1 Approximate Gaussian Process

1.1 Inducing Variables (DIC/FITC)

The unifying view paper [1] gives a pretty good explanation/interpretation of sparse GP methods based on inducing points till circa 2005. The basic idea is to augment existing model $p(\mathbf{y}, \mathbf{f})$ with m additional inducing variables $\mathbf{u} = \{u\}_{i=1}^m$ that are instantiation from the same GP prior, corresponding to a set of inducing locations $X_{\mathbf{u}}$. Inducing variables act as an information channel between training/testing latent function values $\mathbf{f}, \mathbf{f}_*$, alternatively we can think of $\mathbf{u}$ as a sufficient statistics of potentially redundant latent function values $\mathbf{f}$. This translates to a conditional independence assumption $\mathbf{f} \perp\!\!\!\perp \mathbf{f}_* \mid \mathbf{u}$, which implies the following factorization,

$$p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f}, \mathbf{f}_*, \mathbf{u}) = p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{u}) p(\mathbf{f}_* \mid \mathbf{u}) p(\mathbf{u}) \tag{1}$$

Without additional assumptions, we can think of joint priors factorizes akin to a latent variable model,

$$\mathbf{u} \sim \mathcal{N}(0, K_{\mathbf{uu}}) \tag{2}$$

$$\mathbf{f} \mid \mathbf{u} \sim \mathcal{N}(K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} \mathbf{u}, K_{\mathbf{uu}} - Q_{\mathbf{ff}}) \tag{3}$$

where $Q_{\mathbf{ff}} = K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} K_{\mathbf{uf}}$ is Nyström approximation of $K_{\mathbf{ff}}$. Various approximation schemes corresponds to approximations to the conditional prior, $q(\mathbf{f} \mid \mathbf{u}), q(\mathbf{f}_* \mid \mathbf{u})$. For exact GP, inference using the posterior GP $p(\mathbf{f} \mid \mathbf{y})$ requires inversion of a large kernel matrix; The introduction of inducing variables modifies the form that posterior GP $q(\mathbf{f} \mid \mathbf{y})$ takes.

To start, the Subset of Regressor (SoR) or Deterministic Inducing Conditional (DIC) approximation assumes that the latent function values as a linear mapping of inducing variables, e.g. $\mathbf{f} = K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} \mathbf{u}$. Note that the linear projection is one which preserves the mean of conditional prior in Equation (3). This assumption gives to an approximate conditional prior with zero covariance, $q_{\text{SOR}}(\mathbf{f} \mid \mathbf{u}) = \mathcal{N}(K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} \mathbf{u}, 0)$. Because of the linear mapping, $\mathbb{E}_q[\mathbf{f}] = 0$ and $\text{Cov}_q(\mathbf{f}) = K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} K_{\mathbf{uu}} K_{\mathbf{uu}}^{-1} K_{\mathbf{uf}} = Q_{\mathbf{ff}}$, similarly for $q(\mathbf{f}_*)$, and therefore,

$$q_{\text{DIC}}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(0, \begin{bmatrix} Q_{\mathbf{ff}} & Q_{\mathbf{f}*} \\ Q_{\mathbf{f}*} & Q_{**} \end{bmatrix}\right) \tag{4}$$

Inference with the DIC approximation is equivalent to exact inference with a modified covariance function $k_{\text{DIC}}(x_i, x_j) = k(x_i, \mathbf{u}) K_{\mathbf{uu}}^{-1} k(\mathbf{u}, x_j)$. The predictive distribution is then given by

$$q_{\text{DIC}}(\mathbf{f}_* \mid \mathbf{y}) \sim \mathcal{N}\left(Q_{*\mathbf{f}}(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1} \mathbf{y}, Q_{**} - Q_{*\mathbf{f}}(Q_{\mathbf{ff}} + \sigma_n^2 I)^{-1} Q_{\mathbf{f}*}\right) \tag{5}$$

Inversion of a $n \times n$ matrix is replaced with inversion of $m \times m$ matrix, making computation faster.

Sparse Gaussian Process using Pseudo-inputs (SGPP) or Fully Independent Training Conditional (FITC) approximation is another popular approximate GP method [2]. Simply, FITC assumes $\mathbf{f}_i \perp\!\!\!\perp \mathbf{f}_j \mid \mathbf{u}$ for all $i, j \in [n]$, henceforth the conditional priors are fully (conditionally) independent.

$$q_{\text{FITC}}(\mathbf{f} \mid \mathbf{u}) = \prod_{i=1}^{n} p(f_i \mid \mathbf{u}) = \mathcal{N}(K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} \mathbf{u}, \text{diag}\left[K_{\mathbf{uu}} - Q_{\mathbf{ff}}\right]) \tag{6}$$

Compare this to Equation (3), we see that FITC replaces approximate covariances with exact covariances on the diagonal. Equivalently, we can think of $\mathbf{f}$ as a noisy linear map of $\mathbf{u}$ where the noise has a particular form, e.g. $\mathbf{f} = K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} \mathbf{u} + \gamma$ where $\gamma \sim \mathcal{N}(0, \text{diag}\left[K_{\mathbf{uu}} - Q_{\mathbf{ff}}\right])$. We can derive the covariance for the approximate posterior $\text{Cov}_q(\mathbf{f}) = \text{Cov}(K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} \mathbf{u}) + \text{Cov}(\gamma) = Q_{\mathbf{ff}} + \text{diag}\left[K_{\mathbf{uu}} - Q_{\mathbf{ff}}\right]$. Therefore,

$$q_{\text{FITC}}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(0, \begin{bmatrix} Q_{\mathbf{ff}} - \text{diag}\left[Q_{\mathbf{ff}} - K_{\mathbf{ff}}\right] & Q_{\mathbf{f}*} \\ Q_{*\mathbf{f}} & K_{**} \end{bmatrix}\right) \tag{7}$$

Inference with FITC approximation is equivalent to exact inference with a modified covariance function $k_{\text{FITC}}(x_i, x_j) = k_{\text{DIC}}(x_i, x_j) + \delta_{i,j}(k(x_i, x_j) - k_{\text{DIC}}(x_i, x_j))$. The predictive distribution follows,

$$q_{\text{FITC}}(\mathbf{f}_* \mid \mathbf{y}) \sim \mathcal{N}\left(Q_{*\mathbf{f}}(Q_{\mathbf{ff}} + \Lambda)^{-1}\mathbf{y}, Q_{**} - Q_{*\mathbf{f}}(Q_{\mathbf{ff}} + \Lambda)^{-1}Q_{\mathbf{f}*}\right) \tag{8}$$

where $\Lambda = \text{diag}\left[K_{\mathbf{ff}} - Q_{\mathbf{ff}} + \sigma_n^2 I\right]$.

One popular methods for selecting inducing inputs $\mathbf{X}_u$ is simply by maximizing the marginal likelihood

$$q(\mathbf{y} \mid X_{\mathbf{u}}) = \iint p(\mathbf{y} \mid \mathbf{f})q(\mathbf{f} \mid \mathbf{u})p(\mathbf{u} \mid X_{\mathbf{u}})\, d\mathbf{u}\, d\mathbf{f} \tag{9}$$

1.2 Variational Formulation (VFE/SVGP)

Sparse GP with inducing variables is inflicted with the problem that modeling assumptions, e.g. choice of covariance function, is interwined with approximation assumptions. Specifically, we see that inference using approximation schemes like DIC and FITC can be interpreted as performing exact inference on GP model with a modified covariance function, where inducing locations $X_{\mathbf{u}}$ are part of kernel hyperparameters. This is undesirable.

[3] proposes to use variational inference to learning inducing locations. [4] extends the variational bound for stochastic gradient optimization, [5] extends the variational bound to account for non-conjugate likelihood. The accompanying technical report for the 2009 paper [6], the gentle tutorial [7], another report by Thang Bui [8], Mark van der Wilk's PhD thesis [9] are helpful in understanding the derivations.

[3] proposes Variational Free Energy (VFE) which applies variational inference to learn inducing variables in sparse GP, alleviating the above concern. Similar to inducing variable approximations, the model is augmented with inducing variables and assume the same conditional independence relationship so that the joint density factorizes $p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid \mathbf{u})p(\mathbf{u})$. Here $\mathbf{y}$ is observed while $\{\mathbf{u}, \mathbf{f}\}$ are latent variables. In contrast to inducing variable methods which makes approximating assumptions about the conditional priors, the paper introduce a variational density that preserves the true conditional prior with a freely varying prior over the inducing variables, e.g. $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})q(\mathbf{u})$ for some $q(\mathbf{u}) \sim \mathcal{N}(\mu_{\mathbf{u}}, \Sigma_{\mathbf{u}})$ that depends on inducing locations $X_{\mathbf{u}}$ as well as kernel hyperparameters. Together with Equation (3), we can derive an approximate posterior for $\mathbf{f}$,

$$q(\mathbf{f}) = \int p(\mathbf{f} \mid \mathbf{u})q(\mathbf{u})\, d\mathbf{u} = \mathcal{N}(K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\mu_{\mathbf{u}}, K_{\mathbf{ff}} - Q_{\mathbf{ff}} + K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\Sigma_{\mathbf{u}}K_{\mathbf{uu}}^{-1}K_{\mathbf{uf}}) \tag{10}$$

The paper then proposes to find variational parameters $\{\mu_n, \Sigma_n, X_{\mathbf{u}}, \theta\}$ by minimizing the distance between the approximate and true posterior, $q^* = \arg\min_{q(\mathbf{u})} \text{KL}(q(\mathbf{f}, \mathbf{u}) \mid p(\mathbf{f}, \mathbf{u} \mid \mathbf{y}))$. As usual in variational inference, we derive the ELBO to make objective tractable for optimization,

$$\text{KL}(q(\mathbf{f}, \mathbf{u}) \mid p(\mathbf{f}, \mathbf{u} \mid \mathbf{y})) = \int p(\mathbf{f} \mid \mathbf{u})q(\mathbf{u}) \log\left(\frac{p(\mathbf{f} \mid \mathbf{u})q(\mathbf{u})}{p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid \mathbf{u})p(\mathbf{u})}\right) d\mathbf{f}\, d\mathbf{u} + \log p(\mathbf{y}) \tag{11}$$

$$\log p(\mathbf{y}) \geq \int q(\mathbf{u}) \int p(\mathbf{f} \mid \mathbf{u}) \log p(\mathbf{y} \mid \mathbf{f})\, d\mathbf{f}\, d\mathbf{u} - \text{KL}(q(\mathbf{u}) \| p(\mathbf{u})) \tag{12}$$

With $\mathbb{E}[\mathbf{f} \mid \mathbf{u}] = K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\mathbf{u}$ and $\text{Cov}(\mathbf{f} \mid \mathbf{u}) = K_{\mathbf{ff}} - Q_{\mathbf{ff}}$, we can simplify,

$$\langle \log p(\mathbf{y} \mid \mathbf{f}) \rangle_{p(\mathbf{f}|\mathbf{u})} = \int p(\mathbf{f} \mid \mathbf{u}) \left[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\text{tr}\left(\mathbf{y}\mathbf{y}^T - 2\mathbf{y}\mathbf{f}^T + \mathbf{f}\mathbf{f}^T\right)\right] d\mathbf{f} \tag{13}$$

$$= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\text{tr}\left(\mathbf{y}\mathbf{y}^T - 2\mathbf{y}(\mathbb{E}[\mathbf{f} \mid \mathbf{u}])^T + \text{Cov}(\mathbf{f} \mid \mathbf{u}) + \mathbb{E}[\mathbf{f} \mid \mathbf{u}]\,\mathbb{E}[\mathbf{f} \mid \mathbf{u}]^T\right) \tag{14}$$

$$= \log \mathcal{N}(\mathbf{f} \mid K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\mathbf{u}, \sigma^2 I) - \frac{1}{2\sigma^2}\text{tr}\left(K_{\mathbf{ff}} - Q_{\mathbf{ff}}\right) \tag{15}$$

Now with $P := K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}$, $\mathbb{E}_q\left[\mathbf{u}\right] = \mu_{\mathbf{u}}$, and $\mathrm{Cov}_q(\mathbf{u}) = \Sigma_{\mathbf{u}} + \mu_{\mathbf{u}}\mu_{\mathbf{u}}^T$, then

$$\begin{aligned}\langle \log p(\mathbf{y} \mid \mathbf{f})\rangle_{q(\mathbf{f},\mathbf{u})} &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\operatorname{tr}\left(\mathbf{y}\mathbf{y}^T - 2\mathbf{y}(P\mu_{\mathbf{u}})^T + P(\Sigma_{\mathbf{u}} + \mu_{\mathbf{u}}\mu_{\mathbf{u}}^T)P^T + K_{\mathbf{uu}} - Q_{\mathbf{uu}}\right) && (16)\\ &= \log\mathcal{N}(\mathbf{f} \mid K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\mu_{\mathbf{u}}, \sigma^2 I) - \frac{1}{2\sigma^2}\operatorname{tr}\left(K_{\mathbf{ff}} - Q_{\mathbf{ff}}\right) - \frac{1}{2\sigma^2}\operatorname{tr}\left(K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\Sigma_{\mathbf{u}}K_{\mathbf{uu}}^{-1}K_{\mathbf{uf}}\right) && (17)\end{aligned}$$

[3] finds the optimal variational distribution $q(\mathbf{u})$ and substitute back into the lower bound, resulting in a bound where the first term is simply the marginal likelihoood for sparse GP with DIC/DTC approximation while the second term regularizes approximate kernel be close to the true kernel matrix,

$$\mathcal{L}_{\text{VFE}} := \log\mathcal{N}(\mathbf{f} \mid K_{\mathbf{fu}}K_{\mathbf{uu}}^{-1}\mathbf{u}, Q_{\mathbf{ff}} - \sigma^2 I) - \frac{1}{2\sigma^2}\operatorname{tr}\left(K_{\mathbf{ff}} - Q_{\mathbf{ff}}\right) \tag{18}$$

The bound is strictly better with additional inducing variables and becomes tight when $X_{\mathbf{u}} = X$, recovering the full GP marginal likelihood. Maximizing $\mathcal{L}_{\text{VFE}}$ with respect to kernel hyperparameters and inducing locations is tractable as the objective requires $\mathcal{O}(nm^2)$ to compute.

[4] proposes Stochastic Variational Gaussian Process (SVGP) that extends the bound (12) to allow for stochastic variational inference. The key observation is that with $p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^n p(y_i \mid f_i)$, the lower bound factors with respect to data points. Start with Equation (12),

$$\log p(\mathbf{y}) \geq \langle \log p(\mathbf{y} \mid \mathbf{f})\rangle_{q(\mathbf{f})} - \mathrm{KL}\left(q(\mathbf{u}) \| p(\mathbf{u})\right) = \sum_{i=1}^n \langle \log p(y_i \mid f_i)\rangle_{q(f_i)} - \mathrm{KL}\left(q(\mathbf{u}) \| p(\mathbf{u})\right) \tag{19}$$

$$\begin{aligned}\langle \log p(y_i \mid f_i)\rangle_{q(f_i)} &= \langle -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i^2 - 2y_i f_i + f_i^2)\rangle_{q(f_i)} && (20)\\ &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i^2 - 2y_i\mathbb{E}_q\left[f_i\right] + \mathrm{Var}_q(f_i) + \mathbb{E}_q\left[f_i\right]^2) && (21)\\ &= \log\mathcal{N}(y_i \mid K_{f_i\mathbf{u}}K_{\mathbf{uu}}^{-1}\mu_{\mathbf{u}}, \sigma^2) - \frac{1}{2\sigma^2}\left(K_{f_i f_i} - Q_{f_i f_i} - K_{f_i\mathbf{u}}K_{\mathbf{uu}}^{-1}\Sigma_{\mathbf{u}}K_{\mathbf{uu}}^{-1}K_{\mathbf{u}f_i}\right) && (22)\end{aligned}$$

where we have used fact that $q(f_i) \sim \mathcal{N}(P_i\mu_{\mathbf{u}}, K_{f_i f_i} - Q_{f_i f_i} - P_i\Sigma_{\mathbf{u}}P_i^T)$ from (10) and write $P_i := K_{f_i\mathbf{u}}K_{\mathbf{uu}}^{-1}$. This de-coupled term when summed over data points is exactly same as the coupled term in Equation (17). Now we can write down the lower bound which can be optimized with stochastic optimization,

$$\mathcal{L}_{\text{SVGP}} := \sum_{i=1}^n \left(\log\mathcal{N}(y_i \mid P_i\mu_{\mathbf{u}}, \sigma^2) - \frac{1}{2\sigma^2}\left(K_{f_i f_i} - Q_{f_i f_i} - P_i\Sigma_{\mathbf{u}}P_i^T\right)\right) - \mathrm{KL}\left(q(\mathbf{u}) \| p(\mathbf{u})\right) \tag{23}$$

where computation of KL term has $\mathcal{O}(m^3)$ cost

$$\mathrm{KL}\left(q(\mathbf{u}) \| p(\mathbf{u})\right) = \frac{1}{2}\left(\operatorname{tr}\left(K_{\mathbf{uu}}^{-1}\Sigma_{\mathbf{u}}\right) + \mu_{\mathbf{u}}^T K_{\mathbf{uu}}^{-1}\mu_{\mathbf{u}} - m + \log\frac{|K_{\mathbf{uu}}|}{|\Sigma_{\mathbf{u}}|}\right) \tag{24}$$

References

[1] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. "A Unifying View of Sparse Approximate Gaussian Process Regression". In: *The Journal of Machine Learning Research* 6 (Dec. 1, 2005), pp. 1939–1959. ISSN: 1532-4435.

[2] Edward Snelson and Zoubin Ghahramani. "Sparse Gaussian Processes using Pseudo-inputs". In: *Advances in Neural Information Processing Systems* 18 (2005), pp. 1257–1264. URL: https://proceedings.neurips.cc/paper/2005/hash/4491777b1aa8b5b32c2e8666dbe1a495-Abstract.html (visited on 02/05/2021).

[3] Michalis Titsias. "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In: *Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. ISSN: 1938-7228. PMLR, Apr. 15, 2009, pp. 567–574. URL: http://proceedings.mlr.press/v5/titsias09a.html (visited on 02/21/2021).

[4] James Hensman, Nicolo Fusi, and Neil D. Lawrence. "Gaussian Processes for Big Data". In: *arXiv:1309.6835 [cs, stat]* (Sept. 26, 2013). arXiv: 1309.6835. URL: http://arxiv.org/abs/1309.6835 (visited on 03/07/2021).

[5] James Hensman, Alex Matthews, and Zoubin Ghahramani. "Scalable Variational Gaussian Process Classification". In: *arXiv:1411.2005 [stat]* (Nov. 7, 2014). arXiv: 1411.2005. URL: http://arxiv.org/abs/1411.2005 (visited on 02/23/2021).

[6] Michalis K Titsias. "Variational Model Selection for Sparse Gaussian Process Regression". In: (2009), p. 20.

[7] Yarin Gal and Mark van der Wilk. "Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models - a Gentle Tutorial". In: *arXiv:1402.1412 [stat]* (Sept. 29, 2014). arXiv: 1402.1412. URL: http://arxiv.org/abs/1402.1412 (visited on 03/08/2021).

[8] Thang Bui and Richard Turner. "Sparse Approximations for Non-Conjugate Gaussian Process Regression". In: (2014), p. 7. URL: https://thangbui.github.io/docs/reports/tr_sparseNonConj.pdf.

[9] "Sparse Gaussian Process Approximations and Applications". In: (2018), p. 188. URL: https://markvdw.github.io/vanderwilk-thesis.pdf.