FY2022/2022年度

Mathematical Science Team

Kenichi Bannai

数理科学チーム　坂内 健一

数理科学チームでは、幅広い純粋数学・理論物理の研究者の力を借りて、人工知能・機械学習分野における様々な数学的課題に組織的に取り組んでいる。このポスターでは、数理科学チームのE. M. Kıralと近似ベイズ推論チームのT. MöllenhoffとE. Khanの共同研究研究を解説する。

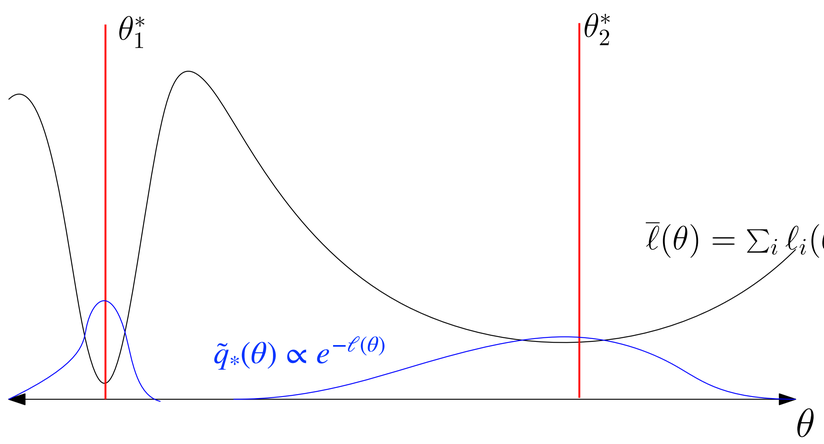
The Lie Group Bayesian Rule

Eren Mehmet Kıral[1,2], Thomas Möllenhoff[1], Emtiyaz Khan[1]

1. RIKEN AIP　2. RIKEN SPDR

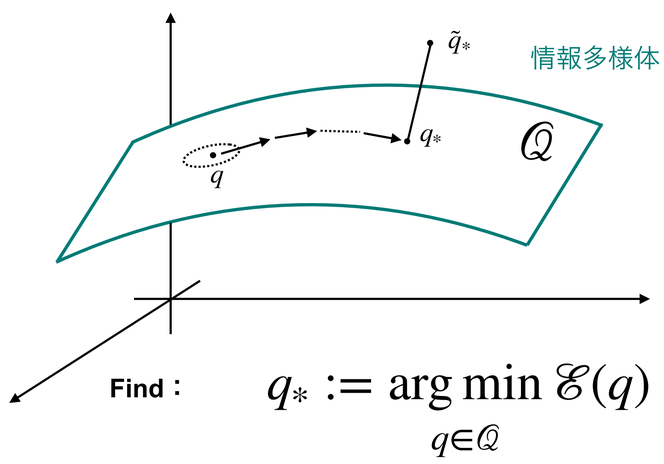
AISTATS 2023

教師付き学習とは、パラメーター付の関数が与えられたとき、データに対する**損失関数**（loss function）が最小となるパラメーターの値を探す問題である。**ベーズ推論**（Bayesian Learning Problem）とは、「値」を「分布」で置き換えて得られる問題である。



損失関数 $\bar{\ell}(\theta)$ の極小値 θ_1^*, θ_2^* は複数ある

エントロピーを極大とする**損失関数の極小を与える分布** $\tilde{q}_*$は一位に存在



Bayesian Update Ruleは、は「良い」分布族を与える情報多様体$\mathcal{Q}$ の中から、最小分布 $\tilde{q}_*$を近似する分布 q_*を見つける手法である。 荒い解から、情報多様体上のニュートン法の類似の手法を用いて、より良い解を見つけ出す。

従来手法であるBayesian Update Ruleでは、２つの制約があった

- 扱える分布族は、基本的にガウス型分布族だけ
- Bayesian Updtae Ruleの適用で、パラメーターが $\mathcal{Q}$ から外れてしまう恐れがある

本研究の**Lie Group Bayesian Rule**は、分布族のパラメーターとして「Lie群」と呼ばれる代数構造をもつ対象を用いるというアイディアに基づいている。Lie群により幅広い分布族を扱えると同時に、Update RuleもLie群の代数構造を用いることにより、パラメーターはLie群の中に収まる形で定義できる。従来手法であるBayesian Update RuleはLie Group Bayesian Ruleの特殊な場合の線形化と解釈できることも示た。様々な分布族やLie群に対して適用でき、従来とは異なる特性を持つ学習アルゴリズムなども見出すことに成功している。

Bayesian Learning Problem

$\ell(\theta)$, a loss function on model parameters $\theta \in \Theta$. We solve

$$q_* \in \arg\min_{q \in \mathcal{Q}} \underbrace{\mathbb{E}_q[\ell] - \tau \mathcal{H}(q)}_{=:\mathcal{E}(q)}$$

for some family of distributions $\mathcal{Q}$ on Θ.

- The expectation $\mathbb{E}_q[\ell] = \int_\Theta \ell(\theta) q(\theta) \mathrm{d}\theta$ prefers regions of ℓ with low loss.
- The entropy $\mathcal{H}(q) = -\int_\Theta q(\theta) \log q(\theta) \mathrm{d}\theta$ prefers a higher spread of q.
- The temperature $\tau > 0$, a balancing term.

Enter Lie Groups

We solve the two issues above by considering families parametrized by a Lie group G. Let G act on Θ, then it acts on a probability distribution by the pushforward action (by reparametrizations). For some distribution q_0

$$\mathcal{Q} := \{q_g = \pi(g) q_0 : g \in G\}$$

where $(\pi(g)q)(\theta) = |\frac{\mathrm{d}(g \cdot \theta)}{\mathrm{d}\theta}|^{-1} q(g^{-1} \cdot \theta)$.

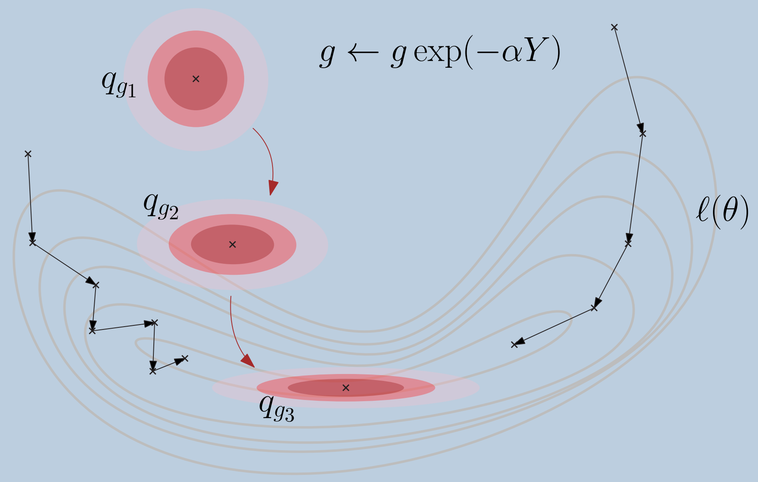
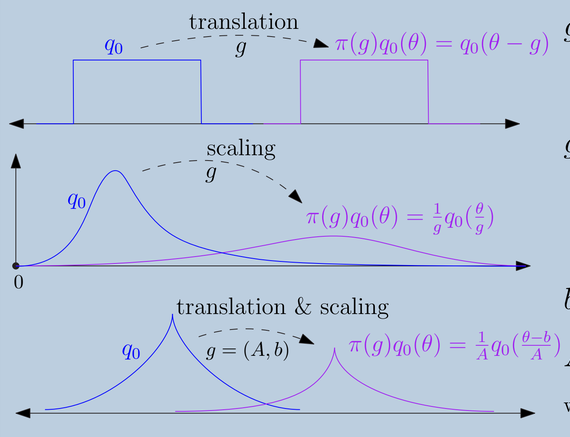


Figure 1: Updating the group element is sufficient to update the distribution. Y is a vector in the Lie algebra (tangent space of G at identity) in the direction of steepest ascent of $\mathcal{E}(q)$ at $q = q_g$ with respect to the Fisher metric. The update always produces a group element g, and therefore a distribution in $\mathcal{Q}$. And we have flexibility in the choice of q_0.

Specific Group Updates



$$g \longleftarrow g - \alpha \mathbb{E}_{q_g}[\partial_\theta \ell]$$

$$g \longleftarrow g \exp\Big(-\alpha \big(\mathbb{E}_{q_g}[\theta \partial_\theta \ell] - \tau \big) \Big)$$

$$b \longleftarrow b + \frac{c_X}{c_y} A \frac{\exp(-\alpha U) - 1}{U} V$$

$$A \longleftarrow A \exp(-\alpha U)$$

where $U = \mathbb{E}_{q_g}[(\theta - b)\partial_\theta \ell] - \tau$

$V = A \mathbb{E}_{q_g}[\partial_\theta \ell]$

In the group setting we turn differentiation w.r.t. distribution parameters to differenting the loss. A family of posteriors with built-in group reparametrizations is the correct setting for this.

Dataset	Metric	Affine uniform	Affine gaussian	Affine laplace	Additive [Or22] uniform	Additive [Or22] gaussian	Additive [Or22] laplace	[SGD]
CIFAR-10	Acc. (↑)	$91.60_{\pm 0.05}$	$91.53_{\pm 0.10}$	$91.87_{\pm 0.04}$	$91.07_{\pm 0.08}$	$91.28_{\pm 0.11}$	$91.14_{\pm 0.12}$	$91.22_{\pm 0.07}$
	NLL (↓)	$0.300_{\pm 0.002}$	$0.294_{\pm 0.004}$	$0.272_{\pm 0.002}$	$0.365_{\pm 0.003}$	$0.328_{\pm 0.008}$	$0.312_{\pm 0.005}$	$0.354_{\pm 0.006}$
	ECE (↓)	$0.040_{\pm 0.001}$	$0.036_{\pm 0.001}$	$0.029_{\pm 0.001}$	$0.052_{\pm 0.001}$	$0.045_{\pm 0.001}$	$0.039_{\pm 0.001}$	$0.050_{\pm 0.001}$
CIFAR-100	Acc. (↑)	$66.08_{\pm 0.12}$	$66.55_{\pm 0.10}$	$66.44_{\pm 0.10}$	$64.29_{\pm 0.09}$	$64.61_{\pm 0.20}$	$64.85_{\pm 0.13}$	$64.19_{\pm 0.14}$
	NLL (↓)	$1.288_{\pm 0.007}$	$1.255_{\pm 0.005}$	$1.247_{\pm 0.006}$	$1.437_{\pm 0.004}$	$1.390_{\pm 0.008}$	$1.359_{\pm 0.006}$	$1.431_{\pm 0.007}$
	ECE (↓)	$0.093_{\pm 0.002}$	$0.079_{\pm 0.002}$	$0.071_{\pm 0.001}$	$0.121_{\pm 0.001}$	$0.107_{\pm 0.001}$	$0.096_{\pm 0.001}$	$0.121_{\pm 0.001}$
TinyImgNet	Acc. (↑)	$51.19_{\pm 0.12}$	$51.13_{\pm 0.16}$	$51.36_{\pm 0.14}$	$49.34_{\pm 0.14}$	$49.62_{\pm 0.15}$	$49.73_{\pm 0.18}$	$49.48_{\pm 0.10}$
	NLL (↓)	$2.099_{\pm 0.005}$	$2.098_{\pm 0.004}$	$2.101_{\pm 0.009}$	$2.234_{\pm 0.008}$	$2.204_{\pm 0.003}$	$2.184_{\pm 0.007}$	$2.231_{\pm 0.004}$
	ECE (↓)	$0.076_{\pm 0.001}$	$0.070_{\pm 0.002}$	$0.065_{\pm 0.001}$	$0.107_{\pm 0.001}$	$0.099_{\pm 0.002}$	$0.089_{\pm 0.001}$	$0.106_{\pm 0.001}$

Sparse and localized behavior of the multiplicative rule

Dataset	Metric	additive	multiplicative
MNIST / MLP	Acc. (↑)	$98.38_{\pm 0.02}$	$98.59_{\pm 0.02}$
	NLL (↓)	$0.083_{\pm 0.001}$	$0.058_{\pm 0.001}$
	ECE (↓)	$0.012_{\pm 0.000}$	$0.006_{\pm 0.000}$
CIFAR–10 / MLP	Acc. (↑)	$58.85_{\pm 0.08}$	$59.19_{\pm 0.07}$
	NLL (↓)	$1.236_{\pm 0.002}$	$1.160_{\pm 0.001}$
	ECE (↓)	$0.085_{\pm 0.001}$	$0.026_{\pm 0.001}$

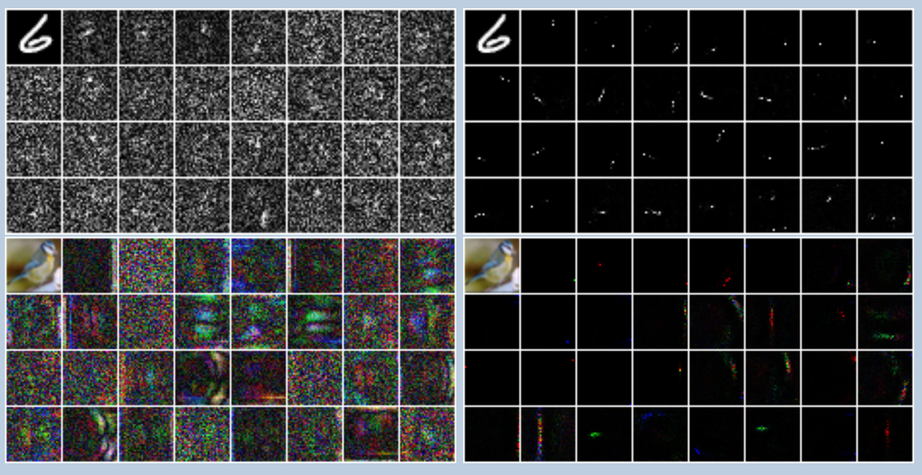


Figure 2: Highest activated first layer filters for the given figure in MNIST (top) or CIFAR10 (bottom); for the additive (left) or multiplicative (right) update rule