Discussion of "Adversarial Bayesian Simulation" by Yuexi Wang and Veronika Ročková

Weining Shen

Department of Statistics, UC Irvine

O'Bayes: September 9, 2022


VERY NICE PAPER
VERONIKA!

Remarks

- "The frontier of simulation-based inference" by Cranmer et al. (2020)
- Traditional simulation-based inference techniques face the following challenges:
- (1) Sample efficiency, (2) Quality of inference, and (3) scalability to large number of observations and new observations.
- Fast development in simulation-based inference recently for three reasons...

Well reflected in the discussed paper...

- (1) "The ML revolution allows us to work with higher-dimensional data, which can improve the quality of inference. Inference methods based on neural network surrogates are directly benefiting from the impressive rate of progress in deep learning."
- (2) "Active learning methods can systematically improve sample efficiency, letting us tackle more computationally expensive simulators."
- (3) "They still treat the simulator as a generative black box that takes parameters as input and provides data as output, with a clear separation between the simulator and the inference engine. A third direction of research is changing this perspective, by opening the black box and integrating inference and simulation more tightly."

Problem setting

- Notation: Parameter θ, observed data $X_0^{(n)}$
- Problem: How to sample from the posterior $\pi(\theta \mid X_0^{(n)}) \propto p_\theta^{(n)}(X_0^{(n)})\pi(\theta)$,
- when the likelihood $p_\theta^{(n)}(X_0^{(n)})$ and prior $\pi(\theta)$ are analytically intractable but easy to draw from?

Combine strengths: ABC and GAN

- ABC: generate fake data and match with the real data to generate posterior samples.

- (1) Generate reference tables $(\theta_j, X_j^{(n)})$, keep θ_j's if their associated summary statistics are close to those of the observed data.

- (2) ABC regression adjustment, improve the match by fitting a weighted regression of θ_j's on summary statistics.

- GAN: directly sample from complex/intractable likelihoods. Generator and Discriminator.

- Remark: at first, I thought it was to incorporate GAN within the ABC framework; but then I realize it's to use ABC within GAN.

Combine strengths: ABC and GAN

- ABC: generate fake data and match with the real data to generate posterior samples.

- (1) Generate reference tables $(\theta_j, X_j^{(n)})$, keep θ_j's if their associated summary statistics are close to those of the observed data.

- (2) ABC regression adjustment, improve the match by fitting a weighted regression of θ_j's on summary statistics.

- GAN: directly sample from complex/intractable likelihoods. Generator and Discriminator.

- Remark: at first, I thought it was to incorporate GAN within the ABC framework; but then I realize it's to use ABC within GAN.

Vanilla GAN to Bayesian GAN

- Vanilla GAN: Given observed data $X_0^{(n)} \sim P_{\theta_0}^{(n)}$, start with noise Z and find a deterministic map $g_\beta : Z \to X$ and $X \sim P_\theta^{(n)}$ such that $d_W(P_\theta^{(n)}, P_{\theta_0}^{(n)})$ is minimized.

- Conditional GAN: the key quantity is no longer X, but $\theta \mid X$.

- Note $\pi_g(X, \theta) = \pi_g(\theta \mid X)\pi(X)$. Fixing the marginal of X, matching joint distribution is the same with matching the conditional distribution.

- In plain words, we need a generator for (X, θ) and a discriminator that decides if a generated (X, θ) is actual data or fake data.

Bayesian GAN

- Wasserstein distance minimization between $\pi_g(X, \theta)$ and $\pi(X, \theta)$:

  $$(g^*, f^*) = \text{argmin}_{g \in \mathcal{G}} \text{argmin}_{f \in \mathcal{F}} |Ef(X, g(Z, X)) - Ef(X, \theta)|.$$

- (1) Estimate critic f and generator g using neural networks
- (2) Use ABC reference tables for empirical approximation of the expectation term.
- Compare between ABC reference table $\{\theta_j, X_j^{(n)})$ and $\{g(Z_j, X_j), X_j\}$ where Z_j's sampled from π_Z.
- Same X_j, marginal of X is kept the same.

Algorithm 1 *B-GAN for Bayesian Simulation (Wasserstein Version).*

Input

Prior $\pi(\theta)$, observed data X_0 and noise distribution $\pi_Z(\cdot)$

Training

Initialize network parameters $\omega^{(0)} = 0$ and $\beta^{(0)} = 0$

Reference Table

For $j = 1, \ldots, T$: Generate (X_j, θ_j) where $\theta_j \sim \pi(\theta)$ and $X_j \sim P_{\theta_j}^{(n)}$.

Wasserstein GAN

For $t = 1, \ldots, N$:

Critic Update (N_{critic} steps): For $k = 1, \ldots, N_{\text{critic}}$

Generate $Z_j \sim \pi_Z(z)$ for $j = 1, \ldots, T$.

Generate $\epsilon_j \overset{\text{iid}}{\sim} U[0,1]$ and set $\bar{\theta}_j = \epsilon_j \theta_j + (1 - \epsilon_j) g_{\beta^{(t-1)}}(Z_j, X_j)$ for $j = 1, \ldots, T$.

Update $\omega^{(t)}$ by applying stochastic gradient descent on (2.5) with the penalty (2.6).

Generator Update (single step)

Generate noise $Z_j \sim \pi_Z(z)$ for $j = 1, \ldots, N$.

Update $\beta^{(t)}$ by applying stochastic gradient descent on (2.5).

Posterior Simulation:

For $i = 1, \ldots, M$: Simulate $Z_i \sim \pi_Z(z)$ and set $\tilde{\theta}_i = g_{\beta^{(N)}}(Z_i, X_0)$.

First refinement for B-GAN

- B-GAN 2step: similarly with query-efficient ABC, generate clever proposals that lead to more efficient/accurate reference tables compared to X_0, then adjust the posterior by importance sampling. Efficiency improvement.
- (1) Generate reference tables using auxiliary proposal $\tilde{\pi}$
- (2) Reweight the samples by using $r(\theta) = \pi(\theta)/\tilde{\pi}(\theta)$, hence the posterior $\tilde{\pi}(\theta|X_0)r(\theta)$ is still proportional to the true posterior.
- (3) The density ratio r can be calculated analytically or approximated using neural networks, or using the probabilities from a classification.

Second refinement for B-GAN

- B-GAN-VB: maximize the evidence lower bound

$$\mathcal{L}(\beta) = -\mathsf{KL}(q_\beta(\theta \mid X_0) || \pi(\theta \mid X_0)) + CD$$

  in terms of β.

- Both the likelihood and posterior are implicit, so they adopt contrast learning for maximizing the evidence lower bound.
- Two contrasting data $\theta \sim \pi(\theta \mid X_0)$ and $\tilde{\theta} \sim q_\beta(\theta \mid X_0)$
- Same fixing-the-marginal and oracle classifier trick applies here:

$$\frac{d^*_{g_\beta}(X,\theta)}{d^*_{g_\beta}(X,\theta)} = \frac{\pi(X,\theta)}{q_\beta(\theta \mid X)\pi(X)}$$

  oracle classifier d_{g_β} to distinguish between $\pi(X,\theta)$ and $q_\beta(\theta \mid X)\pi(X)$.

- Replace aspects of the evidence lower bound with adversarial objectives.

Where is X_0 being used?

- For B-GAN, only in the simulation stage ($\tilde{\theta}_j$'s), not in network training.
- For B-GAN 2step, in the simulation stage ($\tilde{\theta}_j$'s) and proposal calculation, not in network training.
- For B-GAN-VB, in all stages, including network training.

Theory

- Upper bound for the total variational distance between true and approximated posterior measures.
- The error is decomposed into three terms:
  (1) the ability of the critic to tell the true model apart from the approximating model;
  (2) the ability of the generator to approximate the average true posterior;
  (3) the complexity of the (generating and) critic function classes.

Why does B-GAN 2Step work better than B-GAN?

Remark 3. *(2step Motivation) For the proposal distribution $\tilde{\pi}(\theta)$, using similar arguments as in the proof of Theorem 1, the TV distance of the posterior at X_0 (not averaged over $P_{\theta_0}^{(n)}$) can be upper-bounded by*

$$4\, d_{TV}^2\left(\nu(X_0), \mu_{\widehat{\beta}}(X_0)\right) \leq 2\,\mathcal{A}_1(\mathcal{F}, X_0) + \frac{B}{\sqrt{2}}\mathcal{A}_2(\mathcal{G}) + 4\widetilde{C}\, B\sqrt{\frac{\log T \times Pmax}{T}} + A_3(\widetilde{\pi})$$

where $\mathcal{A}_1(\mathcal{F}, X_0) \equiv \inf_\omega \left\| \log \frac{\pi(\theta \,|\, X_0)}{\pi_{\widehat{\beta}}(\theta \,|\, X_0)} - \frac{f_\omega(\theta, X_0)}{r(\theta)} \right]$ *is the discriminability* evaluated at X_0 *(as opposed to (4.4)) and where*

$$A_3(\widetilde{\pi}) = 2\int_{\mathcal{X}} \widetilde{\pi}(X)\left[\|f_\omega(X_0, \theta) - f_\omega(X, \theta)\|_\infty + B\|g_{\widehat{\beta}}(\theta)(X) - g_{\widehat{\beta}}(\theta)(X_0)\|_1\right] \mathrm{d}X$$

and $g_{\widehat{\beta}}(\theta)(X) \equiv \pi(\theta \,|\, X) - \pi_{\widehat{\beta}}(\theta \,|\, X)$. This decomposition reveals how the TV distance can be related to discriminability around X_0 and an average discrepancy between the true and approximated posterior densities relative to their value at X_0 where the average is taken over the marginal $\widetilde{\pi}(X)$. These averages will be smaller the marginal $\widetilde{\pi}(X)$ produces

Question - can we obtain something similar by comparing the error bound between B-GAN and B-GAN-VB?

(scale)	$\theta_1 = 0.01$ bias ($\times 10^{-3}$)	CI width ($\times 10^{-2}$)	$\theta_2 = 0.5$ bias ($\times 10^{-1}$)	CI width	$\theta_3 = 1.0$ bias	CI width	$\theta_4 = 0.01$ bias ($\times 10^{-2}$)	CI width ($\times 10^{-2}$)
B-GAN	4.15	1.89	1.09	0.45	0.24	1.00	0.49	2.18
B-GAN-2S	**0.70**	**0.21 (0.9)**	0.42	**0.10 (0.7)**	**0.11**	0.33 (0.9)	0.13	0.34 (0.8)
B-GAN-VB	1.02	0.25 (0.7)	**0.38**	0.11 (0.9)	**0.11**	**0.29 (0.8)**	**0.12**	**0.29 (0.7)**
SNL	1.05	0.44	0.45	0.17	0.13	0.48	0.15	0.52
SS	9.58	3.80	2.49	0.91	0.49	1.76	0.68	2.72
W2	10.99	4.02 (0.9)	2.42	0.84	0.47	1.73	0.79	2.82

Table 1: Summary statistics of the approximated posteriors under the Lotka-Volterra model (averaged over 10 repetitions). Bold fonts mark the best model of each column. The coverage of the 95% credible intervals are 1 unless otherwise noted in the parentheses.

(scale)	$r = 0.4$ bias ($\times 10^{-1}$)	CI width ($\times 10^{-1}$)	$\kappa = 50$ bias	CI width	$\alpha = 0.09$ bias ($\times 10^{-2}$)	CI width ($\times 10^{-1}$)	$\beta = 0.05$ bias ($\times 10^{-1}$)	CI width
B-GAN	0.44	1.63	2.92	10.78	3.03	1.38	1.22	0.36 (0.8)
B-GAN-2S	0.27	0.79 (0.8)	1.60	5.29 (0.9)	1.06	0.34	1.05	0.26 (0.7)
B-GAN-VB	**0.23**	**0.65 (0.8)**	**1.29**	**4.88 (0.9)**	**0.89**	**0.25 (0.7)**	**1.00**	**0.19 (0.5)**
SNL	0.24	0.93	1.52	5.37	1.01	0.38	1.28	0.39 (0.9)
SS	2.16	8.26	10.60	37.17	15.08	9.18	4.41	0.95
W2	2.59	9.49	10.16	43.20	5.46	2.77	3.92	0.86 (0.6)

Table 2: Summary statistics of the approximated posteriors under the Boom-and-Bust model (averaged over 10 repetitions). Bold fonts mark the best model of each column. The coverage of the 95% credible intervals are 1 unless otherwise noted in the parentheses.

	SS	W2	SNL	B-GAN	B-GAN-2S	B-GAN-VB
Gauss	33.75	221.28	4790.56	2736.93	676.25	726.22
Lotka-Volterra	5846.95	162644.96	3080.96	1610.05	762.21	753.61

Table 6: Computation time of one repetition for each method on Gauss example and Lotka-Volterra (LV) example (in seconds). The time of B-GAN-2S and B-GAN-VB is for computation using the adjusted prior.

Compared to B-GAN, the improvement is significant for both B-GAN 2step and B-GAN-VB, in terms of every aspect.

A few questions

- Compare these two refinements, Which one to use in what scenarios? Is it correct to say B-GAN-VB tends to underestimate uncertainty/CI, but is more accurate for complex models? Some discussions on the scalability would also be helpful.
- Extension to model comparison/model evidence? Streaming data modeling?

Other questions

- Jensen-Shannon divergence and Wasserstein distance. The authors give a nice example of convergence/computational issue for JS divergence. But I wonder what price is paid for using Wasserstein distance, besides computational cost?

- Remark 2 assumes ϵ_n could be $n^{-1/2}$, then the prior concentration condition

$$\Pi(B_n(\theta_0; \epsilon_n)) \geq e^{-C_2 n \epsilon_n^2}$$

needs to be adjusted accordingly.


THANK
YOU
L