A Fresh Look at the Bayes' Theorem from Information Theory

Tan Bui-Thanh

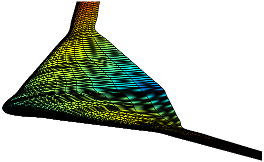
Computational Engineering and Optimization (CEO) Group
Department of Aerospace Engineering and Engineering Mechanics
Institute for Computational Engineering and Sciences (ICES)
The University of Texas at Austin

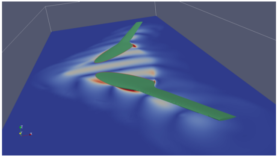
Babuska Series, ICES Sep 9, 2016

Outline

Large-scale computation under uncertainty

Inverse electromagnetic scattering





Randomness

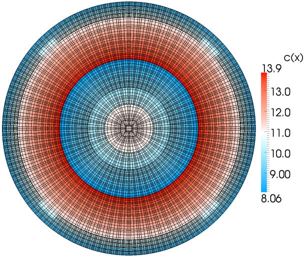
- Random errors in measurements are unavoidable
- Inadequacy of the mathematical model (Maxwell equations)

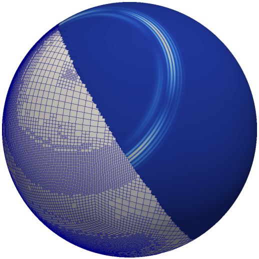
Challenge

How to invert for the invisible shape/medium using computational electromagnetics with $\mathcal{O}\left(10^6\right)$ degree of freedoms?

Large-scale computation under uncertainty

Full wave form seismic inversion





Randomness

- Random errors in seismometer measurements are unavoidable
- Inadequacy of the mathematical model (elastodynamics)

Challenge

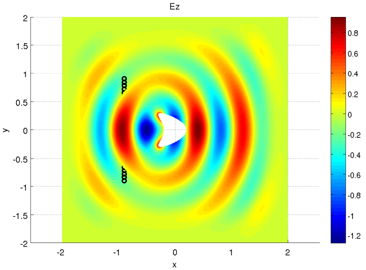
How to image the earth interior using forward computational model with with $\mathcal{O}\left(10^9\right)$ degree of freedoms?

Inverse Shape Electromagnetic Scattering Problem

Maxwell Equations:

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \qquad \text{(Faraday)}$$

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t}, \qquad \text{(Ampere)}$$



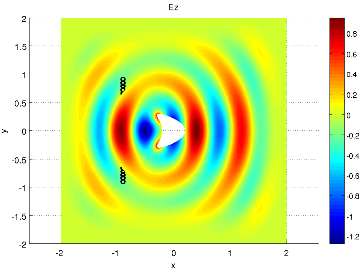
$\mathbf{E}$: Electric field, $\mathbf{H}$: Magnetic field, μ: permeability, ϵ: permittivity

Inverse Shape Electromagnetic Scattering Problem

Maxwell Equations:

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \qquad \text{(Faraday)}$$

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t}, \qquad \text{(Ampere)}$$



$\mathbf{E}$: Electric field, $\mathbf{H}$: Magnetic field, μ: permeability, ϵ: permittivity

Forward problem (discontinuous Galerkin discretization)

$$d = \mathcal{G}(x)$$

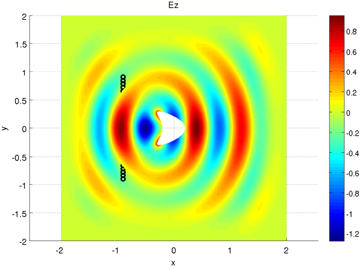
where $\mathbf{G}$ maps shape parameters x to electric/magnetic field d at the measurement points

Inverse Shape Electromagnetic Scattering Problem

Maxwell Equations:

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}, \qquad \text{(Faraday)}$$

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t}, \qquad \text{(Ampere)}$$



$\mathbf{E}$: Electric field, $\mathbf{H}$: Magnetic field, μ: permeability, ϵ: permittivity

Forward problem (discontinuous Galerkin discretization)

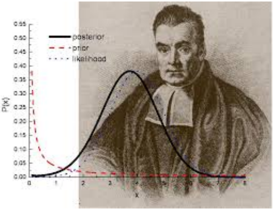
$$d = \mathcal{G}(x)$$

where $\mathbf{G}$ maps shape parameters x to electric/magnetic field d at the measurement points

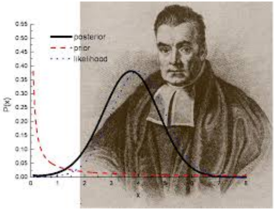
Inverse Problem

Given (possibly noise-corrupted) measurements on d, infer x?

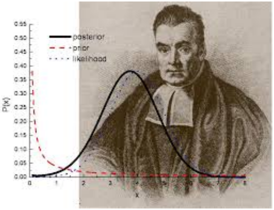
The Bayesian Statistical Inversion Framework



The Bayesian Statistical Inversion Framework



The Bayesian Statistical Inversion Framework



Bayes Theorem

$$\pi_{\text{post}}(x|d) \propto \pi_{\text{like}}(d|x) \times \pi_{\text{prior}}(x)$$

Bayes theorem for inverse electromagnetic scattering

Prior knowledge: The obstacle is smooth:

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\lambda \int_0^{2\pi} r''(x)d\theta\right)$$

Bayes theorem for inverse electromagnetic scattering

Prior knowledge: The obstacle is smooth:

$$\pi_{\mathrm{pr}}(x) \propto \exp\left(-\lambda \int_0^{2\pi} r''(x) d\theta\right)$$

Likelihood: Additive Gaussian noise, for example,

$$\pi_{\mathrm{like}}(d|x) \propto \exp\left(-\frac{1}{2}\left\|\mathcal{G}(x) - d\right\|^2_{C_{\mathrm{noise}}}\right)$$

Outline

Entropy

Definition

We define the uncertainty in a random variable X distributed by $0 \leq \pi(x) \leq 1$ as

$$H(X) = -\int \pi(x) \log \pi(x) \, dx \geq 0$$

Entropy

Entropy

Wiener and Shannon

Kolmogorov

Copied from Sergio Verdu

Entropy

Wiener and Shannon

Kolmogorov

Copied from Sergio Verdu

- Wiener: "...for it belongs to the two of us equally"

Entropy

Wiener and Shannon

Kolmogorov

Copied from Sergio Verdu

- Wiener: "...for it belongs to the two of us equally"
- Shannon: "...a mathematical pun"

Entropy

Wiener and Shannon

Kolmogorov

Copied from Sergio Verdu

- Wiener: "...for it belongs to the two of us equally"
- Shannon: "...a mathematical pun"
- Kolmogorov: "...has no physical interpretation"

Entropy

Entropy of uniform distribution

Entropy

Entropy of uniform distribution

- Let U be a uniform random variable with values in $\mathcal{X}$, and $|\mathcal{X}| < \infty$

Entropy

Entropy of uniform distribution

- Let U be a uniform random variable with values in $\mathcal{X}$, and $|\mathcal{X}| < \infty$
- $\pi(u) := \dfrac{1}{|\mathcal{X}|} \Rightarrow H(U) = \log(|\mathcal{X}|)$

Entropy

Entropy of uniform distribution

- Let U be a uniform random variable with values in $\mathcal{X}$, and $|\mathcal{X}| < \infty$
- $\pi(u) := \dfrac{1}{|\mathcal{X}|} \Rightarrow H(U) = \log(|\mathcal{X}|)$

How uncertain is the uniform random variable?

Entropy

Entropy of uniform distribution

- Let U be a uniform random variable with values in $\mathcal{X}$, and $|\mathcal{X}| < \infty$
- $\pi(u) := \frac{1}{|\mathcal{X}|} \Rightarrow H(U) = \log(|\mathcal{X}|)$

How uncertain is the uniform random variable?

$$H(X) \leq H(U)$$

100 years of uniform distribution

source: Christoph Aistleitner

100 years of uniform distribution

source: Christoph Aistleitner

Hermann Weyl

and Maximum entropy

Maximum entropy distribution

- X with known mean and variance

and Maximum entropy

Maximum entropy distribution

- X with known mean and variance
- $\pi(x)$? with maximum entropy

and Maximum entropy

Maximum entropy distribution

- X with known mean and variance
- $\pi(x)$? with maximum entropy
- $$\max_{\pi(x)} H(X) = -\int \pi(x) \log(\pi(x))\, dx$$

  subject to

  $$\int x\pi(x)\, dx = \mu$$
  $$\int (x-\mu)^2 \pi(x)\, dx = \sigma^2$$
  $$\int \pi(x)\, dx = 1$$

Gaussian and Maximum entropy

Maximum entropy distribution

- X with known mean and variance
- $\pi(x)$? with maximum entropy
- $$\max_{\pi(x)} H(X) = -\int \pi(x) \log(\pi(x))\, dx$$

  subject to

  $$\int x\pi(x)\, dx = \mu$$

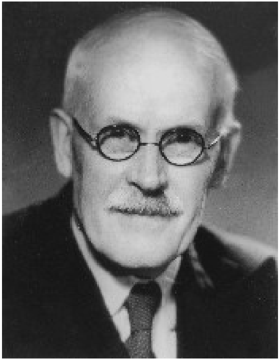
  $$\int (x-\mu)^2 \pi(x)\, dx = \sigma^2$$

  $$\int \pi(x)\, dx = 1$$
- Gaussian distribution: $\pi(x) = \mathcal{N}(\mu, \sigma^2)$

Outline

Relative Entropy

Abraham Wald (1945)



Harold Jeffreys (1945)

$$D\left(\pi || q\right) := \int \pi(x) \log\left(\frac{\pi(x)}{q(x)}\right) dx$$

Kullback-Leibler divergence = Relative Entropy

Solomon Kullback (1951)

Richard Leibler (1951)

$$D\left(\pi || q\right) := \int \pi(x) \log\left(\frac{\pi(x)}{q(x)}\right) dx$$

Kullback-Leibler divergence = Relative Entropy

Solomon Kullback (1951)

Richard Leibler (1951)

$$D(\pi||q) := \int \pi(x) \log\left(\frac{\pi(x)}{q(x)}\right) dx \overset{\text{discrete}}{=} \sum \pi_i \log\left(\frac{\pi_i}{q_i}\right)$$

Information Inequality

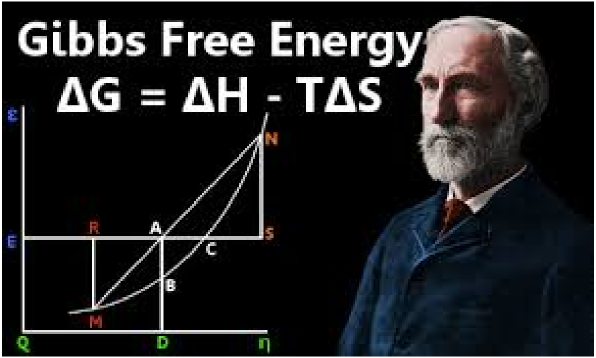
The most important inequality in information theory

$$D\left(\pi || q\right) \geq 0$$

Can we see it easily?

Information Inequality

The most important inequality in information theory



$$D\left(\pi || q\right) \geq 0$$

Can we see it easily?

Outline

From Relative Entropy to Bayes' Theorem

- Toss n times an kth dimensional dice with the prior distribution of each face $\{p_i\}_{i=1}^k$: $\sum_{i=1}^{k} p_i = 1$

From Relative Entropy to Bayes' Theorem

- Toss n times an kth dimensional dice with the prior distribution of each face $\{p_i\}_{i=1}^k$: $\sum_{i=1}^{k} p_i = 1$
- Let n_i is the number of times we see face i: $\frac{n_i}{n} \to p_i$

From Relative Entropy to Bayes' Theorem

- Toss n times an kth dimensional dice with the prior distribution of each face $\{p_i\}_{i=1}^k$: $\sum_{i=1}^{k} p_i = 1$
- Let n_i is the number of times we see face i: $\frac{n_i}{n} \to p_i$
- What is the likelihood that these n faces also distributed by the posterior distrubtion q_i: $\sum_{i=1}^{k} q_i = 1$?

From Relative Entropy to Bayes' Theorem

- Toss n times an kth dimensional dice with the prior distribution of each face $\{p_i\}_{i=1}^k$: $\sum_{i=1}^{k} p_i = 1$
- Let n_i is the number of times we see face i: $\frac{n_i}{n} \to p_i$
- What is the likelihood that these n faces also distributed by the posterior distrubtion q_i: $\sum_{i=1}^{k} q_i = 1$?
- The likelihood of $\{n_i\}_{i=1}^k$ distributed by $\{q_i\}_{i=1}^k$

$$\Pi_{i=1}^k q_i^{n_i}$$

From Relative Entropy to Bayes' Theorem

- Toss n times an kth dimensional dice with the prior distribution of each face $\{p_i\}_{i=1}^k$: $\sum_{i=1}^{k} p_i = 1$
- Let n_i is the number of times we see face i: $\frac{n_i}{n} \to p_i$
- What is the likelihood that these n faces also distributed by the posterior distrubtion q_i: $\sum_{i=1}^{k} q_i = 1$?
- The likelihood of $\{n_i\}_{i=1}^k$ distributed by $\{q_i\}_{i=1}^k$ (Multinomial distribution)

$$L := \frac{n!}{\Pi_{i=1}^k n_i!} \Pi_{i=1}^k q_i^{n_i}$$

From Relative Entropy to Bayes' Theorem

- The likelihood of $\{n_i\}_{i=1}^k$ distributed by $\{q_i\}_{i=1}^k$

$$L := \frac{n!}{\Pi_{i=1}^k n_i!} \Pi_{i=1}^k q_i^{n_i}$$

From Relative Entropy to Bayes' Theorem

- The likelihood of $\{n_i\}_{i=1}^{k}$ distributed by $\{q_i\}_{i=1}^{k}$

$$L := \frac{n!}{\Pi_{i=1}^{k} n_i!} \Pi_{i=1}^{k} q_i^{n_i}$$

- Take the log likelihood

$$\log L = \log(n!) - \sum \log(n_i!) + \sum n_i \log(q_i)$$

From Relative Entropy to Bayes' Theorem

- The likelihood of $\{n_i\}_{i=1}^k$ distributed by $\{q_i\}_{i=1}^k$

$$L := \frac{n!}{\Pi_{i=1}^k n_i!} \Pi_{i=1}^k q_i^{n_i}$$

- Take the log likelihood

$$\log L = \log(n!) - \sum \log(n_i!) + \sum n_i \log(q_i)$$

- Stirling's approximation $\log n! \approx n \log(n) - n$

$$\log L = n \log(n) - \sum n_i \log(n_i) + \sum n_i \log(q_i) + \underbrace{\sum n_i - n}_{0}$$

From Relative Entropy to Bayes' Theorem

- The likelihood of $\{n_i\}_{i=1}^k$ distributed by $\{q_i\}_{i=1}^k$

$$L := \frac{n!}{\Pi_{i=1}^k n_i!} \Pi_{i=1}^k q_i^{n_i}$$

- Take the log likelihood

$$\log L = \log(n!) - \sum \log(n_i!) + \sum n_i \log(q_i)$$

- Stirling's approximation $\log n! \approx n \log(n) - n$

$$\log L = n \log(n) - \sum n_i \log(n_i) + \sum n_i \log(q_i) + \underbrace{\sum n_i - n}_{0}$$

$$\frac{1}{n}(-\log L) = \sum \frac{n_i}{n} \log\left(\frac{n_i/n}{q_i}\right)$$

From Relative Entropy to Bayes' Theorem

- The likelihood of $\{n_i\}_{i=1}^k$ distributed by $\{q_i\}_{i=1}^k$

$$L := \frac{n!}{\Pi_{i=1}^k n_i!} \Pi_{i=1}^k q_i^{n_i}$$

- Take the log likelihood

$$\log L = \log(n!) - \sum \log(n_i!) + \sum n_i \log(q_i)$$

- Stirling's approximation $\log n! \approx n \log(n) - n$

$$\log L = n \log(n) - \sum n_i \log(n_i) + \sum n_i \log(q_i) + \underbrace{\sum n_i - n}_{0}$$

$$\frac{1}{n} (-\log L) = \sum \frac{n_i}{n} \log \left(\frac{n_i/n}{q_i} \right) = \sum p_i \log \left(\frac{p_i}{q_i} \right)$$

From Relative Entropy to Bayes' Theorem

- The likelihood of $\{n_i\}_{i=1}^k$ distributed by $\{q_i\}_{i=1}^k$

$$L := \frac{n!}{\Pi_{i=1}^k n_i!} \Pi_{i=1}^k q_i^{n_i}$$

- Take the log likelihood

$$\log L = \log(n!) - \sum \log(n_i!) + \sum n_i \log(q_i)$$

- Stirling's approximation $\log n! \approx n \log(n) - n$

$$\log L = n \log(n) - \sum n_i \log(n_i) + \sum n_i \log(q_i) + \underbrace{\sum n_i - n}_{0}$$

- Relative entropy = average likelihood

$$\frac{1}{n}(-\log L) = \sum \frac{n_i}{n} \log\left(\frac{n_i/n}{q_i}\right) = \sum p_i \log\left(\frac{p_i}{q_i}\right) = D\left(p||q\right)$$

From Relative Entropy to Bayes' Theorem

Relative entropy = average likelihood

- $$\frac{1}{n}\left(-\log L\right) = D\left(p||q\right)$$

From Relative Entropy to Bayes' Theorem

Relative entropy = average likelihood

- $$\frac{1}{n}(-\log L) = D(p||q)$$
- Write $\sum \to \int$

$$-\int \log(L)p(x)\,dx = \int \log\left(\frac{p}{q}\right) p(x)\,dx$$

From Relative Entropy to Bayes' Theorem

Relative entropy = average likelihood

- $$\frac{1}{n}(-\log L) = D(p||q)$$
- Write $\sum \to \int$

$$-\int \log(L)p(x)\,dx = \int \log\left(\frac{p}{q}\right)p(x)\,dx$$

$$q(x) = L(x)p(x)$$

From Relative Entropy to Bayes' Theorem

Relative entropy = average likelihood → Bayes

- $$\frac{1}{n}\left(-\log L\right) = D\left(p||q\right)$$
- Write $\sum \to \int$
$$-\int \log(L)p(x)\,dx = \int \log\left(\frac{p}{q}\right)p(x)\,dx$$
- Bayes' theorem $q(x) = L(x)p(x)$

From Optimization to Bayes' Theorem

Inverse Problem

- Given observation model

$$d = \mathcal{G}(x) + \varepsilon$$

From Optimization to Bayes' Theorem

Inverse Problem

- Given observation model

$$d = \mathcal{G}(x) + \varepsilon$$

- Inverse task: given d, infer x

From Optimization to Bayes' Theorem

Inverse Problem

- Given observation model

$$d = \mathcal{G}(x) + \varepsilon$$

- Inverse task: given d, infer x
- Statistical inversion: Prior knowledge: $X \sim \pi_{\text{prior}}(x)$. Look for the posterior distribution $\pi_{\text{post}}(x)$ that combines **prior information** and **information from the data**.

From Optimization to Bayes' Theorem

Inverse Problem

- Given observation model

$$d = \mathcal{G}(x) + \varepsilon$$

- Inverse task: given d, infer x
- Statistical inversion: Prior knowledge: $X \sim \pi_{\text{prior}}(x)$. Look for the posterior distribution $\pi_{\text{post}}(x)$ that combines **prior information** and **information from the data**.
- The likelihood: assume $\varepsilon \sim \mathcal{N}(0, C)$

$$\pi_{\text{like}}(x) = \exp\left(-\frac{1}{2}\|d - \mathcal{G}(x)\|_C^2\right)$$

From Optimization to Bayes' Theorem

Prior Elicitation

- Try to get the best prior information = discrepancy relative to the posterior is minimized

From Optimization to Bayes' Theorem

Prior Elicitation

- Try to get the best prior information = discrepancy relative to the posterior is minimized
- Conversely, best prior $\rightarrow$ the **information gained** in the posterior should not be large

From Optimization to Bayes' Theorem

Prior Elicitation

- Try to get the best prior information = discrepancy relative to the posterior is minimized
- Conversely, best prior → the **information gained** in the posterior should not be large
- Equivalently,

$$\pi_{\text{post}} = \underset{\pi(x)}{\arg\min}\, D\left(\pi || \pi_{\text{prior}}\right) = \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right) dx$$

From Optimization to Bayes' Theorem

How about information from the data?

- Want to find x to match the data as well as we can

From Optimization to Bayes' Theorem

How about information from the data?

- Want to find x to match the data as well as we can
- "Equivalently": want to find the posterior distribution such that $\|d - \mathcal{G}(x)\|_C^2$ is minimized!

From Optimization to Bayes' Theorem

How about information from the data?

- Want to find x to match the data as well as we can
- "Equivalently": want to find the posterior distribution such that $\|d - \mathcal{G}(x)\|_C^2$ is minimized!
- One approach: minimize the **mean squared error**

$$\pi_{\text{post}} = \arg\min_{\pi(x)} \int \pi(x) \, \|d - \mathcal{G}(x)\|_C^2 \; dx$$

From Optimization to Bayes' Theorem

How about information from the data?

- Want to find x to match the data as well as we can
- "Equivalently": want to find the posterior distribution such that $\|d - \mathcal{G}(x)\|_C^2$ is minimized!
- One approach: minimize the **mean squared error**

$$\pi_{\text{post}} = \underset{\pi(x)}{\arg\min} \int \pi(x) \, \|d - \mathcal{G}(x)\|_C^2 \, dx = -\int \pi(x) \log\left(\pi_{\text{like}}(x)\right) \, dx$$

From Optimization to Bayes' Theorem

Prior + data information

- From prior

$$\pi_{\text{post}} = \arg\min_{\pi(x)} D\left(\pi || \pi_{\text{prior}}\right) = \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right) dx$$

From Optimization to Bayes' Theorem

Prior + data information

- From prior

$$\pi_{\text{post}} = \arg\min_{\pi(x)} D\left(\pi || \pi_{\text{prior}}\right) = \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right) \, dx$$

- From likelihood

$$\pi_{\text{post}} = \arg\min_{\pi(x)} - \int \pi(x) \log\left(\pi_{\text{like}}(x)\right) \, dx$$

From Optimization to Bayes' Theorem

Prior + data information

- From prior

$$\pi_{\text{post}} = \arg\min_{\pi(x)} D\left(\pi || \pi_{\text{prior}}\right) = \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right) dx$$

- From likelihood

$$\pi_{\text{post}} = \arg\min_{\pi(x)} -\int \pi(x) \log\left(\pi_{\text{like}}(x)\right) dx$$

- A Compromise

$$\pi_{\text{post}} = \arg\min_{\pi(x)} -\int \pi(x) \log\left(\pi_{\text{like}}(x)\right) dx + \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right) dx$$

subject to

$$\int \pi(x)\, dx = 1, \qquad \text{and } \pi(x) \geq 0.$$

From Optimization to Bayes' Theorem

Prior + data information

- A Compromise

$$\pi_{\text{post}} = \arg\min_{\pi(x)} - \int \pi(x) \log\left(\pi_{\text{like}}(x)\right)\, dx + \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right)\, dx$$

subject to

$$\int \pi(x)\, dx = 1, \quad \text{and } \pi(x) \geq 0.$$

From Optimization to Bayes' Theorem

Prior + data information

- A Compromise

$$\pi_{\text{post}} = \arg\min_{\pi(x)} - \int \pi(x) \log\left(\pi_{\text{like}}(x)\right)\, dx + \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right)\, dx$$

subject to

$$\int \pi(x)\, dx = 1, \quad \text{and } \pi(x) \geq 0.$$

- Does it have a solution $\pi_{\text{post}}(x)$? is it unique?

From Optimization to Bayes' Theorem

Prior + data information

- A Compromise

$$\pi_{\text{post}} = \arg\min_{\pi(x)} - \int \pi(x) \log\left(\pi_{\text{like}}(x)\right) \, dx + \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right) \, dx$$

subject to

$$\int \pi(x) \, dx = 1, \quad \text{and } \pi(x) \geq 0.$$

- Does it have a solution $\pi_{\text{post}}(x)$? is it unique?
- How to solve?

From Optimization to Bayes' Theorem

Prior + data information

- A Compromise

$$\pi_{\text{post}} = \arg\min_{\pi(x)} - \int \pi(x) \log\left(\pi_{\text{like}}(x)\right)\, dx + \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right)\, dx$$

subject to

$$\int \pi(x)\, dx = 1, \quad \text{and } \pi(x) \geq 0.$$

- Does it have a solution $\pi_{\text{post}}(x)$? is it unique?
- How to solve?

Lagrangian + calculus of variation

From Optimization to Bayes' Theorem

Prior + data information

- A Compromise

$$\pi_{\text{post}} = \arg\min_{\pi(x)} - \int \pi(x) \log\left(\pi_{\text{like}}(x)\right)\, dx + \int \pi(x) \log\left(\frac{\pi(x)}{\pi_{\text{prior}}(x)}\right)\, dx$$

subject to

$$\int \pi(x)\, dx = 1, \quad \text{and } \pi(x) \geq 0.$$

- Does it have a solution $\pi_{\text{post}}(x)$? is it unique?
- How to solve?

Lagrangian + calculus of variation

- Solution = Bayes' theorem

$$\pi_{\text{post}}\left(x|d\right) = \frac{\pi_{\text{like}}\left(d|x\right) \times \pi_{\text{prior}}\left(x\right)}{\int \pi_{\text{like}}\left(d|x\right) \times \pi_{\text{prior}}\left(x\right)\, dx}$$

Outline

Conclusions

1. Information provide an intuitive and fresh view of Bayes' theorem

Conclusions

1. Information provide an intuitive and fresh view of Bayes' theorem
2. Relative entropy $\rightarrow$ Bayes' theorem

Conclusions

1. Information provide an intuitive and fresh view of Bayes' theorem
2. Relative entropy $\rightarrow$ Bayes' theorem
3. Optimization $+$ information $\rightarrow$ Bayes' theorem