

Amortized Learning. Term Paper. Notes

Urakov Mikhail

October 14, 2025

Contents

| | |
|--|----------|
| 1 Papers Summaries | 3 |
| 1.1 BayesFlow: Learning Complex Stochastic Models With Invertable Neural Networks [Rad+20] | 3 |
| 1.2 Neural Methods for Amortized Inference [ZSH24] | 3 |

1 Papers Summaries

Here I summarize some main thoughts from papers and add questions that arised after reading joint with answers for some. Papers are organized in the order I read them, so there might be dumb questions that have answer in later articles.

1.1 BayesFlow: Learning Complex Stochastic Models With Invertable Neural Networks [Rad+20]

Problem: We have the standart Bayesian setup: model with parameters θ and data x . We want to estimate the posterior $p(\theta | x)$. From Bayes theorem we have

$$p(\theta | x) \propto p(x | \theta)p(\theta)$$

The problem is that in some cases (namely, likelihood-free cases) right-hand side is intractible because we cannot evaluate the $p(x | \theta)$, but we can sample from it, i.e.

$$x_i \sim p(x | \theta) \iff x_i = g(\theta, \xi_i), \xi_i \sim p(\xi)$$

Solution: Introducing normalizing flow that converts prior into Gaussian

$$\theta \sim p(\theta | x) \iff \theta = f_\varphi^{-1}(z; x), z \sim N(z | 0, \mathbb{I})$$

and considering right loss function we can now learn a summary and inference NN and it will work much more faster for all θ 's and x 's

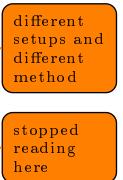
Q&A

1. How does the noise ξ selection affects the result? **Preanswer:** it depends also on simulation we use, and, of course, it does matter what noise we'll choose, because it changes the $p(x)$ and so $p(x | \theta)$
2. Why do we use Gaussian in normalizing flow?
3. Why don't we minimize the reverse KL divergence? **Answer:** it is also an option, which is considered in [Mur12], Chapter 21 and according to [ZSH24]: «*minimizing the reverse KL divergence leads to approximate distributions that are under-dispersed and that tend to concentrate mass on a single mode of the target distribution, whereas minimizing the forward KL divergence leads to ones that are over-dispersed and that cover all modes of the target distribution*». Both approaches are ubiquitous, but forward KL is easier to implement and it is likelihood-free in contrast to reverse KL.

1.2 Neural Methods for Amortized Inference [ZSH24]

They introduce Bayes risk as the common case of loss function in [Rad+20], where it was the KL divergence. blah-blah Minimizing KL divergence vs reverse KL divergence:

Summary networks



Q&A

1. *Average optimality* and what is it, and why do we use it?

References

- [Mur12] Kevin P. Murphy. “Machine learning - a probabilistic perspective”. In: *Adaptive computation and machine learning series*. 2012. URL: <https://api.semanticscholar.org/CorpusID:17793133>.
- [Rad+20] Stefan T. Radev et al. *BayesFlow: Learning complex stochastic models with invertible neural networks*. 2020. arXiv: 2003.06281 [stat.ML]. URL: <https://arxiv.org/abs/2003.06281>.
- [ZSH24] Andrew Zammit-Mangion, Matthew Sainsbury-Dale, and Raphaël Huser. *Neural Methods for Amortized Inference*. 2024. arXiv: 2404.12484 [stat.ML]. URL: <https://arxiv.org/abs/2404.12484>.