A Generalization of the Maximum Entropy Principle (MEP) for curved Statistical Manifolds

Written by Pablo A. Morales and Fernando E. Rosas
Presented by Stash TOMONAGA

What is the Maximum Entropy Principal (MEP)?

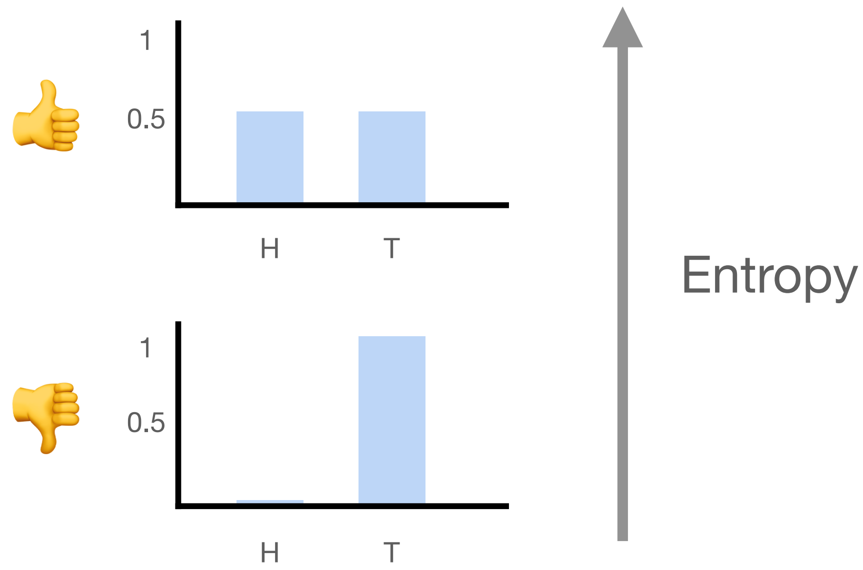
- Entropy: Measurement of Uncertainty
- MEP: Pick a prediction model with Maximum Uncertainty (least-informative) under the circumstance.

Why though…?

- Ex 1: Coin toss

Only know possible outcomes…

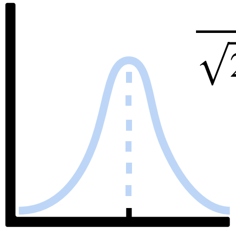
$$\Omega = \{\text{Heads, Tails}\}$$



MEP is easy and great! ⋯ but limited.

- Ex2. What if: we only know the Mean μ and Variance σ^2 ⋯?
  - The Gaussian distribution maximizes entropy

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- MEP can be solved with Numerical Methods (Lagrange Multipliers)

  ⋯MEP is easy and great! ⋯but no.

- MEP is based on Shannon's Entropy, which has restrictions.
  - MEP can't be used often times

Generalizing the Shannon entropy

- MEP is based on Shannon's Entropy.
- Shannon's Entropy limits the range of outputs (to Boltzmann-Gibbs distributions)
- Many people tried to come up with a Generalized Entropy, and force a MEP method, but have had problems.

How SHOULD we come up with a generalized MEP?

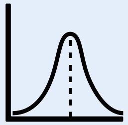
- Instead of generalizing Entropy for the MEP on its own…

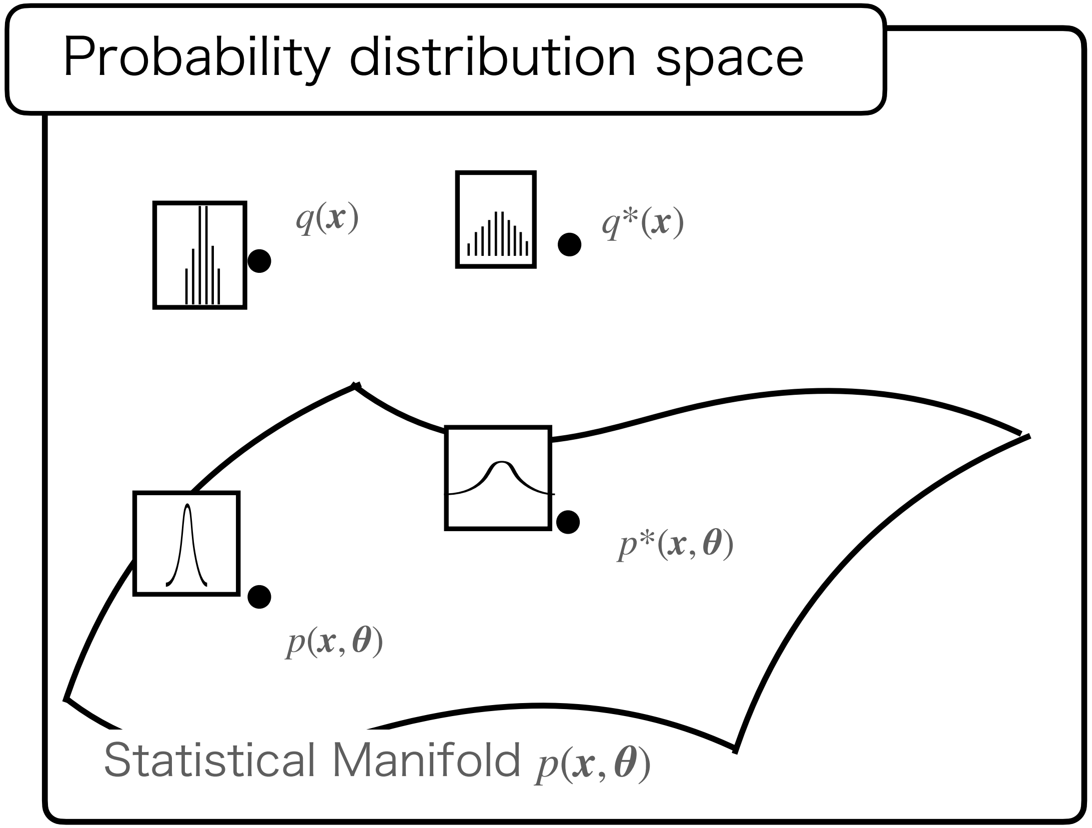
1. Prove that Shannon's MEP is a **Natural Consequence** of the geometry of Flat Statistical Manifolds (cause of limitation).

2. Using the same logic, derive a **Natural** Entropy for Curved Statistical Manifolds.

➡Rényi's Entropy is the Natural Consequence

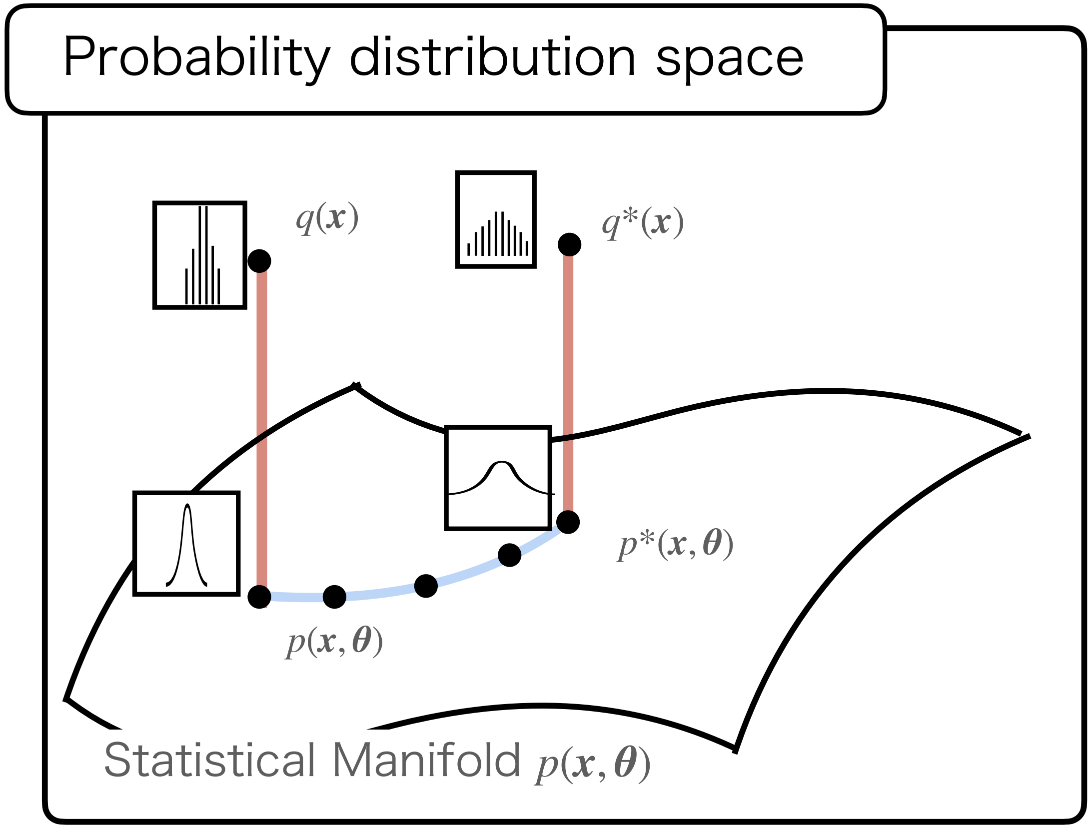
What is a Statistical Manifold···?

- Consider any probability distribution function 
- Consider a space of distributions
- Consider a Parametric Model
- ▸ Space covered by the possible parametric models is the Statistical Manifold
- Relationships between points aren't obvious
- We can define different "rules" for relationships of points



Same manifold, different "rules"

- Consider different "rules" for relationships of points
  - Derivatives (Directions)
  - Metrics (Inner products)
  - Connections (Translations, Affine Transformations)
    - Distances (Divergences)
    - Curvatures



Same manifold, different “rules” #2

Well-studied “Rules” in Geometry

	Metric g + Connection ∇	Dual Structure (g, ∇, ∇^*)
Flat	Euclidian	Dually flat
non-Flat	Riemannian	**Underdeveloped**

Same manifold, different "rules" #2

Well-studied "Rules" in Geometry

	Metric g + Connection ∇	Dual Structure (g, ∇, ∇^*)
Flat	Euclidian	Dually flat
non-Flat	Riemannian	**Underdeveloped**

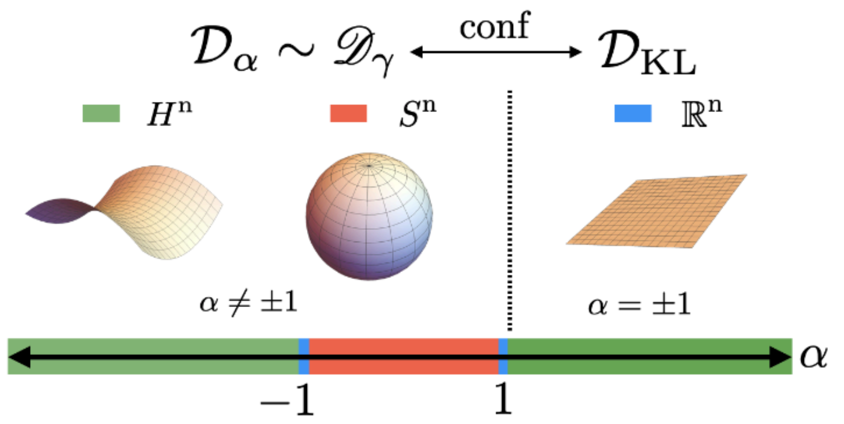
Example of Connection… ∇_α (Flat)

When is the connection flat?

- When $\alpha = \pm 1$, ∇_α gives flat connection
- $\nabla_{\alpha=+1}$ and $\nabla_{\alpha=-1}$ has Dual relationship

(BIG DEAL)

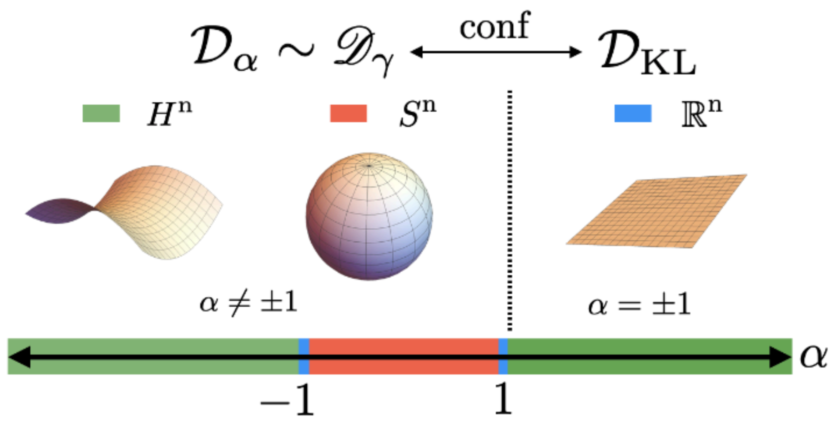
- Comes with Benefits of Flat geometry
- This allows for Shannon's Entropy to function in MEP



Example of Connection⋯ ∇_α (Curved)

When is the connection not flat (curved)?

- When $\alpha \neq \pm 1$, ∇_α gives curved connection
  - Benefits of Flatness break down⋯
  - Can't do MEP⋯



Find a way to Re-cast the "rule" so that the benefits are recovered

➡ Rényi's Entropy is the Natural Consequence of this "rule"

Same manifold, different “rules” #2

Well-studied “Rules” in Geometry

	Metric g + Connection ∇	Dual Structure (g, ∇, ∇^*)
Flat	Euclidian	Dually flat
non-Flat	Riemannian	**Underdeveloped**

Same manifold, different “rules” #2

Well-studied “Rules” in Geometry

		Dual Structure (g, ∇, ∇^*)
Flat		Dually flat
non-Flat	Riemannian	**Underdeveloped**

Shannon’s Entropy

Same manifold, different “rules” #2

Well-studied “Rules” in Geometry

		Dual Structure (g, ∇, ∇^*)
Flat		Dually flat
non-Flat	[illegible]	**Underdeveloped**

Shannon’s Entropy

Rényi’s Entropy

How SHOULD we come up with a generalized MEP?

- Instead of generalizing Entropy for the MEP on its own…

1. Prove that Shannon's MEP is a **Natural Consequence** of the geometry of (Dually) Flat Statistical Manifolds.

2. Using the same logic, derive a **Natural** Entropy for Curved Statistical Manifolds.

➡Rényi's Entropy is the Natural Consequence

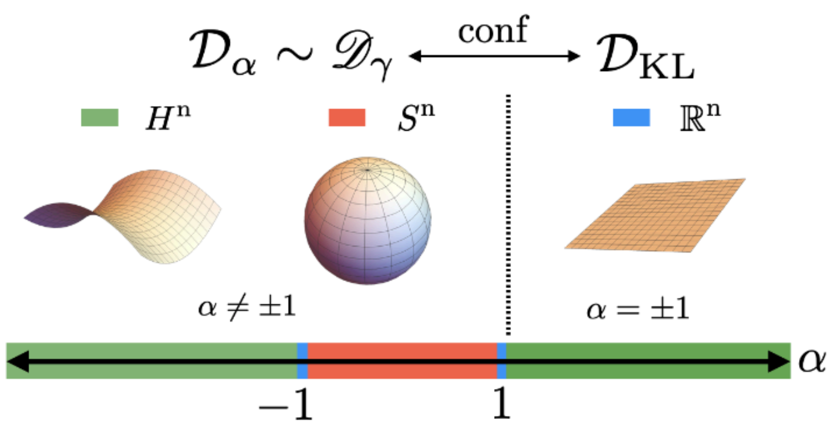
A Generalization of the Maximum Entropy Principle (MEP) for curved Statistical Manifolds

Written by Pablo A. Morales and Fernando E. Rosas
Presented by Stash TOMONAGA

Duality of Manifolds

Dually Flat Geometry

- Big discovery of Information Geometry…
  - Dual Flatness in Statistical Manifolds
  - Many many tools can be used in flatness

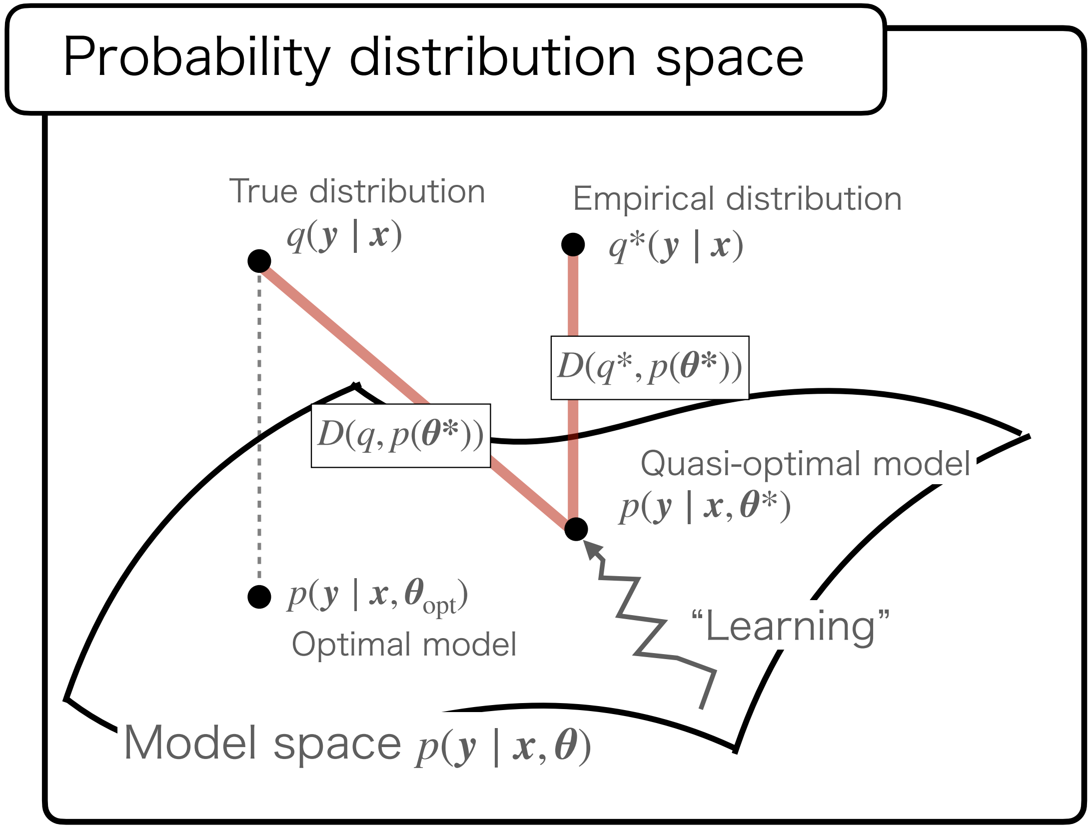


Dually Curved Geometry

- Can't use tools for Flatness
- No consensus on geometry of non-flat (curved) Statistical Manifolds

What is a Statistical Manifold⋯?

- Consider any probability distribution
- Consider a space of distributions

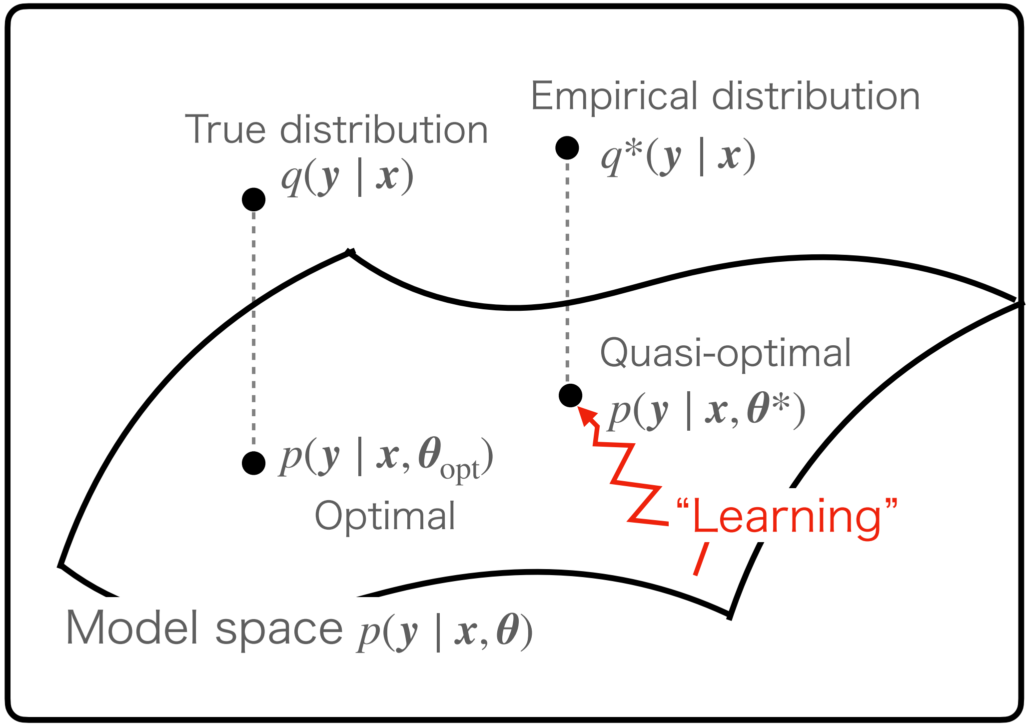


Optimizing to the Empirical Distribution

- Only the Empirical distribution is available
  - Optimum $\boldsymbol{\theta}_{opt}$ is unobtainable
  - Search for Quasi-optimum $\boldsymbol{\theta}*$

$$\boldsymbol{\theta}* \in \arg\min_{\boldsymbol{\theta}} D(q*, p(\boldsymbol{\theta}))$$

$$q*(\boldsymbol{x}, \boldsymbol{y}) \equiv \frac{1}{N} \sum_{i=1}^{N} \delta\left(\boldsymbol{x} - \boldsymbol{x}_i, \boldsymbol{y} - \boldsymbol{y}_i\right)$$



Estimating Generalization Loss in a different way

- We want:

  Generalization Loss $D(q, p(\boldsymbol{\theta}^*))$

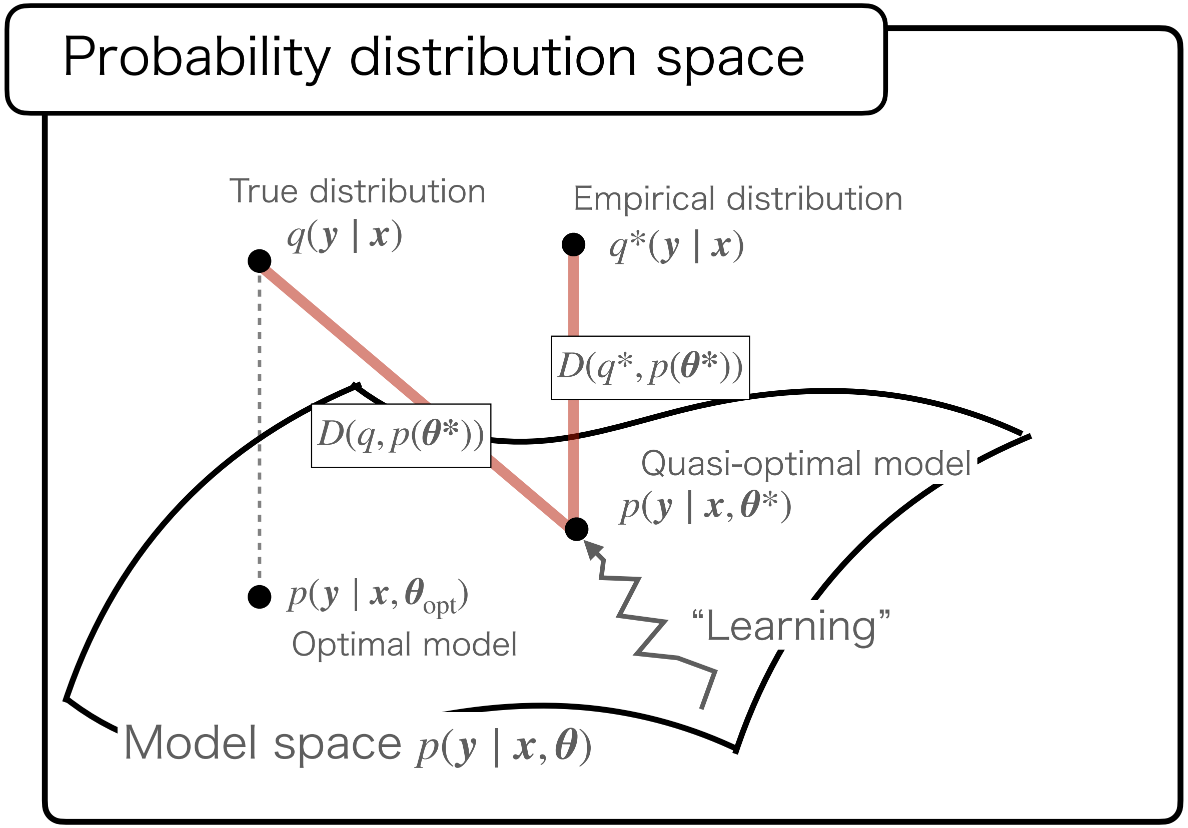
- We have:

  Train Loss $D(q^*, p(\boldsymbol{\theta}^*))$

- Can we estimate:

  Generalization Gap $\mathscr{G}$?

  $$D(q, p(\boldsymbol{\theta}^*)) - D(q^*, p(\boldsymbol{\theta}^*))$$

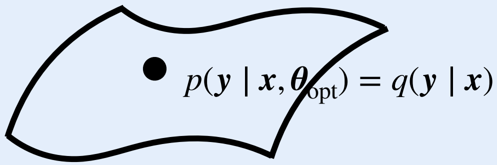


➡Yes, (under certain conditions, using Information Criterion)

Information Criterion

Akaike's Information Criterion

- If there exists θ_{opt} s.t. $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\text{opt}}) = q(\boldsymbol{y} \mid \boldsymbol{x})$



$$D(q, p(\boldsymbol{\theta}^*)) = AIC(p) + D(q^*, p(\boldsymbol{\theta}^*))$$

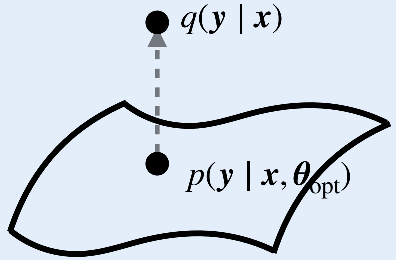
$$AIC(p) = \frac{d}{N}$$

N : Number of samples

d : degrees of freedom (number of parameters) of p

Network Information Criterion

- If not,



$$D(q, p(\boldsymbol{\theta}^*)) = NIC\left(p(\boldsymbol{\theta}^*)\right) + D(q^*, p(\boldsymbol{\theta}^*))$$

$$NIC(p) = \frac{1}{N}\mathrm{Tr}(H^{-1}C)$$

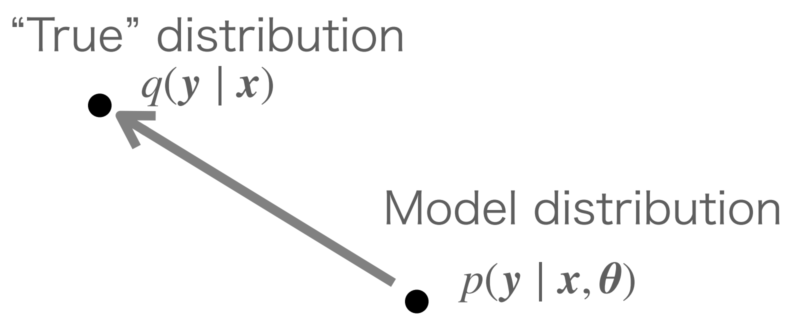
$$H(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \mathbb{E}_q\left[\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}\, \ell\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right)\right]$$

$$C(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \mathbb{E}_q\left[\left(\nabla_{\boldsymbol{\theta}}\, \ell\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right)\right)\left(\nabla_{\boldsymbol{\theta}}\, \ell\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}\right)\right)^{\mathrm{T}}\right]$$

What does Machine Learning Optimize?

Goal: Approximate “true” distribution with a parametric model

- “True” distribution: Stochastic System (Probability Distribution $q(\boldsymbol{y} \mid \boldsymbol{x})$)
- Model distribution: Statistical Model (Probability Distribution $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$)



- Minimize Expected Loss $D(q, p(\boldsymbol{\theta})) = \mathbb{E}_q \left[\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) \right]$

$$\mathbb{E}_q \left[\ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) \right] = \int \ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) q(\boldsymbol{x}) q(\boldsymbol{y} \mid \boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y} = \int \ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) q(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y}$$

$$\boldsymbol{\theta}_{opt} \in \arg\min_{\boldsymbol{\theta}} D(q, p(\boldsymbol{\theta}))$$