Laplace Approximation

Zeel B Patel, Nipun Batra

August 28, 2023

IIT Gandhinagar

Outline

Brook Taylor

Pierre-Simon Laplace

Overall idea

- Posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ might be intractable but we can compute the MAP estimate.
- We know that posterior would be in form: $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}, \boldsymbol{\theta})$, where Z is the normalizing constant.
- We can approximate this posterior using Taylor series expansion around the MAP estimate and it turns out that, after making a few assumptions, the resulting distribution is a Gaussian: $p(\boldsymbol{\theta}|\mathcal{D}) \approx \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}_{MAP}, (\nabla^2 f(\theta_{MAP}))^{-1})$, where f is the negative log joint evaluated at $\boldsymbol{\theta}_{MAP}$ and $\nabla^2 f$ is the Hessian matrix of f.

History

- Wiki article on Taylor's series
- Wiki article on Madhava and Madhava's series

Taylor Series Expansion

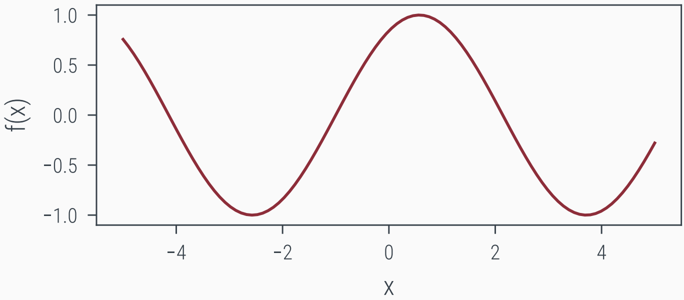
1D Taylor Series

$$\tilde{f}(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \dots$$

Taylor Approximation of a 1D Function

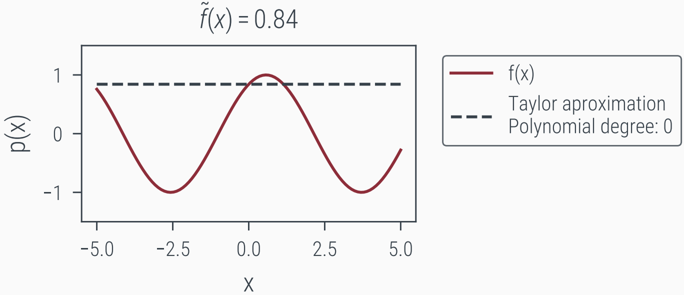
Consider the following function:

$$f(x) = \sin(1 + x)$$



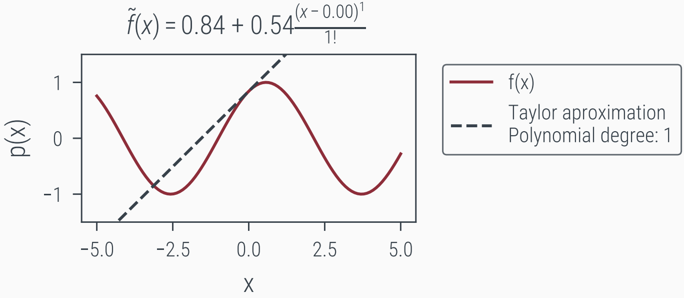
Taylor Approximation of a 1D Function

Taylor approximation at $x_0 = 0$:



Taylor Approximation of a 1D Function

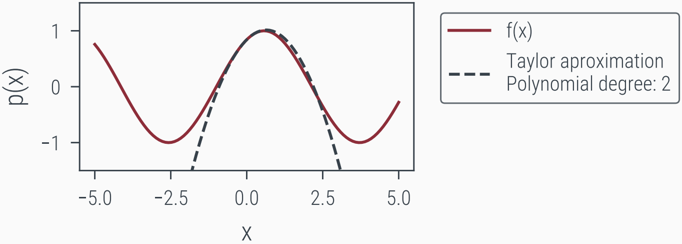
Taylor approximation at $x_0 = 0$:



Taylor Approximation of a 1D Function

Taylor approximation at $x_0 = 0$:

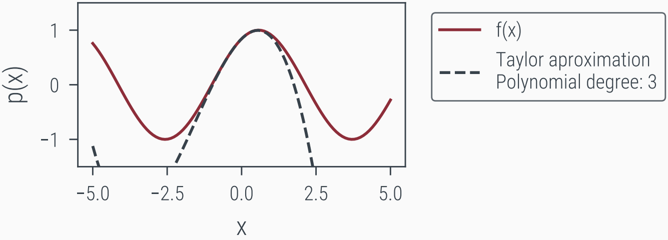
$$\tilde{f}(x) = 0.84 + 0.54\frac{(x-0.00)^1}{1!} - 0.84\frac{(x-0.00)^2}{2!}$$



Taylor Approximation of a 1D Function

Taylor approximation at $x_0 = 0$:

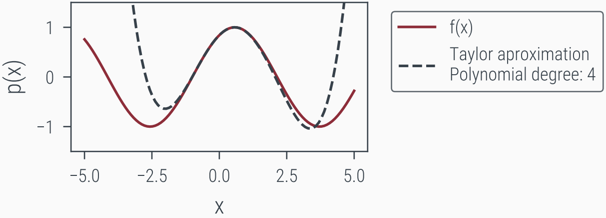
$$\tilde{f}(x) = 0.84 + 0.54\frac{(x-0.00)^1}{1!} - 0.84\frac{(x-0.00)^2}{2!} - 0.54\frac{(x-0.00)^3}{3!}$$



Taylor Approximation of a 1D Function

Taylor approximation at $x_0 = 0$:

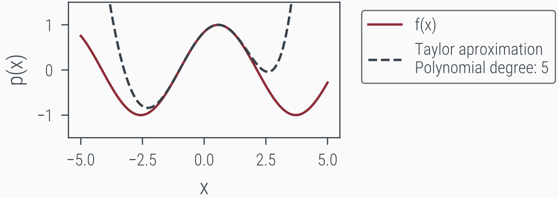
$$\tilde{f}(x) = 0.84 + 0.54\frac{(x-0.00)^1}{1!} - 0.84\frac{(x-0.00)^2}{2!} - 0.54\frac{(x-0.00)^3}{3!} + 0.84\frac{(x-0.00)^4}{4!}$$



Taylor Approximation of a 1D Function

Taylor approximation at $x_0 = 0$:

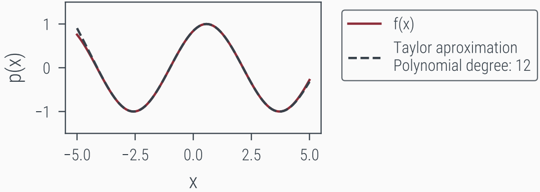
$$\tilde{f}(x) = 0.84 + 0.54\frac{(x-0.00)^1}{1!} - 0.84\frac{(x-0.00)^2}{2!} - 0.54\frac{(x-0.00)^3}{3!} + 0.84\frac{(x-0.00)^4}{4!} + 0.54\frac{(x-0.00)^5}{5!}$$



Taylor Approximation of a 1D Function

Taylor approximation at $x_0 = 0$:

$$\tilde{f}(x) = 0.84 + 0.54\frac{(x-0.00)^1}{1!} - 0.84\frac{(x-0.00)^2}{2!} - 0.54\frac{(x-0.00)^3}{3!} + 0.84\frac{(x-0.00)^4}{4!} + 0.54\frac{(x-0.00)^5}{5!} + \ldots$$



ND Taylor Series

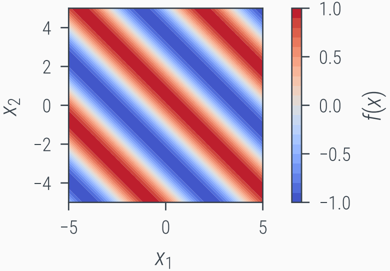
ND Taylor Series

$$\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla^2 f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \dots$$

Approximate a 2d function

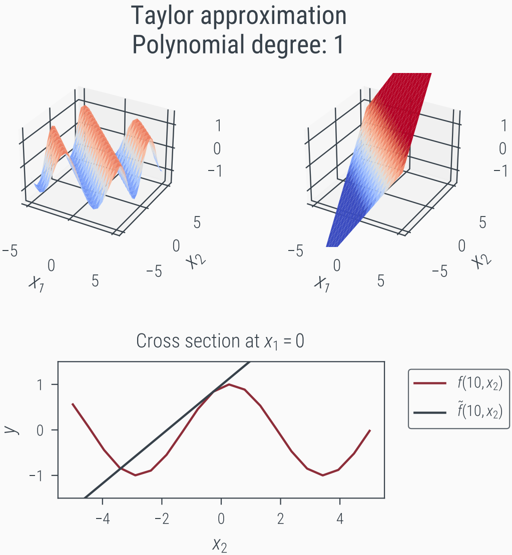
We take the following function:

$$f(x_1, x_2) = \sin(1 + x_1 + x_2)$$



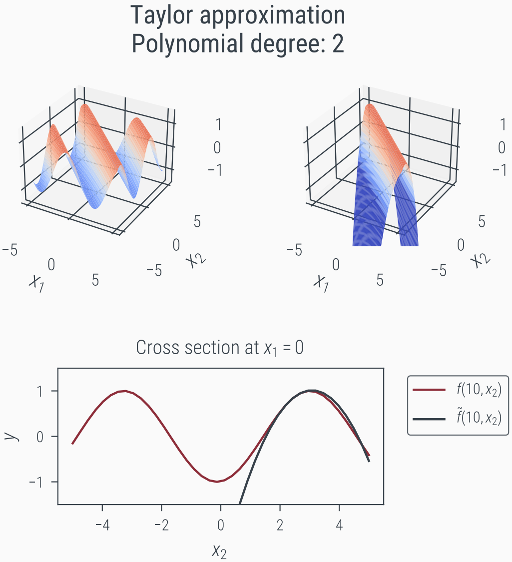
Approximate a 2d function

Taylor approximation at $x_0 = (0, 0)$:



Approximate a 2d function

Taylor approximation at $x_0 = (0, 0)$:



Laplace Approximation

Laplace Approximation

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}, \boldsymbol{\theta})$$

Laplace Approximation

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}, \boldsymbol{\theta})$$

We can rewrite this as:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} e^{-f(\boldsymbol{\theta})}$$

$$f(\boldsymbol{\theta}) = -\log p(\mathcal{D}, \boldsymbol{\theta})$$

Laplace Approximation

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}, \boldsymbol{\theta})$$

We can rewrite this as:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} e^{-f(\boldsymbol{\theta})}$$
$$f(\boldsymbol{\theta}) = -\log p(\mathcal{D}, \boldsymbol{\theta})$$

Note that $f(\boldsymbol{\theta})$ is the negative log joint which is used as a loss function to estimate $\boldsymbol{\theta}_{MAP}$.

Laplace Approximation

- Highest mass is concentrated around $\boldsymbol{\theta}_{MAP}$ and hence it makes sense to get Taylor approximation around that point.

Laplace Approximation

- Highest mass is concentrated around $\boldsymbol{\theta}_{MAP}$ and hence it makes sense to get Taylor approximation around that point.
- In other words, if our approximation is bad where we have low probability mass, it doesn't matter much.

Laplace Approximation

- Highest mass is concentrated around $\boldsymbol{\theta}_{MAP}$ and hence it makes sense to get Taylor approximation around that point.
- In other words, if our approximation is bad where we have low probability mass, it doesn't matter much.
- Thus, we approximate $f(\boldsymbol{\theta})$ as $\tilde{f}(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_{MAP}$ using Taylor series expansion up to second derivative:

$$\begin{aligned}\tilde{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_{MAP}) + \nabla f(\boldsymbol{\theta}_{MAP})^T(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP}) \\ + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})\end{aligned}$$

Laplace Approximation

$$\tilde{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_{MAP}) + \nabla f(\boldsymbol{\theta}_{MAP})^T(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})$$

Laplace Approximation

$$\tilde{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_{MAP}) + \nabla f(\boldsymbol{\theta}_{MAP})^T(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})$$

Since, $\boldsymbol{\theta}_{MAP}$ is minima of $f(\boldsymbol{\theta})$, $\nabla f(\boldsymbol{\theta}_{MAP}) = 0$.

Laplace Approximation

$$\tilde{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_{MAP}) + \nabla f(\boldsymbol{\theta}_{MAP})^T(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP}) \\ + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})$$

Since, $\boldsymbol{\theta}_{MAP}$ is minima of $f(\boldsymbol{\theta})$, $\nabla f(\boldsymbol{\theta}_{MAP}) = 0$.

$$\tilde{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_{MAP}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})$$

Laplace Approximation

$$\tilde{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_{MAP}) + \nabla f(\boldsymbol{\theta}_{MAP})^T(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})$$

Since, $\boldsymbol{\theta}_{MAP}$ is minima of $f(\boldsymbol{\theta})$, $\nabla f(\boldsymbol{\theta}_{MAP}) = 0$.

$$\tilde{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_{MAP}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta} - \boldsymbol{\theta}_{MAP})$$

where $\nabla^2 f(\boldsymbol{\theta}_{MAP})$ is the Hessian matrix of $f(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_{MAP}$.

Laplace Approximation

Plugging this back to the posterior equation:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z}e^{-f(\boldsymbol{\theta})} \quad \text{where } f(\boldsymbol{\theta}) = -\log p(\mathcal{D}, \boldsymbol{\theta})$$

Laplace Approximation

Plugging this back to the posterior equation:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D}) &= \frac{1}{Z} e^{-f(\boldsymbol{\theta})} \quad \text{where } f(\boldsymbol{\theta}) = -\log p(\mathcal{D}, \boldsymbol{\theta}) \\
&\approx \frac{1}{Z} e^{-f(\boldsymbol{\theta}_{MAP})} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})}
\end{aligned}
$$

Laplace Approximation

Plugging this back to the posterior equation:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= \frac{1}{Z} e^{-f(\boldsymbol{\theta})} \quad \text{where } f(\boldsymbol{\theta}) = -\log p(\mathcal{D}, \boldsymbol{\theta}) \\ &\approx \frac{1}{Z} e^{-f(\boldsymbol{\theta}_{MAP})} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})} \\ &= \frac{1}{Z} p(\mathcal{D}, \boldsymbol{\theta}_{MAP}) e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})} \end{aligned}$$

Laplace Approximation

Plugging this back to the posterior equation:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D}) &= \frac{1}{Z} e^{-f(\boldsymbol{\theta})} \quad \text{where } f(\boldsymbol{\theta}) = -\log p(\mathcal{D}, \boldsymbol{\theta}) \\
&\approx \frac{1}{Z} e^{-f(\boldsymbol{\theta}_{MAP})} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})} \\
&= \frac{1}{Z} p(\mathcal{D}, \boldsymbol{\theta}_{MAP}) e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})^T \nabla^2 f(\boldsymbol{\theta}_{MAP})(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})}
\end{aligned}
$$

$$
p(\boldsymbol{\theta}|\mathcal{D}) \approx \mathcal{N}\left(\boldsymbol{\theta}|\boldsymbol{\theta}_{MAP}, \left(\nabla^2 f(\boldsymbol{\theta}_{MAP})\right)^{-1}\right)
$$

$$
Z = p(\mathcal{D}, \boldsymbol{\theta}_{MAP}) \cdot (2\pi)^{D/2} \cdot |\nabla^2 f(\boldsymbol{\theta}_{MAP})|^{-\frac{1}{2}}
$$

Pros and Cons of Laplace Approximation

- Pros:
    - Simple to implement
    - Computationally efficient
    - Can be used to approximate any intractable function

Pros and Cons of Laplace Approximation

- Pros:
  - Simple to implement
  - Computationally efficient
  - Can be used to approximate any intractable function
- Cons:
  - It can give bad approximation when posterior is not unimodal
  - Gaussian assumption can be too restrictive at times
  - Hessian matrix inversion can be numerically unstable and expensive. A diagonal or block-wise approximation can be applied to resolve this. Checkout Laplace-Redux for more details.

Beta-Bernoulli Coin Toss

Let's take Beta-Bernoulli Coin Toss example since we know the closed form posterior for it. Consider the following scenario:

Beta-Bernoulli Coin Toss

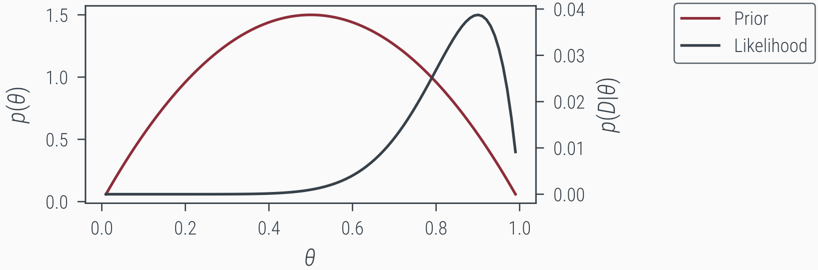
Let's take Beta-Bernoulli Coin Toss example since we know the closed form posterior for it. Consider the following scenario:

- $\mathcal{D} = \{1, 1, 1, 1, 1, 1, 1, 1, 0\}$
- $p(\theta) = \text{Beta}(\alpha = 2, \beta = 2)$
- $\theta = P(H)$
- $p(y|\theta) = \theta^{y}(1 - \theta)^{1-y}$

Beta-Bernoulli Coin Toss

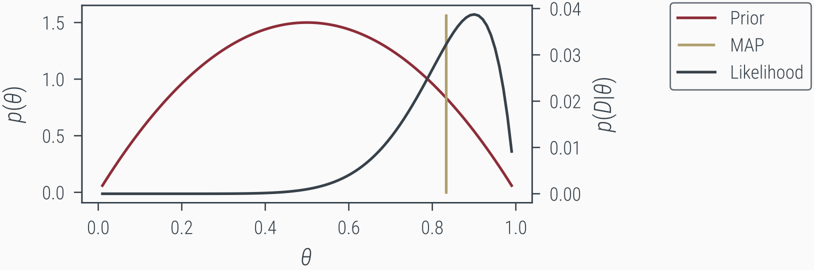
Let's take Beta-Bernoulli Coin Toss example since we know the closed form posterior for it. Consider the following scenario:

- $\mathcal{D} = \{1, 1, 1, 1, 1, 1, 1, 1, 0\}$
- $p(\theta) = \text{Beta}(\alpha = 2, \beta = 2)$
- $\theta = P(H)$
- $p(y|\theta) = \theta^y (1 - \theta)^{1-y}$

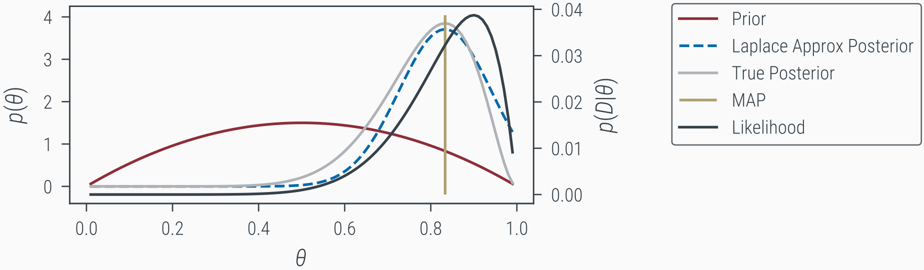


Beta-Bernoulli Coin Toss

MAP estimate:



Laplace Approximation:



Multi-Mode example

Consider a Gaussian Mixture distribution with two modes. We assume that, it is an unnormalized density and we want to get normalized Laplace approximation of it.

Multi-Mode example

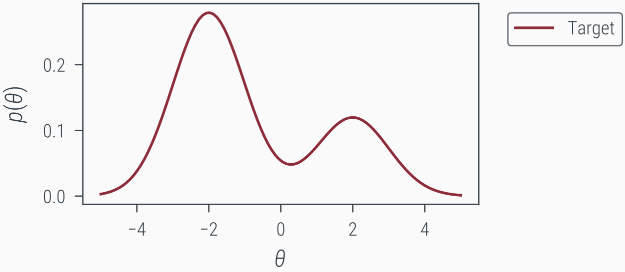
Consider a Gaussian Mixture distribution with two modes. We assume that, it is an unnormalized density and we want to get normalized Laplace approximation of it.

$$p(\theta) = \frac{7}{10}\mathcal{N}(\theta| -2, 1) + \frac{3}{10}\mathcal{N}(\theta|2, 1)$$

Multi-Mode example

Consider a Gaussian Mixture distribution with two modes. We assume that, it is an unnormalized density and we want to get normalized Laplace approximation of it.

$$p(\theta) = \frac{7}{10}\mathcal{N}(\theta| - 2, 1) + \frac{3}{10}\mathcal{N}(\theta|2, 1)$$



Multi-Mode example

Laplace Approximation: