Reconstruction in high-energy Cherenkov neutrino telescopes

Cosmic Neutrino Practicum

Jun 29 2026, Thorsten Glüsenkamp

Overview

- Lecture 1: Introduction and **high-level overview**
  -> Bayes theorem
  -> Frequentist vs Bayesian
  -> Information-theoretic grounding (KL-divergence of the “joint”)
  -> Theme of these lectures: likelihood vs likelihood-free

Overview

- Lecture 1: Introduction and **high-level overview**
  -> Bayes theorem
  -> Frequentist vs Bayesian
  -> Information-theoretic grounding (KL-divergence of the “joint”)
  -> Theme of these lectures: likelihood vs likelihood-free

- Lecture 2: Likelihood methods in detail: **Poisson process likelihoods**
  -> obtain response-PDFs (Analytic / B-Splines)
  -> Tricks for systematics circumvention (1st hit likelihood)
  -> Parametrization: Stochastics
  -> Uncertainties

Overview

- Lecture 1: Introduction and **high-level overview**
  -> Bayes theorem
  -> Frequentist vs Bayesian
  -> Information-theoretic grounding (KL-divergence of the “joint”)
  -> Theme of these lectures: likelihood vs likelihood-free

- Lecture 2: Likelihood methods in detail: **Poisson process likelihoods**
  -> obtain response-PDFs (Analytic / B-Splines)
  -> Tricks for systematics circumvention (1st hit likelihood)
  -> Parametrization: Stochastics
  -> Uncertainties

- Lecture 3: All things **neural networks**
  -> Neural network-based likelihoods
  -> Inverse likelihood-free methods
  -> Encodings and symmetries (Transformers etc.)
  -> Putting it all together

Overview

- Lecture 1: Introduction and **high-level overview**
  -> Bayes theorem
  -> Frequentist vs Bayesian
  -> Information-theoretic grounding (KL-divergence of the “joint”)
  -> Theme of these lectures: likelihood vs likelihood-free

- Lecture 2: Likelihood methods in detail: **Poisson process likelihoods**
  -> obtain response-PDFs (Analytic / B-Splines)
  -> Tricks for systematics circumvention (1st hit likelihood)
  -> Parametrization: Stochastics
  -> Uncertainties

- Lecture 3: All things **neural networks**
  -> Neural network-based likelihoods
  -> Inverse likelihood-free methods
  -> Encodings and symmetries (Transformers etc.)
  -> Putting it all together

Classical statistical tools

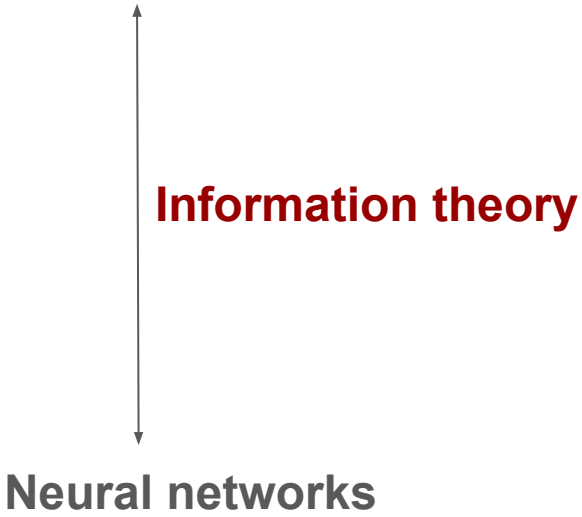
Neural networks

Overview

- Lecture 1: Introduction and **high-level overview**
  -> Bayes theorem
  -> Frequentist vs Bayesian
  -> Information-theoretic grounding (KL-divergence of the “joint”)
  -> Theme of these lectures: likelihood vs likelihood-free

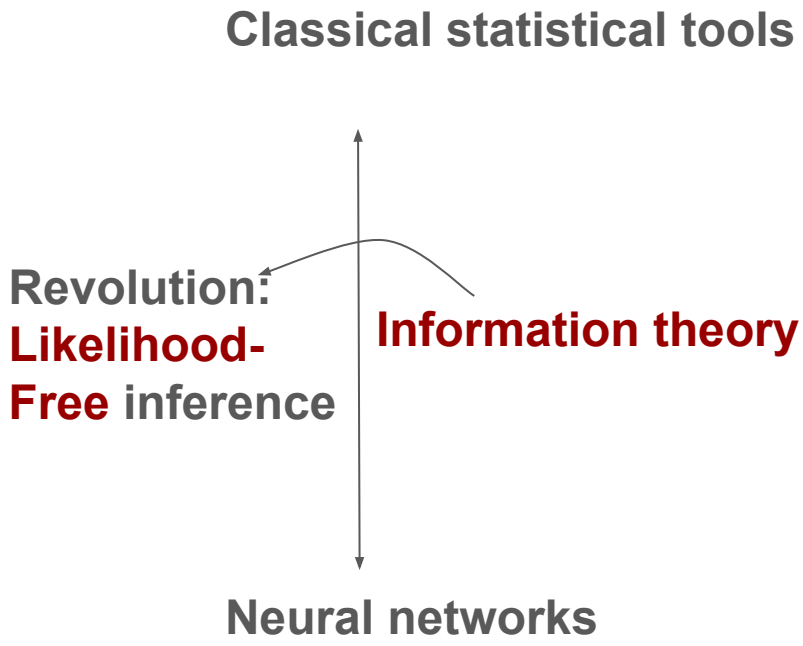
- Lecture 2: Likelihood methods in detail: **Poisson process likelihoods**
  -> obtain response-PDFs (Analytic / B-Splines)
  -> Tricks for systematics circumvention (1st hit likelihood)
  -> Parametrization: Stochastics
  -> Uncertainties

- Lecture 3: All things **neural networks**
  -> Neural network-based likelihoods
  -> Inverse likelihood-free methods
  -> Encodings and symmetries (Transformers etc.)
  -> Putting it all together



Overview

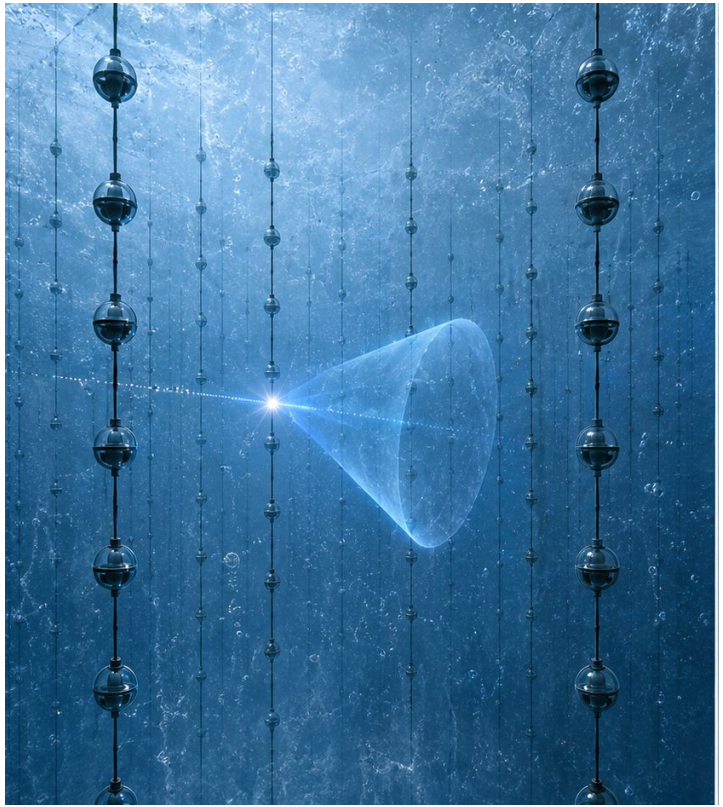
- Lecture 1: Introduction and **high-level overview**
  - **->** Bayes theorem
  - -> Frequentist vs Bayesian
  - -> Information-theoretic grounding (KL-divergence of the “joint”)
  - -> Theme of these lectures: likelihood vs likelihood-free
- Lecture 2: Likelihood methods in detail: **Poisson process likelihoods**
  - -> obtain response-PDFs (Analytic / B-Splines)
  - -> Tricks for systematics circumvention (1st hit likelihood)
  - -> Parametrization: Stochastics
  - -> Uncertainties
- Lecture 3: All things **neural networks**
  - -> Neural network-based likelihoods
  - -> Inverse likelihood-free methods
  - -> Encodings and symmetries (Transformers etc.)
  - -> Putting it all together



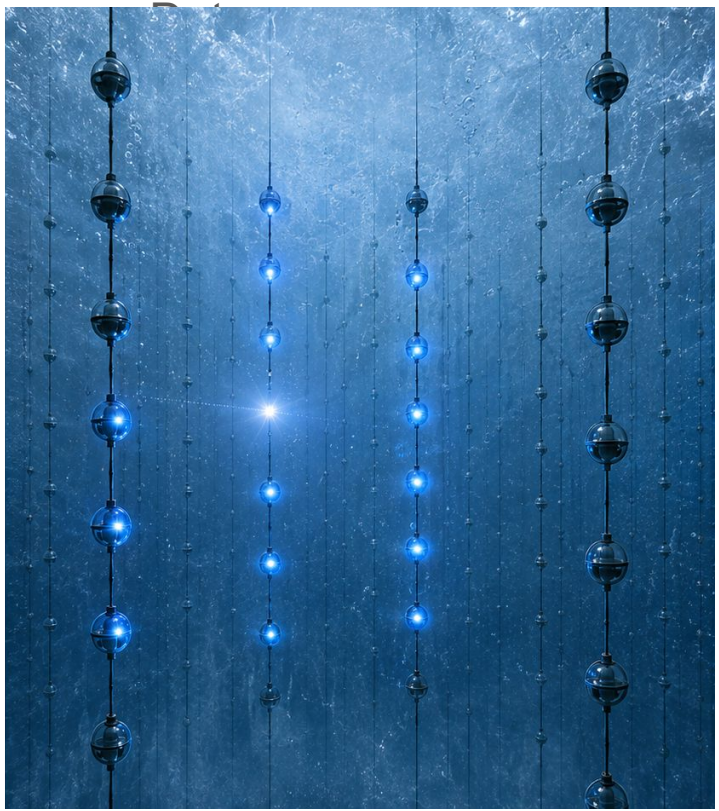
Learning objectives lecture 1:

- Bayes' Theorem and the "joint" distribution
- Frequentist / Bayesian parameter estimation
- Information theory: KL divergence and its interpretation
- likelihood-based vs likelihood-free methods

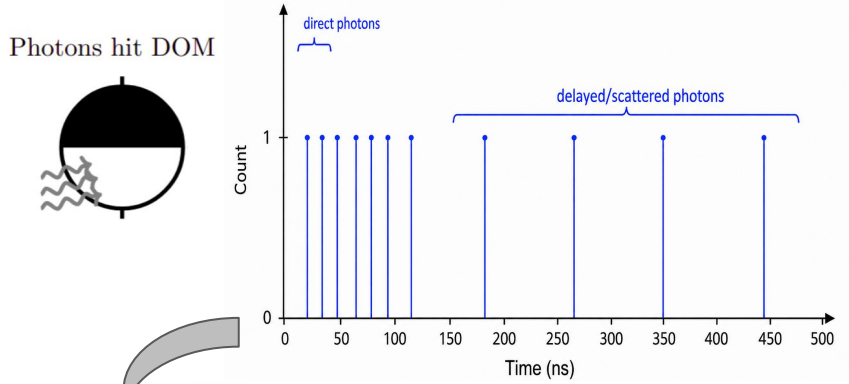
Interaction happens…



What we actually see later

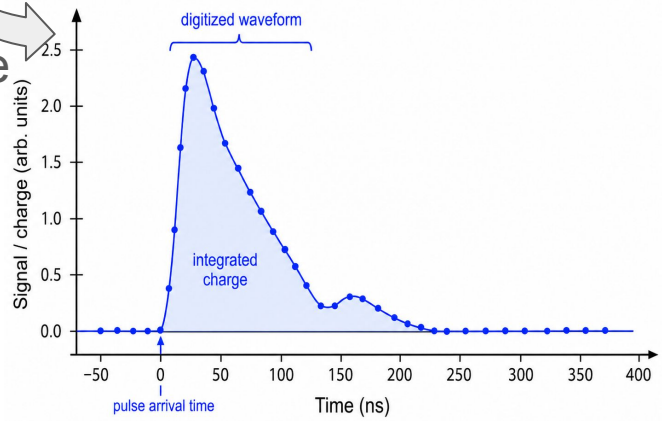


Physical photon hit times



PMT Response

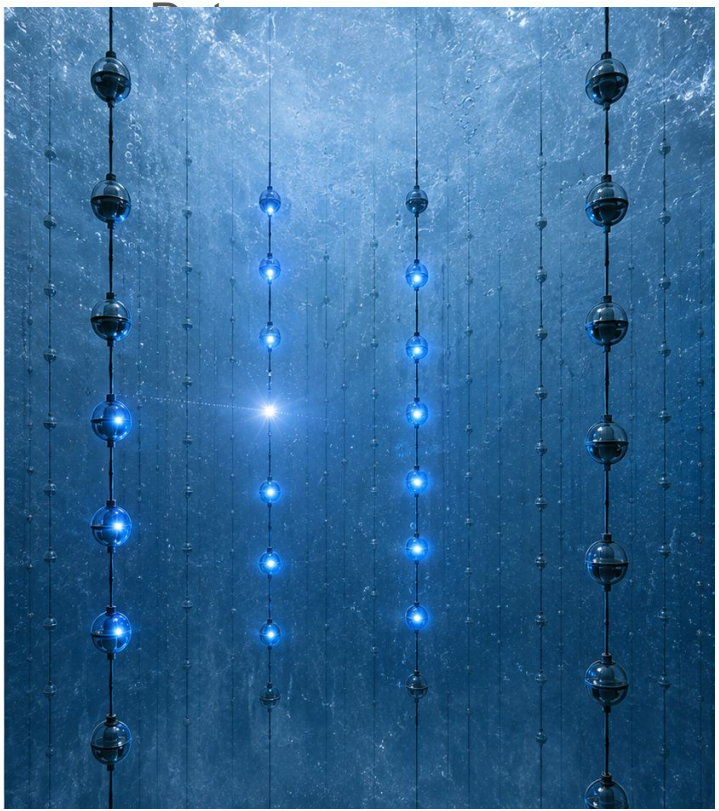
Schematic charge response



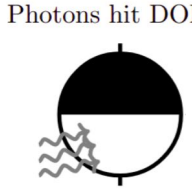
Data “**x**”:

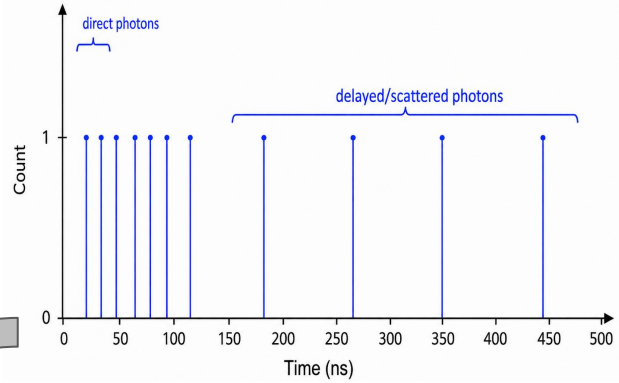
Total charge, charge vs time, leading edge

What we actually see later



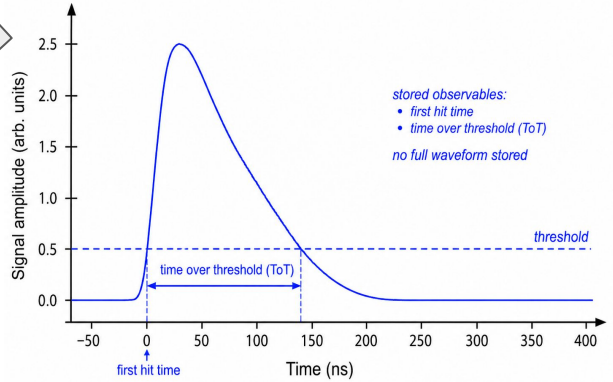
Physical photon hit times





Schematic charge response (KM3NeT)

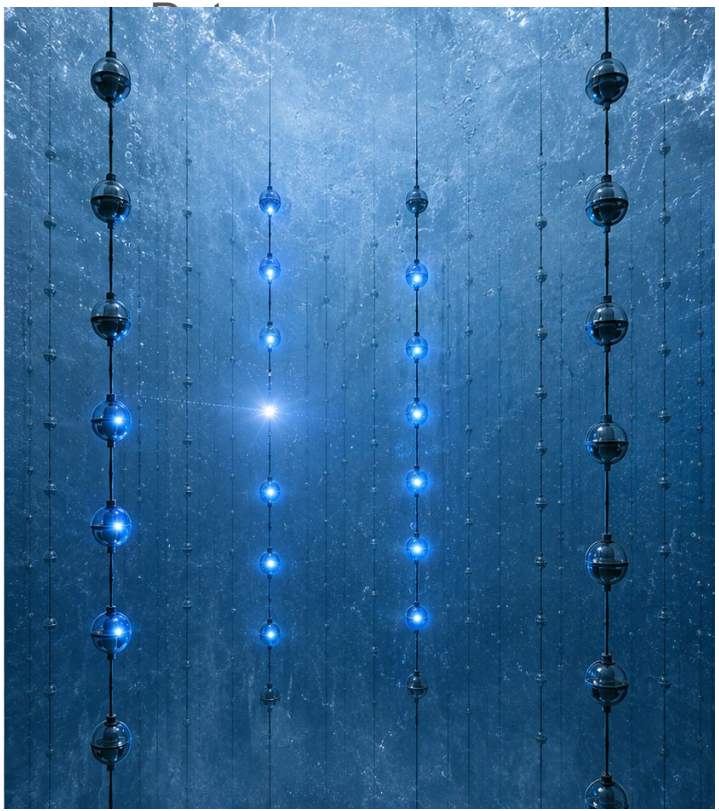
PMT Response



Data “**x**”:

leading edge, ToT

What we actually see later

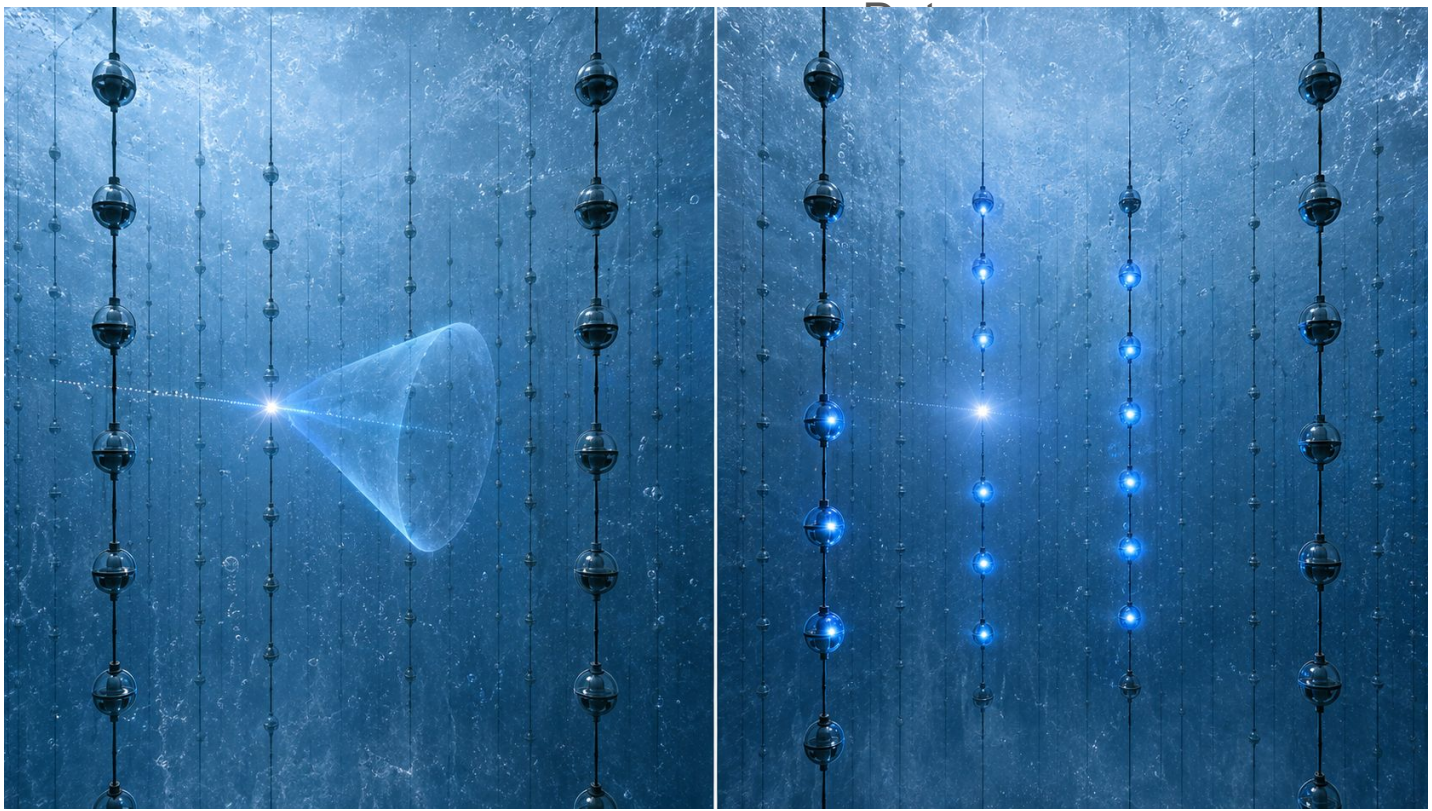


Event parameters θ:

θ={Direction, Energy, Position, Type, anything we record in MC,...}

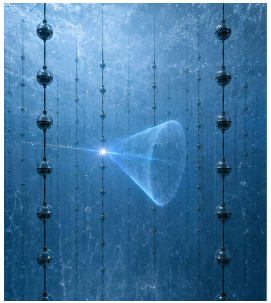
Data “x”

$x=\{x_1, x_2, \ldots, x_N\}$ for N modules



Event parameters θ:

θ={Direction, Energy, Position, Type, anything we record in MC,...}

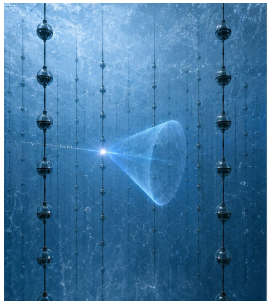


Data “x”

$x=\{x_1,x_2, \ldots, x_N\}$ for N modules

Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

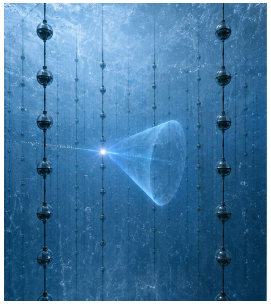
Data “x”
$x=\{x_1,x_2, \ldots, x_N\}$ for N modules



We know direction, energy etc. influence the data
-> correlations exist!
There is a non-trivial joint distribution p(θ,x) !

Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data “x”
x={x_1,x_2, …, x_N} for N modules



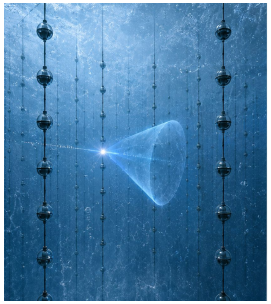
We know direction, energy etc. influence the data
-> correlations exist!
There is a non-trivial joint distribution p(θ,x) !

$$p(\theta, x) = p(\theta|x) \cdot p(x) = p(x|\theta) \cdot p(\theta)$$

Law of conditional probabilities

Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
$x=\{x_1, x_2, \ldots, x_N\}$ for N modules



We know direction, energy etc. influence the data
-> correlations exist!
There is a non-trivial joint distribution p(θ,x) !

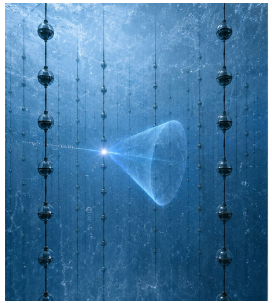
$$p(\theta, x) = p(\theta|x) \cdot p(x) = p(x|\theta) \cdot p(\theta)$$

Bayes' Theorem →

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
x={x_1,x_2, …, x_N} for N modules



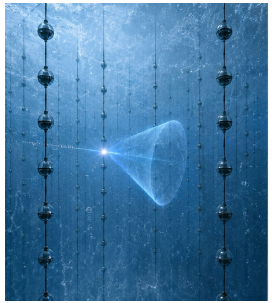
Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
x={x_1,x_2, …, x_N} for N modules

Bayes' Theorem



$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

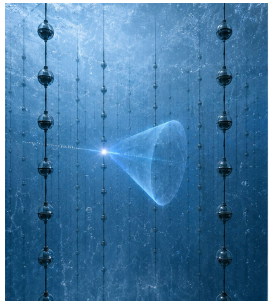
Posterior
(anti-causal / inverse)

Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

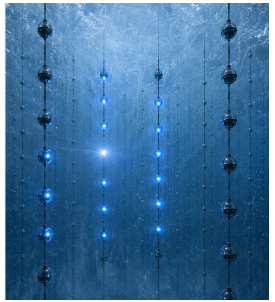
Data "x"
$x=\{x_1, x_2, \ldots, x_N\}$ for N modules

Bayes' Theorem

Prior (apriori parameter assumptions)



$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$



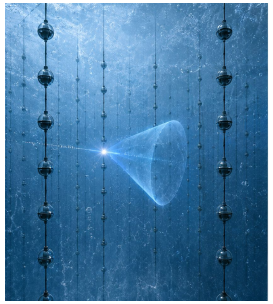
Posterior
(anti-causal / inverse)

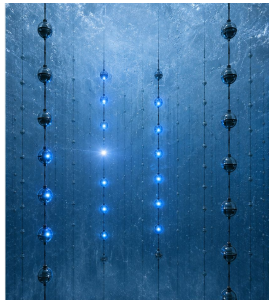
Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
x={x_1,x_2, …, x_N} for N modules

Bayes' Theorem

Prior (apriori parameter assumptions)





$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Posterior
(anti-causal / inverse)

Response function /
Data generation function /
Forward model
(Causal direction)

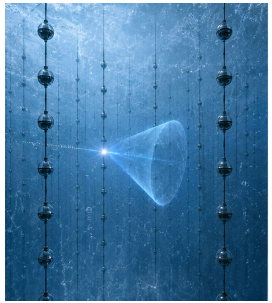
Likelihood $p(x|\theta) \equiv \mathcal{L}(\theta)$

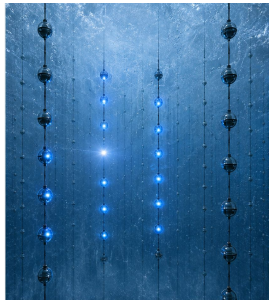
Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
$x=\{x_1, x_2, \ldots, x_N\}$ for N modules

Bayes' Theorem

Prior (apriori parameter assumptions)





$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Posterior
(anti-causal / inverse)

Response function /
Data generation function /
Forward model
(Causal direction)

Likelihood $p(x|\theta) \equiv \mathcal{L}(\theta)$

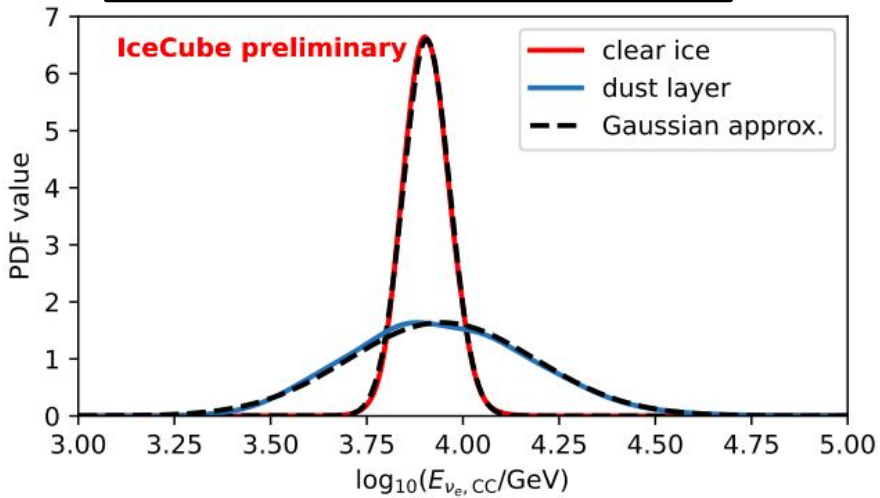
Marginal likelihood / Evidence
(constant w.r.t. θ)

All reconstruction methods:

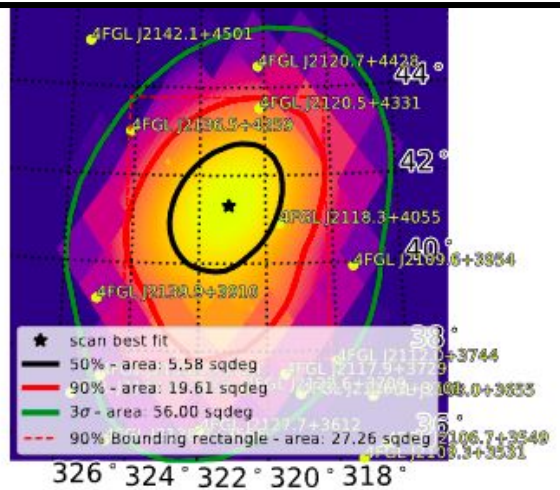
$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Utilize different parts of “Bayes’ Theorem” to construct confidence / credible intervals

Posterior over “log-energy”



Confidence Region over direction (likelihood-scan)



All reconstruction methods:

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Utilize different parts of “Bayes’ Theorem” to construct confidence / credible intervals

Interactive Question 1:

Which part of Bayes’ Theorem do we need for both Frequentist and Bayesian parameter estimation?

Why?

Hint: There are 4 possible parts!

p(x) p(x|θ)

p(θ|x) p(θ)

Learning objectives lecture 1:

- Bayes' Theorem and the "joint" distribution
- **Frequentist / Bayesian parameter estimation**
- Information theory: KL divergence and its interpretation
- likelihood-based vs likelihood-free methods

Event parameters θ:

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

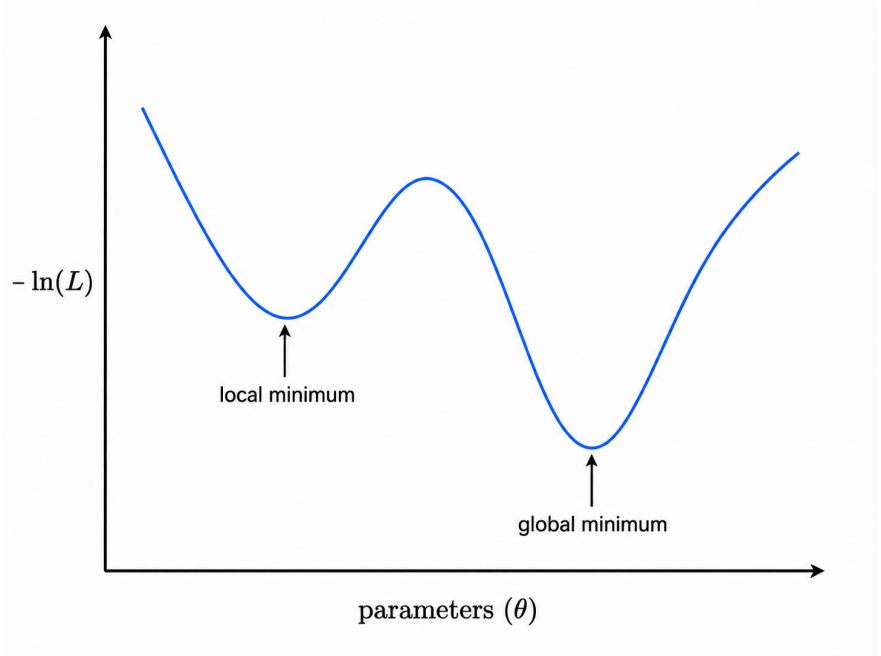
Data "x"
$x=\{x_1, x_2, \ldots, x_N\}$ for N modules

In Practice:
Maximum of log(L)
-> **Minimum of -log(L)**

Maximum Likelihood
(Fisher, Pearson, Wilks', 20s / 30s)

Response function /
Data generation function /
Forward model
(Causal direction)

Likelihood $p(x|\theta) \equiv \mathcal{L} \equiv L$



Event parameters θ:

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

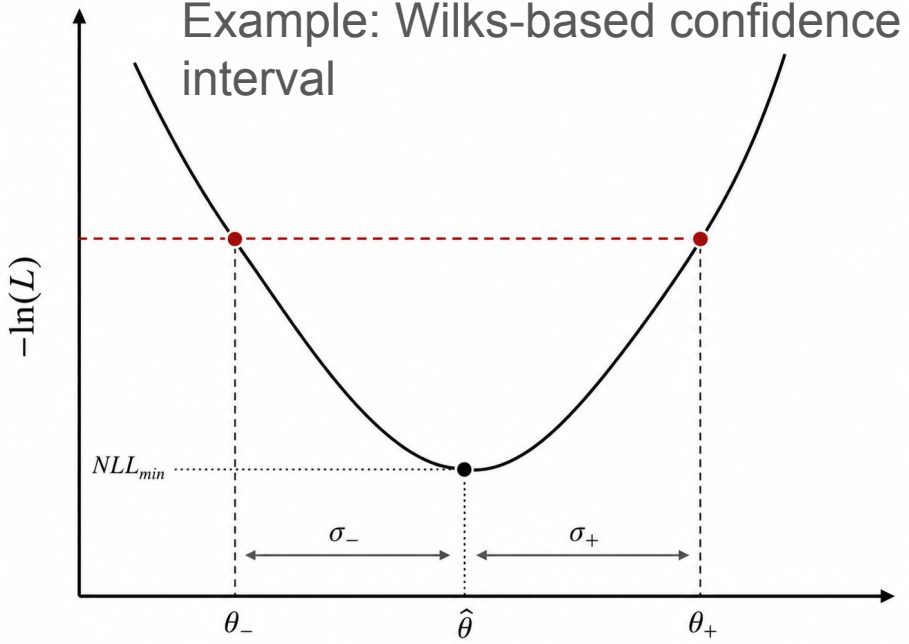
Data "x"
$x=\{x_1, x_2, \ldots, x_N\}$ for N modules

Maximum Likelihood
(Fisher, Pearson, Wilks', 20s / 30s)

Response function /
Data generation function /
Forward model
(Causal direction)

Likelihood $p(x|\theta) \equiv \mathcal{L}(\theta)$

Example: Wilks-based confidence interval



Event parameters θ:

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Data "x"
$x=\{x_1, x_2, \ldots, x_N\}$ for N modules

Maximum Likelihood
(Fisher, Pearson, Wilks', 20s / 30s)

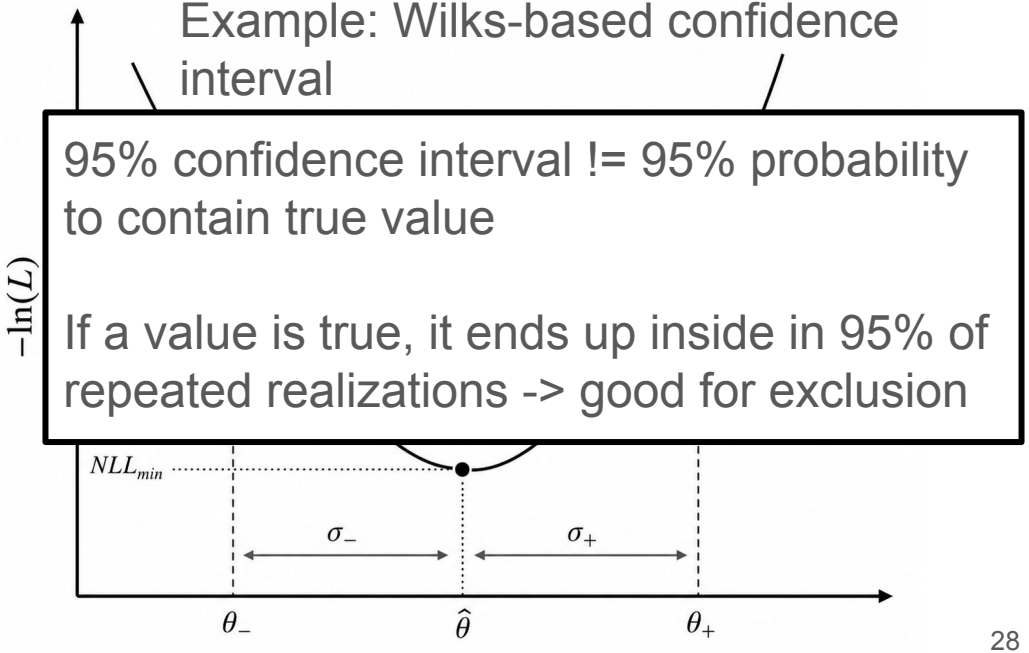
Response function /
Data generation function /
Forward model
(Causal direction)

Likelihood $p(x|\theta) \equiv \mathcal{L}(\theta)$

Example: Wilks-based confidence interval

95% confidence interval != 95% probability to contain true value

If a value is true, it ends up inside in 95% of repeated realizations -> good for exclusion

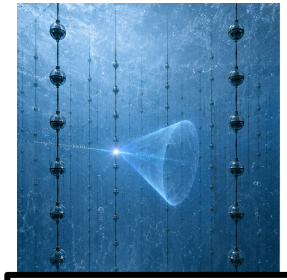


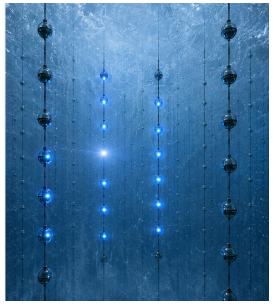
Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
x={x_1,x_2, …, x_N} for N modules

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$



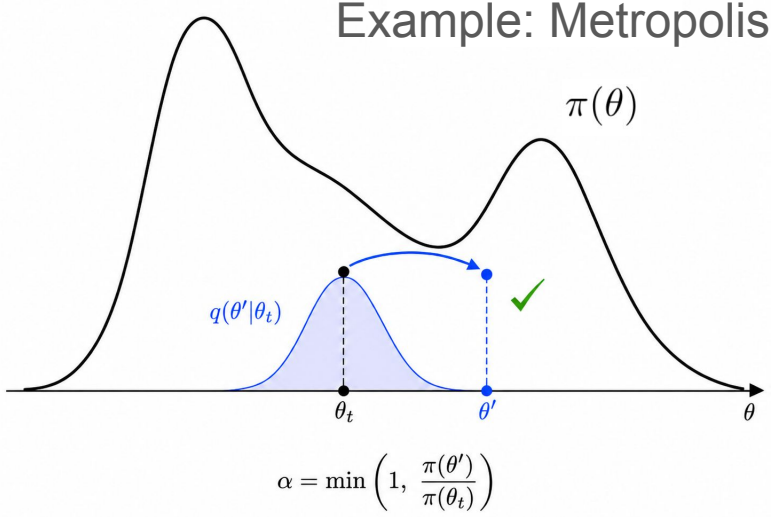


Markov-Chain Monte Carlo
(Bayesian, '50s-today)

Likelihood and Prior

$$p(x|\theta) \cdot p(\theta) \equiv \pi(\theta)$$

Example: Metropolis

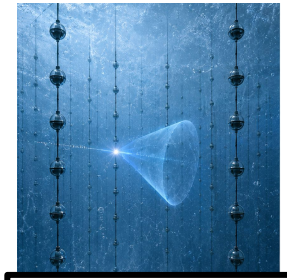


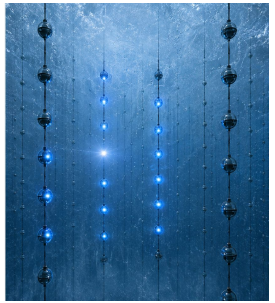
Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
$x=\{x_1, x_2, \ldots, x_N\}$ for N modules

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$





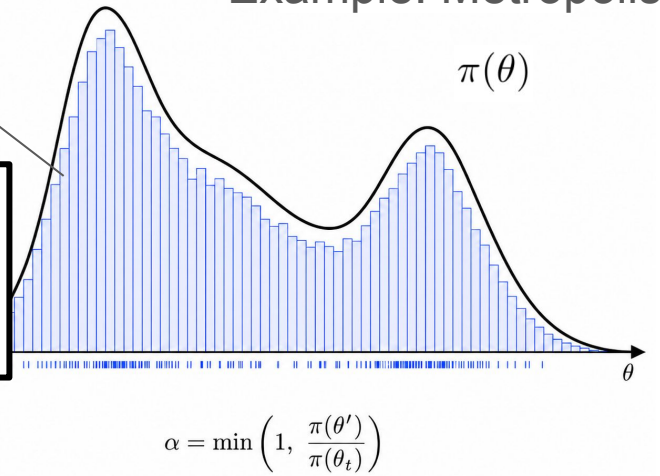
Markov-Chain Monte Carlo
(Bayesian, '50s-today)

Likelihood and Prior

$$p(x|\theta) \cdot p(\theta) \equiv \pi(\theta)$$

Samples From Posterior

Example: Metropolis

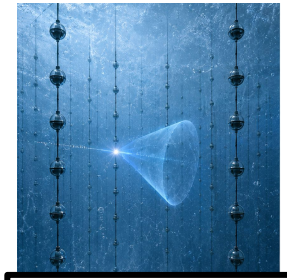


Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
x={x_1,x_2, …, x_N} for N modules

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{\int p(x|\theta) \cdot p(\theta)\, d\theta}$$



95% Credible interval = 95% probability to contain true value under Prior

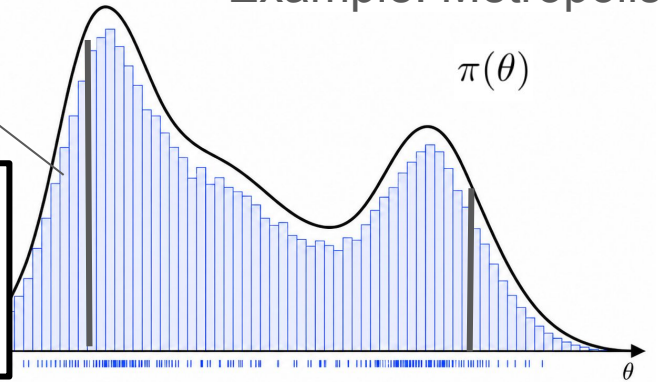
Markov-Chain Monte Carlo (Bayesian, '50s-today)

Likelihood and Prior

$$p(x|\theta) \cdot p(\theta) \equiv \pi(\theta)$$

Samples From Posterior

Example: Metropolis



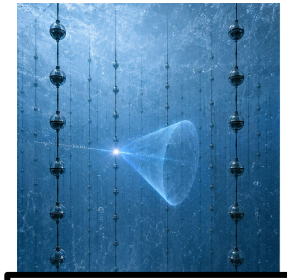
$$\alpha = \min\left(1, \frac{\pi(\theta')}{\pi(\theta_t)}\right)$$

Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
x={x_1,x_2, …, x_N} for N modules

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{\int p(x|\theta) \cdot p(\theta)\, d\theta}$$



95% Credible interval = 95% probability to contain true value under Prior

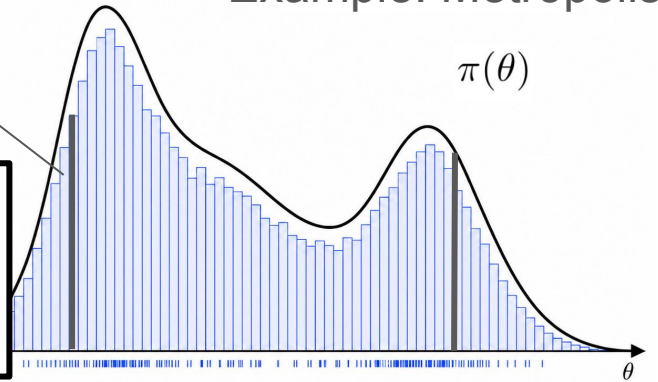
Markov-Chain Monte Carlo (Bayesian, '50s-today)

Likelihood and Prior

$$p(x|\theta) \cdot p(\theta) \equiv \pi(\theta)$$

Samples From Posterior

Example: Metropolis



$$\alpha = \min\left(1, \frac{\pi(\theta')}{\pi(\theta_t)}\right)$$

Event parameters θ:

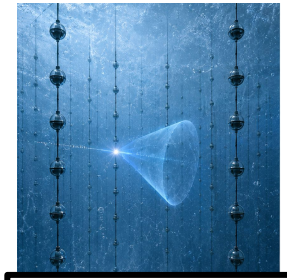
θ={Direction, Energy, Position, Type, anything we record in MC,...}

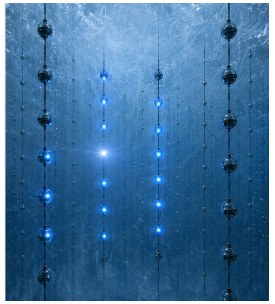
Data "x"

x={x_1,x_2, …, x_N} for N modules

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

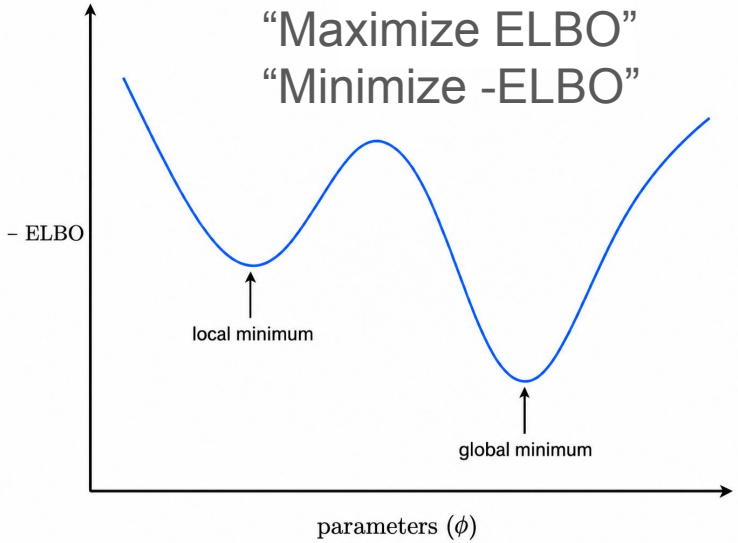




Variational Inference -> Maximize ELBO
(Bayesian, '90s-today)

$$p(x) \geq \underbrace{\int_{\theta} q_\phi(\theta) \cdot \left[\ln p(x;\theta) - \ln \frac{q_\phi(\theta)}{p(\theta)} \right] d\theta}_{\text{evidence lower bound (ELBO)}}$$

"Maximize ELBO"
"Minimize -ELBO"

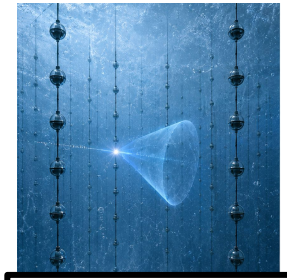


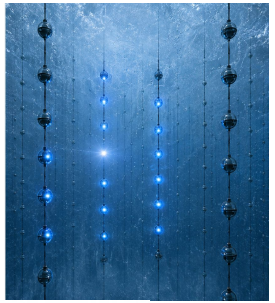
Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
x={x_1,x_2, …, x_N} for N modules

Bayes' Theorem

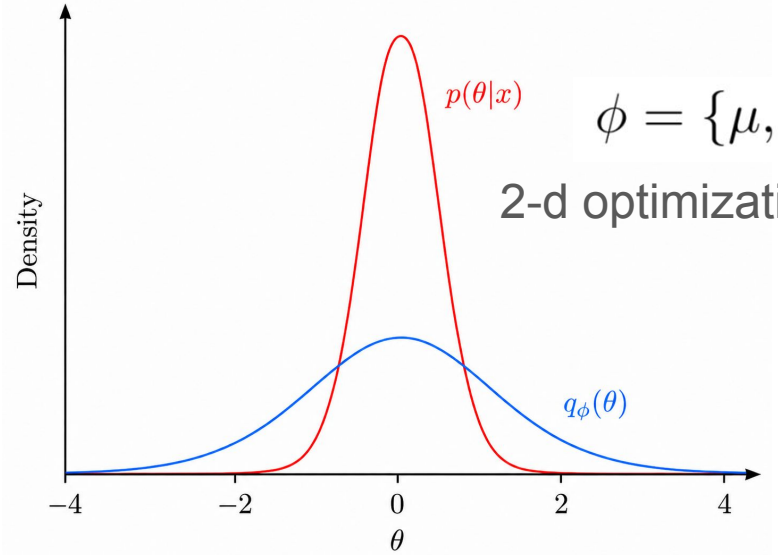
$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$





Variational Inference -> Maximize ELBO (Bayesian, '90s-today)

$$p(x) \geq \underbrace{\int_\theta q_\phi(\theta) \cdot \left[\ln p(x;\theta) - \ln \frac{q_\phi(\theta)}{p(\theta)} \right] d\theta}_{\text{evidence lower bound (ELBO)}}$$

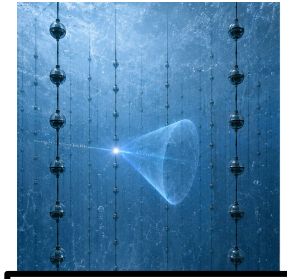


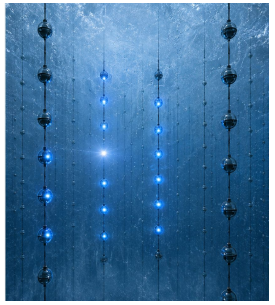
Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
x={x_1,x_2, …, x_N} for N modules

Bayes' Theorem

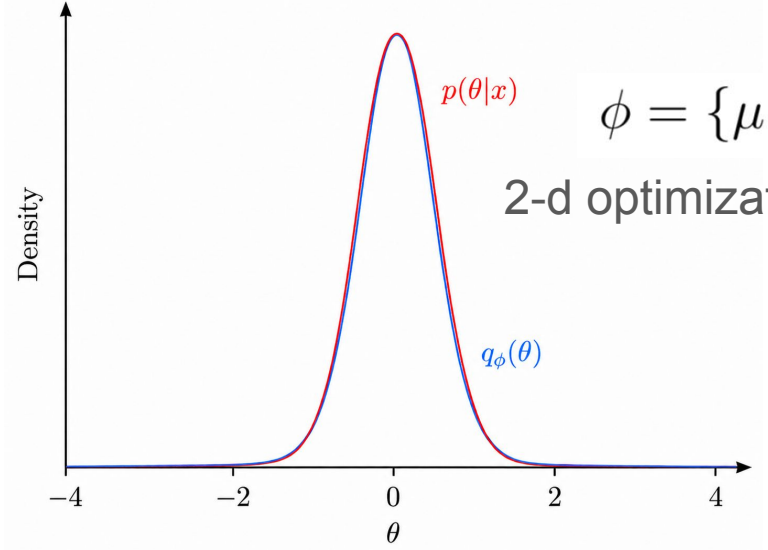
$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$





Variational Inference -> Maximize ELBO (Bayesian, '90s-today)

$$p(x) \geq \underbrace{\int_\theta q_\phi(\theta) \cdot \left[\ln p(x;\theta) - \ln \frac{q_\phi(\theta)}{p(\theta)}\right] d\theta}_{\text{evidence lower bound (ELBO)}}$$



$$\phi = \{\mu, \sigma\}$$

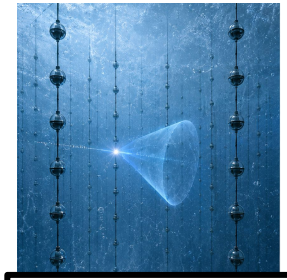
2-d optimization

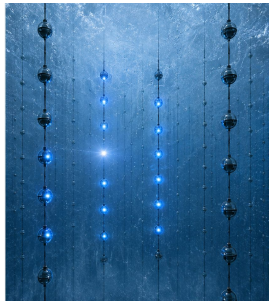
Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
$x=\{x_1, x_2, \ldots, x_N\}$ for N modules

Bayes' Theorem

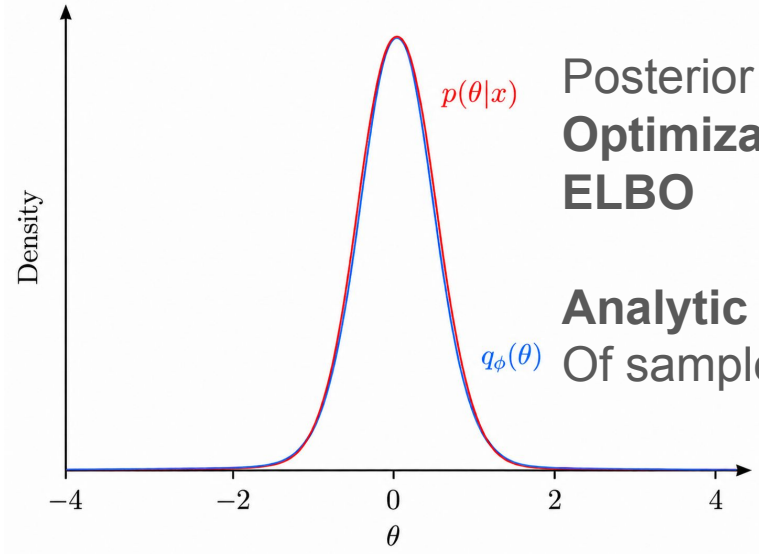
$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$





Variational Inference -> Maximize ELBO
(Bayesian, '90s-today)

$$p(x) \geq \underbrace{\int_\theta q_\phi(\theta) \cdot \left[\ln p(x;\theta) - \ln \frac{q_\phi(\theta)}{p(\theta)}\right] d\theta}_{\text{evidence lower bound (ELBO)}}$$



Posterior via
Optimization of ELBO

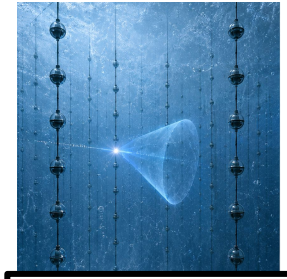
Analytic instead
Of samples

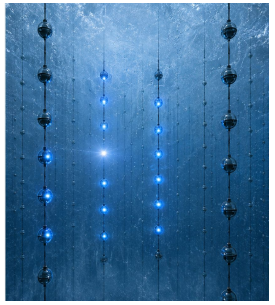
Event parameters θ:
θ={Direction, Energy, Position, Type, anything we record in MC,...}

Data "x"
$x=\{x_1, x_2, \ldots, x_N\}$ for N modules

Bayes' Theorem

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$





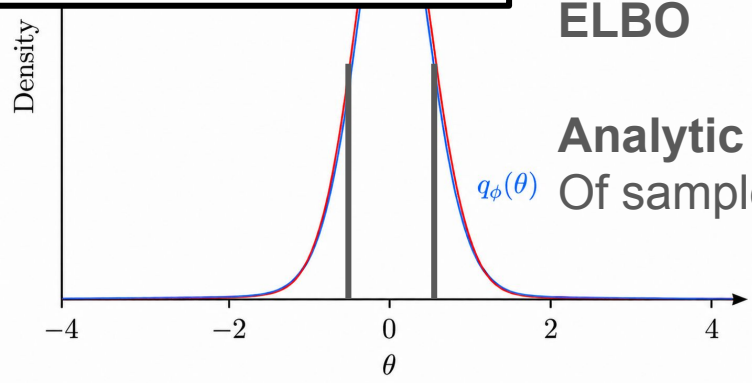
Variational Inference -> M
(Bayesian, '90s-today)

95% Credible interval = 95% probability to contain true value under Prior

Posterior via **Optimization of ELBO**

Analytic instead Of samples

$$p(x) \geq \underbrace{\int_\theta q_\phi(\theta) \cdot \left[\ln p(x;\theta) - \ln \frac{q_\phi(\theta)}{p(\theta)}\right] d\theta}_{\text{evidence lower bound (ELBO)}}$$



Observation: In both Frequentist (Max. likelihood) and traditional Bayesian methods (MCMC / variational inference) we require the likelihood function!

Interactive Question 2:

What are the pros / cons of relying on the likelihood function?

Try to find at least 1 pro / con

Observation: In both Frequentist (Max. likelihood) and traditional Bayesian methods (MCMC / variational inference) we require the likelihood function!

Question: What are the pros / cons of relying on the likelihood function?

Discuss! (5 min)

Pros:

- Well defined and mature statistical tools use it
- Frequentist guarantees (asymptotically optimal, saturating Cramer-Rao bound, etc.)
- Established methods use it: MCMC / variational inference to get posterior in traditional Bayesian analysis

Observation: In both Frequentist (Max. likelihood) and traditional Bayesian methods (MCMC / variational inference) we require the likelihood function!

Question: What are the pros / cons of relying on the likelihood function?

Discuss! (5 min)

Pros:

- Well defined and mature statistical tools use it
- Frequentist guarantees (asymptotically optimal, saturating Cramer-Rao bound, etc.)
- Established methods use it: MCMC / variational inference to get posterior in traditional Bayesian analysis

Cons:

- We need to model it first! Requires likelihood fit for the likelihood function itself!
  -> Numerically challenging! (**next slides, lecture 2**)
- High-dimensional physics hypothesis have many local optima!
  -> Complexity, part 2!
  (**lecture 2**)

1) How to cleanly define “likelihood fit of likelihood?”
2) How could we possibly do likelihood-free modelling?

-> Information-theoretic viewpoint

Learning objectives lecture 1:

- Bayes' Theorem and the "joint" distribution
- Frequentist / Bayesian parameter estimation
- **Information theory: KL divergence and its interpretation**
- likelihood-based vs likelihood-free methods

Information theory

Established in the 40s by Claude Shannon
(probabilities <-> coding theory)

(40s - Shannon):
Shannon-Entropy: (discrete)

$$H(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

(50s - Kullback / Leibler):
Relative Entropy / KL-Divergence:

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Information theory

Established in the 40s by Claude Shannon
(probabilities <-> coding theory)

(40s - Shannon):
Differential Entropy: (continuous)

$$H(p) = -\int p(x) \log p(x)\, dx$$

(50s - Kullback / Leibler):
Relative Entropy / KL-Divergence:

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

Information theory

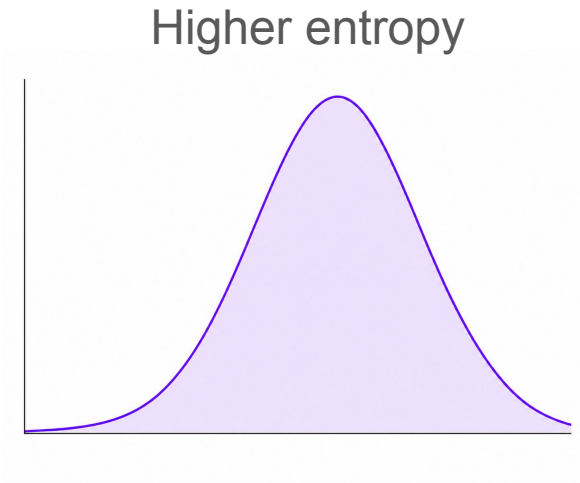
Established in the 40s by Claude Shannon
(probabilities <-> coding theory)

(40s - Shannon):
Differential Entropy: (continuous)

$$H(p) = -\int p(x) \log p(x)\, dx$$

(50s - Kullback / Leibler):
Relative Entropy / KL-Divergence:

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$



Information theory

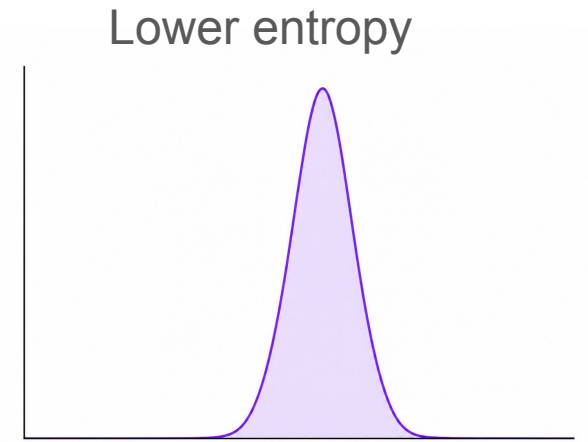
Established in the 40s by Claude Shannon
(probabilities <-> coding theory)

(40s - Shannon):
Differential Entropy: (continuous)

$$H(p) = -\int p(x) \log p(x)\, dx$$

(50s - Kullback / Leibler):
Relative Entropy / KL-Divergence:

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$



Information theory

Established in the 40s by Claude Shannon
(probabilities <-> coding theory)

(40s - Shannon):
Differential Entropy: (continuous)

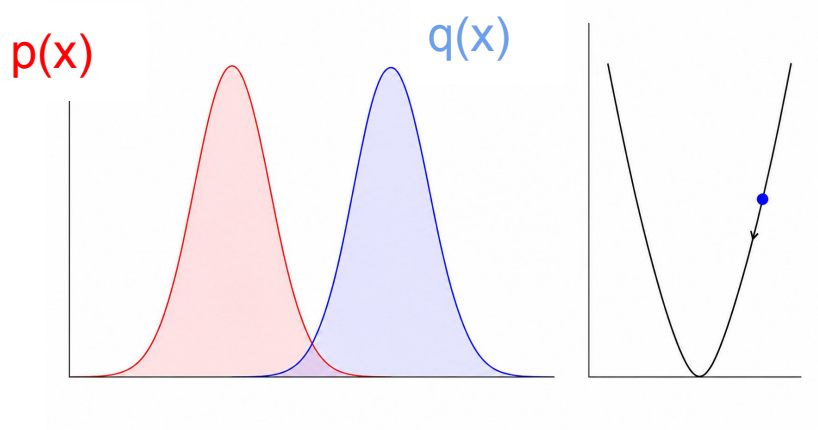
$$H(p) = -\int p(x) \log p(x)\, dx$$

(50s - Kullback / Leibler):
Relative Entropy / KL-Divergence:

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

Notation: q(x) is parametrized
Approximates p(x) / P(x)

Higher KL



Information theory

Established in the 40s by Claude Shannon
(probabilities <-> coding theory)

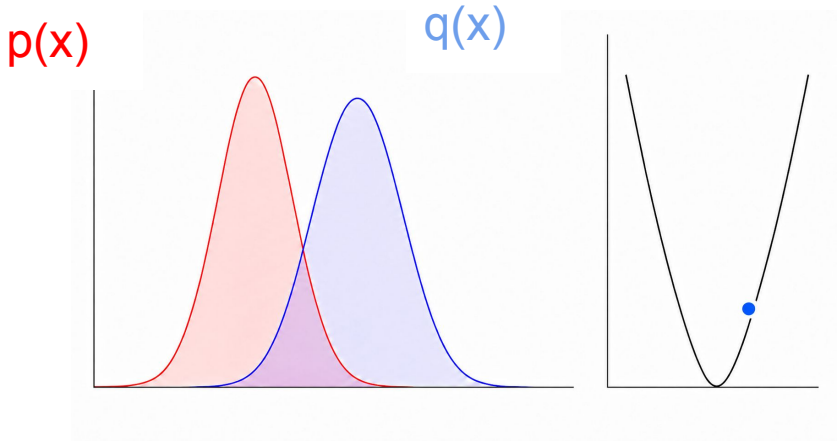
(40s - Shannon):
Differential Entropy: (continuous)

$$H(p) = -\int p(x) \log p(x)\, dx$$

(50s - Kullback / Leibler):
Relative Entropy / KL-Divergence:

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

Lower KL (0 = distributions the same)



Information theory

Established in the 40s by Claude Shann
(probabilities <-> coding theory)

$$-\int p(x)\log q(x)\,dx - \left(-\int p(x)\log p(x)\,dx\right)$$

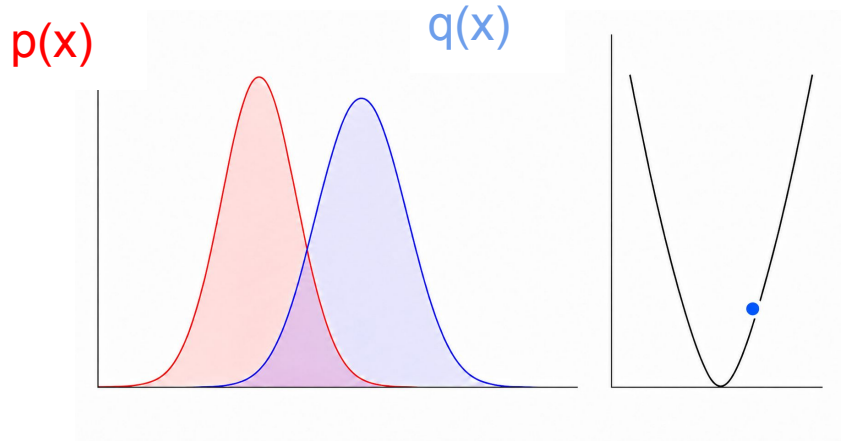
(40s - Shannon):
Differential Entropy: (continuous)

$$H(p) = -\int p(x)\log p(x)\,dx$$

(50s - Kullback / Leibler):
Relative Entropy / KL-Divergence:

$$D_{\mathrm{KL}}(p\parallel q) = \int p(x)\log\frac{p(x)}{q(x)}\,dx$$

Lower KL (0 = distributions the same)



Information theory

Established in the 40s by Claude Shann
(probabilities <-> coding theory)

(40s - Shannon):
Differential Entropy: (continuous)

$$H(p) = -\int p(x) \log p(x)\, dx$$

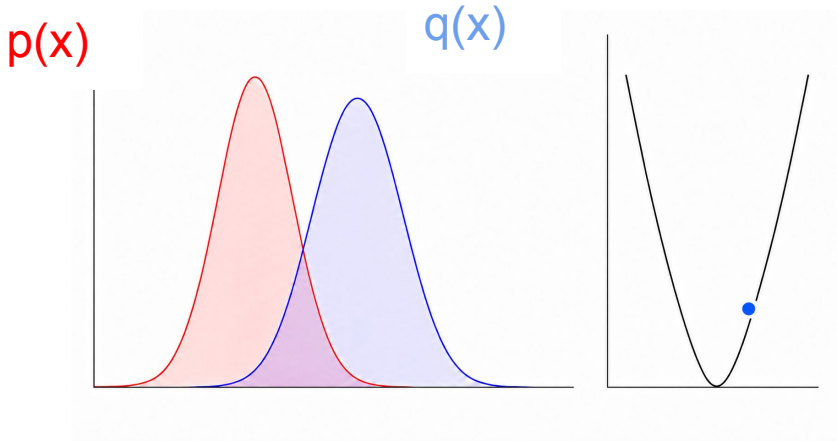
(50s - Kullback / Leibler):
Relative Entropy / KL-Divergence:

$$D_{\mathrm{KL}}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

$$-\int p(x) \log q(x)\, dx - \left(-\int p(x) \log p(x)\, dx\right)$$

cross-entropy — entropy -> const w.r.t q(x)

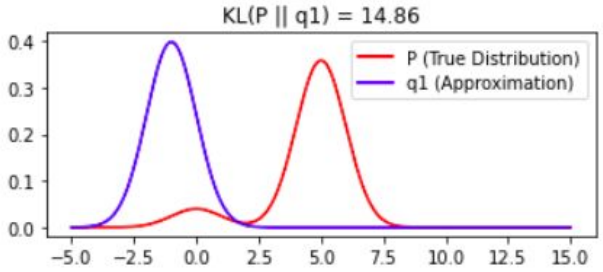
Lower KL (0 = distributions the same)

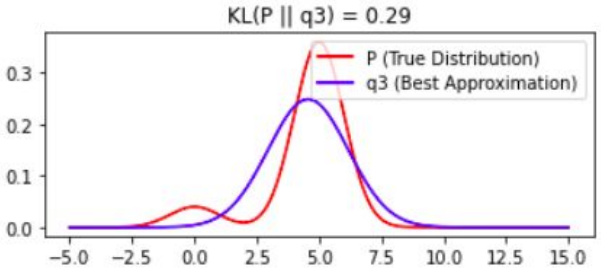


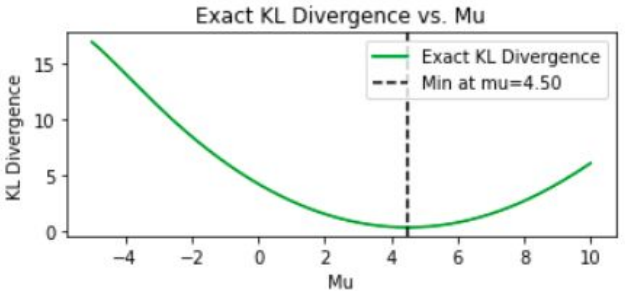
PDF-approximation matching via “forward” KL divergence

True distribution: P(x)
Approximation: q(x)

$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x) \ln\left(\frac{P(x)}{q_\theta(x)}\right) dx$$







PDF-approximation matching via “forward” KL divergence

In practice .. **We do not know P(x) -> we have samples from P(x)**

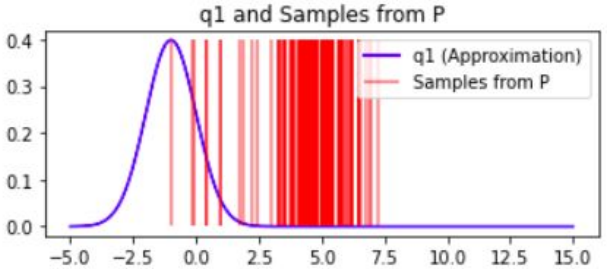
True distribution: P(x)

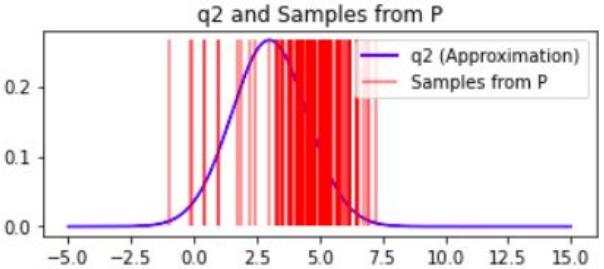
Approximation: q(x)

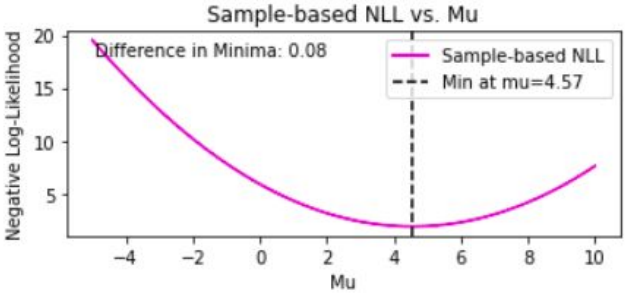
$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x)\,\ln\left(\frac{P(x)}{q_\theta(x)}\right)dx$$

$$D_{\mathrm{KL}}(P|q)[\theta] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i)}{q_\theta(x_i)}\right)$$

→ argmin(θ) $\frac{1}{N}\sum_i \ln\left(\frac{P(x_i)}{q_\theta(x_i)}\right)$ argmin(θ) $\frac{1}{N}\sum_i -\ln\left(q_\theta(x_i)\right)$







Interactive Question 3:

Why do we not use the so-called "reverse" KL divergence?

Hint: See the sample-based approximation?

Forward KL (previous slides):

$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x)\,\ln\left(\frac{P(x)}{q_\theta(x)}\right) dx$$

Reverse KL:

$$D_{\mathrm{KL}}(q|P)[\theta] = \int q_\theta(x)\,\ln\left(\frac{q_\theta(x)}{P(x)}\right) dx$$

P(x): True distribution (unknown) **q(x): Approximation**

Sample-based approximation →

$$D_{\mathrm{KL}}(q|P)[\theta] \approx \frac{1}{N}\sum_i \ln\left(\frac{q_\theta(x_i[\theta])}{P(x_i[\theta])}\right)$$

Question:

Why do we typically not use the "reverse" KL-divergence here?

Forward KL (previous slides):

$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x)\,\ln\left(\frac{P(x)}{q_\theta(x)}\right) dx$$

1) Here we do not have access to P(x)
Evaluate as sample-based expectation value (certain assumptions on q):

$$D_{\mathrm{KL}}(q|P)[\theta] \approx \frac{1}{N}\sum_i \ln\left(\frac{q_\theta(x_i[\theta])}{P(x_i[\theta])}\right)$$

Without P we cannot calculate the gradient

Reverse KL:

$$D_{\mathrm{KL}}(q|P)[\theta] = \int q_\theta(x)\,\ln\left(\frac{q_\theta(x)}{P(x)}\right) dx$$

Question:

Why do we typically not use the "reverse" KL-divergence here?

Forward KL (previous slides):

$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x)\,\ln\left(\frac{P(x)}{q_\theta(x)}\right) dx$$

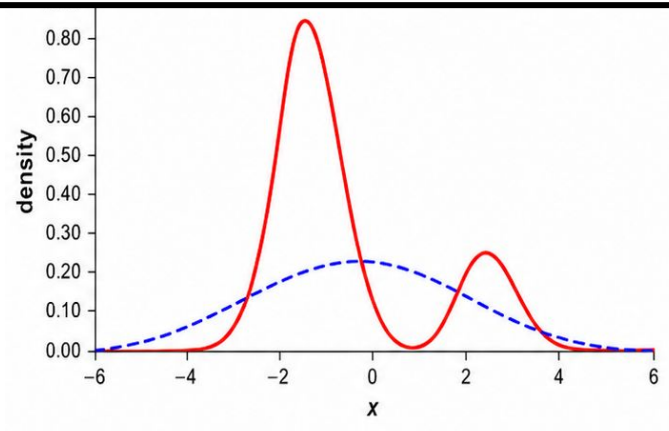
1) Here we do not have access to P(x)

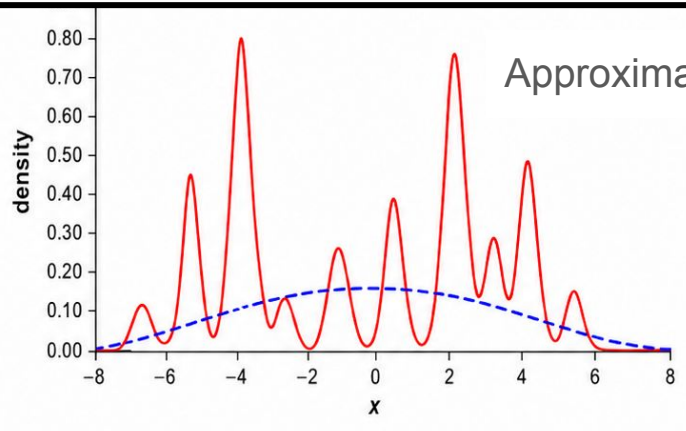
Reverse KL:

$$D_{\mathrm{KL}}(q|P)[\theta] = \int q_\theta(x)\,\ln\left(\frac{q_\theta(x)}{P(x)}\right) dx$$

2) The qualitative approximation properties
are different, and typically worse!

Forward KL: ~ approximation (blue) seeks out probability mass
Approximation = Gaussian: 1st (mean) and 2nd (variance) are matched

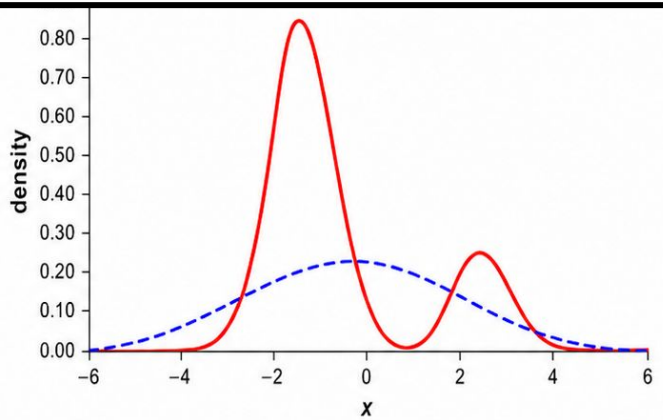


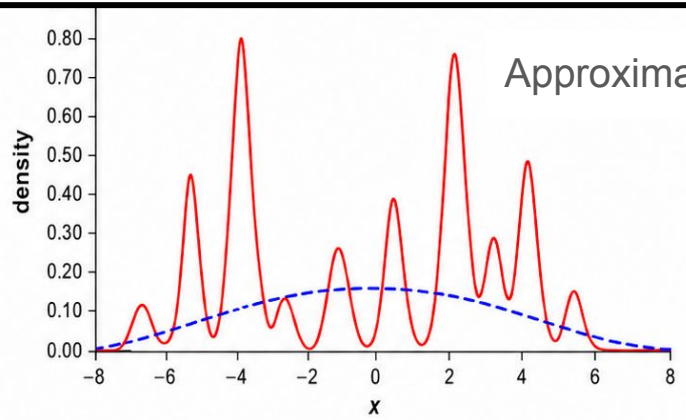


Approximation: Gaussian

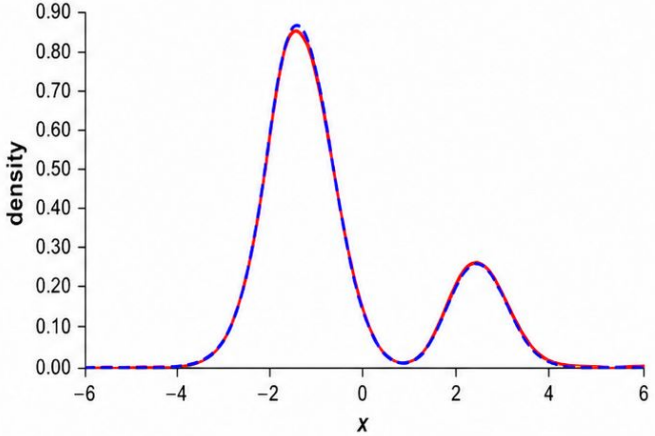
`S

Forward KL: ~ approximation (blue) seeks out probability mass
Approximation = Gaussian: 1st (mean) and 2nd (variance) are matched

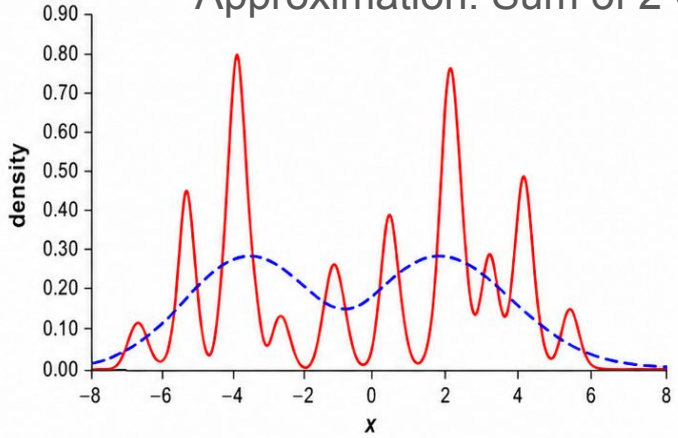




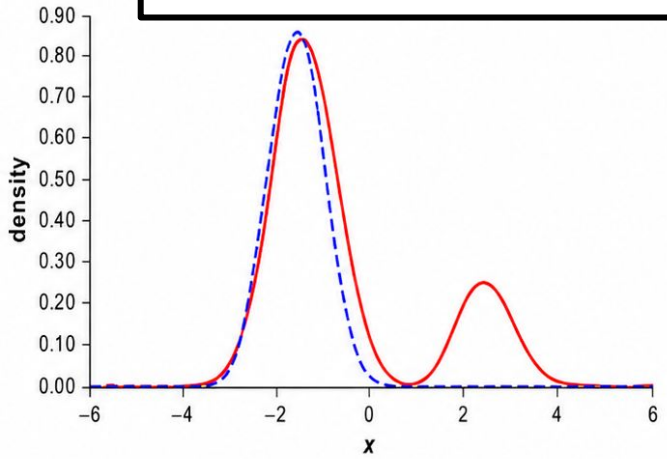
Approximation: Gaussian

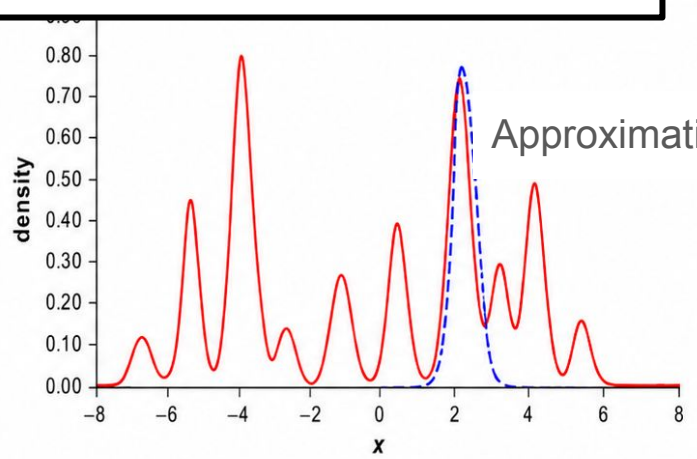


Approximation: Sum of 2 Gaussians



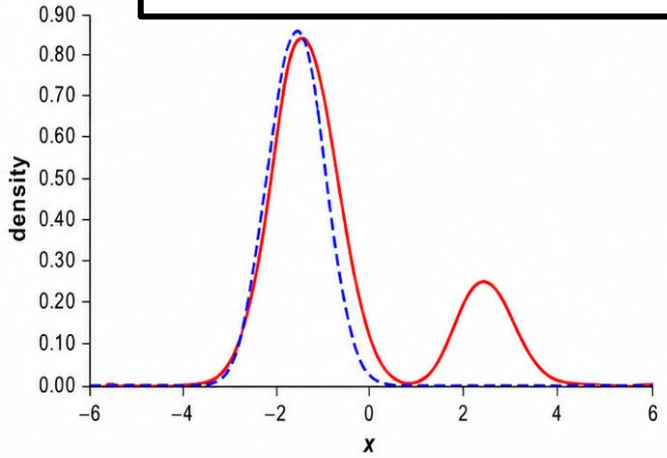
Reverse KL: ~ approximation (blue) seeks out modes

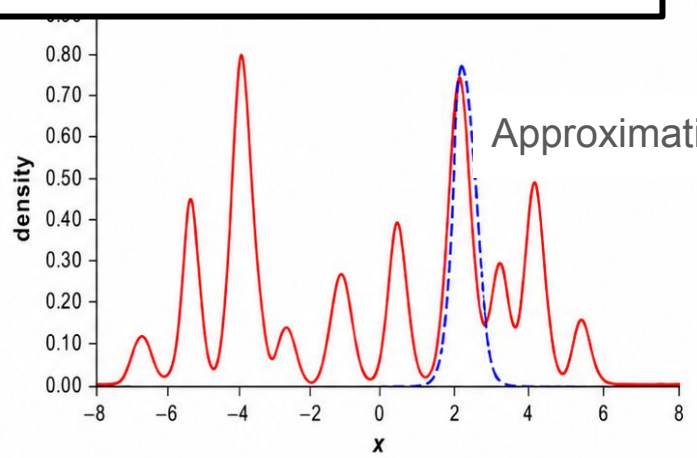




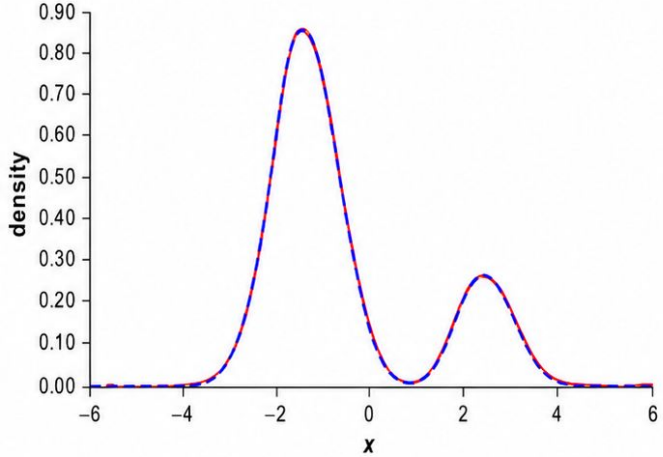
Approximation: Gaussian

Reverse KL: ~ approximation (blue) seeks out modes

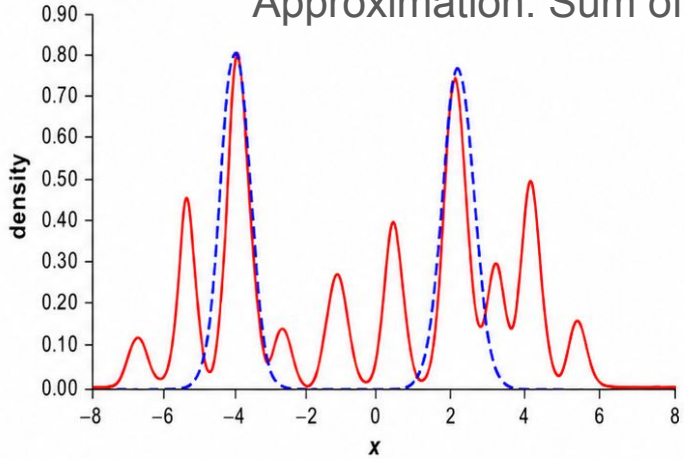




Approximation: Gaussian



Approximation: Sum of 2 Gaussians



PDF-approximation matching via “forward” KL divergence

In practice .. **We do not know P(x) -> we have samples from P(x)**

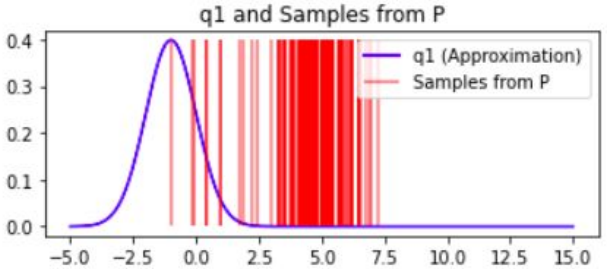
True distribution: P(x)

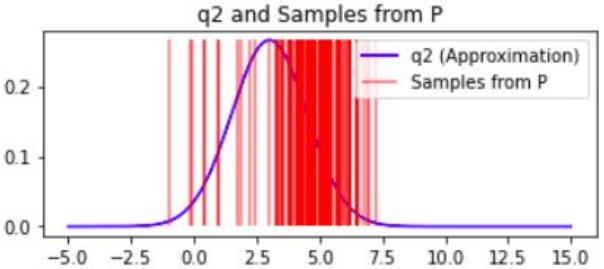
Approximation: q(x)

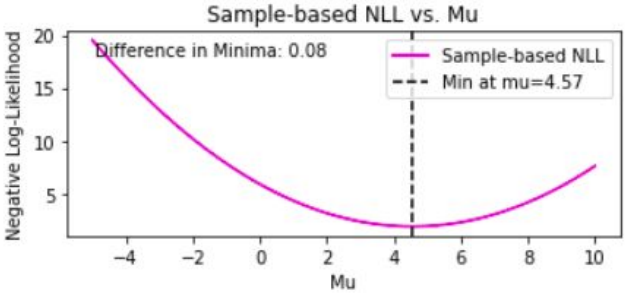
$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x)\,\ln\left(\frac{P(x)}{q_\theta(x)}\right)dx$$

$$D_{\mathrm{KL}}(P|q)[\theta] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i)}{q_\theta(x_i)}\right)$$

→ argmin(θ) $\frac{1}{N}\sum_i \ln\left(\frac{P(x_i)}{q_\theta(x_i)}\right)$ argmin(θ) $\frac{1}{N}\sum_i -\ln\left(q_\theta(x_i)\right)$







PDF-approximation matching via “forward” KL divergence

In practice .. **We do not know P(x) -> we have samples from P(x)**

True distribution: P(x)

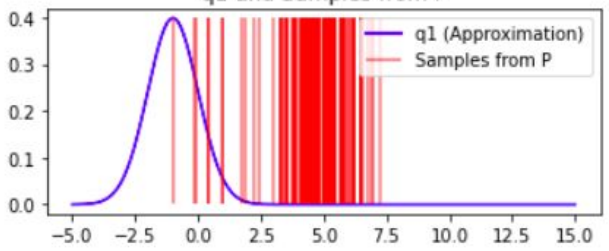
Approximation: q(x)

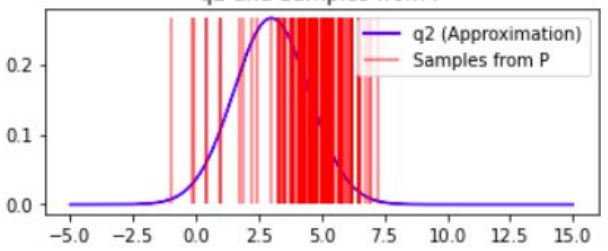
$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x|\theta_{true}) \ln\left(\frac{P(x|\theta_{true})}{q(x|\theta)}\right) dx$$

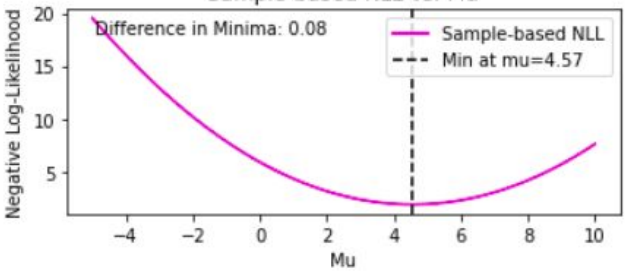
$$D_{\mathrm{KL}}(P|q)[\theta] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_{true})}{q(x_i|\theta)}\right)$$

(Slight rewrite! Same thing..)

⟶ argmin(θ) $\frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_{true})}{q(x_i|\theta)}\right)$ = argmin(θ) $\frac{1}{N}\sum_i -\ln\left(q(x_i|\theta)\right)$







PDF-approximation matching via “forward” KL divergence

True distribution: P(x)
Approximation: q(x)

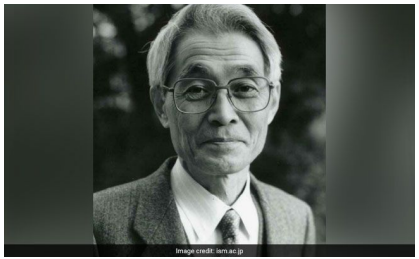
$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x|\theta_{true}) \ln\left(\frac{P(x|\theta_{true})}{q(x|\theta)}\right) dx$$

$$D_{\mathrm{KL}}(P|q)[\theta] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_{true})}{q(x_i|\theta)}\right)$$

⟶ argmin(θ) $\frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_{true})}{q(x_i|\theta)}\right)$ = argmin(θ) $\frac{1}{N}\sum_i -\ln\left(q(x_i|\theta)\right)$

Max. Likelihood for L(θ)=q(x|θ)

1973 (Akaike): **Maximum likelihood derivable from KL divergence**!



$$p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}$$

PDF-approximation matching via “forward” KL divergence

True distribution: P(x)
Approximation: q(x)

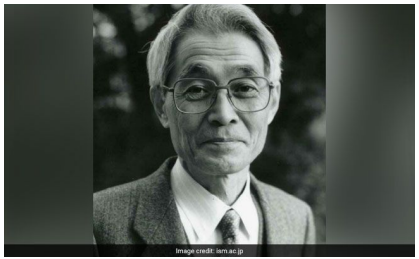
$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x|\theta_{true}) \ln\left(\frac{P(x|\theta_{true})}{q(x|\theta)}\right) dx$$

$$D_{\mathrm{KL}}(P|q)[\theta] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_{true})}{q(x_i|\theta)}\right)$$

⟶ argmin(θ) $\frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_{true})}{q(x_i|\theta)}\right)$ = argmin(θ) $\frac{1}{N}\sum_i -\ln\left(q(x_i|\theta)\right)$

Max. Likelihood for L(θ)=q(x|θ)

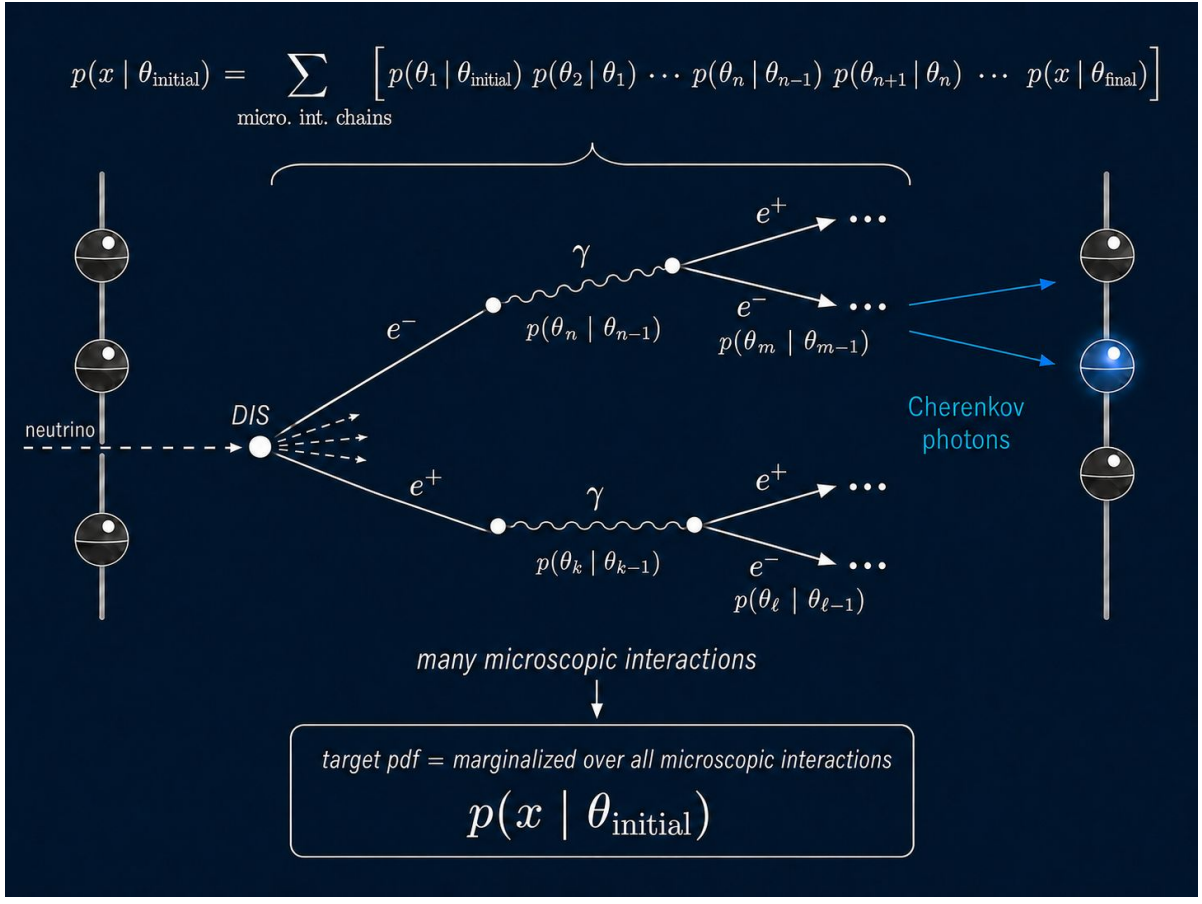
1973 (Akaike): **Maximum likelihood derivable from KL divergence**!



But: We still have to assume a parametrization of the Likelihood! **What is it?**

$$p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)}$$

The KL-divergence using the joint distribution in (x,θ)



Prior
$p(\theta_{initial})$
(energy distribution, zenith distribution,...)

Monte Carlo implements sampling from the joint:

$p(\theta_{initial})\, p(x|\theta_{initial})$

$=p(\theta_{initial}, x)$

$$\omega^\star = \arg\min_{\omega} \; D_{\mathrm{KL}}\Big(P(x,\theta) \;\Big\|\; q(\theta)\, q_\omega(x|\theta) \Big)$$

← “Joint” KL-divergence

$$\omega^\star = \arg\min_\omega \; D_{\mathrm{KL}}\Big(P(x,\theta) \;\Big\|\; q(\theta)\, q_\omega(x|\theta)\Big)$$

“Joint” KL-divergence

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x,\theta)}{q(\theta) q_\omega(x|\theta)}\right]$$

$$\omega^\star = \arg\min_\omega \; D_{\mathrm{KL}}\Big(P(x,\theta) \;\Big\|\; q(\theta)\, q_\omega(x|\theta)\Big)$$

← "Joint" KL-divergence

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x,\theta)}{q(\theta) q_\omega(x|\theta)}\right]$$

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x|\theta)}{q_\omega(x|\theta)}\right] + \mathbb{E}_{P(\theta)}\left[\log \frac{P(\theta)}{q(\theta)}\right]$$

$$\omega^\star = \arg\min_\omega \; D_{\mathrm{KL}}\Big(P(x,\theta) \;\Big\|\; q(\theta)\, q_\omega(x|\theta)\Big)$$

← "Joint" KL-divergence

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x,\theta)}{q(\theta) q_\omega(x|\theta)}\right]$$

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x|\theta)}{q_\omega(x|\theta)}\right] + \mathbb{E}_{P(\theta)}\left[\log \frac{P(\theta)}{q(\theta)}\right]$$

$$= \arg\min_\omega \; \mathbb{E}_{P(\theta)}\left[D_{\mathrm{KL}}\Big(P(x|\theta) \;\Big\|\; q_\omega(x|\theta)\Big)\right] + \mathrm{const}$$

$$\omega^\star = \arg\min_\omega \; D_{\mathrm{KL}}\Big(P(x,\theta) \;\Big\|\; q(\theta)\, q_\omega(x|\theta)\Big)$$

"Joint" KL-divergence

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x,\theta)}{q(\theta) q_\omega(x|\theta)}\right]$$

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x|\theta)}{q_\omega(x|\theta)}\right] + \mathbb{E}_{P(\theta)}\left[\log \frac{P(\theta)}{q(\theta)}\right]$$

data-distribution KL-divergence

$$= \arg\min_\omega \; \mathbb{E}_{P(\theta)}\Big[D_{\mathrm{KL}}\Big(P(x|\theta) \;\Big\|\; q_\omega(x|\theta)\Big)\Big] + \text{const}$$

$$\omega^\star = \arg\min_\omega \; D_{\mathrm{KL}}\Big(P(x,\theta) \;\Big\|\; q(\theta)\, q_\omega(x|\theta)\Big)$$

"Joint" KL-divergence

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x,\theta)}{q(\theta) q_\omega(x|\theta)}\right]$$

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x|\theta)}{q_\omega(x|\theta)}\right] + \mathbb{E}_{P(\theta)}\left[\log \frac{P(\theta)}{q(\theta)}\right]$$

data-distribution KL-divergence

$$= \arg\min_\omega \; \mathbb{E}_{P(\theta)}\Big[D_{\mathrm{KL}}\Big(P(x|\theta) \;\Big\|\; q_\omega(x|\theta)\Big)\Big] + \text{const}$$

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}[-\log q_\omega(x|\theta)] \qquad \text{(drop terms constant in } \omega\text{)}$$

$$\omega^\star = \arg\min_\omega \; D_{\mathrm{KL}}\Big(P(x,\theta) \;\Big\|\; q(\theta)\, q_\omega(x|\theta)\Big)$$

"Joint" KL-divergence

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x,\theta)}{q(\theta) q_\omega(x|\theta)}\right]$$

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}\left[\log \frac{P(x|\theta)}{q_\omega(x|\theta)}\right] + \mathbb{E}_{P(\theta)}\left[\log \frac{P(\theta)}{q(\theta)}\right]$$

data-distribution KL-divergence

$$= \arg\min_\omega \; \mathbb{E}_{P(\theta)}\Big[D_{\mathrm{KL}}\Big(P(x|\theta) \;\Big\|\; q_\omega(x|\theta)\Big)\Big] + \text{const}$$

$$= \arg\min_\omega \; \mathbb{E}_{P(x,\theta)}[-\log q_\omega(x|\theta)] \qquad (\text{drop terms constant in } \omega)$$

$$\approx \arg\min_\omega \; \frac{1}{N}\sum_i^N -\log q_\omega(x_i|\theta_i) \equiv \arg\min_\omega \; \mathcal{L}_{\mathrm{ML}}(\omega).$$

PDF-approximation matching via “forward” KL divergence

True distribution: P(x)
Approximation: q(x)

$$D_{\mathrm{KL}}(P|q)[\theta] = \int P(x|\theta_{true}) \ln\left(\frac{P(x|\theta_{true})}{q(x|\theta)}\right) dx$$

$$D_{\mathrm{KL}}(P|q)[\theta] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_{true})}{q(x_i|\theta)}\right)$$

→ argmin(θ) $\frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_{true})}{q(x_i|\theta)}\right)$ = argmin(θ) $\frac{1}{N}\sum_i -\ln\left(q(x_i|\theta)\right)$

Max. Likelihood for
L(θ)=q(x|θ)

PDF-approximation matching via “forward” KL divergence

True distribution: P(x|θ)
Approximation: q(x|θ)[ω]

$$D_{\mathrm{KL}}(P|q)[\omega] = \int P(x|\theta)\cdot P(\theta)\,\ln\left(\frac{P(x|\theta)}{q_\omega(x|\theta)}\right)dx \;+\; \text{Const}$$

$$D_{\mathrm{KL}}(P|q)[\omega] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) + \text{Const}$$

$$\longrightarrow \text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) = \text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i -\ln\left(q_\omega(x_i|\theta_i)\right)$$

(higher order) Max. Likelihood
“Likelihood fit” of the likelihood function

PDF-approximation matching via “forward” KL divergence

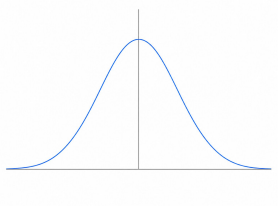
True distribution: P(x|θ)
Approximation: q(x|θ)[ω]

$$D_{\mathrm{KL}}(P|q)[\omega] = \int P(x|\theta)\cdot P(\theta)\,\ln\left(\frac{P(x|\theta)}{q_\omega(x|\theta)}\right)dx \;+\; \text{Const}$$

$$D_{\mathrm{KL}}(P|q)[\omega] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) + \text{Const}$$

$$\longrightarrow \text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) = \text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i -\ln\left(q_\omega(x_i|\theta_i)\right)$$

Event Params:
θ



(higher order) Max. Likelihood
“Likelihood fit” of the likelihood function
“Learn all possible data generating distributions”

$$\begin{pmatrix}\mu\\ \sigma\end{pmatrix} = \vec{f}_\omega(\theta) \qquad q_\omega(x\mid\theta) = \frac{1}{\sqrt{2\pi}\,\sigma_\omega[\theta]}\exp\left(-\frac{(x-\mu_\omega[\theta])^2}{2\sigma_\omega[\theta]^2}\right)$$

PDF-approximation matching via “forward” KL divergence

True distribution: P(x|θ)
Approximation: q(x|θ)[ω]

$$D_{\mathrm{KL}}(P|q)[\omega] =$$

$$D_{\mathrm{KL}}(P|q$$

ame
nimum
ver

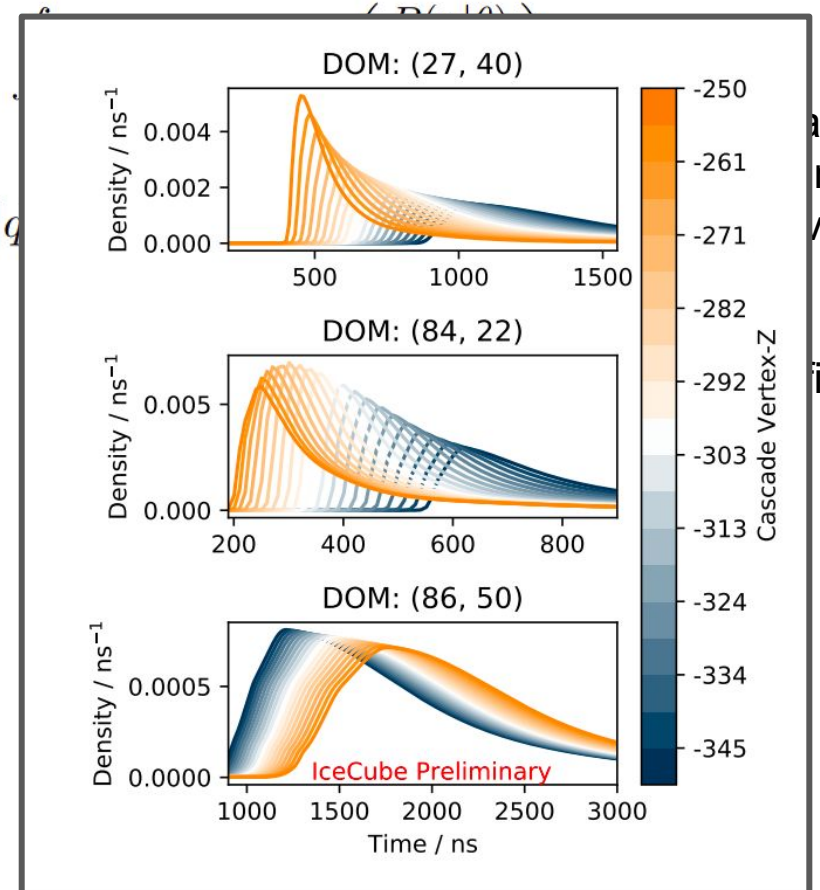
->
finity)

$$\longrightarrow \text{argmin}(\boldsymbol{\omega}) \quad \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right)$$

EventGenerator
ω=Neural Network params

“Learn all possible data generating distributions”

Event Generator (2107.12110)



PDF-approximation matching via “forward” KL divergence

True distribution: P(x|θ)
Approximation: q(x|θ)[ω]

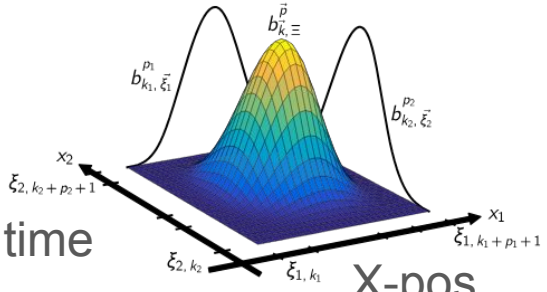
$$D_{\mathrm{KL}}(P|q)[\omega] = \int P(x|\theta)\cdot P(\theta)\,\ln\left(\frac{P(x|\theta)}{q_\omega(x|\theta)}\right)dx \;+\; \text{Const}$$

SplineReco (table fit)
ω=B-Spline coefficients

$$D_{\mathrm{KL}}(P|q)[\omega] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) + \text{Const}$$

$$\longrightarrow \text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) = \text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i -\ln\left(q_\omega(x_i|\theta_i)\right)$$

6-d Tensor-Product Spline (arXiv:1301.2184)



(higher order) Max. Likelihood
“Likelihood fit” of the likelihood function

“Learn all possible data generating distributions”

PDF-approximation matching via “forward” KL divergence

True distribution: P(x|θ)
Approximation: q(x|θ)[ω]

$$D_{\mathrm{KL}}(P|q)[\omega] = \int P(x|\theta)\cdot P(\theta)\,\ln\left(\frac{P(x|\theta)}{q_\omega(x|\theta)}\right)dx \;+\; \text{Const}$$

SplineReco (table fit)
ω=B-Spline coefficients

$$D_{\mathrm{KL}}(P|q)[\omega] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) + \text{Const}$$

$$\text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) = \text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i -\ln\left(q_\omega(x_i|\theta_i)\right)$$

EventGenerator
ω=Neural Network params

(higher order) Max. Likelihood
“Likelihood fit” of the likelihood function

“Learn all possible data generating distributions”

Hint: standard parametrization

Interactive Question 4

Can we switch the order of parametrization?

$$D_{\mathrm{KL}}(P \parallel q) = \int P(x,\theta) \log \frac{P(x,\theta)}{q_\omega(x|\theta)q(\theta)}\, dx$$

q(x|θ) -> q(θ|x)

What might be difficulties?

$$\frac{1}{N}\sum_i -\ln\left(q_\omega(x_i|\theta_i)\right)$$

Would that help us?

Learning objectives lecture 1:

- Bayes' Theorem and the "joint" distribution
- Frequentist / Bayesian parameter estimation
- Information theory: KL divergence and its interpretation
- **likelihood-based vs likelihood-free methods**

PDF-approximation matching via “forward” KL divergence

True distribution: P(x|θ)
Approximation: q(x|θ)[ω]

$$D_{\mathrm{KL}}(P|q)[\omega] = \int P(x|\theta)\cdot P(\theta)\,\ln\left(\frac{P(x|\theta)}{q_\omega(x|\theta)}\right)dx \;+\; \text{Const}$$

$$D_{\mathrm{KL}}(P|q)[\omega] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) + \text{Const}$$

$$\longrightarrow \text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i \ln\left(\frac{P(x_i|\theta_i)}{q_\omega(x_i|\theta_i)}\right) = \text{argmin}(\boldsymbol{\omega})\ \frac{1}{N}\sum_i -\ln\left(q_\omega(x_i|\theta_i)\right)$$

(higher order) Max. Likelihood
“Likelihood fit” of the likelihood function

“Learn all possible data generating distributions”

(Posterior) PDF-approximation matching via "forward" KL divergence

True distribution: P(θ|x)
Approximation: q(θ|x)[ω]

$$D_{\mathrm{KL}}(P|q)[\omega] = \int P(\theta|x) \cdot P(x) \ln\left(\frac{P(\theta|x)}{q_\omega(\theta|x)}\right) dx \ + \text{Const}$$

$$D_{\mathrm{KL}}(P|q)[\omega] \approx \frac{1}{N}\sum_i \ln\left(\frac{P(\theta_i|x_i)}{q_\omega(\theta_i|x_i)}\right) + \text{Const}$$

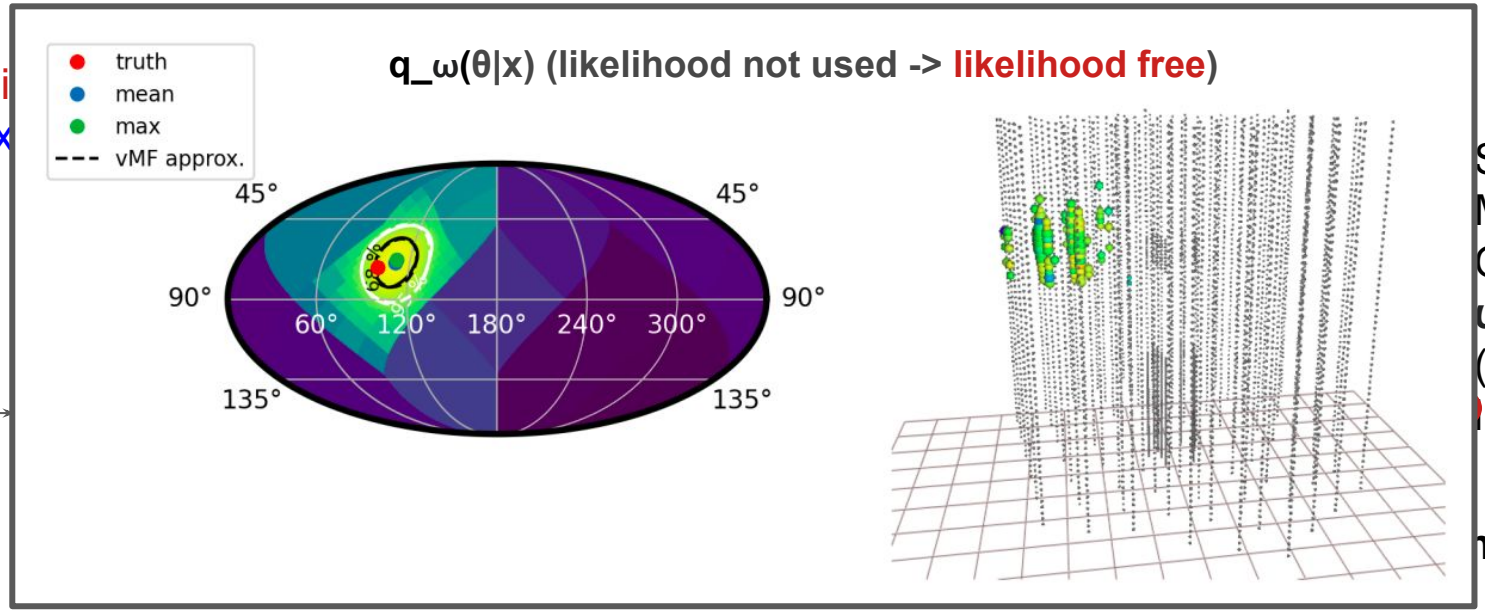
$$\longrightarrow \ \text{argmin}(\boldsymbol{\omega}) \ \frac{1}{N}\sum_i \ln\left(\frac{P(\theta_i|x_i)}{q_\omega(\theta_i|x_i)}\right) = \text{argmin}(\boldsymbol{\omega}) \ \frac{1}{N}\sum_i -\ln\left(q_\omega(\theta_i|x_i)\right)$$

Likelihood-free inference

"Learn all possible posterior distributions"

(Posterior) PDF-approximation matching via “forward” KL divergence

True di
Approx



Same
Minimum
Over
ω
(N ->
nfinity)

nference

(“amortized neural posterior estimation”) Papamakarios (2016)

“Learn all possible posterior distributions”

Neural posterior estimation of IceCube neutrino direction 2604.19846

“Likelihood-free” inference

Neural Posterior Estimation (**NPE**)

$$q_\omega(\theta|x)$$

Neural Ratio Estimation (**NRE**)

$$r_\phi(\theta, x) \approx \frac{p(x \mid \theta)}{p(x)} = \frac{p(\theta \mid x)}{p(\theta)}$$

Bayes’ Theorem

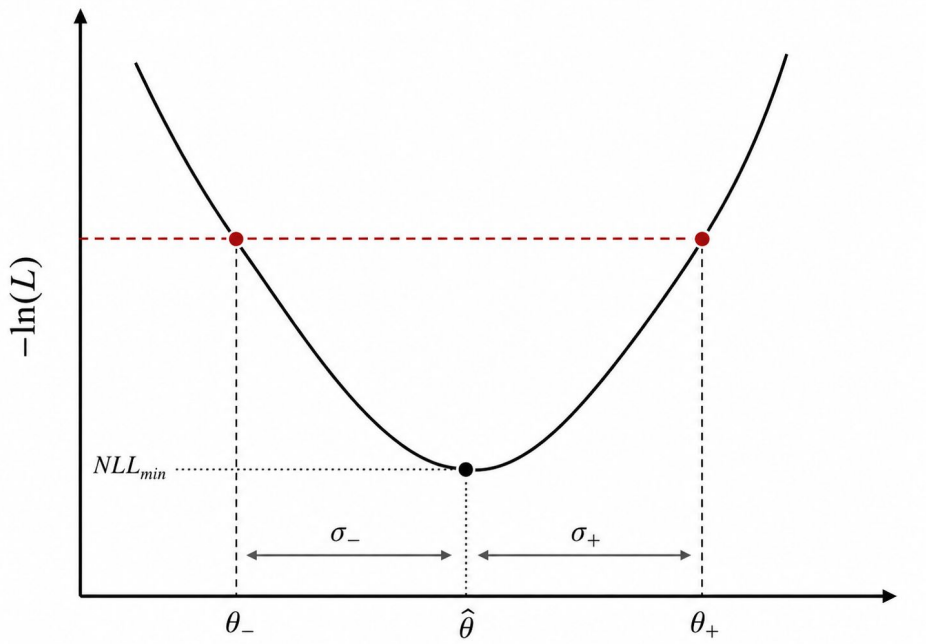
$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

Trained with classifier
Proportional to likelihood, sampling still required

Summary:

Likelihood-based

Max. Likelihood (Frequentist)

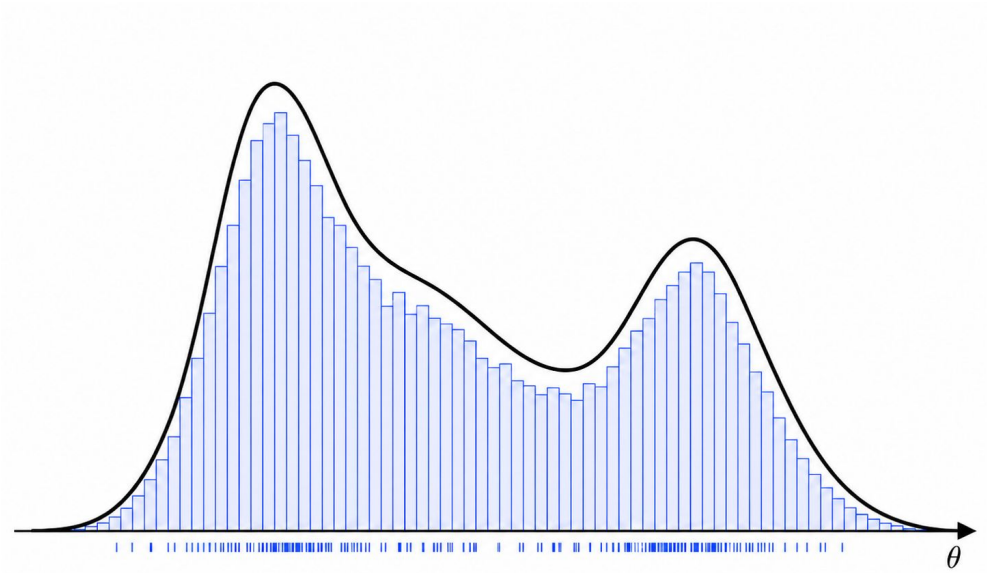


Summary:

Likelihood-based

Max. Likelihood (Frequentist)

MCMC (Bayesian)



$$\alpha = \min\left(1, \frac{\pi(\theta')}{\pi(\theta_t)}\right)$$

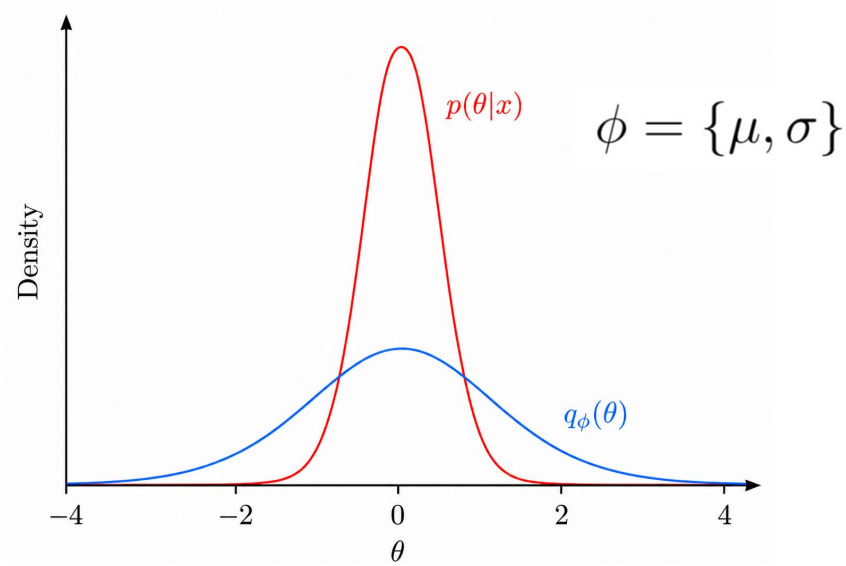
Summary:

Likelihood-based

Max. Likelihood (Frequentist)

MCMC (Bayesian)

ELBO-based
Variational Inference (Bayesian)



$$p(x) \geq \underbrace{\int_\theta q_\phi(\theta) \cdot \left[\ln p(x;\theta) - \ln \frac{q_\phi(\theta)}{p(\theta)}\right] d\theta}_{\text{evidence lower bound (ELBO)}}$$

Summary:

Information Theory!

-> KL divergence of **joint** distribution

$$p(\theta, x) = p(\theta|x) \cdot p(x) = p(x|\theta) \cdot p(\theta)$$

Likelihood-based

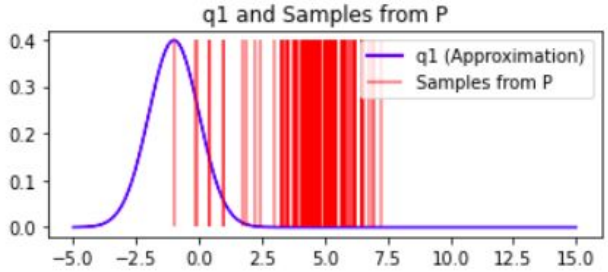
Max. Likelihood (Frequentist)

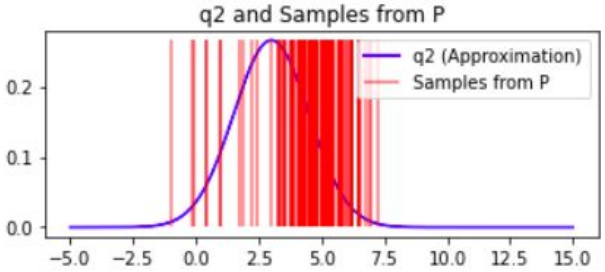
MCMC (Bayesian)

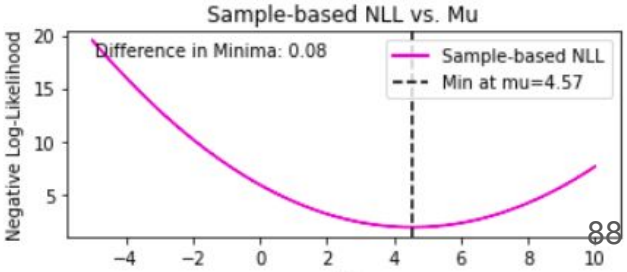
ELBO-based
Variational Inference (Bayesian)

$q_\omega(x|\theta)$

Learn “likelihood”







Summary:

Information Theory!

-> KL divergence of **joint** distribution

$$p(\theta, x) = p(\theta|x) \cdot p(x) = p(x|\theta) \cdot p(\theta)$$

Likelihood-based

Max. Likelihood (Frequentist)

MCMC (Bayesian)

ELBO-based
Variational Inference (Bayesian)

$q_\omega(x|\theta)$

Learn “likelihood”

Per event
Extra numerical step:
Use likelihood to get
Posterior / Confidence Interval

Lecture 2/3

Summary:

Information Theory!

-> KL divergence of **joint** distribution

$$p(\theta, x) = p(\theta|x) \cdot p(x) = p(x|\theta) \cdot p(\theta)$$

Likelihood-based

Max. Likelihood (Frequentist)

MCMC (Bayesian)

ELBO-based
Variational Inference (Bayesian)

$q_\omega(x|\theta)$

Learn "likelihood"

Learn all posteriors

$q_\omega(\theta|x)$

Likelihood-free

Amortized neural
posterior estimation

Per event

Extra numerical step:
Use likelihood to get
Posterior / Confidence Interval

Lecture 2/3

Summary:

Information Theory!

-> KL divergence of **joint** distribution

$$p(\theta, x) = p(\theta|x) \cdot p(x) = p(x|\theta) \cdot p(\theta)$$

Likelihood-based

Max. Likelihood (Frequentist)

MCMC (Bayesian)

ELBO-based
Variational Inference (Bayesian)

$q_\omega(x|\theta)$

Learn "likelihood"

Learn all posteriors

$q_\omega(\theta|x)$

Per event

Extra numerical step:
Use likelihood to get
Posterior / Confidence Interval

Lecture 2/3

Lecture 3

No extra step:
-> direct Posterior
-> Standard NN prediction special case

Likelihood-free

Amortized neural posterior estimation