On Martingale Posteriors, Gaussian Processes & Pseudo Data

Yingzhen Li

yingzhen.li@imperial.ac.uk

References

Discussions based on many papers. To start I recommend reading:

- Fong, Edwin, Chris Holmes, and Stephen G. Walker. "Martingale posterior distributions." *Journal of the Royal Statistical Society Series B: Statistical Methodology 85.5 (2023): 1357-1391.*
- Leibfried, Felix, et al. "A tutorial on sparse Gaussian processes and variational inference." *arXiv preprint arXiv:2012.13962 (2020).*
- Ritter, Hippolyt, et al. "Sparse uncertainty representation in deep learning with inducing weights." *Advances in Neural Information Processing Systems 34 (2021): 6515-6528.*

Two ways to construct posterior

The task:

given data $y_{1:N}$, construct a predictor $p(y_{N+1:\infty}|y_{1:N})$

Solution 1: Explicit Bayesian modelling (de Finetti)

- Build a model:

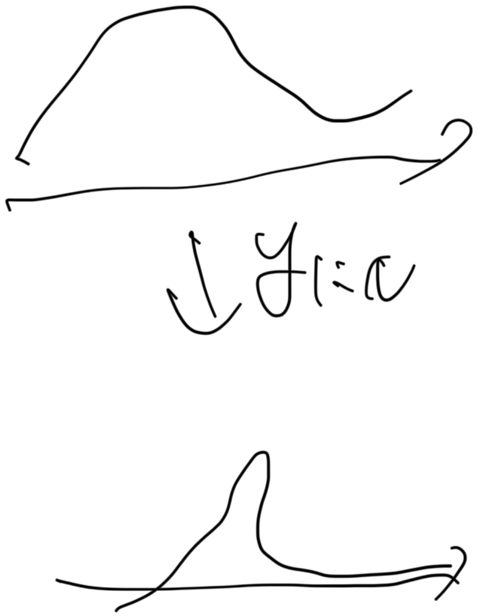
  Prior: $p(\theta)$ Likelihood: $p(y|\theta)$

- Compute posterior:

$$p(\theta|y_{1:N}) \propto \prod_{n=1}^{N} p(y_n|\theta)\, p(\theta)$$

- Bayesian predictive inference:

$$p(y_{N+1:\infty}|y_{1:N}) = \int \prod_{n=N+1}^{\infty} p(y_n|\theta) p(\theta|y_{1:N})\, d\theta$$



Fong et al. "Martingale posterior distributions." JRSSB, 2023

Two ways to construct posterior

The task:

given data $y_{1:N}$, construct a predictor $p(y_{N+1:\infty}|y_{1:N})$

Solution 2: Direct construction of predictive (Doob)

- Define a 1-step predictor (with some conditions):

$$\{p(\cdot\,|y_{1:N})\}_{n>N} \quad \Rightarrow \quad p(y_{N+1:\infty}|y_{1:N}) = \prod_{n=N+1}^{\infty} p(y_n|y_{1:n-1})$$

Fong et al. "Martingale posterior distributions." JRSSB, 2023

Two ways to construct posterior

The task:

given data $y_{1:N}$,   construct a predictor $p(y_{N+1:\infty}|y_{1:N})$

Solution 2: Direct construction of predictive (Doob)

- Define a 1-step predictor (with some conditions):

$$\{p(\cdot\,|y_{1:N})\}_{n>N} \qquad \Rightarrow \qquad p(y_{N+1:\infty}|y_{1:N}) = \prod_{n=N+1}^{\infty} p(y_n|y_{1:n-1})$$

- Consider the following definition:

$$Y_{N+1:M} \sim \{p(\cdot\,|y_{1:N})\}_{n>N}$$

$$p_M(y|Y_{N+1:M}) \coloneqq \hat{p}(y|y_{1:N}, Y_{N+1:M}) \qquad \text{(empirical dist.)}$$

Two ways to construct posterior

The task:

given data $y_{1:N}$,   construct a predictor $p(y_{N+1:\infty}|y_{1:N})$

Solution 2: Direct construction of predictive (Doob)

- Define a 1-step predictor (with some conditions):

  $\{p(\cdot\,|y_{1:N})\}_{n>N} \quad \Rightarrow \quad p(y_{N+1:\infty}|y_{1:N}) = \prod_{n=N+1}^{\infty} p(y_n|y_{1:n-1})$

- Consider the following definition:

  $$Y_{N+1:M} \sim \{p(\cdot\,|y_{1:N})\}_{n>N}$$

  $\{y_{1:N}, Y_{N+1:M}\}$

  $$p_M(y|Y_{N+1:M}) \coloneqq \hat{p}(y|y_{1:N}, Y_{N+1:M}) \quad \text{(empirical dist.)}$$

  $p(y|\theta)$

- Construct the (finite) Martingale posterior $p_M(\theta_M|y_{1:N})$ and take $M \to \infty$:

  $$\theta_M \sim p_M(\theta_M|y_{1:N}) \quad \Leftrightarrow \quad Y_{N+1:M} \sim \{p(\cdot\,|y_{1:N})\}_{n>N},$$
  $$\theta_M = argmin_\theta\, KL[p_M(y|Y_{N+1:M})||p(y|\theta)]$$

Fong et al. "Martingale posterior distributions." JRSSB, 2023

How do they connect to each other?

- Remember the steps for sampling from the Martingale posterior:
  - Specify a predictive model $\{p(\cdot\,|y_{1:N})\}_{n>N}$
  - Sample $Y_{N+1:M} \sim \{p(\cdot\,|y_{1:N})\}_{n>N}$
  - Fit a model $p(y|\theta)$ to dataset $\{y_{1:N}, Y_{N+1:M}\}$ with MLE
  - Consider the distribution of the MLE estimator when $M \to \infty$

Fong et al. "Martingale posterior distributions." JRSSB, 2023

How do they connect to each other?

- Remember the steps for sampling from the Martingale posterior:
  - Specify a predictive model $\{p(\cdot\,|y_{1:N})\}_{n>N}$
  - Sample $Y_{N+1:M} \sim \{p(\cdot\,|y_{1:N})\}_{n>N}$
  - Fit a model $p(y|\theta)$ to $p_M(y|Y_{N+1:M}) := \hat{p}(y|y_{1:N}, Y_{N+1:M})$ with MLE
  - Consider the distribution of the MLE estimator when $M \to \infty$
- Conditions to make this procedure work:
  - The limit exists (a.s., wrt. dist of $Y_{N+1:\infty}$):

$$p_\infty(y|Y_{N+1:\infty}) = \lim_{M\to\infty} p_M(y|Y_{N+1:M}),$$

Fong et al. "Martingale posterior distributions." JRSSB, 2023

How do they connect to each other?

- Remember the steps for sampling from the Martingale posterior:
  - Specify a predictive model $\{p(\cdot\,|y_{1:N})\}_{n>N}$
  - Sample $Y_{N+1:M} \sim \{p(\cdot\,|y_{1:N})\}_{n>N}$
  - Fit a model $p(y|\theta)$ to $p_M(y|Y_{N+1:M}) \coloneqq \hat{p}(y|y_{1:N}, Y_{N+1:M})$ with MLE
  - Consider the distribution of the MLE estimator when $M \to \infty$
- Conditions to make this procedure work:
  - The limit exists (a.s., wrt. dist of $Y_{N+1:\infty}$):

$$p_\infty(y|Y_{N+1:\infty}) = \lim_{M\to\infty} p_M(y|Y_{N+1:M}),$$

  - Bias-free:

$$E[p_\infty(y|Y_{N+1:\infty})] = \hat{p}(y|y_{1:N})$$

Fong et al. "Martingale posterior distributions." JRSSB, 2023

How do they connect to each other?

- Now let's use the Bayesian predictive dist. from Solution 1:

$$p(y_{N+1:M}|y_{1:N}) = \int \prod_{n=N+1}^{M} p(y_n|\theta)p(\theta|y_{1:N})\, d\theta$$

The following two sampling methods are equivalent:

Joint sampling: $\theta \sim p(\theta|y_{1:N})$, $Y_{N+1:M} \sim p(\cdot\,|\theta)$ i.i.d.

Sequential sampling: $Y_m \sim p\left(\cdot\,\middle|Y_{N+1,m-1}, y_{1:N}\right), m = N+1: M$

Fong et al. "Martingale posterior distributions." JRSSB, 2023

How do they connect to each other?

- Now let's use the Bayesian predictive dist. from Solution 1:

$$p(y_{N+1:M}|y_{1:N}) = \int \prod_{n=N+1}^{M} p(y_n|\theta)p(\theta|y_{1:N})\, d\theta$$

The following two sampling methods are equivalent:

Joint sampling: $\theta \sim p(\theta|y_{1:N})$, $Y_{N+1:M} \sim p(\cdot\,|\theta)$ i.i.d.

Sequential sampling: $Y_m \sim p(\cdot\,|Y_{N+1,m-1}, y_{1:N}), m = N+1{:}\,M$

Two solutions are the same:

- If $Y_{N+1:M}$ is generated by $p(y|\theta_0)$ using $\theta_0 \sim p(\theta|y_{1:N})$,
  then under identifiability conditions, the MLE solution converges:
  $\theta_M = argmin_\theta\, KL[p_M(y|Y_{N+1:M})||p(y|\theta)]$, $\theta_M \to \theta_0$ as $M \to \infty$

$P_m(y | Y_{N+1:m})$

$= \frac{1}{m}\sum_{m=1}^{N} \mathbb{1}(y = Y_m)$

$P(\theta | y_{1:N}) \longrightarrow \theta_0 \longrightarrow \{Y_{N+1:\infty}, y_{1:N}\} \longrightarrow \theta_0$

$\sim P(\theta | y_{1:N})$

Fong et al. "Martingale posterior distributions." JRSSB, 2023

How do they connect to each other?

- Now let's use the Bayesian predictive dist. from Solution 1:

$$p(y_{N+1:M}|y_{1:N}) = \int \prod_{n=N+1}^{M} p(y_n|\theta)p(\theta|y_{1:N})\, d\theta$$

The following two sampling methods are equivalent:

Joint sampling: $\theta \sim p(\theta|y_{1:N})$, $Y_{N+1:M} \sim p(\cdot\,|\theta)$ i.i.d.

Sequential sampling: $Y_m \sim p(\cdot\,|Y_{N+1,m-1}, y_{1:N}), m = N+1: M$

Two solutions are the same:

- If $Y_{N+1:M}$ is generated by $p(y|\theta_0)$ using $\theta_0 \sim p(\theta|y_{1:N})$, then under identifiability conditions, the MLE solution converges: $\theta_M = argmin_\theta\, KL[p_M(y|Y_{N+1:M})||p(y|\theta)]$, $\theta_M \to \theta_0$ as $M \to \infty$

- Unbiasedness is satisfied since $Y_{N+1:M}$ are conditionally identically distributed (Berti et al. 2004)

Fong et al. "Martingale posterior distributions." JRSSB, 2023

Two philosophies on Bayesian inference

- Solution 1 (de Finetti): explicit parameter dist. modelling
  - Specify prior $p(\theta)$
  - Specify likelihood $p(y|\theta)$
  - Posterior $p(\theta|y_{1:N})$ from Bayes' rule
- Solution 2 (Doobs): predictive distribution modelling
  - Specify a sequence of predictive dist. $\{p(\cdot\, |y_{1:N})\}_{n>N}$
  - Specify likelihood $p(y|\theta)$
  - Posterior $p(\theta|y_{1:N})$ implicitly defined by the following procedure
    - Sample more data $Y_{N+1:M}$ from the predictive dists.
    - Fit $p(y|\theta)$ to the augmented dataset by MLE
    - Repeat the above two steps multiple times and have a set of samples for θ

Fong et al. "Martingale posterior distributions." JRSSB, 2023

Two philosophies on Bayesian inference

$P(\theta)$

$$f_\theta(y), \pi(\theta) \xrightarrow[\text{Bayes' rule}]{} \pi(\theta \mid y_{1:n}) \xrightarrow[\int f_\theta(y)\, \pi(\theta|y_{1:n})\, d\theta]{\text{posterior predictive}} p(y \mid y_{1:n})$$

$P(y \mid \theta)$

$$\pi(\theta \mid y_{1:n}) \xleftarrow[Y_{n+1:\infty} \sim p(\cdot|y_{1:n})]{\text{Doob's theorem}} p(y \mid y_{1:n}) \xleftarrow[\text{predictive update}]{} p(y)$$

Fong et al. "Martingale posterior distributions." JRSSB, 2023

Two philosophies on Bayesian inference

1-1 correspondence if $\{p(\cdot \mid y_{1:N})\}$ and $f_\theta(y)$ (or in our notation $p(y|\theta)$) are consistent!

$$f_\theta(y), \pi(\theta) \xrightarrow[\text{Bayes' rule}]{} \boxed{\pi(\theta \mid y_{1:n}) \xrightarrow[\int f_\theta(y)\,\pi(\theta|y_{1:n})\,d\theta]{\text{posterior predictive}} p(y \mid y_{1:n})}$$

$$\boxed{\pi(\theta \mid y_{1:n}) \xleftarrow[Y_{n+1:\infty} \sim p(\cdot|y_{1:n})]{\text{Doob's theorem}} p(y \mid y_{1:n})} \xleftarrow[\text{predictive update}]{} p(y)$$

$$\int p(Y_{n+2}=y, Y_{n+1} \mid y_{1:n})\, dY_{n+1}$$

$$= p(Y_{n+1}=y \mid y_{1:n}) \qquad p(Y_{n+\infty} \mid y_{1:n}) \qquad \cancel{q(Y_{\infty}|\theta)}$$

$$= \int_{\infty} \prod p(Y_{\infty}|\theta)\, p(\theta \mid y_{1:n})\, d\theta$$

Fong et al. "Martingale posterior distributions." JRSSB, 2023

Not so surprising to GP researchers…

- Gaussian process modelling:

$$f \sim GP\big(m(\cdot), k(\cdot,\cdot)\big), \qquad y \sim p(y|f(x))$$

Leibfried et al. "A tutorial on sparse Gaussian processes and variational inference." *arXiv:2012.13962*

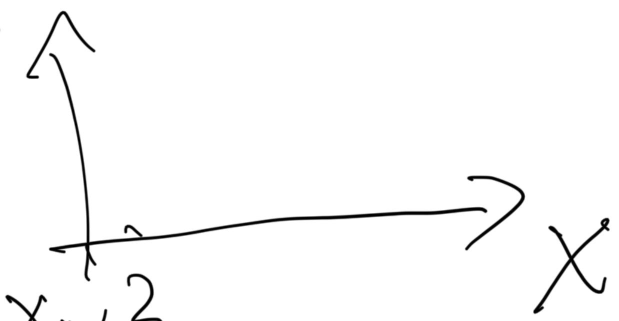
Not so surprising to GP researchers…

- Gaussian process modelling:

$$f \sim GP\big(m(\cdot), k(\cdot,\cdot)\big), \qquad y \sim p(y|f(x))$$

- Important: f is infinite dimensional!
- A prior on f will induce a prior on $Y_{1:\infty}$

$p(y|f(x)) = \mathcal{N}(f(x), \sigma^2 I)$



$\mathbf{X} = \{x_1, \dots, x_N\}$

$\underline{f} = \{f(x_1), \dots, f(x_N)\}$

prior: $p(\underline{f}|\mathbf{X}) = \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_N) \end{bmatrix}, K_{\mathbf{X}\mathbf{X}}\right)$

$K_{\mathbf{X}\mathbf{X}} = \begin{bmatrix} k(x_1,x_1), & k(x_1,x_2) & \cdots \\ & \ddots & \\ & & k(x_N,x_N) \end{bmatrix}$

$N \to \infty$

$p(Y_{1:N}|\mathbf{X}) = \int \prod_n p(Y_n|f(x_n))\, p(\underline{f}|\mathbf{X})\, d\underline{f}$

$= \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \end{bmatrix}, K_{\mathbf{X}\mathbf{X}} + \sigma^2 I\right)$

Leibfried et al. "A tutorial on sparse Gaussian processes and variational inference." *arXiv:2012.13962*

Not so surprising to GP researchers…

- In fact: this idea works for any explicit Bayesian models
- The structure of prior on $Y_{1:\infty}$

Solution 1:

$$p(\theta), \quad \underline{p(y|\theta)}$$

$$p(Y_{1:N}) = \int p(\theta) \prod_n p(Y_n|\theta)\, d\theta$$

$$p(Y_{N+1:M}|y_{1:N}) = \frac{p(Y_{N+1:M}, \underline{y_{1:N}})}{p(\underline{y_{1:N}})}$$

Leibfried et al. "A tutorial on sparse Gaussian processes and variational inference." *arXiv:2012.13962*

GP people is smart about augmenting data

- GP posterior is expensive to compute - $O(N^3)$ time complexity

$$p(f|y_{1:N}) = GP(m_{post}(\cdot), k_{post}(\cdot,\ \cdot))$$

- Sparse GP approximation: use augmented "pseudo data"

$u = f(Z)$

$\rightarrow p(\underline{f}, \underline{u} | X, Z)$ pseudo data

$\rightarrow p(f|y_{1:N}) \approx q(f) \triangleq \int p(\underline{f}|\underline{u}) q(\underline{u}) d\underline{u}$

Leibfried et al. "A tutorial on sparse Gaussian processes and variational inference." *arXiv:2012.13962*

The augmented prior idea goes beyond GPs

- Consider a probabilistic model

Prior: $p(\theta)$ Likelihood: $p(y|\theta)$

- Augmentations for predictive inference:
  - Augment the prior: define $\pi(\theta, U)$ such that

$$\int \pi(\theta, U) dU = p(\theta)$$

Ritter et al. "Sparse uncertainty representation in deep learning with inducing weights." *NeurIPS 2021.*

The augmented prior idea goes beyond GPs

- Consider a probabilistic model

  Prior: $p(\theta)$ Likelihood: $p(y|\theta)$

- Augmentations for predictive inference:
  - Augment the prior: define $\pi(\theta, U)$ such that
    $$\int \pi(\theta, U) dU = p(\theta)$$
  - Now consider the posterior predictive:

$$p(y|y_{1:N}) = \int p(y|\theta) p(\theta|y_{1:N}) d\theta = \frac{p(y, y_{1:N})}{p(y_{1:N})}$$

Ritter et al. "Sparse uncertainty representation in deep learning with inducing weights." *NeurIPS 2021.*

The augmented prior idea goes beyond GPs

- Consider a probabilistic model

  Prior: $p(\theta)$ Likelihood: $p(y|\theta)$

- Augmentations for predictive inference:
  - Augment the prior: define $\pi(\theta, U)$ such that
    $$\int \pi(\theta, U) dU = p(\theta)$$
  - Now consider the posterior predictive:

$$p(y|y_{1:N}) = \int p(y|\theta) p(\theta|y_{1:N}) d\theta = \frac{p(y, y_{1:N})}{p(y_{1:N})}$$

$$p(y_{1:N}) = \int \prod_{n=1}^{N} p(y_n|\theta)\, p(\theta) d\theta = \int \prod_{n=1}^{N} p(y_n|\theta)\, \pi(\theta, U) d\theta dU$$

$$= \int p(y_{1:N}|U) \pi(U) dU$$

$$\int \prod_{n=1}^{N} p(y|\theta) \pi(\theta|U) d\theta$$

$$\pi(\theta, U) = \pi(\theta|U)\, \pi(\theta)$$

Ritter et al. "Sparse uncertainty representation in deep learning with inducing weights." *NeurIPS 2021.*

The augmented prior idea goes beyond GPs

- Consider a probabilistic model

  Prior: $p(\theta)$ Likelihood: $p(y|\theta)$

- Augmentations for predictive inference:
  - Augment the prior: define $\pi(\theta, U)$ such that

    $$\int \pi(\theta, U) dU = p(\theta)$$

  - Now consider the posterior predictive:

    Key: predictive $p(y|y_{1:N})$ remains the same regardless of using $p(\theta|y_{1:N})$ or $p(U|y_{1:N})$

Ritter et al. "Sparse uncertainty representation in deep learning with inducing weights." *NeurIPS 2021.*

The augmented prior idea goes beyond GPs

- Consider a probabilistic model

  Prior: $p(\theta)$ Likelihood: $p(y|\theta)$

- Augmentations for predictive inference:
  - Augment the prior: define $\pi(\theta, U)$ such that
  $$\int \pi(\theta, U) dU = p(\theta)$$
  - Now consider the posterior predictive:

You can then think about how to do variational inference then ;)

$$p(y|y_{1:N}) \approx \int p(y|\theta) q(\theta) d\theta \quad q(\theta) \approx p(\theta|y_{1:N})$$

$$\approx \int p(y|u) q(u) du \quad q(u) \approx p(u|y_{1:N})$$

Ritter et al. "Sparse uncertainty representation in deep learning with inducing weights." *NeurIPS 2021.*

Take aways

- ML people cares about predictions (only?)
- This prompts us to think about more direct approaches
  - Function-space uncertainty (de Finetti)
  - Martingale posterior (Doob)
- Auxiliary variables (e.g., pseudo data) as flexible approach to assist!
  - Gaussian processes
  - General (approximate) Bayesian inference

Take aways

- ML people cares about predictions (only?)
- This prompts us to think about more direct approaches
  - Function-space uncertainty (de Finetti)
  - Martingale posterior (Doob)
- Auxiliary variables (e.g., pseudo data) as flexible approach to assist!
  - Gaussian processes
  - General (approximate) Bayesian inference
- Bonus: other Bayesian methods that involves auxiliary variables
  - Hamiltonian Monte Carlo
  - Polya-Gamma Augmentation
  - …