Variational Bayes: Kullback-Leibler Divergence

Flavio Mejia Morelli

January 16, 2020

Contents

Contents

Terminology: Bayes' theorem

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta) \cdot p(\theta)}{\int_{\Theta} p(y|\theta) \cdot p(\theta) d\theta}$$

Terminology: Bayes' theorem

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta) \cdot p(\theta)}{\int_\Theta p(y|\theta) \cdot p(\theta) d\theta}$$

- $\theta \in \Theta$: a parameter or vector of parameters (e.g. the probability in a binomial distribution) in the parameter space Θ
- y: data
- $p(\theta|y)$: posterior
- $p(y|\theta)$: likelihood function that captures how we are modeling our data stochastically (e.g. y is a binomial variable)
- $p(\theta)$: prior knowledge about the parameters (e.g. a probability can only be between 0 and 1)

Terminology: Bayes' theorem

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta) \cdot p(\theta)}{\int_\Theta p(y|\theta) \cdot p(\theta) d\theta}$$

- $p(y) = \int_\Theta p(y|\theta) \cdot p(\theta) d\theta$: the marginal likelihood or evidence
- $p(y)$ is a constant that normalizes the expression so that the posterior integrates to one.
- Problem: $p(y)$ is **intractable** even for very simple models, and this makes it very hard to calculate the posterior
- Bayesian computation methods get around this intractability by different means

Motivation: Why Variational Bayes?

- One of the main **alternatives** to MCMC
- Estimate different kinds of models much **faster** than MCMC (usually at the expense of precision)
- However, variational inference is **less well understood** than MCMC

MCMC and variational inference

- *Markov Chain Monte Carlo (MCMC)* is one of the most common methods of estimating parameters in a Bayesian model
- Different approaches: Gibbs Sampler, Metropolis-Hastings, Hamiltonian Monte Carlo, NUTS, Sequential MCMC...
- **Pro**: more Monte Carlo samples lead to a more accurate estimate
- **Con**: slow and curse of dimensionality

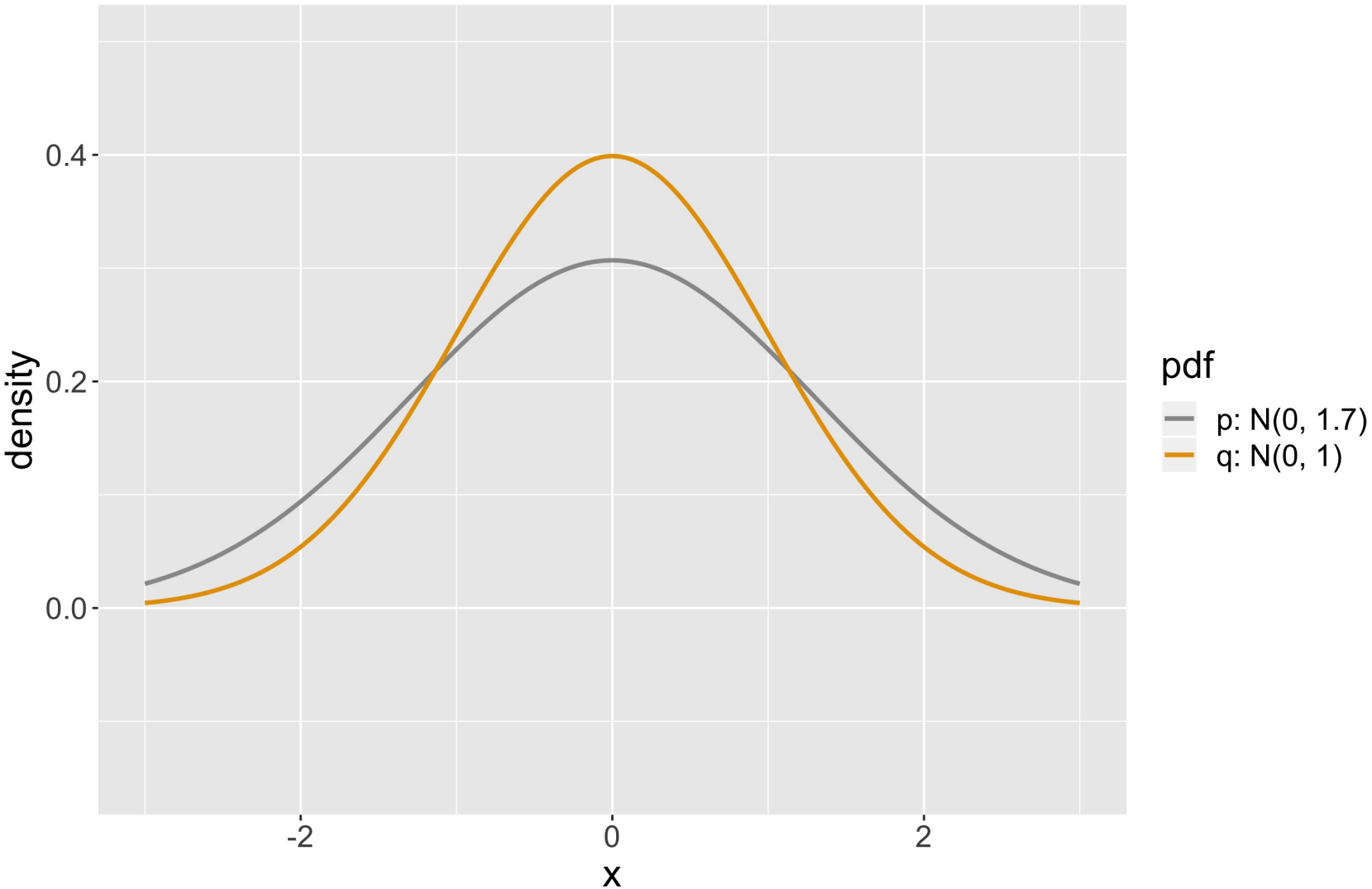
MCMC and variational inference

- Variational inference turn the estimation of the posterior into an **optimization** problem (i.e. maximize or minimize)
- Main idea: find **another probability** function that is easier to work with than the posterior
- **Minimize** the difference between the new probability function an the posterior

How do we measure the difference between probability functions?

Contents

Difference between density functions



Difference between density functions

- Let $p(x)$ and $q(x)$ be two probability densities
- A naive approach for a given x: $\dfrac{q(x)}{p(x)}$

Difference between density functions

- However, if we flip the terms we see that the absolute value of this measure also changes: e.g. $\frac{0.3}{0.1} = 3$, but $\frac{0.1}{0.3} = \frac{1}{3}$
- If we take the **logarithm**, the absolute value stays constant after flipping the probabilities, only the sign changes:

$$log(\frac{q(x)}{p(x)}) = -log(\frac{p(x)}{q(x)})$$

Difference between density functions

- Assume that we are interested mainly on $q(x)$
- Use $q(x)$ as a weight for the difference measure: $q(x)log(\frac{q(x)}{p(x)})$
- When $q(x)$ is low, the difference measure $log(\frac{q(x)}{p(x)})$ does not matter as much, as when $q(x)$ is high!

Difference between density functions

- As a final step, we integrate over all the possible values of x (or sum if the density is not continuous)

$$D_{KL}(q \parallel p) = \int_{\mathcal{X}} q(x)\ log(\frac{q(x)}{p(x)})dx$$

- D_{KL} is called the **Kullback-Leibler divergence** of p from q

Kullback-Leibler Divergence

$$D_{KL}(q \parallel p) = \int_{\mathcal{X}} q(x)\ log(\frac{q(x)}{p(x)})dx$$

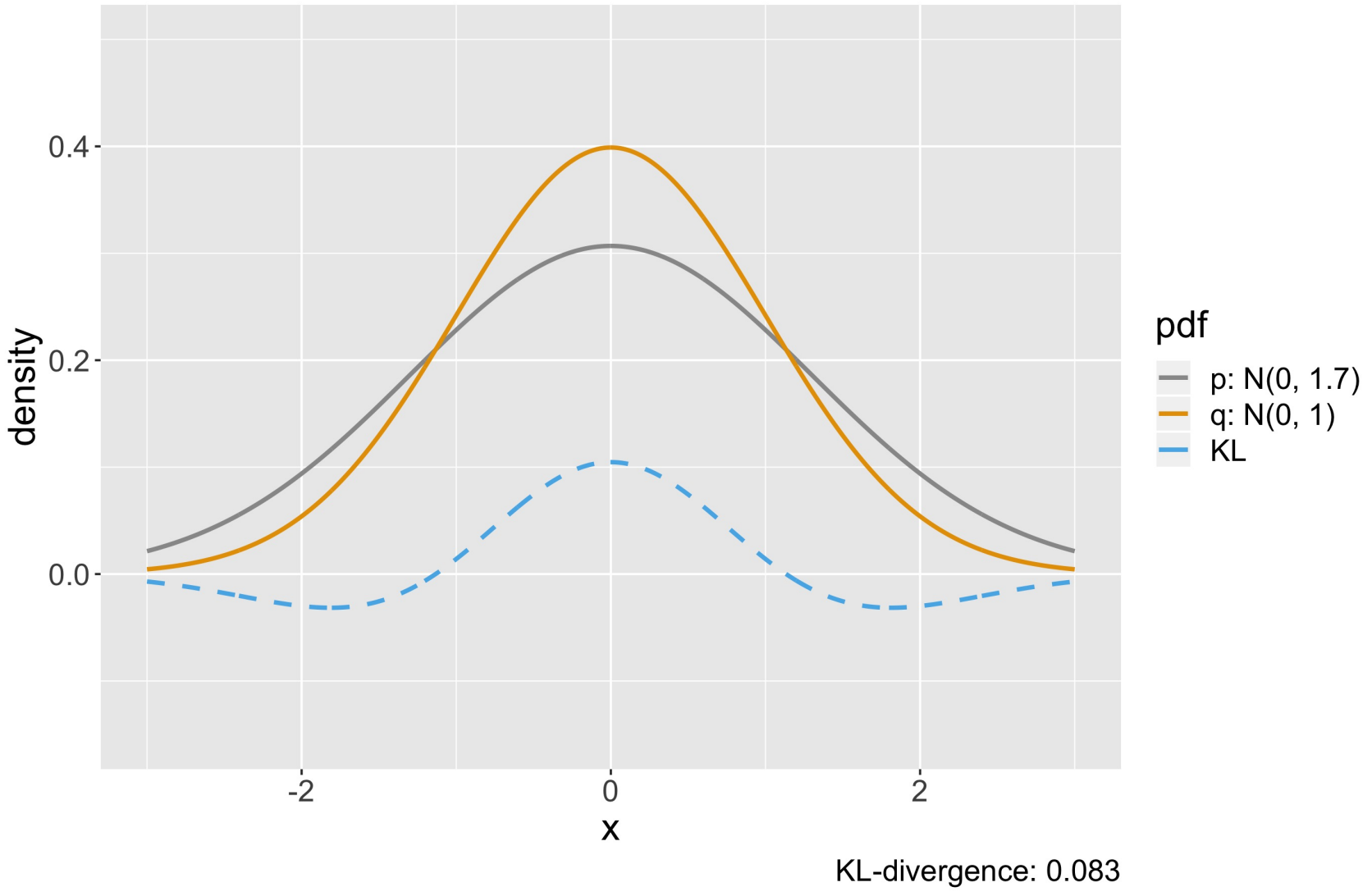
- Common measure for the divergence between two probability densities
- The Kullback-Leibler divergence is **not symmetric**, and thus is cannot be called a "distance": $D_{KL}(q \parallel p) \neq D_{KL}(p \parallel q)$

Kullback-Leibler Divergence

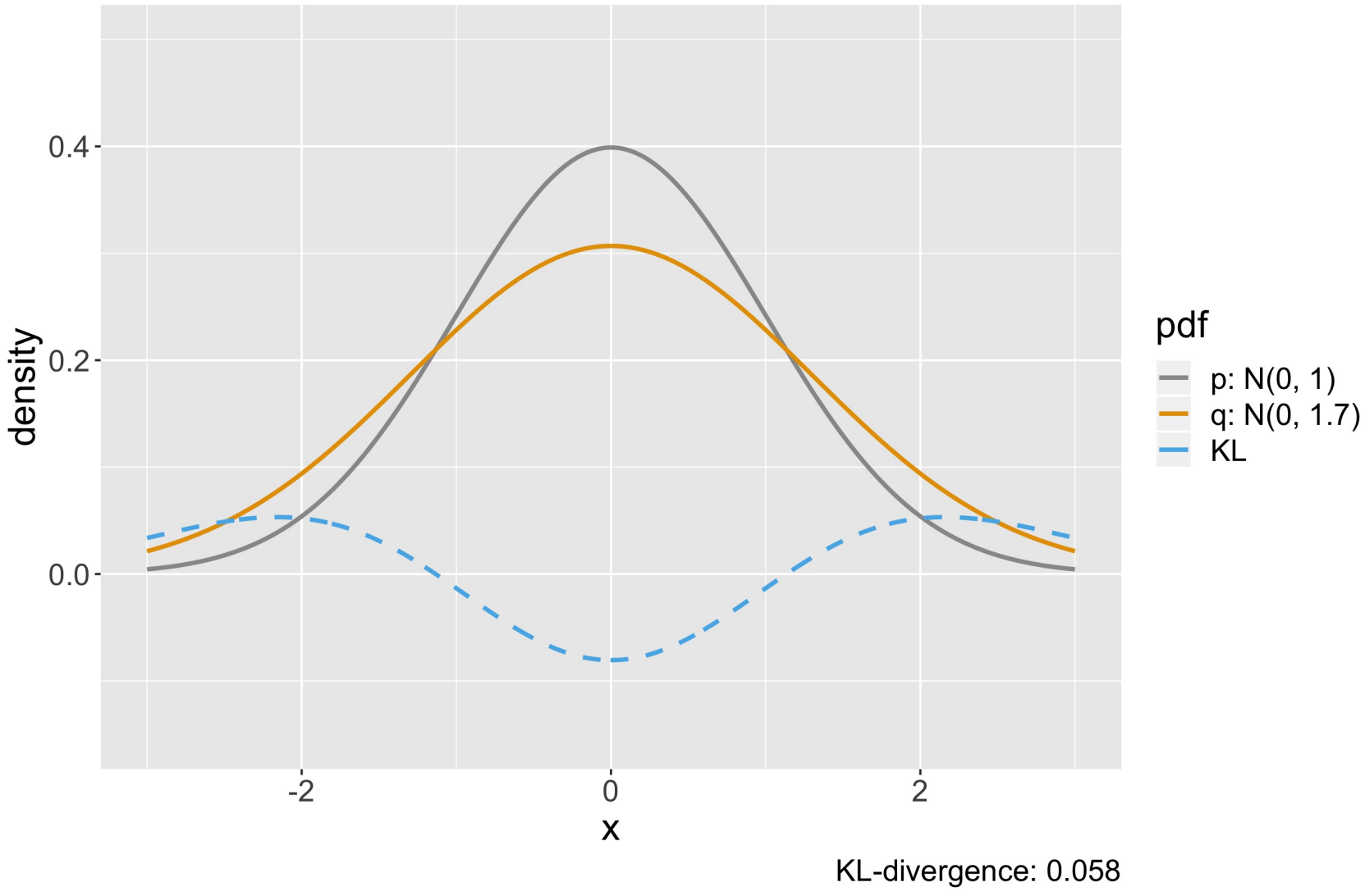
$$\int_{\mathcal{X}} q(x)\ log(\frac{q(x)}{p(x)})dx = -\int_{\mathcal{X}} q(x)\ log(\frac{p(x)}{q(x)})dx$$

- Note that if we flip the densities inside the logarithm, the divergence does not change
- $D_{KL} \geq 0$ for any given probabilities (by Gibbs' inequality)
- $D_{KL} = 0 \implies$ both probabilities are the same at almost each point

Examples of KL-Divergence



Examples of KL-Divergence: interchange p and q



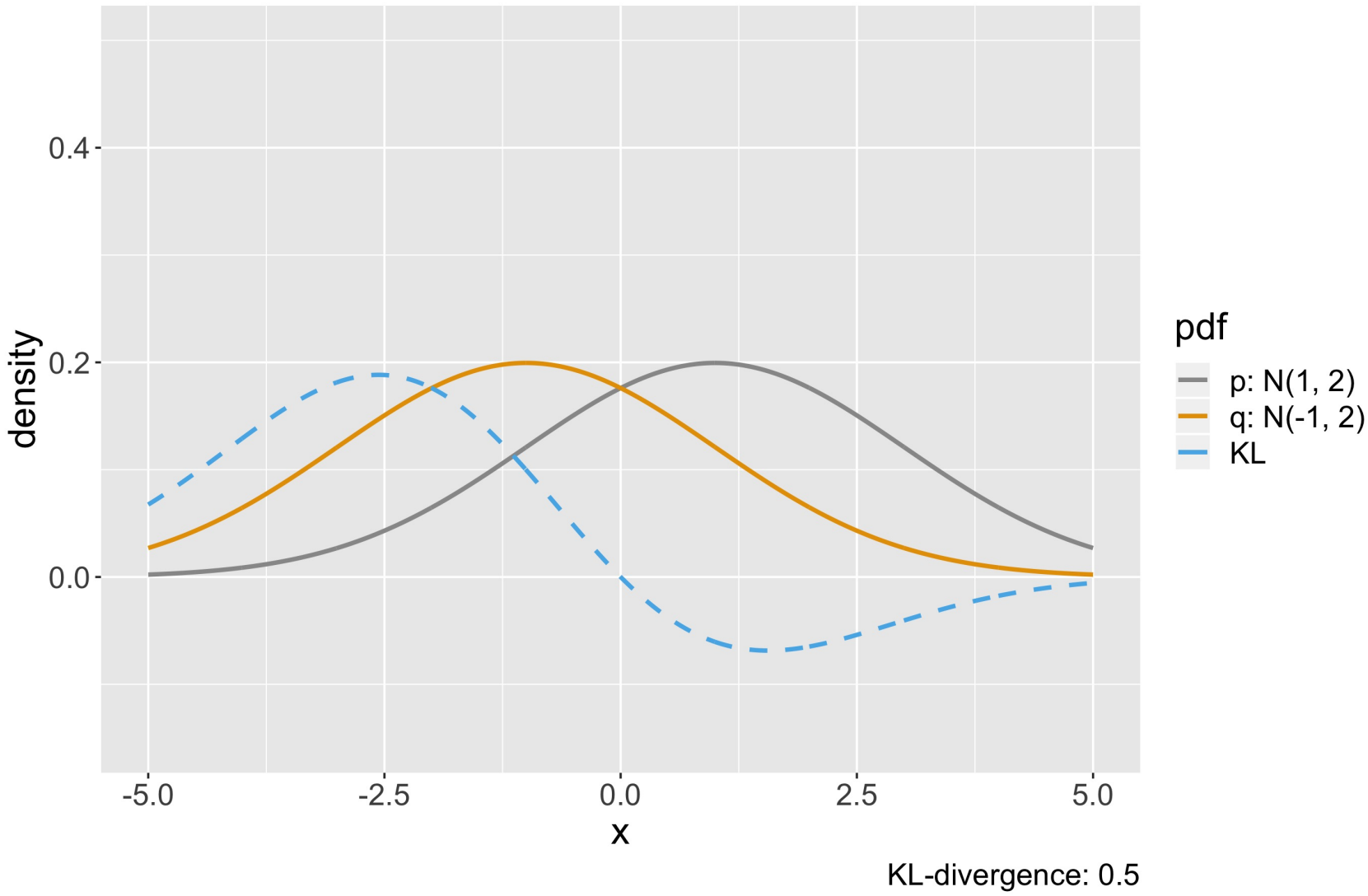
Examples of KL-Divergence: Gaussian

- $q(x) = N(\mu_q, \sigma_q^2)$ and $p(x) = N(\mu_p, \sigma_p^2)$
- $D_{KL}(q||p) = log\left(\frac{\sigma_p}{\sigma_q}\right) + \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} - \frac{1}{2}$
  (Cross-Validated, 2011)
- $(\mu_q - \mu_p)^2$ increases the divergence, as the distributions move away from each other
- We focus on how changes in variance, affect the KL-divergence

Examples of KL-Divergence: Gaussian high variance

- $\mu_q = -1$ and $\mu_p = 1$
- $\sigma_q^2 = \sigma_p^2 = 4$
- $D_{KL}(q||p) = log\left(\frac{2}{2}\right) + \frac{4+4}{2 \cdot 4} - \frac{1}{2} = \frac{1}{2}$

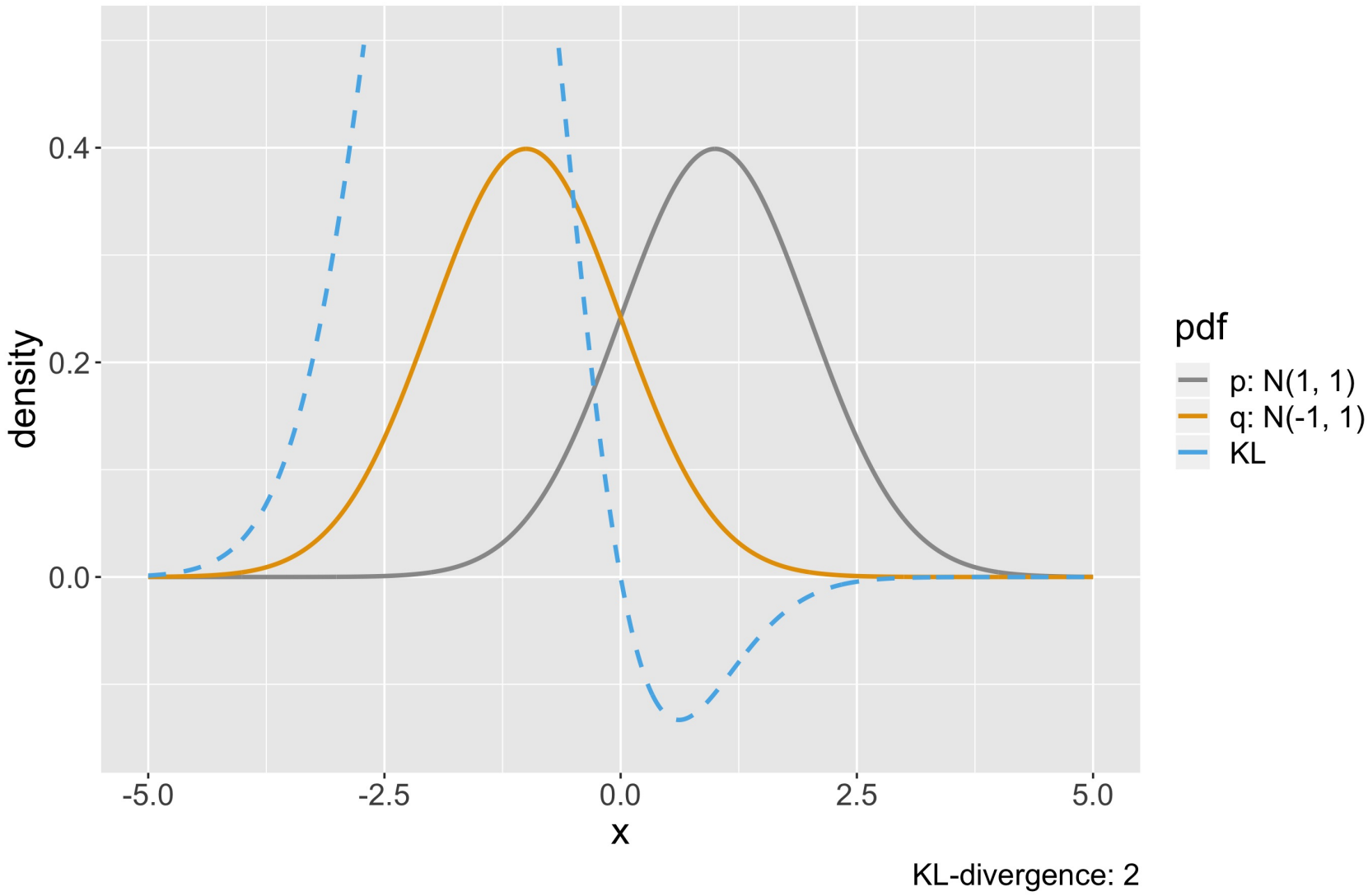
Examples of KL-Divergence: Gaussian high variance



Examples of KL-Divergence: Gaussian low variance

- Keep means constant, reduce variance
- $\mu_q = -1$ and $\mu_p = 1$
- $\sigma_q^2 = \sigma_p^2 = 1$
- $D_{KL}(q||p) = log\left(\frac{1}{1}\right) + \frac{1+4}{2 \cdot 1} - \frac{1}{2} = 2$
- With **lower variance**, the probability masses do not overlap as much, thus **increasing the divergence**

Examples of KL-Divergence: Gaussian low variance



Variational approximation of the posterior

- Find a $q^*(\theta)$ which minimizes the KL divergence between $q(\theta)$ and the posterior $p(\theta|y)$:

$$D_{KL}(q \parallel p) = \int_{\Theta} q(\theta)\ log(\frac{q(\theta)}{p(\theta|y)})$$

- However, in order to calculate the KL divergence we would have to know the posterior $p(\theta|y)$ which is **intractable**.
- We are back to square one...

Side note: What does "variational" mean?

- The term "variational" comes from **variational calculus**
- One of the main topics of calculus is optimization
- Optimization is usually done with respect to a **variable**.
- A common problem in economic is finding an optimum quantity Q^* that maximizes profit given a demand and a cost function
- In contrast, variational calculus optimizes with respect to a **function**. In our case, we are trying to find a $q^*(\theta)$ which minimizes the Kullback-Leibler divergence

Side note: KL-divergence as expected value

- Some papers and textbooks write the KL divergence as an expected value with respect to q
- Because we are weighting by $q(\theta)$, we can express it as an expected value

$$D_{KL}(q \parallel p) = \int_{\Theta} q(\theta) \; log \left(\frac{q(\theta)}{p(\theta|y)} \right) = \mathbb{E}_q \left[log \left(\frac{q(\theta)}{p(\theta|y)} \right) \right]$$

Side note: KL-divergence in information theory

- The KL-divergence can also be derived from an **information theory** perspective
- KL-divergence as the difference between the **cross-entropy** and the **entropy** of a distribution
- I highly encourage you to look at this interpretation of the KL-divergence (Shibuya, 2018)

Contents

Variational approximation of the posterior

- Find a $q^*(\theta)$ which minimizes the KL divergence between $q(\theta)$ and the posterior $p(\theta|y)$:

$$D_{KL}(q \parallel p) = \int_{\Theta} q(\theta) \; log(\frac{q(\theta)}{p(\theta|y)})$$

- However, the posterior is **intractable**

So, what now?

Find an alternative way to optimize the divergence!

ELBO: evidence lower bound

- It can be shown that:

$$log\ p(y) = \underbrace{\int_\Theta q(\theta)\ log\left(\frac{p(y,\theta)}{q(\theta)}\right)}_{\mathcal{L}} + \underbrace{\int_\Theta q(\theta)\ log\left(\frac{q(\theta)}{p(\theta|y)}\right)}_{D_{KL(q\|p)}}$$

- $D_{KL}(q \parallel p)$ is the Kullback-Leibler divergence
- $\mathcal{L}$ is the evidence lower bound
- The Kullback-Leibler divergence is intractable, because it contains the posterior
- On the other hand, it is possible to compute all the terms in $\mathcal{L}$, as $p(y,\theta) = p(\theta)p(y|\theta)$ is known

ELBO: derivation

$$log\ p(y) = log\ p(y) \cdot 1 = log\ p(y) \int_{\Theta} q(\theta) d\theta$$

$$= \int_{\Theta} q(\theta)\ log\ p(y) d\theta = \int_{\Theta} q(\theta)\ log\ \left(\frac{p(y,\theta)}{p(\theta|y)} \right) d\theta$$

$$= \int_{\Theta} q(\theta)\ log \left(\frac{p(y,\theta)}{p(\theta|y)} \cdot \frac{q(\theta)}{q(\theta)} \right) d\theta$$

$$= \int_{\Theta} q(\theta)\ log \left(\frac{p(y,\theta)}{q(\theta)} \right) + \int_{\Theta} q(\theta)\ log \left(\frac{q(\theta)}{p(\theta|y)} \right)$$

Which can be written as (Ormerod & Wand, 2010):

$$log\ p(y) = \mathcal{L} + D_{KL}(q \parallel p)$$

ELBO: importance and optimization

- By rearranging we get:

$$\mathcal{L} = log\ p(y) - D_{KL}(q \parallel p)$$

- As $D_{KL} \geq 0 \Rightarrow log\ p(y) \geq \mathcal{L}$
- Therefore, $\mathcal{L}$ is the lower bound of the logarithm of the evidence $p(y)$
- Hence the name evidence lower bound or ELBO for $\mathcal{L}$

ELBO: importance and optimization

$$\mathcal{L} = log\ p(y) - D_{KL}(q \parallel p)$$

- **Key idea:** Maximizing the ELBO is equivalent to minimizing the Kullback-Leibler divergence
- The idea of maximizing the ELBO is the basis of most variational inference approaches

Making the ELBO tractable

- In theory, we could take any distribution q we like to approximate the posterior
- However, there are usually **restrictions** on q to make the problem more tractable

Making the ELBO tractable

- The most common restrictions are (Ormerod & Wand, 2010):
  - **Mean-field assumption**: $q(\theta)$ factorizes into $\prod_{i=1}^{M} q_i(\theta_i)$ for some partition $\{\theta_1, ..., \theta_M\}$
  - q comes from a **parametric** family of density functions

Contents

Summary: Variational Approximation

- The idea of variational inference is to find a probability distribution that **minimizes** the divergence to the posterior
- The KL-divergence cannot be minimized directly, as it depends on the **intractable** posterior
- Maximizing the ELBO is **equivalent** to minimizing the KL-divergence

Summary: ELBO

- To maximize the ELBO, we have to make **assumptions**
- The most common assumption is the **mean-field assumption**, which treats parameters as independent

Summary: Alternative divergence measures

- The KL-divergence is one of many divergence measures (Blei, Kucukelbir, & McAuliffe, 2017)
- It is possible to use $D_{KL}(p||q)$ instead of $D_{KL}(q||p)$
- Other alternative measures are the α-**divergence** and the f-**divergence**

Summary: Alternative divergence measures

- These alternative measures might offer a **better approximation** of the posterior
- However, this can lead to **higher computational cost**
- In practice, the most popular frameworks (PyMC3, Stan) use the ELBO as a base for the computations
- Moreover, **model misspecification** can be a bigger problem than the approximation error (Wang & Blei, 2019)

Bibliography I

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Cross-Validated. (2011). *KL divergence between two univariate Gaussians.* Retrieved 2020-01-04, from `https://stats.stackexchange.com/questions/7440/kl-divergence-between-two-univariate-gaussians`

Ormerod, J. T., & Wand, M. P. (2010). Explaining Variational Approximations. *The American Statistician*, *64*(2), 140–153.

Shibuya, N. (2018). *Demystifying KL Divergence.* Retrieved 2020-01-05, from `https://medium.com/activating-robotic-minds/demystifying-kl-divergence-7ebe4317ee68`

Wang, Y., & Blei, D. (2019). Variational Bayes under Model Misspecification. In *Advances in Neural Information Processing Systems 32* (pp. 13357–13367).