Amortized Learning. Term Paper. Notes

Urakov Mikhail

December 18, 2025

Contents

1 Goals

Common branch problems that we're trying to solve

1. Evaluate $p(\theta \mid x)$
2. Sample $p(\theta \mid x)$
3. $\mathbb{E}[\theta \mid x], \mathbb{D}[\theta \mid x]$, quantiles
4. $\theta_{MLE} = argmax_{\theta \sim p(\theta)} p(x \mid \theta)$
5. $\theta_{MAP} = argmax_{\theta \sim p(\theta)} p(\theta \mid x)$

2 Projects

1. FMPE, NPSI implementation
2. Summary network experiments (Mamba, CNN, RNN, RevRNN)
3. ABC methods
4. Develop benchmark

3 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.

3.1 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 $p(\theta \mid x)$. From Bayes theorem we have

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

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

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

Solution: Introducing normalizing flow that converts prior into Gaussian

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

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 $p(x \mid \theta)$
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 ubiquitious, but forward KL is easier to implement and it is likelihood-free in contrast to reverse KL.

3.2 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:

different setups and different method

Summary networks

stopped reading here

Q&A

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

3.3 Flow Matching Guide and Code [Lip+24]

Q&A

1. Prove that conditional optimal transport is really linear, i.e.

$$p_t(x) = \int p_{t|1}(x \mid x_1) q(x_1) dx_1$$

where $p_{t|1}(x \mid x_1) = \mathcal{N}(x \mid tx_1, (1-t)^2 I)$ and

$$X_t \sim p_t \iff X_t = tX_1 + (1-t)X_0$$

2. Verify that if u_t generates p_t then they satisfy the Continiuty Equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_t(x) + \operatorname{div}(p_t u_t)(x) = 0$$

Proof. We have $p_t(x) = [\psi_{t\#} p_0](x) = p_0(\psi_t^{-1}(x)) |J_x(\psi_t^{-1})(x)|$. Note that

$$p_{t+s} = [\psi_{s\#} p_t](x)$$

because

$$\begin{aligned}[\psi_{s\#}p_t](x) = p_t(\psi_s^{-1}(x))|J_x(\psi_s^{-1})(x)| = [\psi_{t\#}p_0](\psi_s^{-1}(x))|J_x(\psi_s^{-1})(x)| =\\ = p_0(\psi_t^{-1}(\psi_s^{-1}(x)))|J_x(\psi_t^{-1})(\psi_s^{-1}(x))||J_x(\psi_s^{-1})(x) =\\ = p_0(\psi_{t+s}^{-1}(x))|J_x(\psi_t^{-1}(\psi_s^{-1}))(x)| =\\ = p_0(\psi_{t+s}^{-1}(x))|J_x(\psi_{t+s}^{-1})(x)| = [\psi_{t+s\#}p_0](x)\end{aligned}$$

Then, we can express

$$p_{t+h}(x) - p_t(x) = p_t(\psi_h^{-1}(x))|J_x(\psi_h^{-1})(x)| - p_t(x)$$

Consider $\psi_h(y) = y + \dot{\psi}_h(y)h + o(h)$, take $y = \psi_h^{-1}(x) \implies \psi_h^{-1}(x) = x - \dot{\psi}_h(\psi_h^{-1}(x))h + o(h) = x - u_h(x)h + o(h)$. Then

$$p_t(x) = p_t(x - u_h(x)h + o(h) + u_h(x)h - o(h))$$

$$p_t(x) = p_t(x - u_h(x)h + o(h)) + \langle \nabla_x p_t, u_h(x)h \rangle + o(h)$$

$$p_t(x - u_h(x)h + o(h)) - p_t(x) = -\langle \nabla_x p_t, u_h(x)h \rangle + o(h)$$

$$p_t(\psi_h^{-1}(x)) - p_t(x) = -\langle \nabla_x p_t, u_h(x)h \rangle + o(h)$$

As long as $\psi_h(x) \to id$ as $h \to 0$ we have

$$p_{t+h}(x) - p_t(x) = -\langle \nabla_x p_t, u_h(x)h \rangle + o(h)$$

□

3.4 Approximate Bayesian Computation in Population Genetic, [BZB02]

Q&A

1. Prove the formula for local-linear regression
2. Provide formulas for multidim local-linear regression + code examples
   (a) Make class LLRegression with methods
   (b) Train and test it in various situations (different δ's, try other kernels instead of K_Δ)
   (c) Compare to analytic solutions and MCMC

4 Questions

1. Continious Flows (FMPE, NPSI) recent papers implementation
2. Sequential Methods (SNPE-A,B,C)

References

[BZB02] Mark A Beaumont, Wenyang Zhang, and David J Balding. "Approximate Bayesian Computation in Population Genetics". In: *Genetics* 162.4 (Dec. 2002), pp. 2025–2035. ISSN: 1943-2631. DOI: 10.1093/genetics/162.4.2025. eprint: https://academic.oup.com/genetics/article-pdf/162/4/2025/42049447/genetics2025.pdf. URL: https://doi.org/10.1093/genetics/162.4.2025.

[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.

[Lip+24] Yaron Lipman et al. *Flow Matching Guide and Code.* 2024. arXiv: 2412.06264 [cs.LG]. URL: https://arxiv.org/abs/2412.06264.

[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.