## Gaussian Process Classification

Just as we could use the kernel trick to extend Bayesian linear regression to Gaussian processes for general-purpose nonlinear regression, we may also extend Bayesian linear classification in the same way. In Gaussian process classification, we assume there is a latent function $f\colon \mathcal{X} \to \mathbb{R}$ that is commensurate with the probability of a positive observation; higher latent function values correspond to higher probabilities of positive observations. In Bayesian linear classification, we assumed a parametric (linear) form for this latent function:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}.$$

In Gaussian process classification, rather than choosing a parametric form for $f$, we instead place a Gaussian process prior on $f$:

$$p(f) = \mathcal{GP}(f; \mu, K).$$

Note that a Gaussian prior on the weight vector $\mathbf{w}$ above induces a Gaussian process prior on $f$ with mean function

$$\mu(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\mu}$$

and covariance function

$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x},$$

where $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The Gaussian process formalism allows us to model arbitrary nonlinear classification boundaries by using any desired mean and covariance function for $f$.

### Likelihood

Suppose we have made binary observations at a set of values $\mathbf{X}$, and define $\mathbf{f} = f(\mathbf{X})$ to be the associated set of latent function values. As we did when discussing Bayesian logistic regression, we will assume the following likelihood for a given label observation $y_i$ associated with $\mathbf{x}_i$:

$$p(y_i = 1 \mid f_i) = \sigma(f_i),$$

where $\sigma\colon \mathbb{R} \to (0, 1)$ is a monotonically increasing sigmoid function such as the logisitic function or the standard normal CDF. We again assume the observations are conditionally independent given the latent function values:

$$p(\mathbf{y} \mid \mathbf{f}) = \prod p(y_i \mid f_i).$$

### Inference

Given our prior $p(f) = \mathcal{GP}(f; \mu, K)$ and a set of observations $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, we wish to find the posterior distribution of the latent function values $\mathbf{f} = f(\mathbf{X})$. Note that if we had a Gaussian posterior $\mathbf{f}$, this would induce a GP posterior for the function $f$ given $\mathcal{D}$. The posterior is

$$p(\mathbf{f} \mid \mathcal{D}) = \frac{1}{Z} p(\mathbf{f} \mid \mathbf{X}) p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}\big(\mathbf{f}; \mu(\mathbf{X}), K(\mathbf{X}, \mathbf{X})\big) \prod_i p(y_i \mid f_i).$$

Unfortunately, as we saw, the sigmoid likelihood coupled with the Gaussian prior does not form a tractable posterior. Instead, we must approximate this posterior in some way. Previously we described the Laplace approximation, which approximates the unnormalized log posterior with a second-order Taylor expansion, resulting in a Gaussian approximate posterior centered at the posterior mode

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f} \mid \mathcal{D}).$$

Here we will consider another approximation technique for approximating intractable posterior distributions, which is useful when the likelihood factorizes into one-dimensional terms, as in GP classification when the labels are assumed to be conditionally independent.

## Assumed Density Filtering

Consider a posterior distribution of the form

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{1}{Z} p_0(\boldsymbol{\theta}) \prod_{i=1}^{n} t_i(\boldsymbol{\theta}),$$

where the $t_i$ are typically likelihood terms, for example our $p(y_i \mid f_i)$ above. We assume the prior $p_0(\boldsymbol{\theta})$ has been chosen to be of some nice form, for example a Gaussian, and we will use the Gaussian case to illustrate the idea below.

In **assumed density filtering** (ADF), we assume that the posterior has the same form as the prior $p_0$, and we seek an approximating distribution

$$q(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta} \mid \mathcal{D})$$

from the same family as $p_0$ that approximates the true posterior "as well as possible."

Mechanically, ensuring that our approximate distribution $q$ remains in the same family as $p_0$ is achieved by selecting an (unnormalized) member $\tilde{t}_i$ from the likelihood conjugate to the prior for each of the likelihood terms $t_i$. The result is

$$q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{i=1}^{n} \tilde{Z}_i \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\eta}}_i),$$

and this product will belong to the desired family by conjugacy. Here the constants $\tilde{Z}_i$ and the local parameter vectors $\{\tilde{\boldsymbol{\eta}}_i\}$ are free parameters, called **site parameters**. We will choose the site parameters for each of the approximating distributions $\tilde{t}_i$ to try to improve the fit of the approximating distribution.

For example, the Gaussian distribution is self-conjugate, so the ADF approximation will in this case take the form

$$q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{i=1}^{n} t_i(\boldsymbol{\theta}) \approx \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i=1}^{n} \tilde{Z}_i \mathcal{N}(\boldsymbol{\theta}; \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i),$$

where the site parameters are the constants $\tilde{Z}_i$ (chosen so that the approximation normalizes) as well as the local mean vectors $\tilde{\boldsymbol{\mu}}_i$ and covariance matrices $\tilde{\boldsymbol{\Sigma}}_i$.

Note that in the GP classification case, the likelihood terms in the product are (very conveniently) one-dimensional, because each term simply models a single scalar label, $y_i$, and only depends on a single latent function value $f_i$:

$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i=1}^{n} \tilde{Z}_i \mathcal{N}(f_i; \tilde{\mu}_i, \tilde{\sigma}_i^2).$$

Consider just the first two terms in this product:

$$p(\boldsymbol{\theta} \mid \mathcal{D}_1) \propto p_0(\boldsymbol{\theta}) t_1(\boldsymbol{\theta}).$$

This product will not have the same nice form as the prior, but has only been warped "slightly" away from the prior via a single likelihood term. In many cases, this distribution will be at least partially manageable. Perhaps there will not be a nice closed expression, but we might still be able to compute the normalizing constant or the moments of the posterior.

In assumed density filtering, we will approximate this product with a member of the desired family:

$$q_1(\boldsymbol{\theta}) = \tilde{Z}_1 p_0(\boldsymbol{\theta}) \tilde{t}_1(\boldsymbol{\theta}; \tilde{\boldsymbol{\eta}}_1) \approx p(\boldsymbol{\theta} \mid \mathcal{D}_1) \propto p_0(\boldsymbol{\theta}) t_1(\boldsymbol{\theta}).$$

This approximation is done by matching the moments between the approximation $q_1(\boldsymbol{\theta})$ and the true posterior $p(\boldsymbol{\theta} \mid \mathcal{D}_1)$.

For example, consider $p_0(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. We select the site parameters $\boldsymbol{\eta} = \{\tilde{Z}_1^{-1}, \tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1\}$ such that

$$\begin{aligned} \int q_1(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} &= 1 \\ \mathbb{E}\big[q_1(\boldsymbol{\theta})\big] &= \mathbb{E}\big[p(\boldsymbol{\theta} \mid \mathcal{D}_1)\big] \\ \operatorname{cov}\big[q_1(\boldsymbol{\theta})\big] &= \operatorname{cov}\big[p(\boldsymbol{\theta} \mid \mathcal{D}_1)\big]. \end{aligned}$$

Now the approximate distribution is in the desired density family (Gaussians) and matches the true posterior up to the second moment.

Now consider the first three terms of the product:

$$p(\boldsymbol{\theta} \mid \mathcal{D}_1, \mathcal{D}_2) \propto p_0(\boldsymbol{\theta}) t_1(\boldsymbol{\theta}) t_2(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta} \mid \mathcal{D}_1) t_2(\boldsymbol{\theta}).$$

The idea in assumed density filtering is to substitute in our approximation $q_1(\boldsymbol{\theta})$, giving:

$$p(\boldsymbol{\theta} \mid \mathcal{D}_1, \mathcal{D}_2) \overset{\propto}{\sim} q_1(\boldsymbol{\theta}) t_2(\boldsymbol{\theta}).$$

Now we are in the same situation we were in before! We have a nice "prior" distribution $q_1$ multiplied by a single likelihood term $t_2$. We proceed as before, replacing the true likelihood term $t_2$ with an approximate (conjugate) term $\tilde{Z}_2 \tilde{t}_2(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_2)$, where we again choose the site parameters $(\tilde{Z}_2, \tilde{\boldsymbol{\theta}}_2)$ to match the moments between our new approximation

$$q_2(\boldsymbol{\theta}) = \tilde{Z}_2 q_1(\boldsymbol{\theta}) \tilde{t}_2(\boldsymbol{\theta}; \tilde{\boldsymbol{\eta}}_2) = \tilde{Z}_1 \tilde{Z}_2 p_0(\boldsymbol{\theta}) \tilde{t}_1(\boldsymbol{\theta}; \tilde{\boldsymbol{\eta}}_1) \tilde{t}_2(\boldsymbol{\theta}; \tilde{\boldsymbol{\eta}}_2).$$

and the "less-approximate" posterior $q_1(\boldsymbol{\theta}) t_2(\boldsymbol{\theta})$. We proceed in this fashion until we have processed all the local likelihood terms, resulting in the final approximation

$$q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{i=1}^{n} \tilde{Z}_i \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\eta}}_i).$$

**Moments for Bayesian Probit Regression**

As an example, suppose we use the standard normal CDF as our sigmoid function

$$\sigma(f_i) = \Phi(f_i) = \int_{-\infty}^{f_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, \mathrm{d}x.$$

In this setting, we can compute the site parameters $\tilde{Z}_i$, $\tilde{\mu}_i$, and $\tilde{\sigma}_i^2$ in closed-form. To do so, first observe that if we encode the labels as $y_i \in \{-1, +1\}$, we can succinctly express the likelihood as

$$\Pr(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} \Phi(y_i f_i)$$

where we have made use of the fact that

$$\Pr(y_i = -1 \mid f_i) = 1 - \Pr(y_i = +1 \mid f_i) = 1 - \Phi(f_i) = \Phi(-f_i).$$

Now suppose that the product of the prior and the first $i-1$ likelihood terms has already been approximated as

$$q_{i-1}(\mathbf{f}) \approx \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}^{(i-1)}, \boldsymbol{\Sigma}^{(i-1)})$$

and we are interested in computing the moments of

$$p(\mathbf{f} \mid \mathbf{X}_{1:i}) \propto q_{i-1}(\mathbf{f}) t_i(\mathbf{f})$$

where $t_i(\mathbf{f}) = \Phi(f_i)$. First, consider the normalizing constant

$$\begin{aligned}
\tilde{Z}_i^{-1} = \zeta &= \int q_{i-1}(\mathbf{f}) t_i(\mathbf{f}) \, \mathrm{d}\mathbf{f} \\
&= \int_{-\infty}^{\infty} \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \Phi(y_i f_i) \, \mathrm{d}f_i \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{y_i f_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{a^2}{2}} \frac{1}{\sqrt{2\pi\Sigma_{ii}^{(i-1)}}} e^{-\frac{\left(f_i - \mu_i^{(i-1)}\right)^2}{2\Sigma_{ii}^{(i-1)}}} \, \mathrm{d}a \, \mathrm{d}f_i
\end{aligned}$$

where we have made use of the (again, incredibly convenient) fact that the likelihood factorizes along individual latent function values and the closure property of Gaussians under marginalization. Note that the superscript $(i-1)$ indicates that the mean and variance come from the current approximate belief $q_{i-1}(\mathbf{f})$.

Let us first consider the case where $y_i = +1$. We will define two intermediate quantities:

$$\begin{aligned}
&\alpha = a - f_i + \mu_i^{(i-1)} \rightarrow \mathrm{d}\alpha = \mathrm{d}a \text{ and } a = (+1) f_i \rightarrow \alpha = \mu_i^{(i-1)} \\
&\beta = f_i - \mu_i^{(i-1)} \rightarrow \mathrm{d}\beta = \mathrm{d}f_i \\
&\quad \rightarrow a = \alpha + \beta
\end{aligned}$$

Using these substitutions and swapping the order of the integrals gives

$$\begin{aligned}
\zeta_{y_i=+1} &= \int_{-\infty}^{\mu_i^{(i-1)}} \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{\Sigma_{ii}^{(i-1)}}} e^{-\frac{(\alpha+\beta)^2}{2} - \frac{\beta^2}{2\Sigma_{ii}^{(i-1)}}} \,\mathrm{d}\beta \,\mathrm{d}\alpha \\
&= \int_{-\infty}^{\mu_i^{(i-1)}} \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{\Sigma_{ii}^{(i-1)}}} e^{-\frac{1}{2}\begin{bmatrix}\beta\\ \alpha\end{bmatrix}^T \begin{bmatrix}1+\frac{1}{\Sigma_{ii}^{(i-1)}} & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\beta\\ \alpha\end{bmatrix}} \,\mathrm{d}\beta \,\mathrm{d}\alpha \\
&= \int_{-\infty}^{\mu_i^{(i-1)}} \int_{-\infty}^{\infty} \mathcal{N}\left(\begin{bmatrix}\beta\\ \alpha\end{bmatrix}; \mathbf{0}, \begin{bmatrix}\Sigma_{ii}^{(i-1)} & -\Sigma_{ii}^{(i-1)}\\ -\Sigma_{ii}^{(i-1)} & 1+\Sigma_{ii}^{(i-1)}\end{bmatrix}\right) \,\mathrm{d}\beta \,\mathrm{d}\alpha \\
&= \int_{-\infty}^{\mu_i^{(i-1)}} \mathcal{N}(\alpha; 0, 1+\Sigma_{ii}^{(i-1)}) \,\mathrm{d}\alpha = \Phi\left(\frac{\mu_i^{(i-1)}}{\sqrt{1+\Sigma_{ii}^{(i-1)}}}\right).
\end{aligned}$$

For the case where $y_i = -1$, the procedure is similar except we define the substitution variables as

$$\begin{aligned}
&\alpha = a + f_i - \mu_i \rightarrow \mathrm{d}\alpha = \mathrm{d}a \text{ and } a = (-1)f_i \rightarrow \alpha = -\mu_i \\
&\beta = f_i - \mu_i \rightarrow \mathrm{d}\beta = \mathrm{d}f_i \\
&\quad \rightarrow a = \alpha - \beta
\end{aligned}$$

Using these new substitutions and stepping through the same procedure as above gives an unsurprisingly similar result:

$$\zeta_{y_i=-1} = \Phi\left(\frac{-\mu_i^{(i-1)}}{\sqrt{1+\Sigma_{ii}^{(i-1)}}}\right).$$

Thus, we can collect the two results into a single, concise expression by again using the fact that the labels are encoded as $y_i \in \{-1, +1\}$:

$$\zeta = \Phi\left(\frac{y_i\mu_i^{(i-1)}}{\sqrt{1+\Sigma_{ii}^{(i-1)}}}\right) := \Phi\left(z_i\right)$$

Next, we need to compute the first two moments of the true posterior:

$$\begin{aligned}
\mathbb{E}[f_i] &= \int_{-\infty}^{\infty} f_i \,\frac{1}{\zeta}\mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right)\Phi(y_i f_i)\,\mathrm{d}f_i \\
\mathbb{E}[f_i^2] &= \int_{-\infty}^{\infty} f_i^2 \,\frac{1}{\zeta}\mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right)\Phi(y_i f_i)\,\mathrm{d}f_i
\end{aligned}$$

To compute these, first observe that

$$\begin{aligned}\frac{\partial \zeta}{\partial \mu_i^{(i-1)}} &= \int_{-\infty}^{\infty} \Phi(y_i f_i) \frac{\partial \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right)}{\partial \mu_i^{(i-1)}} \, \mathrm{d}f_i \\ &= \int_{-\infty}^{\infty} \frac{f_i - \mu_i^{(i-1)}}{\Sigma_{ii}^{(i-1)}} \Phi(y_i f_i) \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \, \mathrm{d}f_i \\ &= \frac{1}{\Sigma_{ii}^{(i-1)}} \int_{-\infty}^{\infty} f_i \Phi(y_i f_i) \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \, \mathrm{d}f_i - \frac{\mu_i^{(i-1)}}{\Sigma_{ii}^{(i-1)}} \zeta\end{aligned}$$

and

$$\frac{\partial \zeta}{\partial \mu_i^{(i-1)}} = \frac{\partial}{\partial \mu_i^{(i-1)}} \Phi\left(z_i\right) = \mathcal{N}\left(z_i\right) \frac{\partial z_i}{\partial \mu_i^{(i-1)}} = \mathcal{N}\left(z_i\right) \frac{y_i}{\sqrt{1 + \Sigma_{ii}^{(i-1)}}}$$

where $\mathcal{N}(\cdot)$ is the standard normal PDF. The first term in the top equation is precisely the first moment we wish to compute multiplied by $\frac{\zeta}{\Sigma_{ii}^{(i-1)}}$! Thus, we can derive an expression for the first moment by setting the two (equivalent) expressions for $\frac{\partial \zeta}{\partial \mu_i^{(i-1)}}$ equal to each other and solving:

$$\begin{aligned}&\frac{1}{\Sigma_{ii}^{(i-1)}} \int_{-\infty}^{\infty} f_i \Phi(y_i f_i) \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \, \mathrm{d}f_i - \frac{\mu_i^{(i-1)}}{\Sigma_{ii}^{(i-1)}} \zeta = \mathcal{N}\left(z_i\right) \frac{y_i}{\sqrt{1 + \Sigma_{ii}^{(i-1)}}} \\ &\rightarrow \frac{\zeta}{\Sigma_{ii}^{(i-1)}} \int_{-\infty}^{\infty} f_i \frac{1}{\zeta} \Phi(y_i f_i) \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \, \mathrm{d}f_i = \frac{\mu_i^{(i-1)}}{\Sigma_{ii}^{(i-1)}} \zeta + \mathcal{N}\left(z_i\right) \frac{y_i}{\sqrt{1 + \Sigma_{ii}^{(i-1)}}} \\ &\rightarrow \mathbb{E}[f_i] = \mu_i^{(i-1)} + \mathcal{N}\left(z_i\right) \frac{y_i \Sigma_{ii}^{(i-1)}}{\zeta \sqrt{1 + \Sigma_{ii}^{(i-1)}}} := \mu_i^{(i-1)} + \frac{\mathcal{N}\left(z_i\right) s_i}{\zeta}\end{aligned}$$

Similarly, taking the second derivative of $\zeta$ with respect to $\mu_i$ gives

$$\begin{aligned}
\frac{\partial^2 \zeta}{\partial \mu_i^{(i-1)^2}} &= \frac{1}{\Sigma_{ii}^{(i-1)^2}} \int_{-\infty}^{\infty} \left(f_i^2 - f_i \mu_i^{(i-1)}\right) \Phi(y_i f_i) \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \mathrm{d}f_i \\
&\quad - \frac{1}{\Sigma_{ii}^{(i-1)}} \int_{-\infty}^{\infty} \Phi(y_i f_i) \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \mathrm{d}f_i \\
&\quad - \frac{\mu_i^{(i-1)}}{\Sigma_{ii}^{(i-1)^2}} \int_{-\infty}^{\infty} \left(f_i - \mu_i^{(i-1)}\right) \Phi(y_i f_i) \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \mathrm{d}f_i \\
&= \frac{1}{\Sigma_{ii}^{(i-1)^2}} \int_{-\infty}^{\infty} f_i^2 \Phi(y_i f_i) \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \mathrm{d}f_i \\
&\quad - 2\frac{\mu_i^{(i-1)}}{\Sigma_{ii}^{(i-1)^2}} \int_{-\infty}^{\infty} f_i \Phi(y_i f_i) \mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right) \mathrm{d}f_i - \frac{1}{\Sigma_{ii}^{(i-1)}}\zeta + \frac{\mu_i^{(i-1)^2}}{\Sigma_{ii}^{(i-1)^2}}\zeta
\end{aligned}$$

and

$$\frac{\partial^2 \zeta}{\partial \mu_i^{(i-1)^2}} = -\mathcal{N}(z_i) \frac{z_i}{1 + \Sigma_{ii}^{(i-1)}}$$

Setting these two equal and solving for the (normalized) first term gives

$$\begin{aligned}
&\frac{1}{\Sigma_{ii}^{(i-1)^2}}\zeta \mathbb{E}[f_i^2] - 2\frac{\mu_i^{(i-1)}}{\Sigma_{ii}^{(i-1)^2}}\zeta \mathbb{E}[f_i] - \frac{1}{\Sigma_{ii}^{(i-1)}}\zeta + \frac{\mu_i^{(i-1)^2}}{\Sigma_{ii}^{(i-1)^2}}\zeta = -\mathcal{N}(z_i) \frac{z_i}{1 + \Sigma_{ii}^{(i-1)}} \\
&\to \mathbb{E}[f_i^2] = 2\mu_i^{(i-1)}\mathbb{E}[f_i] + \Sigma_{ii}^{(i-1)} - \mu_i^{(i-1)^2} - \mathcal{N}(z_i) \frac{\Sigma_{ii}^{(i-1)^2} z_i}{\zeta\left(1 + \Sigma_{ii}^{(i-1)}\right)} \\
&:= 2\mu_i^{(i-1)}\mathbb{E}[f_i] + \Sigma_{ii}^{(i-1)} - \mu_i^{(i-1)^2} - \frac{\mathcal{N}(z_i)\, s_i^2 z_i}{\zeta}.
\end{aligned}$$

Finally, the variance of $f_i$ can be computed as

$$\begin{aligned}
\mathrm{var}(f_i) = \mathbb{E}[f_i^2] - \mathbb{E}[f_i]^2 &= 2\mu_i^{(i-1)}\mathbb{E}[f_i] + \Sigma_{ii}^{(i-1)} - \mu_i^{(i-1)^2} - \frac{\mathcal{N}(z_i)\, s_i^2 z_i}{\zeta} \\
&\quad - \mu_i^{(i-1)^2} - 2\mu_i^{(i-1)}\frac{\mathcal{N}(z_i)\, s_i}{\zeta} - \left(\frac{\mathcal{N}(z_i)\, s_i}{\zeta}\right)^2 \\
&= \Sigma_{ii}^{(i-1)} - \frac{\mathcal{N}(z_i)\, s_i^2}{\zeta}\left(z_i + \frac{\mathcal{N}(z_i)}{\zeta}\right)
\end{aligned}$$

Incredibly, we are only part way through the derivation! The next step is to compute the site parameters, $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$, such that the product $q_{i-1}(f_i)\tilde{t}_i(f_i; \tilde{\mu}_i, \tilde{\sigma}_i^2)$ is a Gaussian with the mean and variance that we just computed.

To do so, recall that the product of two independent Gaussian PDFs is an unnormalized Gaussian i.e.

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, )\mathcal{N}(\mathbf{x}; \boldsymbol{\nu}, \mathbf{P}) = \frac{1}{Z}\mathcal{N}(\mathbf{x}; \boldsymbol{\omega}, \mathbf{T}),$$

where

$$\mathbf{T} = (\boldsymbol{\Sigma}^{-1} + \mathbf{P}^{-1})^{-1} \text{ and } \boldsymbol{\omega} = \mathbf{T}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{P}^{-1}\boldsymbol{\nu})$$

In the context of assumed density filtering for Gaussian process classification, we want

$$\mathcal{N}\left(f_i; \mu_i^{(i-1)}, \Sigma_{ii}^{(i-1)}\right)\mathcal{N}(f_i; \tilde{\mu}_i, \tilde{\sigma}_i^2) \propto \mathcal{N}\left(f_i; \mathbb{E}[f_i], \text{var}(f_i)\right).$$

Thus, we can solve for $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ as follows:

$$\text{var}(f_i) = \left(\frac{1}{\Sigma_{ii}^{(i-1)}} + \frac{1}{\tilde{\sigma}_i^2}\right)^{-1} \rightarrow \tilde{\sigma}_i^2 = \left(\frac{1}{\text{var}(f_i)} - \frac{1}{\Sigma_{ii}^{(i-1)}}\right)^{-1}$$
$$\mathbb{E}[f_i] = \text{var}(f_i)\left(\frac{\mu_i^{(i-1)}}{\Sigma_{ii}^{(i-1)}} + \frac{\tilde{\mu}_i}{\tilde{\sigma}_i^2}\right) \rightarrow \tilde{\mu}_i = \tilde{\sigma}_i^2\left(\frac{\mathbb{E}[f_i]}{\text{var}(f_i)} - \frac{\mu_i^{(i-1)}}{\Sigma_{ii}^{(i-1)}}\right).$$

Finally, we can now combine $q_{i-1}(\mathbf{f})$ with our newly computed approximation $\mathcal{N}(f_i; \tilde{\mu}_i, \tilde{\sigma}_i^2)$ to get the next approximation in the sequence, $q_i(\mathbf{f})$.

However, to do so, we now need to "project" our 1-dimensional likelihood term back into the full $n-$dimensional space corresponding to $\mathbf{f}$. We can do this by constructing

- the matrix $\mathbf{S}$, an $n \times n$ matrix of all zeros except for the $i^{\text{th}}$ diagonal element, which is equal to $\frac{1}{\tilde{\sigma}_i^2}$ and
- the vector $\mathbf{m}$, which is an $n$ length vector of all zeros except for the $i^{\text{th}}$ element, which is $\frac{\tilde{\mu}_i}{\tilde{\sigma}_i^2}$.

Then the updated approximation can be expressed as

$$q_i(\mathbf{f}) \propto q_{i-1}(\mathbf{f})\tilde{t}_i(\mathbf{f}; \tilde{\boldsymbol{\eta}}_i) \propto \mathcal{N}\left(\mathbf{f}; \boldsymbol{\mu}^{(i)}, {}^{(i)}\right)$$

where we again apply the closure of Gaussian PDFs to get

$$\boldsymbol{\Sigma}^{(i)} = \left(\boldsymbol{\Sigma}^{(i-1)^{-1}} + \boldsymbol{S}\right)^{-1} \text{ and } \boldsymbol{\mu}^{(i)} = \boldsymbol{\Sigma}^{(i)}\left(\boldsymbol{\Sigma}^{(i-1)^{-1}}\boldsymbol{\mu}^{(i-1)} + \mathbf{m}\right).$$

We would then repeat this process for $i+1, i+2, \ldots$ and so on until all $n$ likelihood terms have been approximated as Gaussians, leading to a final Gaussian approximation for the intractable posterior.