

# Closed-Form Inference

**Mark van der Wilk**

Department of Computing  
Imperial College London

@markvanderwilk  
m.vdwilk@imperial.ac.uk

January 24, 2022

# Computing Posteriors

- ▶ Previous lecture investigated **which** computations we need to perform to find posteriors.
- ▶ Now, we focus on **actually doing them**.
- ▶ We focus on inference problems with two variables, one hidden (**latent**), one observed (**data**).

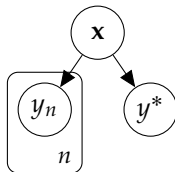
$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \quad (1)$$

- ▶ As before, the model is specified by the full joint, often in terms of tractable densities, i.e.

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \quad (2)$$

# Terminology

$$p(\mathbf{y}, y^*, \mathbf{x}) = p(y^*|\mathbf{x}) \prod_{n=1}^N p(y_n|\mathbf{x})p(\mathbf{x}) \quad (3)$$

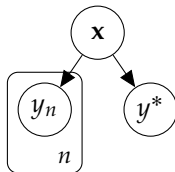


$$\underbrace{p(\mathbf{x}|\mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y}|\mathbf{x})}^{\text{likelihood}} \overbrace{p(\mathbf{x})}^{\text{prior}}}{\underbrace{p(\mathbf{y})}_{\text{marginal likelihood / evidence}}} = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{\int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}} \quad (4)$$

- ▶ When solving an inference problem, what is fixed and what is variable in  $p(\mathbf{x}|\mathbf{y})$ ? ►► observation  $\mathbf{y}$  is fixed,  $\mathbf{x}$  varies.
- ▶ What variable is the likelihood a function of? ►►  $\mathbf{x}$  only!  
We say “likelihood of parameters / latent variable  $\mathbf{x}$ ”.

# Terminology

$$p(\mathbf{y}, y^*, \mathbf{x}) = p(y^*|\mathbf{x}) \prod_{n=1}^N p(y_n|\mathbf{x})p(\mathbf{x}) \quad (5)$$



$$\underbrace{p(y^*|\mathbf{y})}_{\substack{\text{posterior} \\ \text{predictive}}} \stackrel{\text{MA}}{=} \int \underbrace{p(y^*|\mathbf{x})}_{\text{posterior}} \underbrace{p(\mathbf{x}|\mathbf{y})}_{\text{posterior}} d\mathbf{x} = \int p(y^*|\mathbf{x}) \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} d\mathbf{x} \quad (6)$$

- ▶ If we aren't talking about a fixed observed dataset, we can investigate properties of distributions as a function of  $\mathbf{y}$ . If we do so we may refer to  $p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}$  as the *prior predictive distribution*.
- ▶ We use different terminology for these settings to indicate whether  $\mathbf{y}$  is observed and fixed, or whether we investigate how the probability changes for different possible outcomes  $\mathbf{y}$ .

# Example: One-Armed Bandit

- ▶ Each time you run a “one-armed bandit”, you get a random return of  $Y_n$ .
- ▶  $Y_n$  is distributed according to density  $p(y_n|x)$ , with  $\mathbb{E}_{p(y_n|x)}[y_n] = x$ .
- ▶ The mean return is assigned by the manufacturer by sampling from  $p(x)$ .



## Example: One-Armed Bandit

You are interested in computing for example:

- ▶  $p(x|\mathbf{y})$  to understand your belief about the average return. In particular

$$P(X > 0|\mathbf{y}) = \int_0^{\infty} p(x|\mathbf{y})dx. \quad (7)$$

- ▶  $p(y^*|\mathbf{y})$  to understand your belief in your potential return in the next run. In particular

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|x)p(x|\mathbf{y})dx \quad (8)$$

$$P(Y^* > 0|\mathbf{y}) = \int_0^{\infty} p(\mathbf{y}^*|\mathbf{y})d\mathbf{y}^*. \quad (9)$$

- ▶ In general, we are interested in **summary statistics** of posterior distributions ► **integrals**.

# How difficult is Bayesian inference?

- ▶ Let's think about *actually* computing some of these quantities.
- ▶ How difficult is this really?

$$P(X > 0|\mathbf{y}) = \int_0^\infty p(\mathbf{x}|\mathbf{y})d\mathbf{x}, \quad p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}$$

Let's start with the posterior.

- ▶ Computing the numerator of  $p(\mathbf{x}|\mathbf{y})$  is easy: multiplication.
- ▶ Denominator  $p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}$  seems hard.

**Integrals are hard.**

- ▶ Do we really need  $p(\mathbf{y})$ ? It's just a constant... Are relative probabilities not enough?
- ▶ ► No hope of computing  $p(X > 0|\mathbf{y})$  without  $p(\mathbf{y})$ .

## Example: One-Armed Bandit (doing integrals)

Take  $p(y_n|x) = \mathcal{N}(y_n; x, \sigma^2)$  and  $p(x) = \mathcal{N}(x; 0, v)$ .

$$p(x|\mathbf{y}) = \frac{\prod_n \mathcal{N}(y_n; x, \sigma^2) \mathcal{N}(x; 0, v)}{p(\mathbf{y})} \quad (10)$$

$$= \frac{(2\pi\sigma^2)^{-\frac{N}{2}} (2\pi v)^{-\frac{1}{2}}}{p(\mathbf{y})} \exp \left[ -\frac{1}{2\sigma^2} \sum_n (y_n - x)^2 - \frac{1}{2v} x^2 \right] \quad (11)$$

$$= \frac{(2\pi)^{-\frac{N+1}{2}} \sigma^{-N} v^{-\frac{1}{2}}}{p(\mathbf{y})} \exp \left[ -\frac{1}{2\tau} (x - \mu)^2 - \frac{1}{2} \left( \sum_n y_n^2 - \frac{\mu^2}{\tau} \right) \right] \quad (12)$$

$$= c \exp \left[ -\frac{1}{2\tau} (x - \mu)^2 \right] \quad (13)$$

Equate coefficients to obtain

$$\tau = \frac{v\sigma^2}{vN + \sigma^2}, \quad \mu = \frac{v}{vN + \sigma^2} \sum_n y_n. \quad (14)$$



## Example: One-Armed Bandit (doing integrals)

From previous slide we know:

$$p(x|\mathbf{y}) = c \exp\left[-\frac{1}{2\tau}(x - \mu)^2\right], \quad (15)$$

$$\tau = \frac{v\sigma^2}{vN + \sigma^2}, \quad \mu = \frac{v}{vN + \sigma^2} \sum_n y_n. \quad (16)$$

How to find  $c$ ? Two options:

1. We know that  $\int p(x|\mathbf{y})dx = 1$ . Do the integral using  $\int e^{-x^2}dx = \sqrt{\pi}$ .

$$\int c \exp\left[-\frac{1}{2\tau}(x - \mu)^2\right]dx = c \cdot \sqrt{2\pi\tau} = 1 \implies c = \frac{1}{\sqrt{2\pi\tau}} \quad (17)$$

2. Let someone else do the integral, by using knowledge that

$$\mathcal{N}(x; \mu, \tau) = \frac{1}{\sqrt{2\pi\tau}} \exp\left[-\frac{1}{2\tau}(x - \mu)^2\right]. \quad (18)$$

# Why could we compute the posterior?

One reason:

We could integrate the unnormalised posterior.

$$p(x|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \quad (19)$$

$$Z = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad \text{in this case} \quad (20)$$

$$p(x|\mathbf{y}) = \frac{1}{Z}p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \quad \text{in this case} \quad (21)$$

- ▶ This was the case because  $p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$  as a function of  $\mathbf{x}$  implies a Gaussian distribution.
- ▶ We know how to do the integral to normalise a Gaussian.

# Intractable Inference

Example where things don't work out so nicely. Take  $y_n \in \{0, 1\}$ .

$$p(x) = \mathcal{N}(x; 0, v) \quad (22)$$

$$\ell(x) = \frac{1}{1 + e^{-x}} \quad \text{Logistic function} \quad (23)$$

$$p(y_n|x) = \ell(x)^{y_n} \cdot (1 - \ell(x))^{1-y_n}. \quad (24)$$

I.e.  $Y_n$  is Bernoulli distributed with probability  $\ell(x)$ .

$$p(x|\mathbf{y}) = \frac{1}{Z} \frac{e^{-N_1 x - \frac{1}{2v} x^2}}{(1 + e^{-x})^N} \quad (25)$$

$$Z = \int \frac{e^{-N_1 x - \frac{1}{2v} x^2}}{(1 + e^{-x})^N} dx \quad (26)$$

# Intractable Inference

$$Z = \int \frac{e^{-N_1 x - \frac{1}{2v} x^2}}{(1 + e^{-x})^N} dx \quad (27)$$

- ▶ No known “**closed-form**” solution to this integral.
- ▶ Closed-form: Combination of finite number of terms of standard functions (exp, sin, log, sqrt...). Sometimes includes special functions (e.g. Gamma, Bessel...)
- ▶ If no closed-form solution is known, a quantity is also said to be **intractable**.
- ▶ Inference is intractable if it requires computing intractable quantities.

# Tractable Inference

- ▶ In general it is hard to tell when inference is tractable.
- ▶ There is a set of distributions for which you can tell that inference is tractable.

## Definition

A prior and likelihood are **conjugate** if their resulting posterior is of the same family as the prior.

If your prior was tractable,  
then your posterior will be as well!

## Example: Gaussian-Gaussian conjugacy

The Gaussian example we saw earlier was an example of conjugacy.

- ▶ Likelihood formed from Gaussian with unknown mean:

$$L(x) = p(\mathbf{y}|x) = \prod_n \mathcal{N}(y_n; x, \sigma^2) \quad (28)$$

- ▶ Prior from the Gaussian family of distributions:

$$p(x) = \mathcal{N}(x; 0, v) \quad (29)$$

- ▶ Posterior is also Gaussian!

$$p(x|\mathbf{y}) \propto L(x)p(x) \quad (30)$$

$$p(x|\mathbf{y}) = \mathcal{N}\left(x; \frac{v \sum_n y_n}{vN + \sigma^2}, \frac{v\sigma^2}{vN + \sigma^2}\right) \quad (31)$$

# Exponential Family

This is no coincidence. The Gaussian distributions are part of the **exponential family**:

$$p(x|\eta) = h(x) \exp(\eta^\top t(x) - A(\eta)) \quad \eta, t \in \mathbb{R}^D \quad (32)$$

Different  $t(x)$  (and therefore  $A(\eta)$ ), give different distributions.

## Example: Gaussian

$$p(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \quad (33)$$

$$t(x) = [x \quad x^2]^\top, \quad \eta = [\mu/\sigma^2 \quad -\frac{1}{2\sigma^2}]^\top, \quad (34)$$

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2} \log(-2\eta_2), \quad h(x) = (2\pi)^{-\frac{1}{2}}. \quad (35)$$

# Exponential Family

This is no coincidence. The Gaussian distributions are part of the **exponential family**:

$$p(x|\eta) = h(x) \exp(\eta^\top t(x) - A(\eta)) \quad \eta, t \in \mathbb{R}^D \quad (36)$$

Different  $t(x)$  (and therefore  $A(\eta)$ ), give different distributions.

## Example: Bernoulli

$$p(x) = \theta^x \cdot (1 - \theta)^{1-x} \quad x \in \{0, 1\} \quad (37)$$

$$= \exp(x(\log \theta - \log 1 - \theta) + \log 1 - \theta) \quad (38)$$

$$t(x) = x, \quad \eta = \log \frac{p}{1-p}, \quad (39)$$

$$A(\eta) = \log 1 - p, \quad h(x) = 1. \quad (40)$$



# Conjugate Prior for Exponential Family

Exponential families have conjugate priors!  
For the likelihood:

$$\ell(\eta) = p(x|\eta) = h(x) \exp(\eta^\top t(x) - A(\eta)) \quad \eta, t \in \mathbb{R}^D \quad (41)$$

We have the conjugate prior:

$$p(\eta|\tau, n_0) = H(\tau, n_0) \exp(\tau^\top \eta - n_0 A(\eta)) \quad (42)$$

# Exam skills (NOT THIS YEAR)

Previous years:

- ▶ Convert distributions that are exponential families into their **natural form** (i.e. parameterised by  $\eta$ ).
- ▶ Recognise when a likelihood and prior are conjugate, and when they are not.
- ▶ Find conjugate prior to a likelihood in exponential family.

See examples sheet for practice.

# Exam skills (THIS YEAR)

You must be able to:

- ▶ do closed-form inference when distributions are Gaussian,
- ▶ do closed-form inference for discrete distributions,
- ▶ recognise when integrals w.r.t. Gaussians are possible,
- ▶ do integrals if an identity is given.

# Summary

## Inference

The procedure of drawing conclusions from observations.

In Bayesian statistics: Computing some conditional distribution (posterior).

## Closed-form Expressions

A mathematical expression consisting of a finite number of standard operations (pow, exp, log, trig, etc).

See [https://en.wikipedia.org/wiki/Closed-form\\_expression](https://en.wikipedia.org/wiki/Closed-form_expression).

## Closed-form Inference

An inference problem where all relevant quantities (e.g. posteriors) can be computed in closed-form.

# Summary

- ▶ Integrals appear when finding the posterior (normalising constant / marginal likelihood)
- ▶ Integrals appear when making predictions
- ▶ Integrals can only be done in special cases
- ▶ Conjugate models is a (big) family of these special cases, which helps you recognise when you can do the closed-form inference (but this isn't examined this year)

# Reading

Recommended reading:

- ▶ §6.6 of Mathematics for Machine Learning [1].

Further reading:

- ▶ §9.2 of ML: a Probabilistic Perspective [2].

# References I

- [1] M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.
- [2] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.