极客大学机器学习训练营 加餐：AI 理论的可能方向

王然

众微科技 AI Lab 负责人

二○二○年十二月三十一日

大纲

1. Why This Talk
2. Why Optimal Transport: its Advantages and its Disadvantages
3. Mathematical Tools
4. How to proceed
5. References

由于本次讲义涉及到很难的数学内容，所以所有内容都将用英文撰写。

Outline

- Why this talk and why we need theories.
- Why optimal transport: its advantages and its disadvantages.
- What mathematical tools you can use and how you use them.

大纲

1. Why This Talk
2. Why Optimal Transport: its Advantages and its Disadvantages
3. Mathematical Tools
4. How to proceed
5. References

Why this talk

- This is from this **lundberg2017unified** and Lei et al. (2019).
- This has been explained by many as the “pinnacle” of AI mathematics research. People show off by saying they can understand this paper.
- However, there is one part that triggered me.

The part that triggers me

Definition 3.3 (Optimal Transportation Map) The solutions to the Monge's problem is called the optimal transportation map, whose total transportation cost is called the Wasserstein distance between μ and ν, denoted as $\mathcal{W}_c(\mu, \nu)$.

$$\mathcal{W}_c(\mu, \nu) = \min_{T_\# \mu = \nu} \int_X c(x, T(x)) d\mu(x)$$

Why this triggers me

- Basically, with all these fancy theories, by adding so much regularity condition, which does not make sense, we finally arrive at Wasserstein distance.
- Therefore, everything, the optimal transport, Riemmian Geometry is to only introduce Wasserstein distance.
- Thus why not start from the Wasserstein distance already.

Why this triggers me

- What triggers me, even more is that the why we care about Wasserstein distance is because of its connection with W-GAN.
- Basically, we use math as a boosting tool to devise something nobody will use.

Why this triggers me

- Now, it must be said, I am not trying to say the authors are awful people. On the contrary, we respect them a lot.
- The problems lie in how papers are published every day.
- In the past years, mathematical results will cost ten or more years to formulate. That is because important problems need more time.
- Now, if you want to keep a job, you have to squeeze out something every day. Therefore, nobody is daring to do anything anymore. This is even true even for the mathematical community, which has the most patience.

An example

- In 2002, my mentor Prof. van der Vaart was asked to review the work of Prof Le Cam (Vaart et al. 2002), who is the most important statistician ever.
- The publication is on Annals of Statistics, the highest-ranking journal with the most theoretical focus, and in the first draft, he wrote something.
- “If his work will be submitted today, I wonder how many of the reviewers of this journal will even understand and be interested. ”
- You understand why it is deleted.

But why we even need theories?

- A unified theory has the potential of creating a fair race between AI papers.
- A unified theory might be able to help us understand the deep connection between different things. It has been long suspect that GAN, VAE, and RL are the same thing, which is all notoriously difficult to train. Thanks to Levine (2018), this has been (almost) true from the perspective of Variational Bayes. Therefore we can check when something goes wrong, whether there is a corresponding bahevior in other areas and what are their solution.

The problem with optimal transport as it is

- It is perfectly possible to use optimal transport theory to unify the above things.
- However, it is impossible since we impose too much regularity conditions. Conditions for the sake of conditions.
- With its current stage, it will never produce any useful unifying results.

But how about Le Cam

- However, you might want to say, Le Cam already built an empire in Le Cam (2012) and later expanded Bickel et al. (1993). Therefore, why even bother with optimal transport.
- The problem is, when Le Cam, Klaassen, and van der Vaart worked on that field, there is no optimizing difficulty. The reason is that all there is a parametric model, and any reasonable estimation procedure will converge.
- This is not at all the case with deep learning, notwithstanding GAN, VAE, and RL.

大纲

1 Why This Talk

2 Why Optimal Transport: its Advantages and its Disadvantages

3 Mathematical Tools

4 How to proceed

5 References

Optimal Transport as it is

First, let us look at the original optimal transport object.

$$\min_{T_{\#}\mu=\nu} \int_X c(x, T(x))\, d\mu(x) \tag{1}$$

The Intuitive Explanation 1

- We are moving something from one point to another.
- The cost is c.
- T is the transport plan.
- μ is a measure that tells us how easy/difficult it is to move.

Why this is not useful

- The cost has too much restriction. In most cases, it can only be L_1 loss.
- The x usually resides on a space that is too simple. To incorporate everything, x has to lie in the space of measures.
- T is way too restrictive and too simple. The measure-preserving thing only leads to the benefit of useful solutions.

Why this is not useful

- T itself has to be decomposed as least to three operators. One denotes the learning algorithm (such as policy gradient vs. Q-learning), one denotes the model, and one denotes the optimization process (like Adam vs. AdamW).
- One step transportation is useless, since models are trained in multi-step, so we need a sum of step-wise T to measure the whole thing. However, this might be optional.

Why this is not useful

- Now the measure μ is on the measure x, itself becomes a “Random Measure”.

You have had your fun, can we leave now?

- With this setting, there is no way we can find any explicit solution.
- Therefore, the goal is to do to approximation theory.
- I understand even that sounds horrible. If we cannot even figure out abound for Adam, how am I even approaching this problem?
- However, are we solving the right problem?

Going back to Le Cam

- This is the same issue faced by statisticians a century ago.
- There are so many models, each model requires detailed study, so how are we going to do it.
- Nowadays, that will be good news, since there is more paper to publish.
- Back in the day of Le Cam, this is terrible.

The Genius of Le Cam

- Instead of solving the problem itself, let us first solve something much simpler. For example, how about estimating many Gaussian means.
- We create a distance between the original problem and the easy problem. As long as we solve the easy one, the actual performance can be easily measured.
- For a detailed review, see Vaart et al. (2002).

OK, let me the genius play their games ...

极客大学

- However, it is not that hard. In fact, as contradictory as it sounds, it is not that hard to do.
- In fact, the most surprising thing is that this is a theory that potentially link every part of what is considered abstract optimizations.
- And you don't even need to read many books.

大纲

1. Why This Talk
2. Why Optimal Transport: its Advantages and its Disadvantages
3. Mathematical Tools
   - Abstract Ones
   - Inequalities
4. How to proceed
5. References

大纲

1. Why This Talk
2. Why Optimal Transport: its Advantages and its Disadvantages
3. Mathematical Tools
   - Abstract Ones
   - Inequalities
4. How to proceed
5. References

Let us start with Le Cam

- The book by Le Cam himself (Le Cam 2012) has many great ideas. However, it is not easy to follow.
- It builds on two things, General Topological Vector Space Theories and Riesz Lattices.
- The best book for Topological Vector Space is Schaefer (1971), but it is a little bit old so maybe you want to try Bogachev, Smolyanov, and Sobolev (2017). Furthermore, as a foundation, an introduction to uniform space(Bourbaki 1995).

Let us start with Le Cam

- Risez Lattice is a little tricky. It has very nice properties. Basically, you can always think about bands as intervals and then you have weak compactness. Also, the L-space can be identified with L_1 and its dual with similar condition, so you have all the nice properties (like Dunford-Pettis).
- The book is Meyer-Nieberg (2012), but it is a little outdated.

What about the algebraic part?

- You can say that that following book covering most(?) of the geometric part of functional analysis. However, what about the algebraic part.
- Now, here is the thing, The optimal transport theory is built on Rienmanian manifold. The superficial advantages is that you have Geodesics and Curvature to help find the shortest path.
- However, there is something deeper that I seldom see used. It is a Hilbert Space, and the most beautiful part of operator algebras are from operators on Hilbert Space. Now, any operator theory can be applied. A reference can be Conway (2000).

Now about Hilbert Space

- Why not making it a RKHS?
- A reference in RKHS is Berlinet and Thomas-Agnan (2011).
- However, it does not show the power of RKHS, the power can be found in Eggermont, LaRiccia, and LaRiccia (2001), the first real proof.

Hilbert Space is a Banach Space

- Remember the μ? It sounds horrible right? How are going to deal with a measure on something as abstract as a Riemmanian Manifold?
- The it is a Banach Space, therefore, you can use Ledoux and Talagrand (2013). In fact, r.v. defined on Banach Space sometimes has better properties than their counter part.
- However, there is also a book on stochastic analysis on that thing if you want (Hsu 2002), but it does not reveal the full potential of stochastic calculus.

The True Power of Stochastic Calculus

- Can only be fully visible from Jacod and Shiryaev (2013).
- Frankly, it is not modeled on a manifold, but ideas might still prove useful.
- Also, it attacks the same problem Le Cam attacks from a completely different angel.
- A reminder though, instead of using skorohod topology might be replaced by van der Vaart's approach in Van Der Vaart and Wellner (1996).

Now let us go back to Riemannian Geometry

- Now that we are working on measures, can we even build the manifold on an infinite dimensional space?
- Answer is yes. You can do manifold on Banach Space as is pointed out by Lang (2012).
- Don't forget Berger (2012).

Finally Le Cam

- Read the original work (Le Cam 2012).
- If that is too difficult, read Vaart (2000) to understand the mathematics first.
- Then, read the semiparametric book Bickel et al. (1993), that will make you very comfortable with dealing with infinite dimensional space.

大纲

1 Why This Talk

2 Why Optimal Transport: its Advantages and its Disadvantages

3 Mathematical Tools

- Abstract Ones
- Inequalities

4 How to proceed

5 References

Just a few choices

In my opinion, just get familiar with the training techniques is enough, here are some list.

- A book on mathematical inequalities (Pachpatte 2005)
- A book on probabilistic inequalities (Lin and Bai 2011)
- A book on fancy tricks (Giné and Nickl 2016)
- A paper on tricks known as oracle inequalities (Castillo, Schmidt-Hieber, Vaart, et al. 2015).
- Sololev type inequalities can also help (Hebey 2000).

大纲

1. Why This Talk
2. Why Optimal Transport: its Advantages and its Disadvantages
3. Mathematical Tools
4. How to proceed
5. References

It looks scary but ...

You only need to carefully read Bogachev, Smolyanov, and Sobolev (2017), Meyer-Nieberg (2012), Conway (2000), Jacod and Shiryaev (2013), and Lang (2012).
The rest you browse through it, and just remember to find the results when needed. Furthermore, there are always papers you can directly cite.

Questions?

Thank You!

大纲

极客大学

1. Why This Talk
2. Why Optimal Transport: its Advantages and its Disadvantages
3. Mathematical Tools
4. How to proceed
5. References

- Berger, Marcel (2012). *A panoramic view of Riemannian geometry*. Springer Science & Business Media.
- Berlinet, Alain and Christine Thomas-Agnan (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.
- Bickel, Peter J et al. (1993). *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Johns Hopkins University Press Baltimore.
- Bogachev, Vladimir Igorevich, Oleg Georgievich Smolyanov, and VI Sobolev (2017). *Topological vector spaces and their applications*. Springer.
- Bourbaki, Nicolas (1995). *Elements of mathematics: General topology*. Springer.
- Castillo, Ismaël, Johannes Schmidt-Hieber, Aad Van der Vaart, et al. (2015). "Bayesian linear regression with sparse priors". In: *The Annals of Statistics* 43.5, pp. 1986–2018.

- Conway, John B (2000). *A course in operator theory*. American Mathematical Soc.
- Eggermont, Paulus Petrus Bernardus, Vincent N LaRiccia, and VN LaRiccia (2001). *Maximum penalized likelihood estimation*. Vol. 1. Springer.
- Giné, Evarist and Richard Nickl (2016). *Mathematical foundations of infinite-dimensional statistical models*. Vol. 40. Cambridge University Press.
- Hebey, Emmanuel (2000). *Nonlinear Analysis on Manifolds: Sobolev Spaces and Inequalities: Sobolev Spaces and Inequalities*. Vol. 5. American Mathematical Soc.
- Hsu, Elton P (2002). *Stochastic analysis on manifolds*. Vol. 38. American Mathematical Soc.
- Jacod, Jean and Albert Shiryaev (2013). *Limit theorems for stochastic processes*. Vol. 288. Springer Science & Business Media.

- Lang, Serge (2012). *Fundamentals of differential geometry*. Vol. 191. Springer Science & Business Media.
- Le Cam, Lucien (2012). *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.
- Ledoux, Michel and Michel Talagrand (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.
- Lei, Na et al. (2019). “Mode collapse and regularity of optimal transportation maps”. In: *arXiv preprint arXiv:1902.02934*.
- Levine, Sergey (2018). “Reinforcement learning and control as probabilistic inference: Tutorial and review”. In: *arXiv preprint arXiv:1805.00909*.
- Lin, Zhengyan and Zhidong Bai (2011). *Probability inequalities*. Springer Science & Business Media.
- Meyer-Nieberg, Peter (2012). *Banach lattices*. Springer Science & Business Media.

- Pachpatte, Baburao G (2005). *Mathematical inequalities*. Elsevier.
- Schaefer, H (1971). “Topological Vector Spaces [Russian translation]”. In: *Mir, Miseow*.
- Vaart, Aad van der et al. (2002). “The statistical work of Lucien Le Cam”. In: *The Annals of Statistics* 30.3, pp. 631–682.
- Vaart, Aad W Van der (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.
- Van Der Vaart, Aad W and Jon A Wellner (1996). “Weak convergence and empirical processes”. In: *Weak convergence and empirical processes*. Springer, pp. 16–28.