Singular Information Geometry

Alexander Gietelink Oldenziel

University at Krakpot

May 12, 2021

Statistical Learning Theory

Let $q(x) : \Omega \to \mathbb{R}^N$ be a distribution (the 'true' distribution). Let $D_n = \{X_1, \cdots X_n\}$ be a n-sample of $q(x)$, i.e $X_1, \cdots, X_n : \Omega \to \mathbb{R}^n$ are random variables independently distributed according to a distribution $q(x)$.

Definition

A *statistical model* for $q(x)$ is a conditional probability density function $p(x|w) : \Omega \times W \to \mathbb{R}^N$ where $w \in W \subset \mathbb{R}^d$ is a d-dimensional parameter space.

In statistical learning theory we are interested in producing a 'best' distribution $\hat{q}_n(x)$ given the data D_n.

Fischer Information matrix

Definition

Let $p(x|w) = p_w(x)$ be a statistical model, where $w \in \mathbb{R}^d$. The Fischer information matrix is given by

$$I(w) = I_{jk}(w) = \int \frac{\partial}{\partial w_j} \log p(x|w) \cdot \frac{\partial}{\partial w_k} \log p(x|w) dx$$

where $1 \leq j, k \leq d$.

The Fischer information matrix is symmetric and positive semi-definite. We will see that it is not always positive definite however, i.e. it can have zero eigenvalues (singularities!).

Statistical Models

Let $W_0 := \{w \in W : q(x) = p(x|w)\}$

Definition

We say $q(x)$ is realizable if W_0 is nonempty. We will assume our models are realizable.

We say a model $(q(x), p(x|w), W \subset \mathbb{R}$ is identifiable if $w \mapsto p(x|w)$ is injective.

Definition

A model $(q(x), p(x|w)$ is regular if it identifiable and its Fischer matrix $I(w)$ is positive definite. It is singular if it is not regular.

For now let us assume our models are regular.

Likelihood and Kullback-Leibler Divergence

Definition

For a random sample $D_n = \{X_1, \cdots, X_n\}$ and a statistical model $(q(x), p(x|w))$, the likelihood function $L_n(w)$, the Kullback-Leibler divergence and the sample KL-divergence as

- The likelihood $L_n(w) := \Pi_{i=1}^n p(X_i|w)$
- The KL-divergence $KL(q(x), p(x|w)) = K(w) := \int q(x) \log \frac{q(x)}{p(x|w)} dx$. It is also known as the generalization error
- The sample KL-divergence $K_n(w) = \sum_{i=1}^n q(x_i) \log(\frac{q(x_i)}{p(x_i|w)})$. It is also known as the training error.

We want to minimize the KL-divergence.

Remark

An advantage of the MLE is that likelihood does not depend on the 'true' distribution $q(x)dx$.

Maximum Likelihood Estimator

The Maximum Likelihood Estimator principle (MLE) says that we should pick the hypothesis $\hat{q}_{n,MLE} = p(x|w^*)$ with the highest likelihood $L_n(w^*) = \max_{w \in W} L_n(w)$. Note that

$$-\frac{1}{n}\log(L_n(w)) = K_n(w) - \frac{1}{n}\sum_{i=1}^{n}\log q(X_i) = K_n(w) + S_n$$

where S_n denotes the empirical entropy, so maximizing the likelihood means minimizing the sample Kullback-Leibler divergence.

Remark

However(!!) this is not the same as minimizing the Kullback-Leibler divergence - basically because of overfitting. This is the basic reason why statistical learning is not a simple optimization problem.

Example - Two-dimensional Gaussian

Let a parametric probability density function of $(x, y) \in \mathbb{R}^2$ for a given parameter $\mathbb{R}^2$ be defined

$$p(x, y|a, b) = \frac{1}{2\pi} \exp(-\frac{(x-a)^2 + (y-b)^2}{2})$$

For given random samples (x_i, y_i) the likelihood function is

$$L_n(a, b) = \frac{1}{(2\pi)^n} \exp(-\frac{1}{2} \sum_{i=1}^{n} (x_i - a)^2 + (y_i - b)^2)$$

Example - Two-dimensional Gaussian (continued)

If the true distribution $q(x, y) = p(x, y|a_0, b_0)$ then the sample KL-divergence is

$$K_n(a, b) = \frac{a^2 - a_0^2 + b^2 - b_0^2}{2} - (a - a_0)\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - (b - b_0)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

The KL-divergence is

$$K(a, b) = \frac{1}{2}[(a - a_0)^2 + (b - b_0)^2]$$

The Fisher information matrix is everywhere

$$I(a, b) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Fischer Information: Curvature of the KL-divergence

The Fischer information matrix is equal to the Hessian matrix of the Kullback-Leibler distance at the true parameter.

Proposition

Let $w_0 \in W_0 \subset W$ be a true parameter. Then

$$I_{jk}(w_0) = \frac{\partial^2}{\partial w_j \partial w_k} K(w_0)$$

Proof.

Integration by parts. □

Fischer Information: Jeffrey Prior

Suppose we want to pick a prior $\phi(w)$ on a parameter space $W \subset \mathbb{R}^d$. [This will be important when we consider Bayesian learning theory] One important desideratum is that our prior $\phi(w)$ should not depend on how the $p(x|w)$ are parameterized in any essential way.

Jeffrey Prior

The Jeffrey prior is defined as

$$\phi_{Jeffrey}(w) := \sqrt{\det I(w)}$$

The Jeffrey prior has the important property that if $p(x|g(w))$ is another parameterization of the statistical model then the priors $\phi(w), \phi'(g(w))$ are related by the usual change of variables

$$\phi(w) = \phi'(g(w)) \cdot |\frac{dg}{dw}|$$

Fischer Information: Cramer-Rao

Let $p(x|w), w \in W \subset \mathbb{R}^d$ be a parameterized probability distribution. Let D_n be a n-sample from $p(x|w)$.

Cramer-Rao

Let $\hat{w}_n$ be an unbiased estimator of w based on D_n. Then the covariance matrix $Cov(\hat{n}_n)_{jk} := \mathbb{E}[(\hat{w}_j - w_j)(\hat{w}_k - w_k)]$ of $\hat{w}_n$ is bounded from below:

$$Cov(\hat{w}_n) \geq \frac{I(w)^{-1}}{n}$$

Corollary

In particular, if $d = 1$, we have

$$var(\hat{w}) \geq \frac{1}{n \cdot I(w)}$$

Fischer Information: Riemannian metrics

The Fischer Information matrix $I_{jk}(w)$ defines a Riemannian metric on W. Given a path $\gamma : [0, \tau] \to W$ this gives a length

$$L(\gamma) = \int_0^\tau \sqrt{\frac{d\gamma^j(t)}{dt} I_{jk}(\gamma(t)) \frac{d\gamma^k(t)}{dt}} dt$$

Given two points $w_1, w_2 \in W$ we define the metric distance $\mathcal{L}(w_1, w_2)$ to be the length $L(\gamma_0)$ of a shortest path (geodesic) between w_1, w_2.

Theorem

Let $p_1 = p(x|w_1), p_2 = p(x|w_2)$ where $w_1, w_2 \in W \subset \mathbb{R}^d$ be two probability distributions. The Fischer distance equals the symmetrized Kullback-Leibler divergence:

$$\mathcal{L}(p_1, p_2) = \frac{1}{2} K(p_1, p_2) + \frac{1}{2} K(p_1, p_2)$$

Fischer Information: Thermodynamics

We can also think of W as parametrizing thermodynamic macostates, where the parameters $w \in W$ parameterize conjugate variables (temperature, pressure, etc).

Theorem

Let $a, b \in W$ be two thermodynamcis states. The square of the Fischer distance $\mathcal{L}^2$ gives a lower bound on the total entropy production of a thermodynamic transformation in the quasi-static limit[1].

Without going in too much detail, the quasi-static limit this means we start in a thermodynamic state a, change the conjugate variables w_j in very small steps Δw_j and let the system equilibrate after each step Δw_j until we end up at the endpoint b.

Fischer Information: Rate of Evolution

Let's start by assuming we have different kinds of self-replicating entities with populations $P_1, \cdots, P_n$ evolving according to the replicator equation

$$\frac{dP_i(t)}{dt} = f_i P_i(t)$$

after normalizing we get

$$\frac{dp_i}{dt} = (f_i - \langle f \rangle) p_i(t)$$

where $p_i(t) = \frac{P_i(t)}{\sum_{i=1}^{n} P_i(t)}$ and $\langle f \rangle = \sum_{j=1}^{n} f_j p_j(t)$ We call f_i the fitness of species i and $\langle f \rangle$ the mean fitness.

Fischer Information: Rate of Evolution (continued)

Mathematically, we have described a curve $p(t)$ in some ambient parameter space. The following theorem is sometimes described as: "The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time." - but see below

Theorem(Baez-Fisher)

The Fischer information

$$I(t) = |\frac{dp}{dt}|^2 = \sum_i (f_i - \langle f \rangle)^2 p_i = \text{var}(f)$$

Recall also that the Fischer information equals the second derivative of the KL-divergence:
$|\frac{dp}{dt}|^2(t_0) = \frac{d^2}{dt^2} KL(p(t), p(t_0))|_{t=t_0}$.

Neural Networks are Singular!

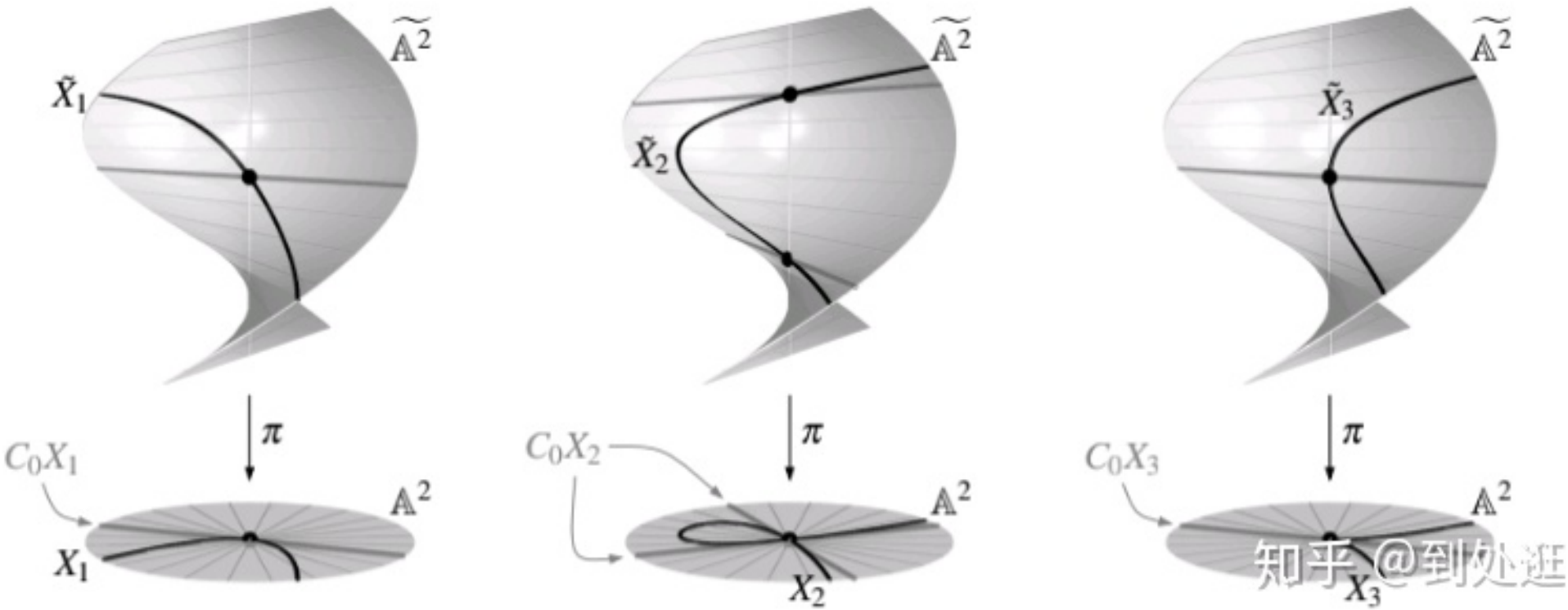
Let $p(x, y|a, b) = q_0(x)\frac{1}{\sqrt{2\pi}}\exp(\frac{1}{2}(y - a \cdot \tanh(bx))^2$ where $q_0(x)$ is a constant probability density function of x, $x \in \mathbb{R}^1$, $y \in \mathbb{R}^1$ and $(a, b) \in \mathbb{R}^2$. This is the simplest three-layer neural network.
One can check that if $ab = 0$ the Fisher Information matrix is degenerate.

Remark

In fact, almost every statistical model is singular! Layered neural networks, normal mixtures, Boltzmann machines, Bayes networks, Hidden Markov models, etc etc are all generically singular.

Blow-ups

Algebraic geometers know what to do with singularities: blow them up!



By using succesive blowups we can resolve (Hironaka's resolution of singularities) a singular variety W into a smooth manifold $b : M \to W$.

Bayesian Learning Theory

Definition

A Bayesian statistical model is a statistical model $(q(x), p(x|w)), W \subset \mathbb{R}^d$ together with a distribution $\phi(w)$, the 'prior', on W.

The Bayesian predictive distribution is defined as

$$\hat{q}_{n,Bayes}(x) = \int p(x|w)p(w|D_n)dw$$

Birational Geometry

Let $(q(x), p(x|w), \phi(x)), w \in W \subset \mathbb{R}^d$ be a Bayesian statistical model. Assume it is realizable. Using resolution of singularities we can define a birational invariant λ.

Theorem

- If $(q(x), p(x|w), \phi(x))$ is regular then $\lambda = \frac{d}{2}$.
- If $(q(x), p(x|w), \phi(x))$ is singular then $\lambda < \frac{d}{2}$.

Generalization Error

Let $\hat{q}_n(x)$ be some estimate of the true unknown distribution $q(x)$ based on the dataset D_n.

Definition

The generalisation error of the predictor $\hat{q}_n(y|x)$ is

$$K(\hat{q}_n) :== \int q(x) \log \frac{q(x)}{\hat{q}_n(x)} dx.$$

The average generalisation error over a sample D_n is denoted $\mathbb{E}_n K(\hat{q}_n)$.

Theorem (Watanabe)

Let $\hat{q}_n(x) = \hat{q}_{n,Bayes}$ be the Bayesian predictive distribution. Then

$$\mathbb{E}_n G(n, \hat{q}_n) = \frac{\lambda}{n} + o(\frac{1}{n})$$

if $\hat{q}_n$ is the Bayes predictive distribution[2].

In other words the birational invariant λ is the *learning coefficient*!

MLE and Bayes predictive distribution

Let $\hat{q}_{n,MLE}(x)$ be the MLE estimator

Theorem (Watanabe)

There is a constant C such that

$$\mathbb{E}_n K(\hat{q}_{n,MLE}) = \frac{C}{n} + o(\frac{1}{n})$$

if the statistical model $(q(x), p(x|w), \phi(x))$ is regular then $C = \frac{d}{2}$.

In general $C > \lambda$ so in singular situations the Bayesian predictive distribution outperforms MLE.

Why do overparameterized models work so well in ML?

The learning coefficient λ is generally much smaller than $\frac{d}{2}$. It seems that existing techniques in ML are able somehow able to effectively approximate the Bayesian predictive distribution.

Conclusion and further questions

Singular learning theory and Information geometry forms a powerful framework for machine learning and artificial intelligence!

- In practice it is hard to calculate the Bayesian predictive distribution $\hat{q}_{n,Bayes}(x)$. Approximation techniques exist (variational Bayes / mean field approximation...) but it is an open question if similar generalization error bounds hold for them.
- It is hard to calculate the birational invariant λ for large neural networks. *Can Compositionality help in calculating λ for large neural networks?*


Gavin E. Crooks.
Measuring thermodynamic length.
Physical Review Letters, 99(10), Sep 2007.


Daniel Murfet, Susan Wei, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella.
Deep learning is singular, and that's good, 2020.