Variational Prediction & Transductive Learning

Mark van der Wilk

March 28, 2024

UNIVERSITY OF OXFORD
Department of COMPUTER SCIENCE

Variational Prediction

Bayesian Models

In ML, we only care about the **predictive distribution**:

$$p(\boldsymbol{y}^* \mid \boldsymbol{y}).$$

Impossible to specify directly (name one case where possible).

Usually easier to specify a *generative model* $p(\boldsymbol{y}, \boldsymbol{y}^*)$:

$$p(\boldsymbol{y}^*|\boldsymbol{y}) = \frac{p(\boldsymbol{y}^*, \boldsymbol{y})}{p(\boldsymbol{y})}$$

Usually easier to specify with parameters (exchangeable, de Finetti's theorem):

$$p(\boldsymbol{y}, \boldsymbol{y}^*) = \int \left[\prod_{y_i \in (\boldsymbol{y}, \boldsymbol{y}^*)} p(y_i|\theta) \right] p(\theta) \, \mathrm{d}\theta$$

Bayesian Models: Take-homes

If we reparameterise θ, will $p(\boldsymbol{y}^*, \boldsymbol{y})$ change?

I.e. $\theta' = t(\theta), \quad P_{\theta'}(B) = P_\theta(t^{-1}(B))$.

If we reparameterise θ, will $p(\boldsymbol{y}^* | \boldsymbol{y})$ change?

If we reparameterise θ, will $p(\boldsymbol{y})$ change?

- Specific parameterisation doesn't matter to *observables*.
- We don't really care about any properties of parameters, they are simply a **means to an end**.

Variational Inference

Find $q(\theta) \approx p(\theta|\boldsymbol{y})$ by

$$\arg\min_{q \in Q} \mathrm{KL}[q(\theta) \parallel p(\theta|\boldsymbol{y})].$$

Find $p(\boldsymbol{y}^*|\boldsymbol{y})$ as

$$p(\boldsymbol{y}^*|\boldsymbol{y}) \approx q(\boldsymbol{y}^*) = \int p(\boldsymbol{y}^*|\theta)q(\theta)\,\mathrm{d}\theta.$$

⚠ **This is a pain, needs Monte Carlo.**

Can we not find $q(\boldsymbol{y}^*) \approx p(\boldsymbol{y}^*|\boldsymbol{y})$ directly?

We want to avoid:

- costly MC integration to find predictive $p(\boldsymbol{y}^*|\boldsymbol{y})$.
- computation wasted on parameters, and focus on prediction.

Variational Prediction

Want to minimise

$$\mathrm{KL}\left[q_{\boldsymbol{y}^*} \| p_{\boldsymbol{y}^*|\boldsymbol{y}}\right] = \int q(\boldsymbol{y}^*) \log \frac{q(\boldsymbol{y}^*)}{p(\theta|\boldsymbol{y})} \,\mathrm{d}\boldsymbol{y}^*$$

$$= \int q(\boldsymbol{y}^*) \log \frac{q(\boldsymbol{y}^*) p(\boldsymbol{y})}{\int p(\boldsymbol{y}^*|\theta) p(\boldsymbol{y}|\theta) p(\theta) \,\mathrm{d}\theta} \,\mathrm{d}\boldsymbol{y}^*$$

So, sadly, the usual variational inference trick doesn't apply, since the integral prevents us from getting expectations over tractable densities (which allows low-variance MC estimation in VI).

Any ideas?

- Jensen's inequality over $\int \ldots \mathrm{d}\theta$?

Tractable Variational Prediction

We *can* instead minimise

$$\mathrm{KL}\left[q_{\boldsymbol{y}^*,\theta} \| p_{\boldsymbol{y}^*,\theta|\boldsymbol{y}}\right] = \mathrm{KL}\left[q_{\boldsymbol{y}^*} \| p_{\boldsymbol{y}^*|\boldsymbol{y}}\right] + \underbrace{\mathbb{E}_{q_{\boldsymbol{y}^*}}\left[\mathrm{KL}\left[q_{\theta|\boldsymbol{y}^*} \| p_{\theta|\boldsymbol{y},\boldsymbol{y}^*}\right]\right]}_{\geq 0}$$

$$\therefore \mathrm{KL}\left[q_{\boldsymbol{y}^*,\theta} \| p_{\boldsymbol{y}^*,\theta|\boldsymbol{y}}\right] \geq \mathrm{KL}\left[q_{\boldsymbol{y}^*} \| p_{\boldsymbol{y}^*|\boldsymbol{y}}\right]$$

This *does* give a MC-tractable ELBO [1]:

$$\mathrm{KL}\left[q_{\boldsymbol{y}^*,\theta} \parallel p_{\boldsymbol{y}^*,\theta|\boldsymbol{y}}\right] = \int q(\boldsymbol{y}^*, \theta) \log \frac{q(\boldsymbol{y}^*, \theta) p(\boldsymbol{y})}{p(\boldsymbol{y}^*|\theta) p(\boldsymbol{y}|\theta) p(\theta)} \, \mathrm{d}\boldsymbol{y}^* \, \mathrm{d}\theta$$

$$\therefore \log p(\boldsymbol{y}) - \mathrm{KL}\left[q \parallel p_{\boldsymbol{y}^*,\theta|\boldsymbol{y}}\right] = \underbrace{\int q(\boldsymbol{y}^*, \theta) \log \frac{p(\boldsymbol{y}^*|\theta) p(\boldsymbol{y}|\theta) p(\theta)}{q(\boldsymbol{y}^*, \theta)} \, \mathrm{d}\boldsymbol{y}^* \, \mathrm{d}\theta}_{\mathcal{L}}$$

Tractable Variational Prediction

Putting the bound in another form:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{q_{\boldsymbol{y}^*}}\left[\mathbb{E}_{q_{\theta|\boldsymbol{y}^*}}[\log p(\boldsymbol{y}|\theta) + \log p(\boldsymbol{y}^*|\theta)]\right] + \\ -\mathbb{E}_{q_{\boldsymbol{y}^*}}\left[\mathrm{KL}\left[q_{\theta|\boldsymbol{y}^*} \parallel p_\theta\right]\right] + \\ \mathcal{H}[q(\boldsymbol{y}^*)]\end{aligned}$$

This is very similar to the familiar variational bound.

A. A. Alemi and B. Poole [1] suggest to parameterise $q(\boldsymbol{y}^*, \theta)$ by taking

$$\begin{aligned} q_{\boldsymbol{y}^*} &\in Q_p \\ q_{\theta|\boldsymbol{y}^*} &\in Q_c \qquad \text{NB: Conditionals!}\end{aligned}$$

When is this Useful?

Remember our goals!

- Definitely useful when we want to obtain $q(\boldsymbol{y}^*) \approx p(\boldsymbol{y}^*|\boldsymbol{y})$ *at training time.*

? What is an example of a model where this is useful?

Diffusion models? Good to amortise generation cost at training?

As an aside: I worked on a kind of variational prediction years ago. Not for amortisation, but instead for finding closed-form approximations of intractable predictive distributions [2].

When does VP work?

What does “work” mean?

$\Rightarrow$ We obtain low $\mathrm{KL}\left[q_{\boldsymbol{y}^*} \parallel p_{\boldsymbol{y}^*|\boldsymbol{y}}\right]$.

Remember:

$$\mathrm{KL}\left[q_{\boldsymbol{y}^*,\theta} \parallel p_{\boldsymbol{y}^*,\theta|\boldsymbol{y}}\right] = \mathrm{KL}\left[q_{\boldsymbol{y}^*} \parallel p_{\boldsymbol{y}^*|\boldsymbol{y}}\right] + \mathbb{E}_{q_{\boldsymbol{y}^*}}\left[\mathrm{KL}\left[q_{\theta|\boldsymbol{y}^*} \parallel p_{\theta|\boldsymbol{y},\boldsymbol{y}^*}\right]\right]$$

- Sufficient: $\mathrm{KL}\left[q_{\boldsymbol{y}^*,\theta} \parallel p_{\boldsymbol{y}^*,\theta|\boldsymbol{y}}\right]$ is small.
- $\mathrm{KL}\left[q_{\theta|\boldsymbol{y}^*} \parallel p_{\theta|\boldsymbol{y},\boldsymbol{y}^*}\right]$ is constant over $\boldsymbol{y}^*$, and our parameterisation of $q(\boldsymbol{y}^*)$ is flexible.

Transductive Learning

Defining (Bayesian) Transductive Learning

When can we say that transductive learning has taken place?

Transductive Learning

We want the predictions *that we care about* to be better, *without* our inductive learning capability getting better.

For transductive learning to have taken place, we need:

$$\mathrm{KL}\left[q_{\theta}^{\mathrm{VI}} \parallel p_{\theta|\boldsymbol{y}}\right] \leq \mathrm{KL}\left[q_{\theta}^{\mathrm{VP}} \parallel p_{\theta|\boldsymbol{y}}\right]$$

$$\mathrm{KL}\left[q_{\boldsymbol{y}^*}^{\mathrm{VI}} \parallel p_{\boldsymbol{y}^*|\boldsymbol{y}}\right] \geq \mathrm{KL}\left[q_{\boldsymbol{y}^*}^{\mathrm{VP}} \parallel p_{\boldsymbol{y}^*|\boldsymbol{y}}\right]$$

Can we prove that VP can/cannot do transductive learning?

I don’t know, happy to chat.

Bayesian Transductive Learning

We only know

$$\begin{aligned}\mathrm{KL}\left[q^{\mathrm{VP}}_{\boldsymbol{y}^*,\theta} \| p_{\boldsymbol{y}^*,\theta|\boldsymbol{y}}\right] &= \mathrm{KL}\left[q^{\mathrm{VP}}_{\boldsymbol{y}^*} \| p_{\boldsymbol{y}^*|\boldsymbol{y}}\right] + \mathbb{E}_{q_{\boldsymbol{y}^*}}\left[\mathrm{KL}\left[q^{\mathrm{VP}}_{\theta|\boldsymbol{y}^*} \| p_{\theta|\boldsymbol{y},\boldsymbol{y}^*}\right]\right] \\ &= \mathrm{KL}\left[q^{\mathrm{VP}}_{\theta} \| p_{\theta|\boldsymbol{y}}\right] + \mathbb{E}_{q_\theta}\left[\mathrm{KL}\left[q^{\mathrm{VP}}_{\boldsymbol{y}^*|\theta} \| p_{\boldsymbol{y}^*|\theta,\boldsymbol{y}}\right]\right]\end{aligned}$$

If we assume that $Q_M \subseteq Q$ the implied $q^{\mathrm{VP}}_\theta \in Q_M$, then we have

$$\mathrm{KL}\left[q^{\mathrm{VP}}_{\boldsymbol{y}^*,\theta} \| p_{\boldsymbol{y}^*,\theta|\boldsymbol{y}}\right] \geq \mathrm{KL}\left[q^{\mathrm{VI}}_{\theta} \| p_{\theta|\boldsymbol{y}}\right]$$

We can also find (but of limited help):

$$\mathrm{KL}\left[q^{\mathrm{VI}}_{\theta} \| p_{\theta|\boldsymbol{y}}\right] > \mathrm{KL}\left[q^{\mathrm{VI}}_{\boldsymbol{y}} \parallel p_{\boldsymbol{y}|\boldsymbol{y}^*}\right] \qquad \text{DPI}$$

Data Processing Inequality

Given a conditional $p(\boldsymbol{y}|\theta)$, and marginals

$$p(\theta) \qquad \Rightarrow \qquad p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\theta)p(\boldsymbol{y})\,\mathrm{d}\theta$$

$$q(\theta) \qquad \Rightarrow \qquad q(\boldsymbol{y}) = \int p(\boldsymbol{y}|\theta)q(\theta)\,\mathrm{d}\theta$$

Then,

$$\mathrm{KL}[q_\theta \parallel p_\theta] \geq \mathrm{KL}\left[q_{\boldsymbol{y}} \parallel p_{\boldsymbol{y}}\right].$$

Data Processing Inequality

Any processing cannot make distributions easier to distinguish from one another.

Variational Prediction
for
Sparse Gaussian Processes

Sparse Gaussian Processes

They are a great testbed for inference methods, because:

- You can control for many variables (e.g. control for optimisation behaviour by finding variational dists in closed-form)
- You can mathematically characterise/understand the true posterior (closed-form, but *computationally* intractable) [3]
- It is actually possible to get to the very accurate regime [4], [5]
- Parameters *are* predictions (specifically relevant for this case)

Transductive learning in approx GPs should concentrate inducing points around prediction areas. Board.

Variational Prediction for Sparse GPs

VP tells us to minimise $\mathrm{KL}\left[q^{\mathrm{VP}}_{\boldsymbol{y}^*,\theta} \| p_{\boldsymbol{y}^*,\theta|\boldsymbol{y}}\right]$.

For Sparse GPs, $\theta = (\boldsymbol{f}, \boldsymbol{u})$, $\boldsymbol{y} = \boldsymbol{f}^*$, so

$$\mathrm{KL}\left[q^{\mathrm{VP}}_{\boldsymbol{f}^*,\boldsymbol{f},\boldsymbol{u}} \| p_{\boldsymbol{f}^*,\boldsymbol{f},\boldsymbol{u}|\boldsymbol{y}}\right].$$

We choose the usual special posterior, but we need an arbitrary joint between $\boldsymbol{f}^*$ and $\boldsymbol{u}$:

$$q(\boldsymbol{f}^*, \boldsymbol{f}, \boldsymbol{u}) = q(\boldsymbol{f}^*, \boldsymbol{u})p(\boldsymbol{f}|\boldsymbol{u}, \boldsymbol{f}^*)$$

This is a normal inducing point approximation

The targeted distribution is just the normal *full* posterior over functions.

Conclusion

Conclusion

- You can train a predictive distribution with variational inference A. A. Alemi and B. Poole [1].
  - They haven't managed to get it to work at large scale.
  - My guess is that the goal is to speed up generation in diffusion models.
- Can *also* be thought of as a way to do Bayesian transductive learning.
- Not clear whether it actually can.
  - Can *any* Bayesian method do transductive learning? Or are we forced to do inference over everything, and be hampered in performance by the poorest part?
- In GPs, it just becomes the usual method, approximating the whole posterior.

Bibliography

[1] A. A. Alemi and B. Poole, "Variational Prediction," *arXiv preprint arXiv:2307.07568*, 2023.

[2] A. D. Ialongo, M. van der Wilk, and C. E. Rasmussen, "Closed-form inference and prediction in Gaussian process state-space models," *arXiv preprint arXiv:1812.03580*, 2018.

[3] M. Bauer, M. Van der Wilk, and C. E. Rasmussen, "Understanding probabilistic sparse Gaussian process approximations," *Advances in neural information processing systems*, vol. 29, 2016.

[4] D. R. Burt, C. E. Rasmussen, and M. van der Wilk, "Rates of Convergence for Sparse Variational Gaussian Process Regression," in *Proceedings of the 36th International Conference*

on Machine Learning, in Proceedings of Machine Learning Research. 2019.

[5] D. R. Burt, C. E. Rasmussen, and M. van der Wilk, "Convergence of Sparse Variational Inference in Gaussian Processes Regression," *Journal of Machine Learning Research*, 2020.