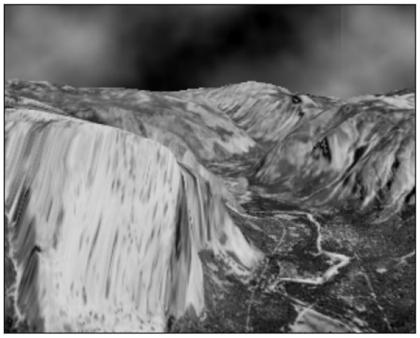
Finally, we derive the posterior distribution $\boldsymbol{F}|\tilde{\boldsymbol{F}}$ (Rasmussen & Williams, 2006):

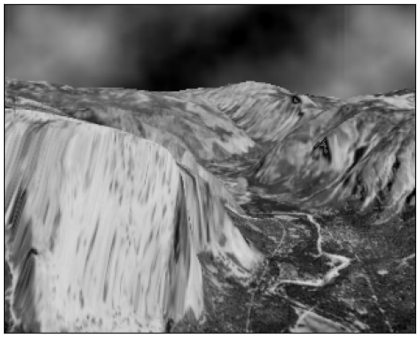
$$\boldsymbol{F}|\tilde{\boldsymbol{F}} \sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{K}\left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)^{-1}\left(\tilde{\boldsymbol{F}} - \boldsymbol{\mu}\right), \boldsymbol{K} - \boldsymbol{K}\left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)^{-1}\boldsymbol{K}\right)$$

See Appendix B for a full derivation. This posterior gives us our updated optical flow estimates and uncertainties. It is computed after kernel hyperparameters are optimized via empirical Bayes (Rasmussen & Williams, 2006) using gradient descent. See Appendix C for mathematical details.

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:

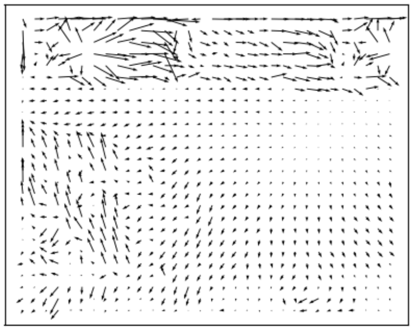




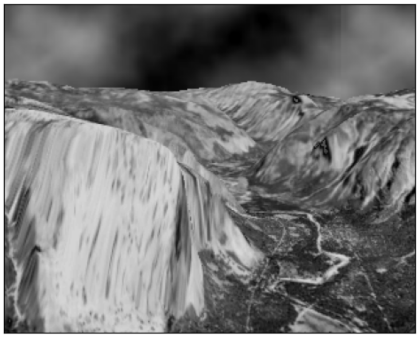
The images are 252 × 316 RGB images, which we first grayscaled. Spatial and temporal derivatives were respectively computed using Sobel filters and forward differencing. Lucas-Kanade and Wang-Orquiza computation was performed using 15×15 windows. GPR was performed using PyTorch, with a constant mean and RBF covariance kernel. Gradient descent was performed using Adam optimization (Kingma & Ba, 2014) (learning rate $\eta = 0.1$).

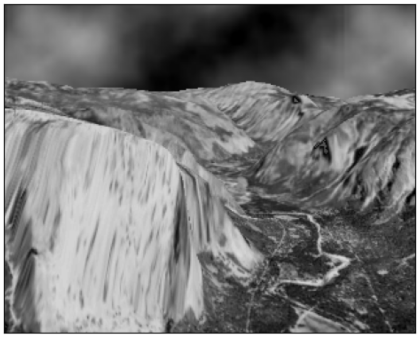
4 Results

Below is our Lucas-Kanade optical flow field, as a baseline. We see more turbulent estimates in the sky (top) and cliff (bottom-left) sections, which correspond to their low texture in the image pair. Contrast this with the high-texture valley (bottom-right), which already has much smoother estimates. We see our motivation for a quantification of confidence.

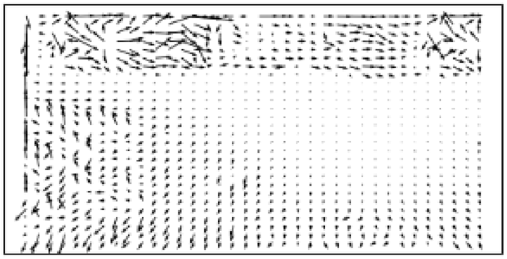


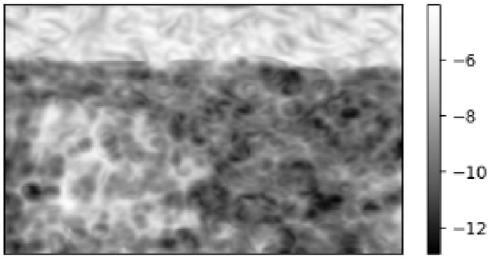
Yosemite images reproduced for convenience:



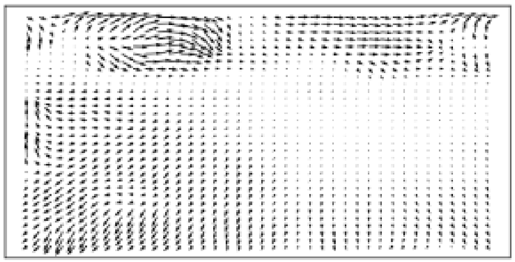


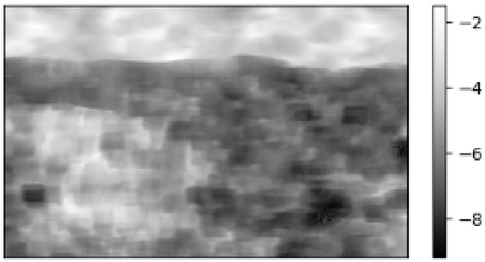
Below is our Wang-Orquiza optical flow field (left), along with our error covariance visualization (right). Each pixel has a (2×2) covariance matrix, and we plot its larger eigenvalue (log-scle), which represents that pixel's maximal variance over all directions. We confirm out intuition, and our model is more uncertain in estimating flows for the sky and cliff regions.





Below are our GPR posterior mean (left) and covariance eigenvalues (right). Our estimates and uncertainties have both effectively been denoised, and our resultant smoothed flows and variances, due to our Gaussian prior, are much more realistic.





5 Discussion

Our novel optical flow estimation methodology combining our Wang-Orquiza reformulation or Lucas-Kanade with spatial Gaussian process regression proved effective in enforcing spatial smoothness between optical flows and producing estimates of prediction confidence. There are a myriad of directions in which to extend this work, and we list the most interesting of these below.

- Identify if certain kernel formulations better suit certain image sequences, through theoretical exploration or empirical comparison. For example, when is Gaussian or Laplace covariance preferred over the other?
- 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.
- 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 many applications. This work has demonstrated the potential for probabilistic optical flow computation through Gaussian processes, to offer a small contribution to this fascinating intersection of computer vision and Bayesian statistics.

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.

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 we prove $\tilde{\boldsymbol{F}}|\boldsymbol{F} \sim \mathcal{N}\left(\boldsymbol{F}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)$ and $\boldsymbol{F} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{K}\right)$ implies $\tilde{\boldsymbol{F}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{K}\right)$. For simplicity of derivations, we assume $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}} = \sigma_{\boldsymbol{\eta}}^2 \boldsymbol{I}$:

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

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

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

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

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

This yields:

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

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

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

We use the intermediate result:

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

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

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

Substituting this into the above expression yields:

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

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

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

Using this, the previous expression becomes:

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

We again complete the square:

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

This yields:

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

The middle term in the second exponent is:

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

Thus we achieve the final expression for the density:

$$p\left(\tilde{\boldsymbol{F}}\right) = \frac{1}{\sqrt{2\pi \det\left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}} \right)}} \exp\left(-\frac{1}{2} \left(\tilde{\boldsymbol{F}} - \boldsymbol{\mu} \right)^T \left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}} \right)^{-1} \left(\tilde{\boldsymbol{F}} - \boldsymbol{\mu} \right) \right)$$

and subsequently the target distribution:

$$\tilde{\boldsymbol{F}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}} \right)$$

Appendix B

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

The quadratic terms must equate:

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

The linear terms must also equate:

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

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

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

Thus, we have:

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

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:

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

where Schur complement $\boldsymbol{S} \coloneqq \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$. We invert:

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

For $\boldsymbol{M} = \boldsymbol{\Sigma}$, 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 $\boldsymbol{\mu}, \boldsymbol{\Sigma}$:

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

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

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

$$\boldsymbol{F}|\tilde{\boldsymbol{F}} \sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{K}\left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)^{-1}\left(\tilde{\boldsymbol{F}} - \boldsymbol{\mu}\right), \boldsymbol{K} - \boldsymbol{K}\left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)^{-1}\boldsymbol{K}\right)$$

Appendix C

Here we show how to maximize the marginal likelihood of $\tilde{\boldsymbol{F}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)$:

$$p\left(\tilde{\boldsymbol{F}}\right) = \frac{1}{\sqrt{(2\pi)^{2N} \det\left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)}} \exp\left(-\frac{1}{2}\left(\tilde{\boldsymbol{F}} - \boldsymbol{\mu}\right)^T \left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)^{-1} \left(\tilde{\boldsymbol{F}} - \boldsymbol{\mu}\right)\right)$$

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

$$-\log p\left(\tilde{\boldsymbol{F}}\right) = \frac{1}{2}\left(\tilde{\boldsymbol{F}} - \boldsymbol{\mu}\right)^T \left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right)^{-1} \left(\tilde{\boldsymbol{F}} - \boldsymbol{\mu}\right) + \frac{1}{2}\log\det\left(\boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}\right) + N\log 2\pi$$

Let $\mathcal{L} = -\log p\left(\tilde{\boldsymbol{F}}\right)$, $\boldsymbol{y} = \tilde{\boldsymbol{F}} - \boldsymbol{\mu}$, $\boldsymbol{K}_y = \boldsymbol{K} + \boldsymbol{\Sigma}_{\tilde{\boldsymbol{f}}}$, and φ be a parameter of $\boldsymbol{m}$:

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

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

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

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

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

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

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

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

These extend naturally to the partials below:

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

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

These derivatives compose the gradient vector at each descent step.