# Derivation of Bayes Rule
# — A Variational Approach —

Prof. James Richard Forbes

McGill University, Department of Mechanical Engineering

# Bayes Rule

- Previously showed that Bayes Rule (or Bayes Theorem),

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

  can be derived from the Marginalization and Conditioning Theorems, applied in sequence.
- Often Bayes Rule is said to be “optimal”, but in what sense?
- Following [1] and [2], Bayes Rule will be derived.

# Differential Entropy (Shannon Information)

- Given $p(\mathbf{x})$, the differential entropy, or Shannon information, is

$$H_p(\mathbf{x}) = -E[\ln p(\mathbf{x})] = -\int_{-\infty}^{\infty} p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \tag{2}$$

  where the integration bounds are assumed to be $\mathbf{a} = -\infty$ and $\mathbf{b} = \infty$.
- Differential entropy is a measure of average surprise or unpredictability.
  - When $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the differential entropy is

$$H_p(\mathbf{x}) = \tfrac{1}{2} \ln((2\pi e)^{n_x} \det \boldsymbol{\Sigma}). \tag{3}$$

  - The larger $\det \boldsymbol{\Sigma}$ is, meaning the Gaussian is "short and fat", the larger $H_p(\mathbf{x})$ is, representing more average surprise in the distribution $p(\mathbf{x})$.
  - The smaller $\det \boldsymbol{\Sigma}$ is, meaning the Gaussian is "tall and skinny" like a delta function, the smaller $H_p(\mathbf{x})$ is, representing less average surprise in the distribution $p(\mathbf{x})$.

# How “Close” are Two Densities?

- A *divergence*, $\mathrm{D}(q||p)$, is a measure of the distance between two pdfs, such as $q(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$.
- A divergence is not a metric because a divergence
  - does not have to be symmetric (i.e., $\mathrm{D}(q||p) \neq \mathrm{D}(p||q)$),
  - does not have to to satisfy the triangle inequality.
- A very popular divergence is the *Kullback-Leibler* (KL) divergence,

$$\mathrm{D}_{\mathrm{KL}}(q||p) = -\int_{-\infty}^{\infty} q(\mathbf{x}) \ln\left(\frac{p(\mathbf{y}|\mathbf{x})}{q(\mathbf{x})}\right) \mathrm{d}\mathbf{x} = -\mathbb{E}_q\left[\ln\left(\frac{p(\mathbf{y}|\mathbf{x})}{q(\mathbf{x})}\right)\right]. \quad (4)$$

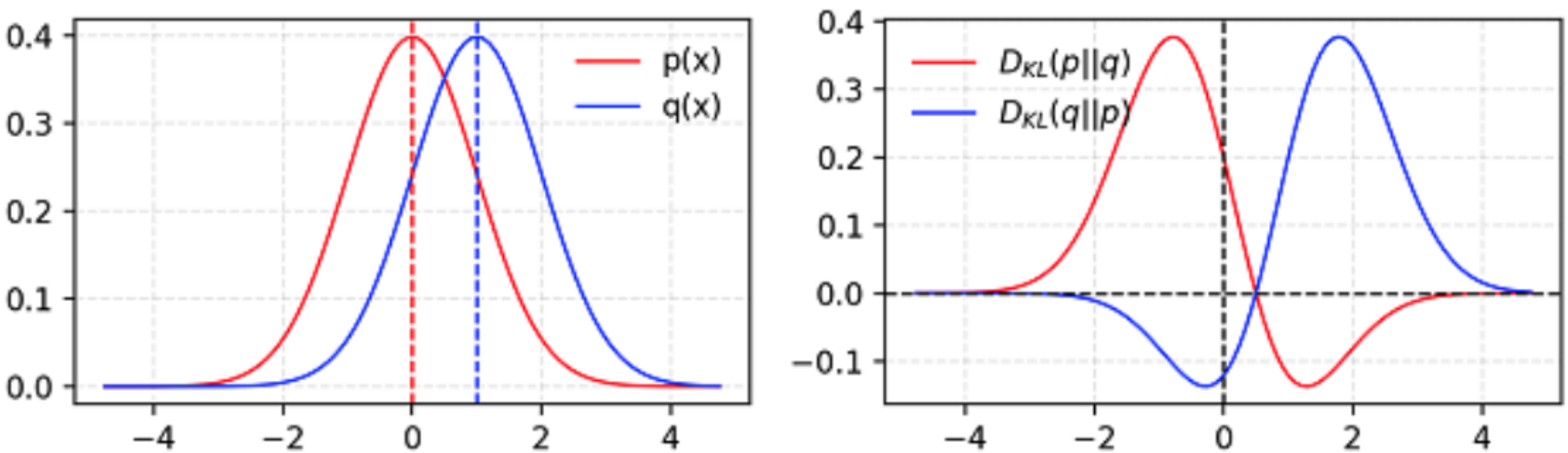


Figure: Left: two pdfs. Right: two KL divergences. (github.com/apachecn/HackCV-Translate)

- $\mathsf{D}_{\mathsf{KL}}(q||p) = 0$ means $q(\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$.

- The KL divergence can be written as

$$\begin{aligned}\mathsf{D}_{\mathsf{KL}}(q||p) &= -\int_{-\infty}^{\infty} q(\mathbf{x}) \ln\left(\frac{p(\mathbf{y}|\mathbf{x})}{q(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \\ &= -\mathbb{E}_q\left[\ln p(\mathbf{y}|\mathbf{x})\right] + \mathbb{E}_q\left[\ln q(\mathbf{x})\right] \\ &= -\mathbb{E}_q\left[\ln p(\mathbf{y}|\mathbf{x})\right] - H_q(\mathbf{x}),\end{aligned}$$

  where the definition of differential entropy in (2) has been used.
- The term $\ln p(\mathbf{y})$ is not a function of $\mathbf{x}$.

## Problem Set-up

- There are two pieces of information available:
  - an *a priori* density (a prior), $p_{\mathrm{prior}}(\mathbf{x})$, and
  - measurement data $\mathbf{y}$.
- The task at hand is to find the *a posteriori* density (the posterior),

$$p_{\mathrm{post}}(\mathbf{x}|\mathbf{y}),$$

  given the prior and the measurement data.

## Possible Solutions

- Find $p_{\text{post}}(\mathbf{x}|\mathbf{y})$ that is "close to", in some sense, the prior $p_{\text{prior}}(\mathbf{x})$.
  - A good idea.
  - Use KL divergence to minimize the difference between $p_{\text{post}}(\mathbf{x}|\mathbf{y})$ and $p_{\text{prior}}(\mathbf{x})$.
  - The problem is that the measurement data, $\mathbf{y}$, isn't use.
  - Also, in the absence of an additional cost to minimize, $p_{\text{post}}(\mathbf{x}|\mathbf{y}) = p_{\text{prior}}(\mathbf{x})$ is the optimal solution.
- Find $\mathbf{x}$, and then $p_{\text{post}}(\mathbf{x}|\mathbf{y})$ if possible/tractable, that best fits the data.
  - Also a good idea.
  - Leads to maximum likelihood estimation (MLE).
  - The problem is that the prior, $p_{\text{prior}}(\mathbf{x})$, isn't use.

## Solution

- Do both: find $p_{\text{post}}(\mathbf{x}|\mathbf{y})$ that is "close to", in some sense, the prior $p_{\text{prior}}(\mathbf{x})$ *and* best fits the measurment data.

# Fitting to the Prior

- To find $q(\mathbf{x}|\mathbf{y}) = p_{\text{post}}(\mathbf{x}|\mathbf{y})$ that is "close to" the prior $p_{\text{prior}}(\mathbf{x})$, use KL divergence.
- To this end, minimize

$$
\begin{aligned}
J(q(\mathbf{x}|\mathbf{y})) &= \mathsf{D}_{\mathsf{KL}}(q||p_{\text{prior}}) \\
&= -\int_{-\infty}^{\infty} q(\mathbf{x}) \ln\left(\frac{p_{\text{prior}}(\mathbf{x})}{q(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \qquad (5) \\
&= -\int_{-\infty}^{\infty} q(\mathbf{x}) \left(\ln p_{\text{prior}}(\mathbf{x}) - \ln q(\mathbf{x})\right) \mathrm{d}\mathbf{x} \qquad (6)
\end{aligned}
$$

- In the absence of an additional cost to minimize, $q(\mathbf{x}|\mathbf{y}) = p_{\text{post}}(\mathbf{x}|\mathbf{y}) = p_{\text{prior}}(\mathbf{x})$ is the optimal solution.

# Maximum Likelihood Estimation (MLE)

- Consider a likelihood function,

$$L(\mathbf{x}) = p_{\text{like}}(\mathbf{y}|\mathbf{x}). \tag{7}$$

  - For example, consider the measurement model

$$\mathbf{y} = \mathbf{g}(\mathbf{x}) + \mathbf{n}, \tag{8}$$

    where $\mathbf{y} \in \mathbb{R}^{n_y}$ is the measurement data, $\mathbf{g} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ is the measurement model, and $\mathbf{x} : \mathbb{R}^{n_x}$ is the state to be inferred from the data, and $\mathbf{n} \in \mathbb{R}^{n_y}$ is noise with pdf $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$.
  - In this case, the likelihood function is

$$L(\mathbf{x}) = p_{\text{like}}(\mathbf{y}|\mathbf{x}) = \eta \exp\left(-\tfrac{1}{2}(\mathbf{y} - \mathbf{g}(\mathbf{x}))^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{g}(\mathbf{x}))\right). \tag{9}$$

- To find the most likely $\mathbf{x}$, maximize $L(\mathbf{x}) = p_{\text{like}}(\mathbf{y}|\mathbf{x})$ given measurement data $\mathbf{y}$.
  - In practise, rather than maximizing $L(\mathbf{x})$, the negative log likelihood,

$$-\ln L(\mathbf{x}) = -\ln p_{\text{like}}(\mathbf{y}|\mathbf{x}) \tag{10}$$

    is minimized.
  - When (9) is the likelihood function, the negative log likelihood is

$$-\ln L(\mathbf{x}) = -\ln p_{\text{like}}(\mathbf{y}|\mathbf{x}) = \tfrac{1}{2}(\mathbf{y} - \mathbf{g}(\mathbf{x}))^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{g}(\mathbf{x})) - \ln \eta. \tag{11}$$

- Once $\mathbf{x}$ is found, compute $p_{\text{post}}(\mathbf{x}|\mathbf{y})$ (e.g., the covariance in the Gaussian case) if possible/tractable.

# A Variational Approach

- Consider the cost function

$$J(q(\mathbf{x}|\mathbf{y})) = \alpha_1 \underbrace{\mathsf{D}_{\mathsf{KL}}(q||p_{\text{prior}})}_{J_{\text{I}}(q(\mathbf{x}|\mathbf{y}))} + \alpha_2 \underbrace{\int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \left(-\ln p_{\text{like}}(\mathbf{y}|\mathbf{x})\right) \, \mathrm{d}\mathbf{x}}_{J_{\text{II}}(q(\mathbf{x}|\mathbf{y}))}, \tag{12}$$

  where $q(\mathbf{x}|\mathbf{y}) = p_{\text{post}}(\mathbf{x}|\mathbf{y})$ and $\alpha_1, \alpha_2 > 0$ are weights.
- Note that $q(\mathbf{x}|\mathbf{y})$ is subject to two constraints:
  - $q(\mathbf{x}|\mathbf{y}) \geq 0$, and

$$\int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x} - 1 = 0. \tag{13}$$

- The first constraint will be ignored (but, after the fact, can be verified).
- The second constraint can be enforced using the method of Lagrange multipliers.
- Augmenting (12) with a Lagrange multiplier times the constraint (13) results in

$$I(q(\mathbf{x}|\mathbf{y}), \lambda) = \alpha_1 \underbrace{\mathsf{D}_{\mathsf{KL}}(q||p_{\text{prior}})}_{J_{\text{I}}(q(\mathbf{x}|\mathbf{y}))} + \alpha_2 \underbrace{\int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \left(-\ln p_{\text{like}}(\mathbf{y}|\mathbf{x})\right) \, \mathrm{d}\mathbf{x}}_{J_{\text{II}}(q(\mathbf{x}|\mathbf{y}))}$$

$$+ \lambda \left( \int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x} - 1 \right). \tag{14}$$

- Minimizing (12) subject to the constraint (13) is a *variational problem* because

$$J : \mathcal{D} \to \mathbb{R},$$

  where the domain of the cost function $\mathcal{D}$ is the set of all pdfs that satisfy (13) and are positive
- Put “in” a pdf and get “out” a real number.

# Cost Function Interpretations

- Minimizing (extremizing) the cost function $J_{\mathrm{I}}(q(\mathbf{x}|\mathbf{y}))$ in (14) results in a $q(\mathbf{x}|\mathbf{y})$ that is "close to" $p_{\mathrm{prior}}(\mathbf{x})$ in the sense that KL divergence is minimized.
- Minimizing the cost function $J_{\mathrm{II}}(q(\mathbf{x}|\mathbf{y}))$ in (14) results in a $q(\mathbf{x}|\mathbf{y})$ with the following properties.
  - The density $q(\mathbf{x}|\mathbf{y})$ must be close to zero when $-\ln p_{\mathrm{like}}(\mathbf{y}|\mathbf{x})$ is large, but is permitted to be large when $-\ln p_{\mathrm{like}}(\mathbf{y}|\mathbf{x})$ is small.
  - To make this more concrete, consider again the likelihood given in (9), with negative log likelihood given by

$$-\ln L(\mathbf{x}) = -\ln p_{\mathrm{like}}(\mathbf{y}|\mathbf{x}) = \tfrac{1}{2}(\mathbf{y}-\mathbf{g}(\mathbf{x}))^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{g}(\mathbf{x})) - \ln \eta. \qquad (11)$$

    - When $-\ln p_{\mathrm{like}}(\mathbf{y}|\mathbf{x})$ is large, meaning $\mathbf{x}$ values don't result in accurate $\mathbf{y}$ predictions, $q(\mathbf{x}|\mathbf{y})$ must be small.
    - When $-\ln p_{\mathrm{like}}(\mathbf{y}|\mathbf{x})$ is small, meaning $\mathbf{x}$ values result in accurate $\mathbf{y}$ predictions, $q(\mathbf{x}|\mathbf{y})$ is permitted to be large.
- $q(\mathbf{x}|\mathbf{y})$ cannot be small everywhere because $J_{\mathrm{I}}(q(\mathbf{x}|\mathbf{y}))$ would not be minimized, and the constraint (13) would be violated.
- Minimizing the cost function $I(q(\mathbf{x}|\mathbf{y}), \lambda)$ in (14) trades off finding a posterior density that is not too different than the prior (the $J_{\mathrm{I}}(q(\mathbf{x}|\mathbf{y}))$ term), but that "matches the data" (the $J_{\mathrm{II}}(q(\mathbf{x}|\mathbf{y}))$ term), subject to the constraint (13).

# An Alternative Form of the Cost Function

- By adding and subtracting $H_q(\mathbf{x}) = -\int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \ln q(\mathbf{x}|\mathbf{y})\, \mathrm{d}\mathbf{x}$ to $J_{\mathrm{II}}(q(\mathbf{x}|\mathbf{y}))$, $J_{\mathrm{II}}(q(\mathbf{x}|\mathbf{y}))$ can also be written as

$$
\begin{aligned}
J_{\mathrm{II}}(q(\mathbf{x}|\mathbf{y})) &= \int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \left(-\ln p_{\mathrm{like}}(\mathbf{y}|\mathbf{x})\right) \mathrm{d}\mathbf{x} + H_q(\mathbf{x}) - H_q(\mathbf{x}) \qquad (15)\\
&= -\int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \ln p_{\mathrm{like}}(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x} + \int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \ln q(\mathbf{x}|\mathbf{y})\, \mathrm{d}\mathbf{x} + H_q(\mathbf{x})\\
&= -\int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \ln \left( \frac{p_{\mathrm{like}}(\mathbf{y}|\mathbf{x})}{q(\mathbf{x}|\mathbf{y})} \right) \mathrm{d}\mathbf{x} + H_q(\mathbf{x})\\
&= \mathsf{D}_{\mathsf{KL}}(q||p_{\mathrm{like}}) + H_q(\mathbf{x}). \qquad (16)
\end{aligned}
$$

- Thus, the cost function (14) can alternatively be written as

$$I(q(\mathbf{x}|\mathbf{y})) = \alpha_1 \underbrace{\mathsf{D}_{\mathsf{KL}}(q||p_{\text{prior}})}_{J_{\text{I}}(q(\mathbf{x}|\mathbf{y}))} + \alpha_2 \Big( \underbrace{\mathsf{D}_{\mathsf{KL}}(q||p_{\text{like}})}_{J^D_{\text{II}}(q(\mathbf{x}|\mathbf{y}))} + \underbrace{H_q(\mathbf{x})}_{J^S_{\text{II}}(q(\mathbf{x}|\mathbf{y}))} \Big) + \lambda \left( \int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x} - 1 \right) \tag{17}$$

- Minimizing the alternative form of $I(q(\mathbf{x}|\mathbf{y}))$ in (17) can be interpreted as follows.
  - Minimization of $J_{\text{I}}(q(\mathbf{x}|\mathbf{y})$ forces $q(\mathbf{x}|\mathbf{y}))$ to be "similar to" the $p_{\text{prior}}(\mathbf{x})$ in the sense that KL divergence is minimized.
  - Minimization of $J^D_{\text{II}}(q(\mathbf{x}|\mathbf{y})$ forces $q(\mathbf{x}|\mathbf{y}))$ to be "similar to" the $p_{\text{like}}(\mathbf{y}|\mathbf{x})$ in the sense that KL divergence is minimized.
  - Minimization of $J^S_{\text{II}}(q(\mathbf{x}|\mathbf{y})) = H_q(\mathbf{x})$ forces $q(\mathbf{x}|\mathbf{y})$ to have minimal average surprise (i.e., $q(\mathbf{x}|\mathbf{y})$ shouldn't be too "short and fat", but should be more "tall and skinny" like a delta function).

# The Details

- Compute the derivative of $I(q(\mathbf{x}|\mathbf{y}), \lambda)$ with respect to $\lambda$ and set the result to zero. To this end,

$$\frac{\mathrm{d}I(q(\mathbf{x}|\mathbf{y}), \lambda)}{\mathrm{d}\lambda} = \int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y})\,\mathrm{d}\mathbf{x} - 1 = 0. \tag{18}$$

- Next, compute the first variation of $I(q(\mathbf{x}|\mathbf{y}), \lambda)$ and set the result to zero.
- Consider the first variation of $J_{\mathrm{I}}(q(\mathbf{x}|\mathbf{y}))$, that being

$$\begin{aligned}
\delta J_{\mathrm{I}}(q(\mathbf{x}|\mathbf{y})) &= -\int_{-\infty}^{\infty} \delta q(\mathbf{x}|\mathbf{y}) \left(\ln p_{\mathrm{prior}}(\mathbf{x}) - \ln q(\mathbf{x}|\mathbf{y})\right)\,\mathrm{d}\mathbf{x} \\
&\quad - \int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \left(-\frac{1}{q(\mathbf{x}|\mathbf{y})}\delta q(\mathbf{x}|\mathbf{y})\right)\,\mathrm{d}\mathbf{x} \\
&= \int_{-\infty}^{\infty} \left(1 - \left(\ln p_{\mathrm{prior}}(\mathbf{x}) - \ln q(\mathbf{x}|\mathbf{y})\right)\right) \delta q(\mathbf{x}|\mathbf{y})\,\mathrm{d}\mathbf{x}.
\end{aligned} \tag{19}$$

- Next, consider the first variation of $J_{\mathrm{II}}(q(\mathbf{x}|\mathbf{y}))$, that being

$$\delta J_{\mathrm{II}}(q(\mathbf{x}|\mathbf{y})) = -\int_{-\infty}^{\infty} \ln p_{\mathrm{like}}(\mathbf{y}|\mathbf{x}) \delta q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x}. \tag{20}$$

- Finally, the first variation of the constraint is

$$\delta \left[ \lambda \left( \int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x} - 0 \right) \right] = \lambda \int_{-\infty}^{\infty} \delta q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x}. \tag{21}$$

- Thus, the first variation of $I(q(\mathbf{x}|\mathbf{y}), \lambda)$ is

$$\begin{aligned}
&\delta I(q(\mathbf{x}|\mathbf{y}), \lambda) \\
&= \alpha_1 \int_{-\infty}^{\infty} (1 - (\ln p_{\text{prior}}(\mathbf{x}) - \ln q(\mathbf{x}|\mathbf{y}))) \, \delta q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x} \\
&\quad - \alpha_2 \int_{-\infty}^{\infty} \ln p_{\text{like}}(\mathbf{y}|\mathbf{x}) \delta q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x} + \lambda \int_{-\infty}^{\infty} \delta q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x} \\
&= \int_{-\infty}^{\infty} \big(\alpha_1 (1 - (\ln p_{\text{prior}}(\mathbf{x}) - \ln q(\mathbf{x}|\mathbf{y}))) - \alpha_2 \ln p_{\text{like}}(\mathbf{y}|\mathbf{x}) + \lambda\big) \delta q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x}.
\end{aligned} \tag{22}$$

- Setting $\delta I(q(\mathbf{x}|\mathbf{y}), \lambda)$ equal to zero, and noting that $\delta q(\mathbf{x}|\mathbf{y})$ is an arbitrary variation, results in the requirement that

$$\alpha_1 (1 - (\ln p_{\text{prior}}(\mathbf{x}) - \ln q(\mathbf{x}|\mathbf{y}))) - \alpha_2 \ln p_{\text{like}}(\mathbf{y}|\mathbf{x}) + \lambda = 0, \tag{23}$$

$$\alpha_1 - \alpha_1 \ln p_{\text{prior}}(\mathbf{x}) + \alpha_1 \ln q(\mathbf{x}|\mathbf{y}) - \alpha_2 \ln p_{\text{like}}(\mathbf{y}|\mathbf{x}) + \lambda = 0. \tag{24}$$

- Rearranging (24) and solving for $q(\mathbf{x}|\mathbf{y})$,

$$\ln\left(\frac{p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\alpha_2}p_{\text{prior}}(\mathbf{x})^{\alpha_1}}{q(\mathbf{x}|\mathbf{y})^{\alpha_1}}\right) = (\alpha_1+\lambda), \tag{25}$$

$$\frac{p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\alpha_2}p_{\text{prior}}(\mathbf{x})^{\alpha_1}}{q(\mathbf{x}|\mathbf{y})^{\alpha_1}} = \exp(\alpha_1+\lambda), \tag{26}$$

$$q(\mathbf{x}|\mathbf{y})^{\alpha_1} = \exp(-\alpha_1-\lambda)p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\alpha_2}p_{\text{prior}}(\mathbf{x})^{\alpha_1}, \tag{27}$$

$$q(\mathbf{x}|\mathbf{y}) = \exp\left(-\frac{\alpha_1+\lambda}{\alpha_1}\right)p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\frac{\alpha_2}{\alpha_1}}p_{\text{prior}}(\mathbf{x}). \tag{28}$$

- The expression for $q(\mathbf{x}|\mathbf{y})$ given in (28) is not normalized. To find the normalization constant, use the constraint (13), the axiom of total probability.
- To this end,

$$\int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y})\,\mathrm{d}\mathbf{x} = 1, \tag{29}$$

$$\exp\left(-\frac{\alpha_1+\lambda}{\alpha_1}\right)\int_{-\infty}^{\infty} p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\frac{\alpha_2}{\alpha_1}}p_{\text{prior}}(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1, \tag{30}$$

$$\exp\left(-\frac{\alpha_1+\lambda}{\alpha_1}\right) = \left(\int_{-\infty}^{\infty} p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\frac{\alpha_2}{\alpha_1}}p_{\text{prior}}(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)^{-1}. \tag{31}$$

- Therefore,

$$q(\mathbf{x}|\mathbf{y}) = \left( \int_{-\infty}^{\infty} p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\frac{\alpha_2}{\alpha_1}} p_{\text{prior}}(\mathbf{x})\, \mathrm{d}\mathbf{x} \right)^{-1} p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\frac{\alpha_2}{\alpha_1}} p_{\text{prior}}(\mathbf{x}),$$

$$p_{\text{post}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\frac{\alpha_2}{\alpha_1}} p_{\text{prior}}(\mathbf{x})}{\int_{-\infty}^{\infty} p_{\text{like}}(\mathbf{y}|\mathbf{x})^{\frac{\alpha_2}{\alpha_1}} p_{\text{prior}}(\mathbf{x})\, \mathrm{d}\mathbf{x}}, \tag{32}$$

  where $q(\mathbf{x}|\mathbf{y}) = p_{\text{post}}(\mathbf{x}|\mathbf{y})$.
- It can be verified that $p_{\text{post}}(\mathbf{x}|\mathbf{y}) \geq 0$.
- Notice that when $\alpha_1 = \alpha_2 = 1$,

$$p_{\text{post}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\text{like}}(\mathbf{y}|\mathbf{x}) p_{\text{prior}}(\mathbf{x})}{\int_{-\infty}^{\infty} p_{\text{like}}(\mathbf{y}|\mathbf{x}) p_{\text{prior}}(\mathbf{x})\, \mathrm{d}\mathbf{x}}, \tag{33}$$

  and Bayes Rule is recovered!

# Bayes Rule is Optimal

Bayes Rule is optimal in the sense that the cost function

$$J(q(\mathbf{x}|\mathbf{y})) = \mathsf{D}_{\mathsf{KL}}(q||p_{\text{prior}}) + \int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \left(-\ln p_{\text{like}}(\mathbf{y}|\mathbf{x})\right) \, \mathrm{d}\mathbf{x} \tag{34}$$

$$= \mathsf{D}_{\mathsf{KL}}(q||p_{\text{prior}}) + \mathsf{D}_{\mathsf{KL}}(q||p_{\text{like}}) + H_q(\mathbf{x}) \tag{35}$$

subject to the constraint

$$\int_{-\infty}^{\infty} q(\mathbf{x}|\mathbf{y}) \, \mathrm{d}\mathbf{x} = 1 \tag{13}$$

is minimized.

# References

These slides are based on [1–4].


[1] T. Bui-Thanh and O. Ghattas, "Bayes is Optimal," *The Institute for Computational Engineering and Sciences, The University of Texas at Austin*, 2015, ICES Report 15-04.

[2] K. Tuggle and R. Zanetti, "Automated Splitting Gaussian Mixture Nonlinear Measurement Update," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 41, no. 3, pp. 725–734, 2018.

[3] T. D. Barfoot, *State Estimation for Robotics*. New York, New York: Cambridge University Press, 2017.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.