Machine Learning Talk II

Information Geometry & Learning

Axel G. R. Turnquist

NJIT Department of Mathematical Sciences

October 2, 2020

The Case for Probability Theory

- Probabilistic methods avoid the **curse of dimensionality**
- Gives access to probabilistic inequalities (useful in **ergodic dynamical systems**, analyzing **rare events**)
- Can solve problem by projecting into higher space. Helpful for optimization. Idea:

$$\min f(x) \leq \int f(x) d\mu \tag{1}$$

  Optimization + noise gives **Euler-Lagrange equations** in some higher dim. space. ODE $\rightarrow$ PDE, PDE $\rightarrow$ master eqns.
- Seems useful in proof-building, e.g.:

$$A \implies B, a.s., \text{in probability}, \text{in distribution} \tag{2}$$

- Phenomenon is fundamentally **stochastic**. How can you test for this? Idea: run **time/space correlation**.

Guiding Questions for This Talk

1. What is entropy and what is its role in learning?
2. Is a statistical manifold infinite dimensional?
3. What is the role of the prior in learning? Do we need it?
4. What is the rate of learning?
5. Can't I just run an ordinary gradient descent?
6. What's the connection with the Wasserstein distance?

The Case for Information Manifolds

Why do we want to use manifolds to describe statistical inference and learning? Manifolds allow for us to perform calculus, define vectors, etc. This allows us more easily to perform:

- **Asymptotic analysis**. Where second-order asymptotics will necessarily involve the concept of **curvature**.
- **Projection theorems**. By defining orthogonality.
- **Gradient descent**. Performing a gradient in parameter space (low-dimensional).

Entropy

In the information theoretical context, information is known as **Shannon information**. Key points:

- Low probability event is highly informative.
- High probability event, being predictable imparts little information.
- The information from event A *and* event B should be additive, i.e. $I(A \cap B) = I(A) + I(B)$, for events A and B (in the σ-algebra)

Choose: $I(x) = -\log(p(x))$. But, rare events happen rarely. So, the **Shannon entropy** from the system is weighted by how likely the event is (the "average" info. given by a prob. distribution):

$$S = -\int p(x) \log(p(x))\, dx \tag{3}$$

Statistical Learning vs. Regression

Statistical learning involves a **statistical model** $M = \{p(x, \xi)\}$, parametrized by ξ, which is typically from a subset of $\mathbb{R}^n$. The problem is then, given data, find the "best" ξ. Can do this by:

1. Find f, such that $\hat{\xi} = f(x_1, \ldots, x_N)$ is a good estimator of ξ.
2. Formulating it as an optimization problem and proceeding directly to find the optimal ξ (**natural gradient**)

Blackbox **regression** merely seeks to fit the unknown function, as in machine learning. Ideal: robust regression, easily analyzable. This depends on the data. Overfitting, predictability issues.

Discussion: other reasons to continue with the statistical learning paradigm?

Statistical Learning Point-of-View

We can extract **information** from comparing one distribution to another and playing a game. Say, take a Gaussian prior to learn the mean and variance. The **KL-divergence** tells you how quickly you realize you may be sampling from the wrong distribution:

$$\mathbb{P}[\hat{p}; p] = e^{-ND_{KL}[\hat{p}:p]} \tag{4}$$

where, KL-divergence is:

$$D_{KL}[p : q] = -\int \log\left(q(x)\right) p(x)dx + \int \log\left(p(x)\right) p(x)dx \tag{5}$$

which is the diff. between the average info. of q assuming that it is p and the average info. of p. This is **asymmetric**, which reflects that learning from different dist. affects how quickly one learns.

Rate of Learning

We try to learn:

$$\hat{\xi} = f(x_1, \ldots, x_N) \tag{6}$$

given an appropriate f such that $\hat{\xi} \to \xi$. How far away are we from the real parameter? Defining,

$$V_{ij} = \mathbb{E}\left[(\hat{\xi}_i - \xi_i)(\hat{\xi}_j - \xi_j)\right] \tag{7}$$

we find that:

$$V \geq \frac{1}{N} G^{-1} \tag{8}$$

where G is the Fisher information matrix, which is the Hessian of the KL-divergence calculated at a point in the statistical manifold.

Rate of Learning

To estimate well, we need a good **estimator** and a good **statistical model**. Given the **maximum likelihood estimator** and the **exponential family** (which includes the Gaussians), we find the CR-bound achieved:

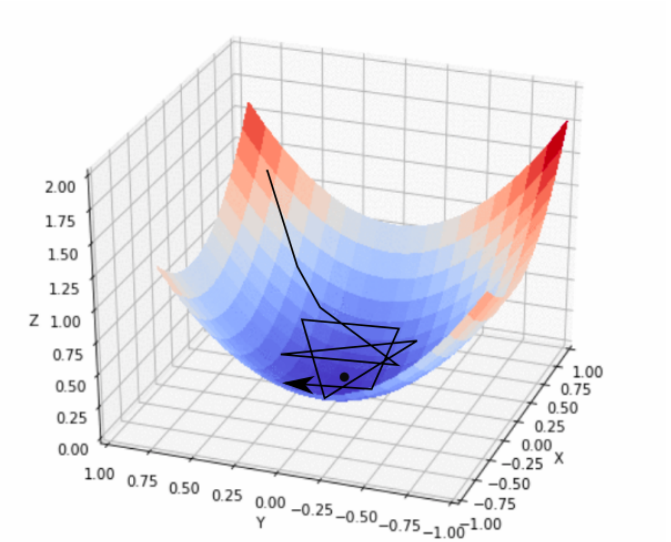
$$V = \frac{1}{N} G^{-1} \tag{9}$$

Note: further asymptotic analysis via geometry reveals that deviations from second-order convergence comes from deviation from the exponential family (**statistical curvature**) and the curvature of the inverse image of the estimator f.

Natural Gradient Approach

Q: Can't we just do a steepest descent in the Riemannian parameter space?
A: Yes, this is called the **natural gradient approach**.

- **Advantages**: Overcomes issue of critical slowdown in deep learning. Bigger neural network is not necessarily better.
- **Disadvantages**: Slow?



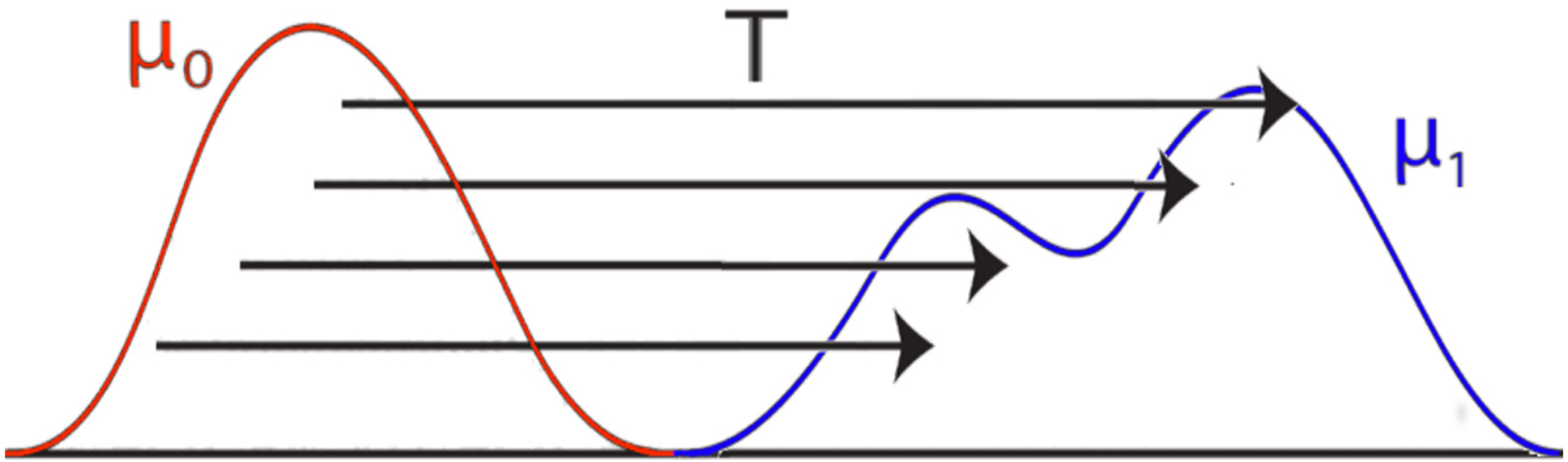
Infinite-Dimensional Point of View — Speculation

Claim: Exponential families are good to use when we *a priori* know very little, and for **supervised** learning problems.

Infinite-Dimensional Perspective:
Can make an infinite-dimensional manifold of probability measures using **Fisher-Rao metric**.

- What if one wants to learn the statistical models? I.e. do **unsupervised learning**? Then maybe it makes more sense to work in the full space.
- What if one is just learning via a neural network and does not really understand the parameter space?

Q: What contributions has optimal transport made to learning?



N. Papadakis, Optimal Transport for Image Processing, habilitation à diriger des recherches, Université de Bordeaux, Dec. 2015

A: Stay tuned for my next lecture...

Questions?

Some Useful Resources

- "Information Geometry and Its Applications" Amari, Shun-ichi.
- "The Concentration of Measure Phenomenon" Ledoux, Michel.
- "High-Dimensional Probability" Vershynin, Roman.
- "Pattern Recognition and Machine Learning" Christopher M. Bishop

Future Talks

Further potential topics:

- Adversarial attacks
- Data augmentation
- ???

Oct 9: Binan Gu