GP JOURNAL CLUB

On Sparse Variational Methods and the KL Between Stochastic Processes

Wessel Bruinsma

Machine Learning Group, University of Cambridge

18 Dec 2020

d. G. Matthews, A. G., Hensman, J., Turner, R. E., & Ghahramani, Z. (2016). On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In A. Singh & J. Zhu (Eds.), *Proceedings of the 22nd international conference on artificial intelligence and statistics* (Vol. 54). Proceedings of Machine Learning Research, Proceedings of Machine Learning Research. eprint: `https://arxiv.org/abs/1504.07027`

Introduction

- Titsias (2009) introduced trick to efficiently approximate posteriors of GPs using fewer *inducing points*.
- Very popular, but remained unclear whether Titsias' trick targets *posterior process*.
- d. G. Matthews et al. (2016) established a general condition under which this is the case.
- Puts massive body of work on solid theoretical grounding.

Today

Titsias' Trick and the Posterior Process

Titsias' Trick

- Consider prior $f \sim \mathcal{GP}(0, k)$ and observations $D = (\mathbf{x}, \mathbf{y})$.
- Goal: Efficiently approximate $p(f \mid D)$.

1. Augment model with inducing points $\mathbf{u}$:

$$p(f, \mathbf{u}) = \underbrace{p(f)}_{\text{original prior}}\overbrace{p(\mathbf{u} \mid f)}^{\text{augmentation}}.$$

2. Define approximate posterior:

$$q(f, \mathbf{u}) = \overbrace{p(f \mid \mathbf{u})}^{\text{from augmented prior}}\underbrace{q(\mathbf{u})}_{\text{variational variable}}.$$

3. Optimise approximation:

$$q^*(\mathbf{u}) = \operatorname*{arg\,max}_{q(\mathbf{u}) \in \mathcal{P}_{\mathsf{G}}} \mathsf{ELBO}(q(\mathbf{u}))$$

✓ Profit!

Today's Questions

implicitly depends on $q(\mathbf{u})$: $q(f) = \int p(f \mid \mathbf{u}) q(\mathbf{u}) \, \mathrm{d}\mathbf{u}$

1. Is it true that

$$q^*(\mathbf{u}) = \operatorname*{arg\,min}_{q(\mathbf{u}) \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(q(f), p(f \mid D)) \text{ ?}$$

- This is a KL between *processes*!

2. Can we formally make sense of "$p(f)$"?

Analysis with the ∞-Dimensional Lebesgue Measure

Disclaimer

- Thm: The ∞-dimensional Lebesgue measure does *not* exist.
- For now, think
$$f = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$$
for some large $n \gg 1$.
- We will later assign meaning to "$p(f)$".

Approach

1. Is it true that

$$q^*(\mathbf{u}) = \underset{q(\mathbf{u}) \in \mathcal{P}_{\mathsf{G}}}{\arg\min}\ \mathrm{KL}(q(f), p(f \mid D))\,?$$

- Strategy: Decompose

$$\mathrm{KL}(q(f, \mathbf{u}), p(f, \mathbf{u} \mid D))$$

in two ways.

First Decomposition

- Note that

$$\frac{p(f,\mathbf{u}\,|\,D)}{p(f,\mathbf{u})} = \frac{p(D\,|\,f)}{p(D)} \quad \text{and} \quad \frac{q(f,\mathbf{u})}{p(f,\mathbf{u})} = \frac{q(\mathbf{u})}{p(\mathbf{u})}.$$

- Then

$$\begin{aligned}
&\mathrm{KL}(q(f,\mathbf{u}), p(f,\mathbf{u}\,|\,D)) \\
&\quad = \int \log \frac{q(f,\mathbf{u})/p(f,\mathbf{u})}{p(f,\mathbf{u}\,|\,D)/p(f,\mathbf{u})} q(f,\mathbf{u})\,\mathrm{d}f\,\mathrm{d}\mathbf{u} \\
&\quad = \int \log \frac{q(\mathbf{u})/p(\mathbf{u})}{p(D\,|\,f)/p(D)} q(f,\mathbf{u})\,\mathrm{d}f\,\mathrm{d}\mathbf{u} \\
&\quad = \log p(D) - \underbrace{\Big(\mathbb{E}_q[\log p(D\,|\,f)] - \mathrm{KL}(q(\mathbf{u}), p(\mathbf{u}))\Big)}_{\mathsf{ELBO}(q(\mathbf{u}))}.
\end{aligned}$$

Second Decomposition

- Chain rule:

$$
\begin{aligned}
&\mathrm{KL}(q(f,\mathbf{u}), p(f,\mathbf{u} \mid D)) \\
&\quad = \underset{\substack{\uparrow \\ \text{target!}}}{\mathrm{KL}(q(f), p(f \mid D))} + \underset{\substack{\uparrow \\ ☹}}{\mathbb{E}_q[\mathrm{KL}(q(\mathbf{u} \mid f), p(\mathbf{u} \mid f, \cancel{D}))]}.
\end{aligned}
$$

Equating the Decompositions

- Result so far:

$$\log p(D) - \mathsf{ELBO}(q(\mathbf{u})) \\ = \mathrm{KL}(q(f), p(f \mid D)) + \mathbb{E}_q[\mathrm{KL}(q(\mathbf{u} \mid f), p(\mathbf{u} \mid f))].$$

- Therefore, if $q(\mathbf{u} \mid f) = p(\mathbf{u} \mid f)$, then

$$\operatorname*{arg\,max}_{q(\mathbf{u}) \in \mathcal{P}_{\mathsf{G}}} \mathsf{ELBO}(q(\mathbf{u})) = \operatorname*{arg\,min}_{q(\mathbf{u}) \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(q(f), p(f \mid D))\,!$$

- Thm (d. G. Matthews et al., 2016):

$$p(\mathbf{u} \mid f) = \delta(\mathbf{u} - T(f)) \implies q(\mathbf{u} \mid f) = \delta(\mathbf{u} - T(f)).$$

deterministic transform of f

Equating the Decompositions (2)

- Example 1 (inducing points):

$$T(f) = (f(z_1), \ldots, f(z_m)).$$

- Example 2 (interdomain transform):

$$T(f) = z \mapsto \int h(z, x) f(x) \, \mathrm{d}x.$$

- Proof (ish):

$$\begin{aligned} q(\mathbf{u} \,|\, f) &= \frac{q(\mathbf{u})}{q(f)} q(f \,|\, \mathbf{u}) = \frac{q(\mathbf{u})}{q(f)} p(f \,|\, \mathbf{u}) \\ &= \frac{q(\mathbf{u})}{q(f)} \frac{p(f)}{p(\mathbf{u})} p(\mathbf{u} \,|\, f) = \cancel{\frac{q(\mathbf{u})}{q(f)} \frac{p(f)}{p(\mathbf{u})}} \delta(\mathbf{u} - T(f)). \end{aligned}$$

Formalising the Argument

Setup of Formalism

- Consider

$$X = \text{all functions } f\colon \mathcal{X} \to \mathbb{R},$$
$$Z = \text{all functions } g\colon \mathcal{Z} \to \mathbb{R},$$

and assume that these spaces are Polish. ← complete, separable, metrisable

- We will consider probability measures on

$$(X \times Z, \mathcal{B}(X) \otimes \mathcal{B}(Z)).$$

Borel σ-algebra

- For such $\mathbb{P}$,

$$\text{marginals:} \quad \mathbb{P}_X \text{ and } \mathbb{P}_Z,$$
$$\text{conditionals:} \quad \mathbb{P}_{X \mid Z} \text{ and } \mathbb{P}_{Z \mid X}.$$

- Generally denote $A \in \mathcal{B}(X)$ and $B \in \mathcal{B}(Z)$.

General Kullback–Leibler Divergence

$$p(\mathbf{x}) = 0 \implies q(\mathbf{x}) = 0$$

- Def. ($\mathbb{R}^n$): For $q(\mathbf{x})$ and $p(\mathbf{x})$ such that $q(\mathbf{x}) \ll p(\mathbf{x})$, define

$$\mathrm{KL}(q(\mathbf{x}), p(\mathbf{x})) = \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \, \mathrm{d}\mathbf{x}.$$

- Def. (general): For $\mathbb{Q}$ and $\mathbb{P}$ such that $\mathbb{Q} \ll \mathbb{P}$, define

$$\mathrm{KL}(\mathbb{Q}, \mathbb{P}) = \int \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu} \log \frac{\mathrm{d}\mathbb{Q}/\mathrm{d}\mu}{\mathrm{d}\mathbb{P}/\mathrm{d}\mu} \, \mathrm{d}\mu$$

plays role of $\mathrm{d}\mathbf{x}$

where μ is any measure such that $\mathbb{P} \ll \mu$.

Setup of Titsias' Trick

- Augmented model:

augmentation

$$\mathbb{P}(A \times B) = \int_A \mathbb{P}_{Z \mid X}(B \mid f)\, \mathrm{d}\mathbb{P}_X(f).$$

original prior

- Approximate posterior:

from augmented prior

$$\mathbb{Q}(A \times B) = \int_B \mathbb{P}_{X \mid Z}(A \mid g)\, \mathrm{d}\mathbb{Q}_Z(g).$$

variational variable

- Exact posterior:

$$\frac{\mathrm{d}\mathbb{P}(\,\cdot \mid D)}{\mathrm{d}\mathbb{P}}(f, g) = \frac{p(D \mid f)}{\mathbb{E}_{\mathbb{P}}[p(D \mid f)]}.$$

FORMALISE: First Decomposition

WHAT WE WROTE

Observations:

$$\frac{p(f,\mathbf{u}\,|\,D)}{p(f,\mathbf{u})} = \frac{p(D\,|\,f)}{p(D)},$$

$$\frac{q(f,\mathbf{u})}{p(f,\mathbf{u})} = \frac{q(\mathbf{u})}{p(\mathbf{u})}.$$

ELBO:

$$\begin{aligned}
&\mathrm{KL}(q(f,\mathbf{u}), p(f,\mathbf{u}\,|\,D)) \\
&= \int \log \frac{q(f,\mathbf{u})/p(f,\mathbf{u})}{p(f,\mathbf{u}\,|\,D)/p(f,\mathbf{u})}\, \overset{q(f,\mathbf{u})\,\mathrm{d}f\,\mathrm{d}\mathbf{u}}{\overset{\big\downarrow}{\mathrm{d}q}} \\
&= \int \log \frac{q(\mathbf{u})/p(\mathbf{u})}{p(D\,|\,f)/p(D)}\, \mathrm{d}q \\
&= \log p(D) - \mathsf{ELBO}(q(\mathbf{u})).
\end{aligned}$$

WHAT WE MEANT

$$\frac{\mathrm{d}\mathbb{P}(\,\bullet\,|\,D)}{\mathrm{d}\mathbb{P}}(f,g) = \frac{p(D\,|\,f)}{\mathbb{E}_{\mathbb{P}}[p(D\,|\,f)]},$$

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(f,g) = \frac{\mathrm{d}\mathbb{Q}_Z}{\mathrm{d}\mathbb{P}_Z}(g).$$

$$\begin{aligned}
&\mathrm{KL}(\mathbb{Q}, \mathbb{P}(\,\bullet\,|\,D)) \\
&= \int \log \frac{\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}(\,\bullet\,|\,D)/\mathrm{d}\mathbb{P}}\, \mathrm{d}\mathbb{Q} \\
&= \int \log \frac{\mathrm{d}\mathbb{Q}_Z/\mathrm{d}\mathbb{P}_Z}{p(D\,|\,f)/p(D)}\, \mathrm{d}\mathbb{Q} \\
&= \log p(D) - \mathsf{ELBO}(\mathbb{Q}_Z).
\end{aligned}$$

FORMALISE: Second Decomposition

- Nothing to do! Chain rule works for general KL.

- Result so far:

$$\begin{aligned}&\log \mathbb{E}_{\mathbb{P}}[p(D \mid f)] - \mathsf{ELBO}(\mathbb{Q}_Z) \\ &\quad = \mathrm{KL}(\mathbb{Q}_X, \mathbb{P}_X(\,\cdot \mid D)) + \mathbb{E}_{\mathbb{Q}}[\mathrm{KL}(\mathbb{Q}_{Z \mid X}, \mathbb{P}_{Z \mid X})].\end{aligned}$$

- Therefore, if $\mathbb{Q}_{Z \mid X} = \mathbb{P}_{Z \mid X}$, then

$$\operatorname*{arg\,max}_{\mathbb{Q}_Z \in \mathcal{P}_{\mathsf{G}}} \mathsf{ELBO}(\mathbb{Q}_Z) = \operatorname*{arg\,min}_{\mathbb{Q}_Z \in \mathcal{P}_{\mathsf{G}}} \mathrm{KL}(\mathbb{Q}_X, \mathbb{P}_X(\,\cdot \mid D))\,!$$

- Thm (d. G. Matthews et al., 2016):

$$\mathbb{P}_{Z \mid X}(\,\cdot \mid f) = \delta_{T(f)} \implies \mathbb{Q}_{Z \mid X}(\,\cdot \mid f) = \delta_{T(f)}.$$

- Proof: More involved. See paper.

Wrap-Up

Conclusion

- Titsias' trick targets the posterior process if *inducing function* is a deterministic transform of f (d. G. Matthews et al., 2016).

$\Rightarrow$ Formally justifies inducing points and interdomain transforms.

- Although ∞-dimensional Lebesgue measure does not exist, manipulations can directly translate to formal manipulations with $p(f)\,\mathrm{d}f = \mathrm{d}\mathbb{P}$.

Appendix

References

d. G. Matthews, A. G., Hensman, J., Turner, R. E., & Ghahramani, Z. (2016). On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In A. Singh & J. Zhu (Eds.), *Proceedings of the 22nd international conference on artificial intelligence and statistics* (Vol. 54). Proceedings of Machine Learning Research, Proceedings of Machine Learning Research. eprint: `https://arxiv.org/abs/1504.07027`

Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk & M. Welling (Eds.), *Proceedings of the 12th international conference on artificial intelligence and statistics* (Vol. 12, pp. 567–574). Proceedings of Machine Learning Research. Proceedings of Machine Learning Research. Retrieved from `http://proceedings.mlr.press/v5/titsias09a/titsias09a.pdf`