CS 8501 Advanced Topics in Machine Learning

Lecture 03: Bayesian Statistics

Yangfeng Ji
Information and Language Processing Lab
Department of Computer Science
University of Virginia
https://yangfengji.net/

Introduction

Interpretation of Probability

- Frequentist: probability is the long-run frequency of repeatable experiments
  - Flip a coin
  - Toss a dice
- Bayesian: probability is a degree of (personal) belief
  - The probability of a nuclear war (or betting on any non-repeatable future event)
- Machine learning researchers use both of them, depending what they want to do

Frequentist Statistics

- Assume an underlying true data distribution p^*
- Data sampled from this distribution is called sampling distribution $\mathcal{D} \sim p^*$
- Estimating the parameter based on the sampling distribution as

$$\hat{\theta} = \mathbf{argmax}_\theta \, p(\mathcal{D} \mid \theta)$$

Bayesian Statistics

- In Bayesian approach, we treat the parameter θ as a random variable
- Using the posterior distribution to summarize the information of θ is at the core of Bayesian statistics.

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\int_\theta p(\mathcal{D} \mid \theta)p(\theta)}$$

Bayesian Statistics (II): Posterior Distribution

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{\int_\theta p(\mathcal{D} \mid \theta)p(\theta)}$$

- Here, we assume θ is a continuous random variable
- $p(\mathcal{D})$ is called marginal likelihood or evidence
- The computation of $p(\mathcal{D})$ is usually the key challenge of Bayesian inference

Bayesian Statistics (III): Predictive Probability

To make a prediction on a new data point x, we have

$$p(x \mid \mathcal{D}) = \int_\theta p(x \mid \theta) p(\theta \mid \mathcal{D}) d\theta$$

Which consider all the possible values of θ based on the probability distribution $p(\theta \mid \mathcal{D})$

Summarizing Posterior Distribution

- The issues of MAP
- The advantage of inference with posterior distributions

Maximum A Posteriori (MAP)

It is a point estimate from the posterior distribution

$$\hat{\theta}_{MAP} \leftarrow \operatorname{argmax}_\theta p(\theta \mid \mathcal{D})$$

Some alternative formulations

- $\hat{\theta} \leftarrow \operatorname{argmax}_\theta \log p(\theta \mid \mathcal{D})$
- $\hat{\theta} \leftarrow \operatorname{argmax}_\theta \log p(\theta) + \log p(\mathcal{D} \mid \theta)$
- It is not necessary to actually compute the posterior distribution

MAP Estimation

MAP estimation gives a point estimation of θ

$$\hat{\theta} = \mathrm{argmax}_{\theta} p(\theta \mid \mathcal{D})$$

Or equivalently

$$\hat{\theta} = \mathrm{argmax}_{\theta} p(\mathcal{D} \mid \theta) \cdot p(\theta) = \mathrm{argmax}_{\theta} \{\log p(\mathcal{D} \mid \theta) + \log p(\theta)\}$$

which avoid the computation of $p(\mathcal{D})$, but this is not a full Bayesian method.

MAP with a Gaussian Prior

Consider the log form:

$$\hat{\theta} = \mathrm{argmax}_\theta \{\log p(\mathcal{D} \mid \theta) + \log p(\theta)\}$$

If $p(\theta)$ is defined as the Gaussian distribution $\mathcal{N}(\theta; 0, 1/\lambda)$, then MAP is equivalent to learning with ℓ_2 regularization

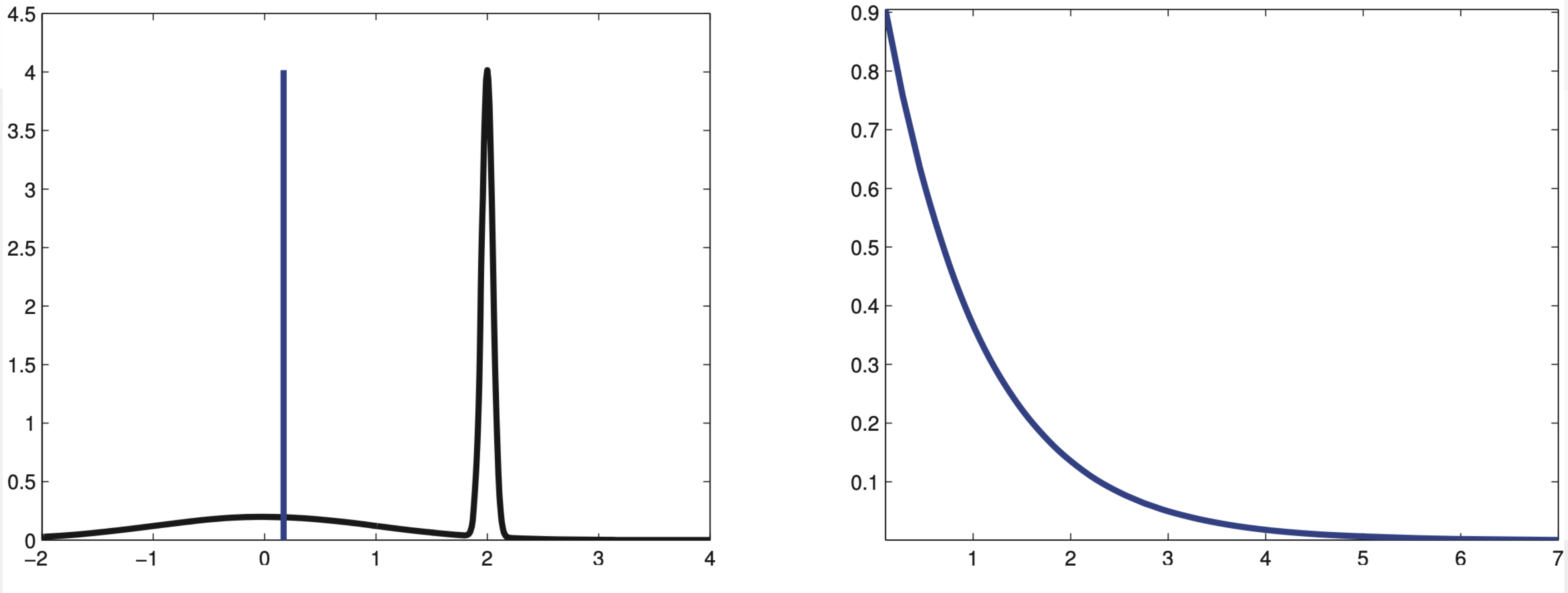
$$\hat{\theta} = \mathrm{argmax}_\theta \{\log p(\mathcal{D} \mid \theta) - \lambda \|\theta\|_2^2\}$$

where λ is the regularization parameter.

Issues with MAP

The mode is an untypical point

$$\hat{\theta}_{MAP} \leftarrow \mathrm{argmax}_{\theta} p(\theta \mid \mathcal{D})$$



(a) (b)

Issues with MAP (II)

No measure of uncertainty

$$\hat{\theta}_{MAP} \quad \text{v.s.} \quad p(\theta \mid \mathcal{D})$$

Considering the uncertainty of θ, expectation is the best way of summarizing the randomness

$$p(x \mid \mathcal{D}) = \int_{\theta} p(x \mid \theta) p(\theta \mid \mathcal{D}) = \mathbb{E}_{p(\theta \mid \mathcal{D})}[p(x \mid \theta)]$$

Issues of MAP (III)

Plugging in the MAP estimate can result in overfitting

- Predictive distribution can be over-confident if not modeling the uncertainty of parameters, which is a common problem of any point estimate methods
- Example:

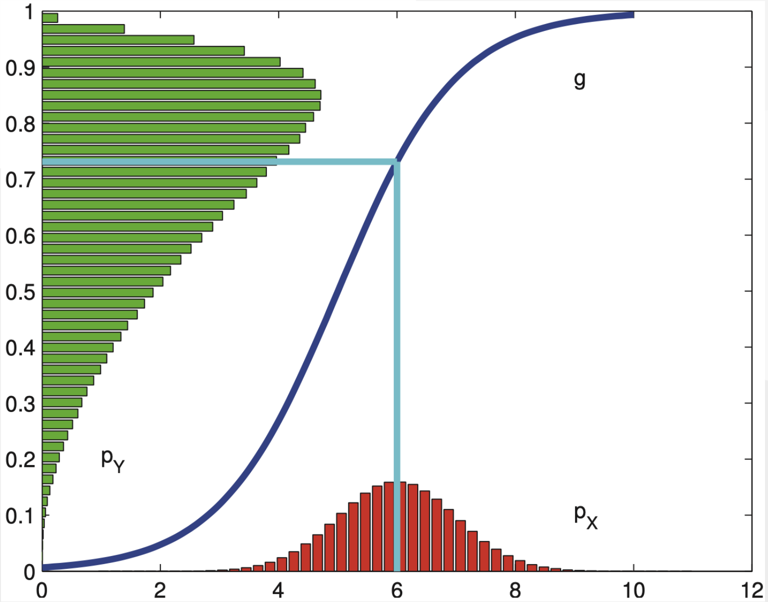
$$\hat{\theta} = \mathrm{argmax}_\theta \{\log p(\mathcal{D} \mid \theta) - \lambda \|\theta\|_2^2\}$$

The limitation of ℓ_2 regularization for avoiding overfitting.

Issues of MAP (IV)

MAP estimation is not invariant to reparametrization: consider

- X follows a Gaussian distribution
- $f : X \to Y$ is a nonlinear function



Finding the mode of X may not help finding the model of Y

Inference with Posterior

From (Murphy, 2012):

"*For example, suppose you are about to buy something from Amazon.com, and there are two sellers offering it for the same price. Seller 1 has 90 positive reviews and 10 negative reviews. Seller 2 has 2 positive reviews and 0 negative reviews. Who should you buy from?*"

Inference with MLE

Let θ_1 and θ_2 be the unknown reliabilities of the two sellers

- Seller 1: 90 positive reviews; 10 negative reviews
- Seller 2: 2 positive reviews; 0 negative reviews

MLE of θ

- $\theta_{1,MLE} = 0.9$
- $\theta_{2,MLE} = 1.0$

Inference with MAP

- Assume the uniform prior $\theta_i \sim \mathrm{Beta}(1,1)$
- The posterior of each θ_i
  - $p(\theta_1 \mid \mathcal{D}_1) = \mathrm{Beta}(91, 11)$
  - $p(\theta_2 \mid \mathcal{D}_2) = \mathrm{Beta}(3, 1)$
- MAP, also the mode of each Beta posterior
  - $\theta_{1,MAP} = \frac{\alpha-1}{\alpha+\beta-2} = 0.9$
  - $\theta_{2,MAP} = 1.0$
  - The results are not surprising

Inference with the Whole Posterior Distribution

- Consider θ_1 and θ_2 are both random variables, both of them can pick values between 0 and 1
- The question of seller 1 is better than seller 2 is formulated as

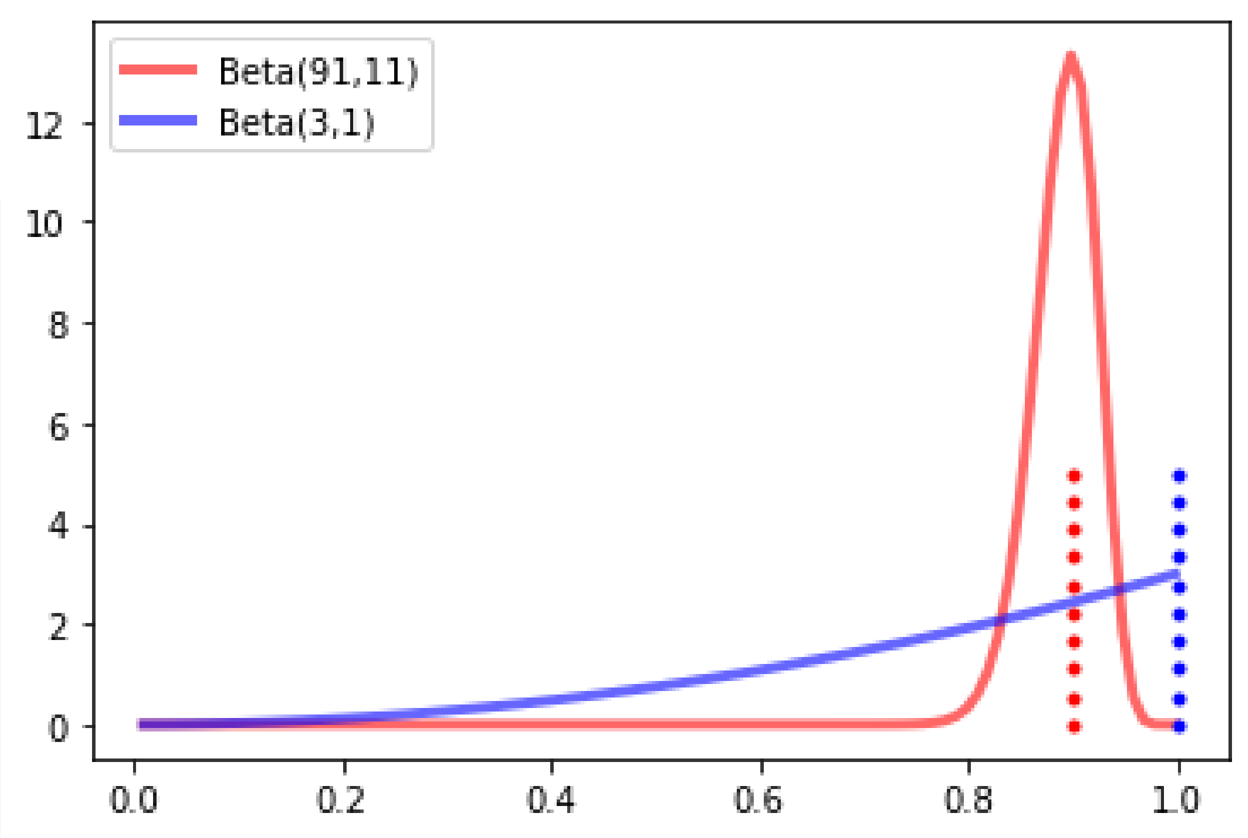
$$p(\theta_1 > \theta_2 \mid \mathcal{D})$$

- Compute $p(\theta_1 > \theta_2 \mid \mathcal{D})$ as

$$p(\theta_1 > \theta_2 \mid \mathcal{D}) = \iint I(\theta_1 > \theta_2) p(\theta_1 \mid \mathcal{D}_1) p(\theta_2 \mid \mathcal{D}_2) = 0.710$$

Inference with the Whole Posterior Distribution (II)

Why even a uniform prior can help?



Why Bayesian Approach?

Observations:

- Three methods: MLE, MAP, and Bayesian approach
- Two of them agree with each other

Then

- Why we prefer the Bayesian approach?

Another Interpretation

- Assume we only have sufficient data for seller 1 to use the frequentist approach
- The reliability of seller 1 is

$$\hat{\theta} = 0.9$$

- Assume all reviews are independent, then what is the chance that seller 1 have at least one negative reviews in the first two reviews?

Why Bayesian Approach?

Arguments for Bayesian Approach

Exchangeable A sequence of random variable $(x_1, x_2, \dots)$ is infinitely exchangeable, if for any n the joint probability $p(x_1, \dots, x_n)$ is invariant to permutation of the indices

$$p(x_1, \dots, x_n) = p(x_{\pi_1}, \dots, x_{\pi_n})$$

- Consider a set of images $(x_1, \dots, x_n)$ with a common background x_0
  - $(x_1 + x_0, \dots, x_n + x_0)$ are not independent
  - $(x_1 + x_0, \dots, x_n + x_0)$ are exchangeable

Arguments for Bayesian Approach (II)

De Finetti's theorem

A sequence of random variable $(x_1, x_2, \ldots,)$ is infinitely exchangeable if and only if, for all n, we have

$$p(x_1, \ldots, x_n) = \int p(\theta) \prod_{i=1}^{n} p(x_i \mid \theta) d\theta$$

where θ is some hidden common random variable (possibly infinite dimensional). That is, $\{x_i\}$ are iid conditional on θ.

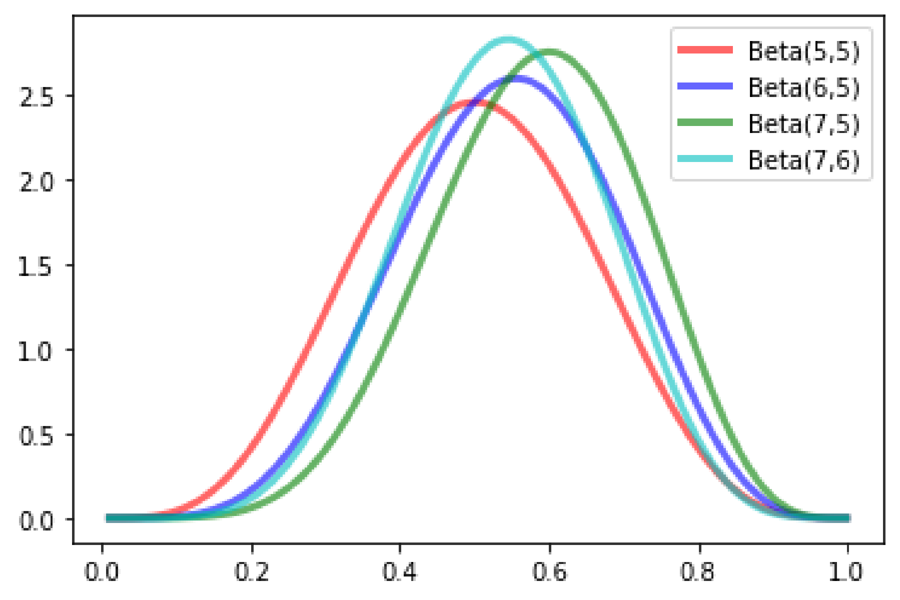
- The existence of a hidden variable θ

Arguments for Bayesian Approach (III)

Online Learning

The posterior can be further updated with new datasets, which provides an approach to continual learning

$$p(\theta \mid \mathcal{D}_{1:t}) \propto p(\mathcal{D}_t \mid \theta) \cdot p(\theta \mid \mathcal{D}_{1:t-1})$$



Priors

Part of the content is selected from Chapter 03 of The Bayesian Choice (2007) by Christian Robert.

Difficulty of Selecting Priors

[Robert, 2007]: Undoubtedly, the most **critical** and most **criticized** point of Bayesian analysis deals with the choice of the prior distribution, since, once this prior distribution is known, inference can be led in an almost mechanic way by

- minimizing posterior losses,
- computing higher posterior density regions, or
- integrating out parameters to find the predictive distribution.

Difficulty of Selecting Priors (II)

[Robert, 2007]: the systematic use of

- parameterized distributions (like the normal, gamma, beta, etc.) and
- the further reduction to conjugate distributions

cannot be justified at all times, since they *trade* an improvement in the analytical treatment of the problem *for* the subjective determination of the prior distribution and may therefore ignore part of the prior information.

Difficulty of Selecting Priors (III)

[Robert, 2007]

- Ungrounded prior distributions produce unjustified posterior inference
  - It is always possible to choose a prior distribution that gives the answer one wishes
- There is no such thing as *the* prior distribution, except for very special settings

Justification of Selecting Priors

- Conjugate priors
- Maximum entropy priors
- Non-informative priors

Conjugate Priors

- A prior $p(\theta) \in \mathcal{F}$ is a conjugate prior for a likelihood function $p(\mathcal{D} \mid \theta)$ if the posterior is in the same parameterized family as the prior, i.e., $p(\theta \mid \mathcal{D}) \in \mathcal{F}$.
- In previous discussion, we have seen two examples of conjugate priors
  - Beta distribution for the binomial model
  - Dirichlet distribution for the multinomial model

Non-informative Priors

- Derive the priors from the sample distribution (aka, the data)
- Laplace's prior: give the same likelihood to each value of the parameter (Principle of Insufficient Reasoning)
- The Jeffreys prior: based on the likelihood function

$$\pi(\theta) \propto (I(\theta))^{\frac{1}{2}}$$

where

$$I(\theta) = \mathbb{E}[(\frac{\partial \log p(X \mid \theta)}{\partial \theta})^2]$$

Maximum Entropy Priors

Entropy of a (prior) distribution $p(\theta)$ is defined as

$$H(p) = -\sum_{\theta} p(\theta) \log p(\theta)$$

or

$$H(p) = -\int_{\theta} p(\theta) \log p(\theta) d\theta$$

Maximum entropy distributions

- Discrete random variable: uniform distribution
- Continuous random variable with a given variance σ^2: Gaussian

Maximum Entropy Priors (II)

Assume θ is a binary random variable, then its entropy is defined as

$$H(p) = -\theta \log \theta - (1-\theta) \log(1-\theta)$$

With $\frac{dH(p)}{d\theta} = 0$, we have

$$\theta = \frac{1}{2}$$

Hierarchical Bayes

Hierarchical Prior

$$\eta \to \theta \to \mathcal{D}$$

- the prior distribution of θ is $p(\theta \mid \eta)$
- $p(\eta)$ also has its own parameters (usually, pre-defined hyper-parameters)
- $p(\eta)$ can just be an non-informative prior

Modeling Related Cancer Rates

Consider the problem of predicting cancer rates in several cities

- N_i: number of people in city i
- x_i: number of people in city i who died of cencer

$$x \sim \mathrm{Bin}(N_i, \theta_i)$$

Two Simple Estimation

- Estimate θ_i individually for each city
  - Probably not enough data
- Estimate all θ_i's as one single value $\tilde{\theta}$

$$\tilde{\theta} = \frac{\sum_i x_i}{\sum_i N_i}$$

Modeling with a Hierarchical Prior

$$p(\mathcal{D}, \theta, \eta \mid N) = p(\eta) \prod_{i=1} \{\mathrm{Bin}(x_i \mid N_i, \theta_i)\mathrm{Beta}(\theta_i \mid \eta)\}$$

where $\eta = (a, b)$

For example,

- $p(\eta) = p(a) \cdot p(b)$
- each of them could be a Gamma distribution

Meta Learning

- Based on Grant et al., 2018

Meta Learning

- A family of tasks $\mathcal{T}$
- A dataset $\mathcal{D}$ collected for the tasks $\mathcal{T}$
- The tasks share some common structure such that learning to solve a single task has the potential to aid in solving another

MAML

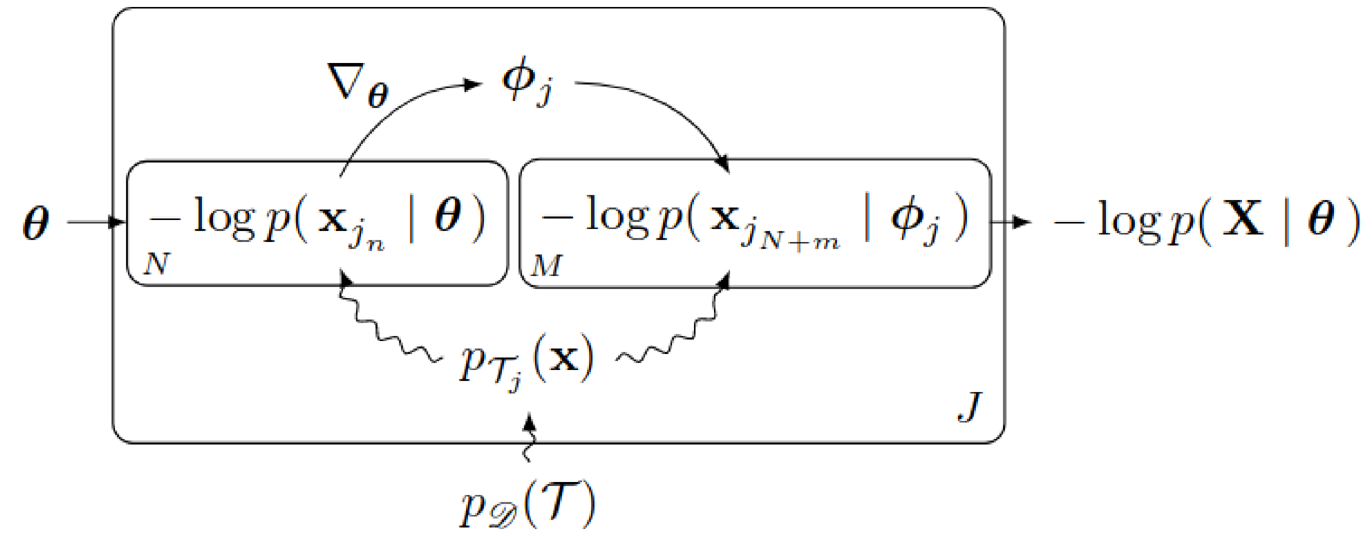
The MAML (Model-Agnostic Meta-Learning) updateing rule

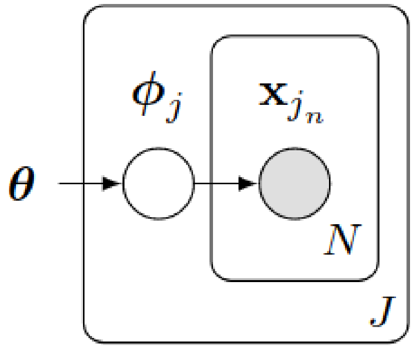
$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{J} \sum_{j} \left[\frac{1}{M} \sum_{m} -\log p\big(\mathbf{x}_{j_{N+m}} \mid \underbrace{\boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n} -\log p\left(\mathbf{x}_{j_n} \mid \boldsymbol{\theta} \right)}_{\boldsymbol{\phi}_j} \big) \right]$$

where

- $x_{j_1}, \ldots, x_{j_N} \sim p_{\mathcal{T}_j}(x)$: a small sample of data from task j
- $x_{j_{N+1}}, \ldots, x_{j_{N+M}} \sim p_{\mathcal{T}_j}(x)$: another sample of data from the same task

Graphical Representations





Mathematical formulation:

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_j \left(\int p(\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_N} \mid \boldsymbol{\phi}_j)\, p(\boldsymbol{\phi}_j \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\phi}_j \right)$$

Approximation

Hierarchical Bayesian model:

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_j \left(\int p(\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_N} \mid \boldsymbol{\phi}_j)\, p(\boldsymbol{\phi}_j \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\phi}_j \right)$$

MAML as an approximation:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{J} \sum_j \left[\frac{1}{M} \sum_m -\log p\big(\mathbf{x}_{j_{N+m}} \mid \underbrace{\boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \frac{1}{N} \sum_n -\log p(\mathbf{x}_{j_n} \mid \boldsymbol{\theta})}_{\boldsymbol{\phi}_j}\big) \right]$$

Thank You!