02 人工智能的数理基础

伊利诺伊大学香槟分校孙若愚：从全局景观的角度分析并改进 GAN

整理：智源社区 杨埔

孙若愚本次演讲的主题为《Global Landscape of GANs: Analysis and Improvement —how two lines of code change makes difference》。

孙若愚，伊利诺伊大学香槟分校（UIUC）助理教授，北京大学数学系理学学士，明尼苏达大学电气工程专业博士。曾经是斯坦福大学管理科学与工程系博士后研究员，Facebook 人工智能研究中心客座研究员。孙若愚在深度学习的优化、机器学习的非凸优化、大规模优化中都做出了重要贡献。个人主页：https://ruoyus.github.io/

在演讲中，首先，对于目前常用的 GAN（Generative Adversarial Network，生成式对抗网络），孙若愚指出了在概率空间上研究的缺陷并提出在函数空间上研究。然后在函数空间上，通过观察函数的景观（landscape），他发现传统的 JS-GAN 有“盆地”，这很有可能导致训练陷入严格的局部最小点无法逃逸，使得模型崩溃。而 RS-GAN（Relativistic Standard GAN）没有这样的盆地，也就是有更好的景观。这样的现象在理论上得到了证明，并且在实验中也验证了 RS-GAN 比 JS-GAN 表现得更好。这项工作让我们对 GAN 有了更好的理论上的理解，它是由孙若愚与 Tiantian Fang, Alex Schwing 合作完成，原始论文见参考文献（《On Understanding the Global Landscape of Generative Adversarial Nets》）。

一、传统 GAN 模型的背景介绍

GAN 是一种非常流行的生成模型。随着时间的推移，GAN 的功能逐渐强大，生成的目标越来越接近真实目标，主要应用于还原受损的图片、风格迁移、生成图像、生成视频等等。

GAN 面临两个非常困难的挑战： 第一，是 GAN 调参特别困难，需要花费大量时间调参才能让模型起作用； 第二，是 GAN 模型非常庞大，需要花费大量的计算资源。这两点导致了 GAN 难以训练，所以人们希望通过对 GAN 的理论理解去设计更好的算法。

传统的 JS-GAN 本质上是让生成样本的概率分布和真实数据的概率分布尽量接近，数学化的表达就是一个 min-max 优化问题:

- The problem is $\min_{p_g} \phi(p_g, p_{\text{data}})$, (1)

where $\phi(p_g, p_{\text{data}}) = \max_D E_{x\sim p_{\text{data}}, y\sim p_g} \log(D(x)) + \log(1 - D(y))$.

- Equivalent to min max L(p_g , D), for certain L.
- **Sanity check:** Loss $\phi(p_g, p_{\text{data}})$ is minimized iff $p_g = p_{\text{data}}$.
- **Math subject:** min-max optimization, game theory, probability

图 1：传统的 JS-GAN

对 GAN 的理论研究分为两类。一是从统计的角度，例如关于 JS 距离（参考文献 Goodfellow et al’14）、W–GAN（参考文献 Arjovsky & Bottou, 2017）、f–GAN（参考文献 Nowozin et al.’16）以及泛化边界（参考文献 Arora, Ge, Liang, Ma, and Zhang, 2017）等等。二是从优化的角度，例如考虑收敛到局部最小点或者稳定点（参考文献 Daskalakis et al., 2018; Daskalakis & Panageas, 2018; Azizian et al., 2019; Gidel et al., 2019; Mazumdar et al.; Yazıcı et al., 2019; Jin et al., 2019; Sanjabi et al., 2018）。

孙若愚这里用了一个非常形象的图来讲述了研究模型理论由易到难的三步：S1 概率空间（pdf space），S2 生成函数空间（generator function space），S3 参数空间（parameter space）。理论和实际的差距就在于，在实际中我们都是在参数空间，而理论上通常只能研究概率空间。在优化理论上有四步：01 需要全局最小点吗，02 存在坏局部最小点吗，03 如何收敛到全局最小点，04 收敛速度。

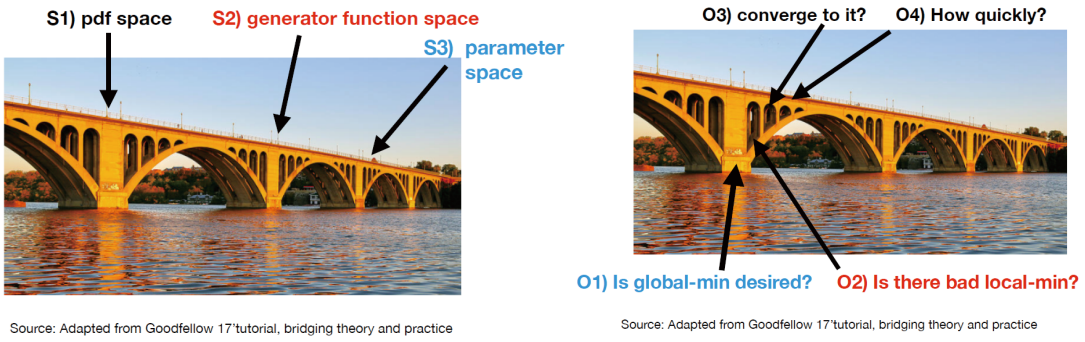


图 2：从简单的理论模型过渡到复杂的理论模型

他这项工作最大的进步就在于从概率空间上的理论研究进步到了生成函数空间和参数空间上。优化理论是在 01 和 02 上。接下来的部分，孙若愚重点研究了生成函数空间和局部最小点空间的情况，因为这种情况最有启发性，能帮助我们理解 GAN。研究清楚这种情况之后，其余三种情况都能够很容易地得到对应的结论。

	(S1) pdf space	(S2) G function space	(S3) parameter space
(O1) Sanity check	[Goodfellow et al. 14]	This work	This work
(O2) Local-min are good?	[Goodfellow et al. 14]	This work	This work
(O3,4) Convergence to local-min	Nagarajan & Kolter, 2017;		Mescheder et al. '18 (linear D), Sanjabi et al.'18, Jin et al.'19, Chu et al. '20, Daskalakis et al.'18,Yazıcı et al.'19, Gidel et al.'19

图 3：GAN 的优化分析

二、经验损失（Empirical Loss）和泛化损失（Population Loss）

为什么要考虑生成函数空间呢?

考虑这个 Min–Max 优化问题上: $\min_{p_g}\max_{D} E_{x\sim p_{data},y\sim p_g}\log D(x)+\log\left(1-D(y)\right)$，目标函数是期望的形式，称为泛化损失 (Polulation Loss)。参考文献《Generative Adversarial Nets》作为研究 GAN 的第一篇文献，就提到了目标函数 $\phi_{JS}\left(p_g,p_{data}\right)$ 对于 p_g 是凸的，也就是说可以收敛到全局最小值 $p_g = p_{data}$。许多传统的分析方法也是在概率空间上来做的，包括理论的 (参考文献 Chu, Blanchet and Glynn’19, Johnson and Zhang’19) 和实验的 (参考文献 Gong et al’19, TAC–GAN)。虽然这样的方法可以在参数空间上将优化问题“凸化”，但事实上这对于所有的目标函数都是对的。具体来说，概率密度的任何线性函数对于这个概率密度都是凸的。注意到这是一个函数空间的函数，所以它与传统的优化问题有所不同，要优化的变量本身就是一个函数。可以总结为这样的定理: 对于任意函数 f，$\mathrm{E}_{y-p_g}f(y)$ 对于 p_g 总是凸的。这一方面说明对于这样的目标函数来优化 p_g，优化问题总是比较容易的; 但另一方面则告诉我们，概率空间的观点根本不会利用 GAN 的结构特性。按照这样的理论，对于任何的函数都能转化为一个 p_g 的凸函数，都能够很容易地优化，这显然与我们的经验不太相符。

因此，孙若愚自然地关注经验损失 (Empirical Loss) 而非上面的泛化损失，从函数空间的角度来考虑。也就是说，真实数据分布 p_{data} 现在是 n 个固定的真实样本 $x_1,x_2,\ldots,x_n\in\mathbb{R}^{\mathrm{d}}$，目标生成分布 p_g 也不再是概率空间，而是样本 $Y=\left(y_1,y_2,\ldots,y_n\right)\in\mathbb{R}^{n\times d}$ 的函数空间，并且要使得 Y_i 与 X_i 尽量接近。有人可能会对泛化性产生质疑: 如果总是让生成的样本与原始样本相似，会不会导致过拟合? 结论是不会的，参考文献 (Arora et al’18) 对 GAN 泛化的界做了相关的理论分析，而他也得到类似的结论。大致的思路如下图所示:

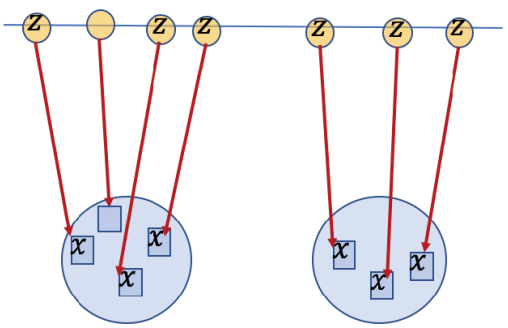


图 4: 泛化是可能的

三、JS–GAN 和 RSGAN 的分析

孙若愚在这一部分给出了一些直观解释。

他举了下图这个简单的例子。假设有两个真实样本 x_1 和 x_2，现在要生成两点 y_1 和 y_2 使得分别与真实样本相似。但经过图中的步骤，生成器趋向于两点均生成为与 x_1 相似的样本，于是 GAN 失效了。这种现象称之为模型崩溃 (mode collapse)，意思是生成器总是生成一种类型的图片，不具备多样性。从优化的角度来看，这就是一个很差的局部最小点。

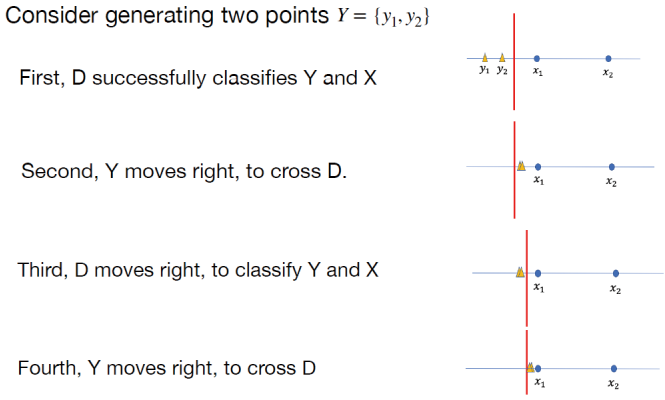


图 5：为什么 GAN 会失败?

孙若愚认为导致这种问题的原因是所有的生成样本共用一个标准。他用一个形象的例子来说明：想象在一门课中有若干学生，如果仅仅规定 60 分及格，那么所有的学生都会在达到 60 分之后就懈怠了；但是如果给出不同的激励措施，比如前 20% 的同学达到 90 分能去更好的学校，后 80% 的同学达到 60 分能够及格，那么学生们的潜力往往能被更大的激发出来。

孙若愚下面正式定义了 h-GAN 和 R-h-GAN。两者主要的区别在于损失函数，h-GAN 的真实样本和生成样本是分开的，定义如下:

$$\textbf{h-GAN: } \min_X \phi_h(Y,X) \text{ , where } \phi_h(Y,X) = \max_f \frac{1}{2n}\sum_{i=1}^{n} h(f(x_i)) + \sum_{i=1}^{n} h(-f(y_i)).$$

$$\text{Example: in JS-GAN, } h(u) = \log(\frac{1}{1+e^{-f(u)}})$$

(注：这里有一个笔误，外层应该是优化生成样本 Y 使得目标函数最小化，而不是固定的真实样本 X)

而 R-h-GAN 对应的真实样本和生成样本在一个 h 函数中，定义如下:

$$\textbf{Relativistic GAN: } \min_Y \phi_{h,\mathrm{R}}(Y,X) \text{ where } \phi_{h,\mathrm{R}}(Y,X) = \max_f \frac{1}{2n}\sum_{i=1}^{n} h(f(x_i) - f(y_i)).$$

$$\textbf{Example: } \text{in relativistic standard GAN (RS-GAN), } h(u) = \log(\frac{1}{1+e^{-f(u)}})$$

孙若愚强调了 R-GAN 这种配对的思想。事实上他本来想把它命名为”coupled-GAN”，但参考文献 (Jolicoeur-Martineau’2019) 已经提到过了 R-GAN。然而他们的动机是不同的：参考文献 (Jolicoeur-Martineau’2019) 展现了 R-GAN 在实验上优越的效果，而孙若愚希望打破上面提到的局部最小。孙若愚的动机来源于 W-GAN (Wasserstein GAN):

$$\phi_{\mathrm{W}}(Y,X) = \max_{|f|_L \le 1} \frac{1}{n}\sum_i \left[f(x_i) - f(y_i) \right]$$

W–GAN 与 JS–GAN 有两点区别： 一是把逻辑回归损失转化为了线性的，二是自动地将真实样本和生成样本配对了，所以 W–GAN 也是一种特殊的 R–h–GAN。他猜测配对是至关重要的，他影响了生成函数的景观，但是把逻辑回归损失转化为线性并没有什么帮助。所以可以推测： 如果保持逻辑回归损失不变，并且能够配对，那么这样的 GAN 应该会有更好的效果。而事实上，这就是 RS–GAN。

四、全局景观的分析：正式结论

上一部分都是一些直观的感受和想法，孙若愚在这一部分给出正式的结论。

首先还是考虑两个点的情况。会出现以下四种情形:

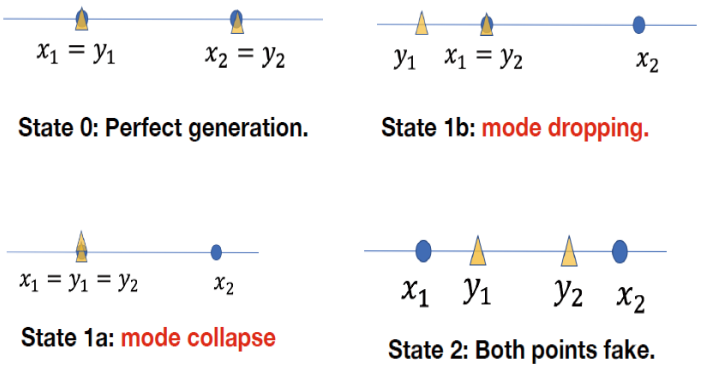


图 6：两点样本

通过如下简单的计算:

$$\phi_{\mathrm{JS}}(Y,X)=\begin{cases}-\log 2\approx -0.6931 & \text{if } \{x_1,x_2\}=\{y_1,y_2\}\\ -\log 2/2\approx -0.3467 & \text{if } |\{x_1,x_2\}\cap\{y_1,y_2\}|=1,\\ \frac{1}{4}(2\log 2-3\log 3)\approx -0.4774, & \text{if } y_1=y_2\in\{x_1,x_2\},\\ 0 & \text{if } |\{x_1,x_2\}\cap\{y_1,y_2\}|=\emptyset.\end{cases}\qquad \phi_{\mathrm{RS}}(Y,X)=\begin{cases}-\log 2\approx -0.6931, & \text{if } \{x_1,x_2\}=\{y_1,y_2\}\\ -\frac{1}{2}\log 2\approx -0.3466, & \text{if } |\{i:x_i=y_i\}|=1\\ 0 & \text{otherwise.}\end{cases}$$

可以知道，上述情形 1a，即 $(y_1,y_2)=(x_1,x_2)$，在 JS–GAN 中是一个严格的局部极小点； 而 RS–GAN 不存在严格的局部极小点。下图更加直观，JS–GAN 的模型崩溃（情形 1a）导致了一个盆地，而 RS–GAN 更加平滑，没有盆地。“盆地”一词十分形象，他在后面对其给出了一个非严格的定义： 一个没有到全局最小点的非增路径的区域。优化的知识告诉我们，非严格的局部极小点是容易逃逸出去的，而严格的极小点附近是一个盆地，在训练过程中很难跳出去。

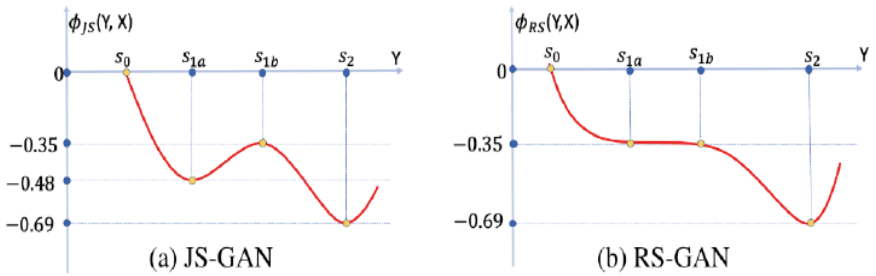


图 7：JS–GAN 的模型崩溃会导致一个盆地，而 RS–GAN 会更加平滑

有了两个点的例子作为基础，孙若愚下面给出更一般的——h–GAN 和 R–GAN 在 n 个真实样本上的结论:

Assumption 1: $\sup_t h(t) = 0; h(0) < 0$; h is concave.

Recall: $\phi_h(Y, X) = \max_f \frac{1}{2n} \sum_{i=1}^{n} h(f(x_i)) + \sum_{i=1}^{n} h(-f(y_i))$.

Theorem 1 If all $y_i \in \{x_1, x_2, \ldots, x_n\}$ but some x_i is not in the generated data set, then Y is a sub-optimal strict local-min of $\phi_h(Y, X)$.

- **In words:** "mode-collapse" = "bad basin"
- $(n^n - n!)$ basins in h-GAN (e.g. JS-GAN) landscape.

图 8：h–GAN 在真实样本上的结论

$$\phi_{h,\mathrm{R}}(Y, X) = \max_f \frac{1}{2n} \sum_{i=1}^{n} h(f(x_i) - f(y_i)).$$

Global-min-reachable (GMR): If from any point u, there is a continuous path from u to a global minimum of F such that F is non-increasing along the path, we say F satisfies GMR.

- **Theorem 2:** Y is a global-min of $g(Y) = \phi_{h,\mathrm{R}}(Y, X)$ iff $\{x_1, x_2, \ldots, x_n\} = \{y_1, y_2, \ldots, y_n\}$. In addition, **g is GMR.**

- This implies: R-GAN (including RS-GAN) does not have bad basins.

图 9：R–GAN 在真实样本上的结论

简单来说，h–GAN 一定存在严格的局部最小点，也就是盆地。换句话说，模型崩溃就等价于盆地。并且这样的盆地在 h–GAN 的全局景观中是非常多的。而 R–GAN 具有良好的性质，不仅没有盆地，而且可以证明一个更强的结论——生成器具有全局最小点可达性（GMR）——从任意初始点出发都存在一条非增路径到达全局最小点。

孙若愚接着考虑参数空间的情形。如果分别将生成器和鉴别器通过神经网络参数化为 $G_w(z)$ 和 $f_\theta(u)$，上面的结论就能够扩展到参数空间中，在一定的前提条件下得到如下类似的结论:

Assumption 1 (informal): Both $G_w(z)$ and $f_\theta(u)$ have enough representation power.

$$\min_w \varphi_h(w) \quad \text{where} \quad \varphi_h(w) = \max_\theta \frac{1}{2n} \sum_{i=1}^{n} h(f_\theta(x_i) - f_\theta(G_w(z_i))).$$

Proposition 1 (informal) The loss function $\varphi_h(w)$ is NOT global-min-reachable.

$$\min_w \varphi_{h,\mathrm{R}}(w) \quad \text{where} \quad \varphi_{h,\mathrm{R}}(Y, X) = \max_\theta \frac{1}{2n} \sum_{i=1}^{n} h(f_\theta(x_i) - f_\theta(G_w(z_i))).$$

Proposition 2 (informal) The loss function $\varphi_{h,\mathrm{R}}(w)$ is global-min-reachable.

图 10：参数空间中的结果

证明用到了他从前的一篇关于神经网络性质的工作（参考文献 Li, Ding, Sun'2019），其中证明了足够宽的神

经网络没有差的盆地，并且这个“足够宽”是在实际训练中能够达到的。这一神经网络的性质很容易地推广到 GAN 中。

以上结论的证明涉及到一些数学技巧，尤其是定理 2 有一定的难度，它需要我们搞清楚目标函数在每个点处的取值并建立一条到全局极小点的非增路径，主要用到了图理论（Graph theory）。

五、实验结果

孙若愚在这一部分展示了一些有趣的实验结果。

4.1 两个例子

孙若愚首先展示了只有两个真实样本的 GAN 的训练结果。

可以看到，在训练过程中，JS–GAN 的所有生成样本很长一段时间都是一起在两点之间来回跑，而 RS–GAN 的生成样本则迅速的分成两个部分并分别聚集在两点附近。这说明 RS–GAN 比 JS–GAN 训练得更快。

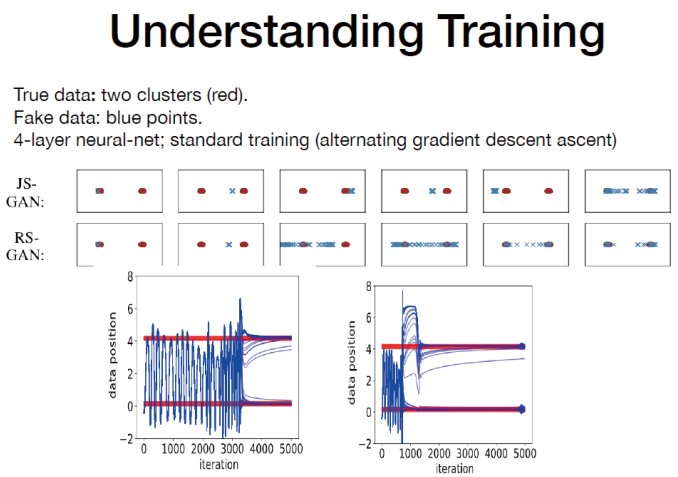
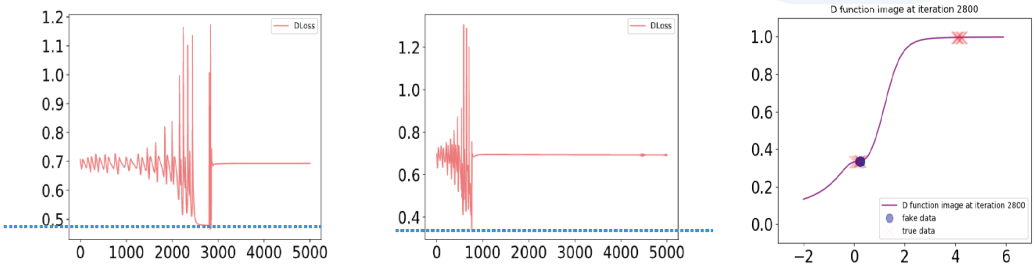


图 11：对 JS–GAN 和 RS–GAN 进行训练，分析其结果

孙若愚在观察鉴别器损失（DLoss）时发现了更有意思的现象，如下图 (a)(b)。蓝色虚线表示训练过程中鉴别器损失的最小值。可以看到，JS–GAN 中间有很长一段时间都停留在了最小值（0.48）附近，这个最小值恰好就是我们前面的情形 1a 模型崩溃计算得到的值，这说明它陷入了一个比较宽的盆地，难以跳出； 而 RS–GAN 的最小值（0.35）虽然也触及到了局部最小点，但是之后迅速跳出。下图 (c) 是 JS–GAN 在训练 2800 轮时的鉴别器函数，这验证了之前关于盆地的理论确实在实际中也发生了。



(a) JS-GAN (b) RS-GAN (c) JS-GAN, D image.

图 12：鉴别器损失

孙若愚接下来通过一个动画来形象的帮助我们理解 JS–GAN 和 RS–GAN 训练过程的区别。

4.2 真实数据

孙若愚认为在数学上要分析 n 个真实样本的非线性的训练过程过于困难了。所以他希望确定平衡点，暂时忽略训练过程的细节。

他作出了以下三个预测:

1. RS–GAN 比 JS–GAN 效果更好，有时候它们的差距非常巨大;
2. 对于窄的网络而言，这种差距更大了；(这是因为宽网络的景观更好，这帮助 JS–GAN 逃离盆地)
3. 存在一些差的初始点，导致 JS–GAN 的训练失败。

然后验证了它们:

1. 在 CIFAR–10 上训练，得到如下结果:

FID score: lower better. **IS**: higher better.

	CIFAR-10		
	Inception Score ↑	FID ↓	Model size
Real Dataset	11.24±0.19	5.18	
Standard CNN			
JS-GAN	6.27±0.10	49.13	100%
WGAN-GP	6.68±0.06	39.66	100%
RS-GAN	7.02±0.07	33.79	100%
JS-GAN+ SN	7.42±0.08	28.07	100%
RS-GAN+ SN	7.32±0.08	27.16	100%

Gap: 15.3

图 13：JS–GAN 和 RS–GAN 在 CIFAR–10 上训练，结果对比

他补充解释了为什么训练过程中加了 SN (spectral normalization) 之后缩小了它们的差距，这是因为 SN 导致网络的景观更好，帮助 JS–GAN 逃离盆地。

2. 在不同宽度的网络上进行训练，得到如下结果：

	CIFAR-10	
	IS ↑	FID ↓
ResNet + Hinge Loss		
JS^{hinge}	7.92±0.08	21.30
JS^{hinge}+GD channel/2	7.63±0.05	27.21
JS^{hinge}+GD channel/4	6.79±0.09	37.51
JS^{hinge}+BottleNeck	7.16±0.10	33.24
R^{hinge_HL}	8.03±0.09	19.07
R^{hinge_HL}+GD channel/2	7.69±0.10	22.79
R^{hinge_HL}+GD channel/4	7.11±0.06	32.35
R^{hinge_HL}+BottleNeck	7.52±0.05	24.07

图 14：JS–GAN 和 RS–GAN 在不同宽度的网络上进行训练，结果对比

很明显，窄网络差距有 9.2，远大于宽网络的 1.2。

3. 在 MNIST 上只取一个初始点进行训练，得到如下结果：

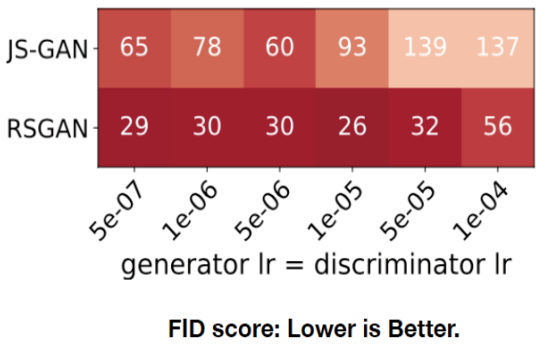


图 15：JS–GAN 和 RS–GAN 在 MNIST 上训练，结果对比

好的景观对于初始点具有鲁棒性。之前的结果都是细致的调整了初始点，取的训练的最优结果，所以最终结果表现得好。但如果只取一个初始点，它们的差距往往会更大。

六、结语

孙若愚在这次演讲中讨论了一些关于 GAN 的理论。从直观、理论和实验三个方面说明了：JS–GAN 有差的盆地，这导致 JS–GAN 训练速度慢、效果不好、容易失败；而 RS–GAN 的函数景观非常好，所以各个方面都优于 JS–GAN。最后，他认为未来有两个研究的方向，理论上真正地理解 GAN 的行，实际中设计更小的、更容易训练的 GAN。他的演讲让我们对 GAN 有了进一步的理解，他从全局景观研究 GAN 的这种分析方法有重要的启发意义。

Johannes Schmidt-Hieber：从偏差 - 方差下界看过参数化和隐式正则效应

整理：智源社区 王惠远

在第二届北京智源大会“数理基础”专题论坛上，Twente 大学教授 Johannes Schmidt-Hieber 作了题为《Overparametrization and the bias-variance dilemma》的演讲。Johannes Schmidt-Hieber 是 Twente 大学的教授，也是 Annals of Statistics 与 Bernoulli 的副主编。

传统的统计学方法认为，模型的表现能力可以分解成偏差 (bias) 的平方与方差 (variance) 的和的形式，且偏差与方差之间存在着一种折衷：即一个模型的偏差和方差不可能同时都被控制的很小，随着其中一项的减小，另外一项会不可避免的变大。另一方面，经典的理论认为模型的偏差和方差与参数个数有关，随着参数个数的增加，模型的偏差会逐渐减少而方差会较快的增大。为了保证模型有一个较好的泛化能力，模型参数个数通常会比样本量少 (图 1 左)。

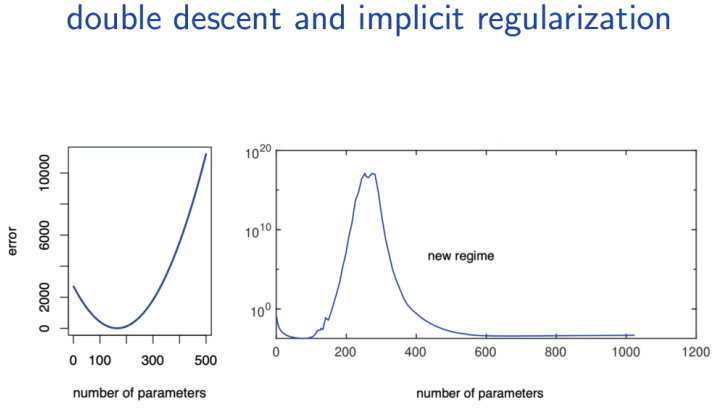


图 1：过参数化与隐式正则效应

随着神经网络在这几年的兴起，人们发现过参数化 (overparametrization) 的神经网络的泛化能力也会很好。随着参数的增加，一些模型如神经网络和随机森林的泛化水平会呈现先减再增，到达一个临界值再减的模式 (图 1 右)。这种模式被称为“双 U 下降” (double U descent)。在机器学习和统计学领域，人们猜测在过参数化的情形模型自身存在某种隐式正则效应来控制方差。什么时候会出现“双 U 下降”模式，“双 U 下降”模式的内在机制是如何作用的？这些都是值得探讨的问题。

Schmidt-Hieber 教授从统计学最大最小下界的角度出发，发展出一个可以给出一些常见函数空间和模型的偏差 – 方差折衷 (bias-variance trade-off) 的下界的方法，在一定程度上给出以上问题的解释。这样的下界可以告诉我们对于某个模型，是否存在一种方法可以避免偏差 – 方差折衷现象，于是可以更进一步地表明经典的“U 型曲线”成立的范围。

Schmidt–Hieber 教授主要考虑了三个经典模型，分别是高斯白噪声模型的逐点估计问题，高维情形下的高斯序列模型和 L2 损失下的高斯白噪声模型。

1. 高斯白噪声模型的逐点估计问题

pointwise estimation

Gaussian white noise model: We observe $(Y_x)_x$ with

$$dY_x = f(x)\,dx + n^{-1/2}\,dW_x$$

- estimate $f(x_0)$ for a fixed x_0
- $\mathscr{C}^\beta(R)$ denotes ball of Hölder β-smooth functions
- for any estimator $\widehat{f}(x_0)$, we obtain the bias-variance lower bound

$$\inf_{\widehat{f}} \sup_{f\in\mathscr{C}^\beta(R)} \left|\mathrm{Bias}_f\left(\widehat{f}(x_0)\right)\right|^{1/\beta} \sup_{f\in\mathscr{C}^\beta(R)} \mathrm{Var}_f\left(\widehat{f}(x_0)\right) \gtrsim \frac{1}{n}$$

- bound is attained by most estimators
- generates U-shaped curve

图 2：高斯白噪声模型的逐点估计问题

高斯白噪声模型是一个经典的统计学非参模型。Schmidt–Hieber 教授和他的合作者假设目标函数在一个足够光滑的函数空间（Hölder 空间）里，且只估计这个函数在某个任意给定的点的函数值。通过使用一个期望换元不等式，Schmidt–Hieber 教授得到了高斯白噪声模型的逐点估计问题的偏差 – 方差下界。通过这个下界我们可以发现，无论我们用什么方法来解决这个问题，我们估计值的偏差和方差一致地展现出反比关系，且反比例系数约为样本量的倒数。

2. 高维情形下的高斯序列模型

high-dimensional models

Gaussian sequence model:

- observe independent $X_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \ldots, n$
- $\Theta(s)$ the space of s-sparse vectors (here: $s \le \sqrt{n}/2$)
- bias-variance decomposition

$$E_\theta[\|\widehat{\theta} - \theta\|^2] = \underbrace{\|E_\theta[\widehat{\theta}] - \theta\|^2}_{B^2(\theta)} + \textstyle\sum_{i=1}^n \mathrm{Var}_\theta(\widehat{\theta}_i)$$

- **bias-variance lower bound:** if $B^2(\theta) \le \gamma s \log(n/s^2)$, then,

$$\sum_{i=1}^n \mathrm{Var}_0\left(\widehat{\theta}_i\right) \gtrsim n\left(\frac{s^2}{n}\right)^{4\gamma}$$

- bound is matched (up to a factor in the exponent) by soft thresholding
- bias-variance trade-off more extreme than U-shape
- results also extend to high-dimensional linear regression

图 3：高维情形下的高斯序列模型

高斯序列问题可以被设定成一个高维参数估计问题。使用类似的方法，Schmidt–Hieber 教授得到了关于高斯序列问题的偏差 – 方差下界。通过这个下界我们得知，如果我们控制了估计的偏差，那么这个估计的方差会有一个至少随样本量线性增长的上界。以下为 L2 损失下的高斯白噪声模型。

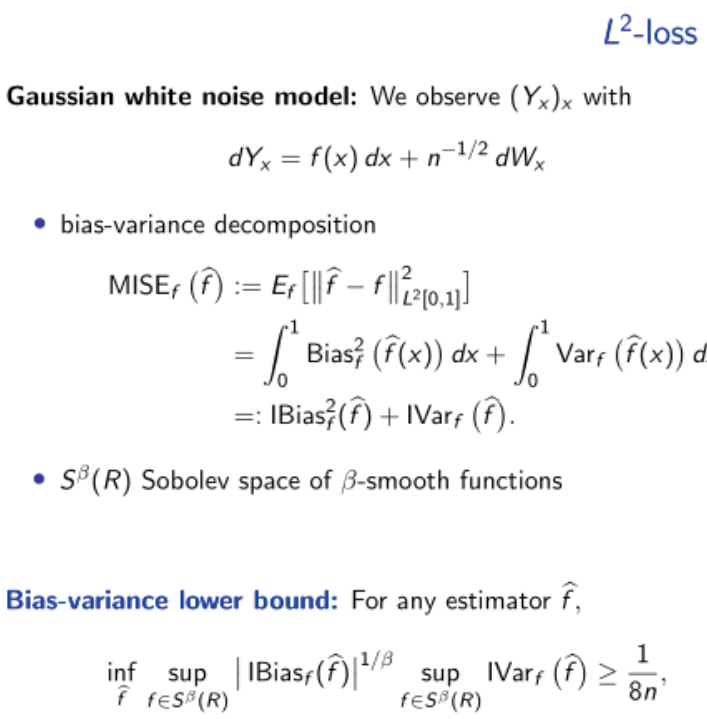


图 4：L2 损失下的高斯白噪声模型

L2 损失下的高斯白噪声模型与逐点收敛下的高斯白噪声模型相比，不是关注某一个固定点处对目标函数的估计，而是关于自变量的测度对于损失函数进行积分。通过简化处理，Schmidt–Hieber 教授同样地得到了与逐点收敛类似的结论：无论用什么方法来解决这个问题，估计值的偏差和方差一致地展现出反比关系，且反比例系数约为样本量的倒数。

上面的结论不仅有其独特的统计学意义，还为最近几年的热门问题“双 U 下降曲线”和隐式正则化提供了一个新的切入点。如果我们认为随着参数增加到充分大（如“过参数化”的情形），偏差会逐渐变小直至消失，那么根据上面的结论，我们可以得出双 U 下降曲线至少在特定的模型里（如高斯白噪音模型）不会出现。由此可知，双 U 曲线的出现预示着尽管参数个数是自由增长的，但参数个数的增长不会导致偏差的任意缩小，反而会使偏差稳定在一个方差也不太大的合理范围内，这种额外的偏差可能由某种隐式的正则效应引入。关于隐式正则效应的思考一直在持续，希望在未来能有一个完整的刻画。

中国科学院研究员戴彧虹：约束极小化极大优化的优化条件

整理：智源社区 罗丽

在第二届北京智源大会“人工智能的数理基础”专题论坛中，中国科学院数学与系统科学研究院研究员、智源学者戴彧虹做了主题为《Optimality Conditions for Constrained Minimax Optimization》(约束极小化极大优化的优化条件）的报告。

报告中，戴彧虹介绍了极小化极大优化的相关研究现状，值函数微分在雅可比唯一性约束条件下的定义、Karush–Kuhn–Tucker 系统的强规律性，以及极小化极大优化的一阶、二阶必要最优条件及其二阶充分最优条件。

一、极小化极大优化的背景

1.1 极小化极大问题是什么?

极小化极大问题主要包括两个方面，一是使目标函数 $f:x\times y\to R$ 最小化，另一个是使目标函数 $f:x\times y\to R$ 最大化。极小化极大问题是数学、生物学、社会科学以及经济学的重要研究领域[1]，其广泛的应用和丰富的数学结构使得该问题的研究已超过数十年[2]。近几年的研究表明，极小化极大优化在机器学习领域中具有重要作用，例如，Generative Adversarial Networks[3](GANs)（生成对抗网络），Adversarial Training[4]（对抗训练）以及 Multi–agent Reinforcement Learning[5]（多智能体强化学习）。

1.2 极小化极大问题的研究方法

(1) GDA 方法：GDA(Gradient descent ascent，GDA 算法及其交替格式如下： 梯度下降上升) 是研究极小化极大问题的经典方法，该算法在 x 梯度下降步长和 y 梯度上升步长之间交替[6]:

$$x_{t+1}=x_t-\alpha\nabla_x f(x_t,y_t),\quad y_{t+1}=y_t+\alpha\nabla_y f(x_t,y_t)$$
$$x_{t+1}=x_t-\alpha\nabla_x f(x_t,y_t),\quad y_{t+1}=y_t+\alpha\nabla_y f(x_{t+1},y_t)$$

但 GDA 方法也存在一些缺点，比如，即使在简单的双线性博弈中，GDA 方法也可能不会收敛。

(2) Consensus Optimization: 通过把梯度的二范数做为罚项加入到目标函数中，得到共识优化（consensus optimization）的思想：$||\nabla_x f(x_t,y_t)||^2+||\nabla_y f(x_t,y_t)||^2=c^2$:

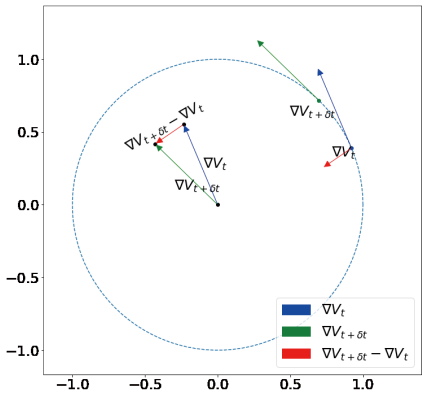
$$x_{t+1}=x_t-\alpha\nabla_x(f(x_t,y_t)+\gamma||\nabla_x f(x_t,y_t)||^2)$$
$$y_{t+1}=y_t-\alpha\nabla_y(-f(x_t,y_t)+\gamma||\nabla_y f(x_t,y_t)||^2)$$

(3) Symplectic Gradient Adjustment(SGA): 辛梯度调节方法[8]修改了关联的矢量场，以引导迭代器越过微分博弈的哈密顿分量的卷曲。

(4) 其他算法：Subgradient methods for saddle point problems[9] (鞍点问题的次梯度方法)；Optimistic mirror descent methods[10] (乐观镜像下降)；Proximal point algorithms[11] 和 Extragradient algorithms[12] (近点算法、超梯度算法)；Non-convex Min-Max Optimization: Applications, Challenges, & Recent Theoretical Advances (非凸最小最大优化：应用，挑战和最新理论进展)。

(5) 戴彧虹等提出 **SCA 方法：**根据“以匀速圆周运动的物体向心加速度的方向朝向圆心”的理论，彭伟、戴彧虹等人提出同时向心加速度 (SCA) 方法[13]，该方法与 Consensus Optimization 方法类似，其优点是不需要计算 Jacobi 矩阵。

$$
\begin{aligned}
G_x &= \nabla_x f(x_t, y_t) + \frac{\beta_1}{\alpha_1}(\nabla_x f(x_t, y_t) - \nabla_x f(x_{t-1}, y_{t-1})),\\
x_{t+1} &= x_t - \alpha_1 G_x;\\
G_y &= \nabla_y f(x_t, y_t) + \frac{\beta_2}{\alpha_2}(\nabla_y f(x_t, y_t) - \nabla_y f(x_{t-1}, y_{t-1})),\\
y_{t+1} &= y_t + \alpha_2 G_y.
\end{aligned}
$$



1.3 最优性理论的定义

(1) 纳什均衡 Nash equilibrium。纳什均衡 (Nash equilibrium) 是众所周知的最优性理论，在许多的研究中，研究人员提出了局部 Nash 均衡的概念。

定义 1：点(x^*, y^*)是函数$f(x, y)$的一个纳什均衡，在$X \times Y$中对任意的(x, y)，存在

$$f(x^*, y) \le f(x^*, y^*) \le f(x, y^*).$$

点(x^*, y^*)是函数$f(x, y)$的一个局部纳什均衡，如果存在$\delta > 0$，那么，对任意的(x, y)满足$\| x - x^* \| < \delta$且$\| y - y^* \| < \delta$，则存在

$$f(x^*, y) \le f(x^*, y^*) \le f(x, y^*).$$

戴彧虹同时指出，纳什均衡也存在如下缺点：

- 实际上，纳什均衡的概念并不能反映最小玩家和最大玩家之间的顺序。

- 相比之下，机器学习中的大多数应用（包括 GAN 和对抗训练）都与序列博弈有关，在序列博弈中，在一个玩家行动后另一个玩家会再行动。
- 当 f 函数为非凸 – 非凹时，$min\ x\ max\ y\ f$ 通常不等于 $max\ y\ min\ x\ f$，因此玩家行动的顺序对问题的研究至关重要。

综上研究表明，在极小化极大优化的大多数机器学习的应用中，局部纳什均衡的概念并不适用。

(2) 斯坦克尔伯格均衡 Stackelberg equilibrium。当给第一个玩家给定动作 x，Stackelberg 均衡是第二个玩家 $f(x,y)$ 的最大化者，并获得最大值 $\phi(x) := \max_{y\in Y} f(x,y)$，在此称其为全局极小化极大点。

定义 2：点 (x^*,y^*) 是全局极小化极大点，在 $X\times Y$ 中对任意的 (x,y)，存在

$$f(x^*,y) \le f(x^*,y^*) \le \max_{y'\in Y} f(x,y').$$

与纳什均衡不同，由于极值定理，即使函数 f 为非凸 – 非凹函数，全局极小化极大点也始终存在。

(3) 局部极小化极大点。2019 年，Jin，Netrapalli 和 Jordan[14] 在研究中提出了无约束极小化极大优化的局部极小化极大点的正确定义。

定义 3：点 (x^*,y^*) 是 f 的局部极小化极大点，如果 $\delta_0>0$，且当 $\delta\to 0$ 时函数 $\mathrm{h}(\delta)\to 0$，那么对任意的 $\delta\in(0,\ \delta_0]$ 和任意 (x,y) 满足 $\|x-x^*\|<\delta$ 且 $\|\mathrm{y}-y^*\|<\delta$，则存在:

$$f(x^*,y) \le f(x^*,y^*) \le \max_{y':\,\|y'-y^*\|<h(\delta)} f(x,y').$$

该文章还研究了梯度下降上升（GDA）的渐近行为，并指出 GDA 的所有稳定极限点恰好都是局部极小化极大点至某些退化点。

(4) 约束极小化极大问题。约束极小化极大问题

$$\min_{x\in\Phi}\ \max_{y\in Y(x)}\ f(x,y), \qquad (1)$$

其中，$f:\Re^n\times\Re^n\to\Re,\ \Phi\subset\Re^n$ 是由以下项定义的决策变量 x 的可行集

$$\Phi=\{x\in\Re^n : H(x)=0, G(x)\le 0\} \qquad (2)$$

且 $Y:\Re^n \rightrightarrows \Re^m$ 是由以下项定义的集值映射

$$Y(x)=\{y\in\Re^m : h(x,y)=0, g(x,y)\le 0\} \qquad (3)$$

其中，给定函数 $h:\Re^n\times\Re^m\to\Re^{m_1}$，$g:\Re^n\times\Re^m\to\Re^{m_2}$，$H:\Re^n\to\Re^{n_1}$ 且 $G:\Re^n\to\Re^{n_2}$。

1.4 研究动机

戴彧虹表示，极小化极大优化问题本质上是一个 Bi-level Programming Problem[15][16][17][18][19][20][21][22] (双层规划问题)。对于无约束的连续非凸 – 非凹，Jin，Netrapalli 和 Jordan[23] 在研究中提出一个基本问题，即极小化极大问题的局部最优的正确定义是什么，并提出了局部最优的定义——局部极小化极大。在该研究的基础上，戴彧虹等对约束极小化极大优化问题 (1) 的局部极小化极大点的定义进行了扩展，并提出了约束极小化极大优化问题的局部极小化极大点的必要最优条件和充分最优条件。

1.5 扩展的局部极小化极大点的定义

定义 4: 点 $(x^*, y^*) \in \Re^n \times \Re^m$ 是问题 (1) 的局部极小化极大点，如果存在 $\delta_0 > 0$ 且函数 $\eta : (0, \delta_0] \to \Re_+$ 满足当 $\delta \to 0$ 时函数 $\eta(\delta) \to 0$，那么，对任意的 $\delta : (0, \delta_0]$ 且对任意 $(x, y) \in [B_\delta(x^*) \cap \Phi] \times [Y(x^*) \cap B_\delta(y^*)]$，存在:

$$f(x^*, y) \le f(x^*, y^*) \le \max\{f(x, z) : z \in Y(x) \cap \mathbf{B}_{\eta(\delta)}(y^*)\}$$

二、值函数的微分

2.1 雅可比唯一性下的值函数的微分

(1) 参数化极小化极大问题: 令 $(x^*, y^*) \in \Re^n \times \Re^m$ 为一个点，f, g, h 是在 (x^*, y^*) 附近的两次连续微分，用 (P_x) 表示以下问题:

$$\begin{aligned} \max_{z \in \Re^m} \quad & f(x, z) \\ \text{s.t.} \quad & h(x, z) = 0, \\ & g(x, z) \le 0. \end{aligned} \tag{4}$$

问题 (P_x) 的拉格朗日定义为:

$$\mathcal{L}(x; z, \mu, \lambda) = f(x, z) + \mu^T h(x, z) - \lambda^T g(x, z).$$

(2) 雅可比唯一性条件

定义 5: 令 $(y^*, \mu^*, \lambda^*) \in \Re^{m_1} \times \Re^{m_2}$ 为一个点，问题 (P_{x^*}) 在 (y^*, μ^*, λ^*) 处的雅可比唯一性条件满足:

a) 点 (y^*, μ^*, λ^*) 是问题 (P_{x^*}) 的 Karush–Kuhn–Tucker 点，即:

$$\begin{aligned} & \nabla_y \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) = 0, \\ & h(x^*, y^*) = 0, \\ & 0 \le \lambda^* \perp g(x^*, y^*) \le 0. \end{aligned}$$

b) 线性独立约束条件成立于 y^*；即向量集

$$\{\nabla_y h_1(x^*,y^*),\ldots,\nabla_y h_{m_1}(x^*,y^*)\}\cup\{\nabla_y g_i(x^*,y^*):i\in I_{x^*}(y^*)\}$$

是线性独立的，其中

$$I_{x^*}(y^*)=\{i:g_i(x^*,y^*)=0,i=1,\ldots,m_2\}.$$

c) 对于 λ^*，在 y^* 处的严格互补条件为

$$\lambda_i^*-g_i(x^*,y^*)>0,\quad i=1,\ldots,m_2.$$

d) 在 $\left(y^*,\mu^*,\lambda^*\right)$ 处的二阶充分最优条件为

$$\langle\nabla_{yy}^2\mathcal{L}(x^*;y^*,\mu^*,\lambda^*)d_y,d_y\rangle<0\quad\forall d_y\in\mathcal{C}_{x^*}(y^*),$$

其中 $C_{x^*}\left(y^*\right)$ 是在 y^* 时问题 $\left(P_{x^*}\right)$ 的临界锥

$$\mathcal{C}_{x^*}(y^*)=\left\{d_y\in\Re^m:\begin{array}{l}\nabla_y g_i(x^*,y^*)d_y\le 0,i\in I_{x^*}(y^*);\\ \mathcal{J}_y h(x^*,y^*)d_y=0;\nabla_y f(x^*,y^*)d_y\le 0\end{array}\right\}.$$

(3) 存在的解路径

引理 6： 令 $\left(x^*,y^*\right)\in\Re^n\times\Re^m$ 是 f,g,h 在 $\left(x^*,y^*\right)$ 附近的两次连续可微的点。令 $\left(\mu^*,\lambda^*\right)\in\Re^{m_1}\times\Re^{m_2}$，使得问题 $\left(P_{x^*}\right)$ 在 $\left(y^*,\mu^*,\lambda^*\right)$ 处满足雅可比唯一条件。然后满足 $\delta_0>0$，$\varepsilon_0>0$，且存在一个两次连续微分映射 $(y,\mu,\lambda):B_{\delta_0}\left(x^*\right)\to B_{\varepsilon_0}\left(y^*\right)\times B_{\varepsilon_0}\left(\mu^*\right)\times B_{\varepsilon_0}\left(\lambda^*\right)$，使得当 $x\in B_{\delta_0}\left(x^*\right)$ 时，问题 $\left(P_{x^*}\right)$ 在 $\left(y^*,\mu^*,\lambda^*\right)$ 处满足雅可比唯一条件。

(4) 值函数的导数。定义（最佳）值函数

$$\varphi(x)=f(x,y(x)),\quad x\in\mathbf{B}_{\delta_0}(x^*),\qquad(5)$$

命题 2.1： 如果满足引理 6 的假设且 φ 由公式 (5) 定义，那么

$$\nabla_x\phi(x)=\nabla_x\mathcal{L}(x;y(x),\mu(x),\lambda(x))$$

且

$$\begin{aligned}\nabla^2\phi(x)=\ &\nabla_{xx}^2\mathcal{L}(x;y(x),\mu(x),\lambda(x))\\ &-N(x)^TK(x)^{-1}N(x).\end{aligned}$$

$$
\begin{aligned}
N(x) &= \begin{bmatrix} \nabla^2_{yx}\mathcal{L}(x;y(x)\mu(x),\lambda(x)) \\ 0 \\ \mathcal{J}_x h(x,y(x)) \\ \mathcal{J}_x g(x,y(x)) \end{bmatrix}, \quad K(x)=[K_1(x)\ K_2(x)]; \\
K_1(x) &= \begin{bmatrix} \nabla^2_{yy}\mathcal{L}(x;y(x),\mu(x),\lambda(x)) & 0 \\ 0 & -2\mathrm{Diag}(\lambda(x)) \\ \mathcal{J}_y h(x,y(x)) & 0 \\ \mathcal{J}_y g(x,y(x)) & 2\mathrm{Diag}\left(\sqrt{-g(x,y(x))}\right) \end{bmatrix}; \\
K_2(x) &= \begin{bmatrix} \mathcal{J}_y h(x,y(x))^T & \mathcal{J}_y g(x,y(x))^T \\ 0 & 2\mathrm{Diag}\left(\sqrt{-g(x,y(x))}\right) \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.
\end{aligned} \tag{6}
$$

2.2 Karush–Kuhn–Tucker 系统的强规律性

(1) 假设 A

定义 7： 令$\left(x^*,y^*\right)\in\Re^n\times\Re^m$是满足$\wedge_{x^*}\left(y^*\right)\neq\varnothing$的点，问题$\left(P_{x^*}\right)$在$y^*$时的**强二阶充分最优条件**为：

$$
\sup_{(\mu,\lambda)\in\Lambda_{x^*}(y^*)}\langle\nabla^2_{yy}\mathcal{L}(x^*;y^*,\mu,\lambda)d_y,d_y\rangle<0\quad\forall d_y\in\mathrm{aff}\,C_{x^*}(y^*)\setminus\{0\},
$$

其中，$C_{x^*}\left(y^*\right)$是在y^*时问题$\left(P_{x^*}\right)$的临界锥。

定义 8： 令$\left(x,y\right)\in\Re^n\times\Re^m$为点，如果满足$\wedge_x\left(y\right)\neq\varnothing$，则问题$\left(P_{x^*}\right)$在$y\in Y(x)$处的假设 A 成立，在 y 处其线性独立约束条件和强二阶充分最优条件成立。

(2) 存在的解路径

引理 9： 令$\left(x^*,y^*\right)\in\Re^n\times\Re^m$是$f,g,h$在$\left(x^*,y^*\right)$附近的两次连续可微的点。对问题$\left(P_{x^*}\right)$，假设假设 A 在$y^*$处成立。则存在$\delta_0>0$，$\varepsilon_0>0$，且存在 Lipschitz 局部连续映射$(y,\mu,\lambda):B_{\delta_0}\left(x^*\right)\to B_{\varepsilon_0}\left(y^*\right)\times B_{\varepsilon_0}\left(\mu^*\right)\times B_{\varepsilon_0}\left(\lambda^*\right)$满足$\left(y\left(x^*\right),\eta\left(x^*\right),\lambda\left(x^*\right)\right)=\left(y^*,\mu^*,\lambda^*\right)$且

$$
\begin{aligned}
&\nabla_y\mathcal{L}(x;y(x),\mu(x),\lambda(x))=0,\\
&h(x,y(x))=0,\\
&g(x,y(x))-\Pi_{\Re^{m_2}_-}(\lambda(x)+g(x,y(x)))=0
\end{aligned}
$$

此时，$x\in B_{\delta_0}\left(x^*\right)$，且当$x\in B_{\delta_0}\left(x^*\right)$时，在$y\left(x\right)$处问题$\left(P_{x^*}\right)$的假设 A 成立。

注释： 线性操作$W:\Re^{m_1}\to\Re^{m_2}$定义为

$$\mathcal{A}(x,W)=\begin{bmatrix}\nabla^2_{yy}\mathcal{L}(x;y(x),\mu(x),\lambda(x)) & \mathcal{J}_y h(x,y(x))^T & \mathcal{J}_y g(x,y(x))^T\\ \mathcal{J}_y h(x,y(x)) & 0 & 0\\ (I-W)\mathcal{J}_y g(x,y(x)) & 0 & W\end{bmatrix},$$

集值映射 $\mathrm{A}_B:\Re^n\Rightarrow\Re^{m_2\times m_2}$ 定义为

$$\mathbb{A}_B(x)=\left\{\mathcal{A}(x,W):W\in\partial_B\Pi_{\Re^{m_2}_-}(\lambda(x)+g(x,y(x)))\right\},$$

集值映射 $\mathrm{A}_C:\Re^n\Rightarrow\Re^{m_2\times m_2}$ 定义为

$$\mathbb{A}_C(x)=\left\{\mathcal{A}(x,W):W\in\partial\Pi_{\Re^{m_2}_-}(\lambda(x)+g(x,y(x)))\right\}.$$

(3) $\left(\mathrm{A}_C(x)\right)$ 的非奇异性

命题 2.2： 令 $\left(x^*,y^*\right)\in\Re^n\times\Re^m$ 是 f,g,h 在 $\left(x^*,y^*\right)$ 附近的两次连续可微的点。对问题 $\left(P_{x^*}\right)$，假设假设 A 在 y^* 处成立。令 $\delta_0>0$ 在引理 9 中给出，则集值映射 $\left(\mathrm{A}_B(x)\right)$，$\left(\mathrm{A}_C(x)\right)$ 在 x^* 处为上半连续，并且对于 $\delta\in(0,\delta_0)$，满足当 $x\in B\left(x^*,\delta\right)$ 时，$\left(\mathrm{A}_B(x)\right)$ 和 $\left(\mathrm{A}_C(x)\right)$ 中的每个元素都是非奇异的。

(4) 解映射微分

命题 2.3： 令 $\left(x^*,y^*\right)\in\Re^n\times\Re^m$ 是 f,g,h 在 $\left(x^*,y^*\right)$ 附近的两次连续可微的点。对问题 $\left(P_{x^*}\right)$，假设假设 A 在 y^* 处成立。令 $\delta_0>0$，$\varepsilon_0>0$，且 $\left(y(\cdot),\eta(\cdot),\lambda(\cdot)\right)$ 由引理 9 给出，对 $x\in B_{\delta_0}\left(x^*\right)$，满足

a) $\left(y(\cdot),\eta(\cdot),\lambda(\cdot)\right)$ 在 x 处的方向导数满足

$$\begin{pmatrix}y'(x;d_x)\\ \mu'(x;d_x)\\ \lambda'(x;d_x)\end{pmatrix}\in\left\{H(x,W)d_x:W\in\partial_B\Pi_{\Re^{m_2}_-}(\lambda(x)+g(x,y(x)))\right\}$$

b) $\left(y(\cdot),\eta(\cdot),\lambda(\cdot)\right)$ 在 x 处的 B 次微分满足

$$\partial_B\begin{pmatrix}y\\ \mu\\ \lambda\end{pmatrix}(x)\in\left\{H(x,W):W\in\partial_B\Pi_{\Re^{m_2}_-}(\lambda(x)+g(x,y(x)))\right\}$$

c) $\left(y(\cdot),\eta(\cdot),\lambda(\cdot)\right)$ 在 x 处的 **Clarke generalized Jacobian** 满足

$$\partial\begin{pmatrix}y\\ \mu\\ \lambda\end{pmatrix}(x)\in\left\{H(x,W):W\in\partial\Pi_{\Re^{m_2}_-}(\lambda(x)+g(x,y(x)))\right\}$$

其中

$$H(x,W)=-\mathcal{A}(x,W)^{-1}\begin{pmatrix}\nabla^2_{yx}\mathcal{L}(x;y(x),\mu(x),\lambda(x))\\ \mathcal{J}_x h(x,y(x))\\ (I-W)\mathcal{J}_x g(x,y(x))\end{pmatrix}$$

(5) 值函数微分

推论 10： 令$\left(x^*,y^*\right)\in\Re^n\times\Re^m$是$f,g,h$在$\left(x^*,y^*\right)$附近的两次连续可微的点。对问题$\left(P_{x^*}\right)$，假设假设 A 在$y^*$处成立。令$\delta_0>0$，$\varepsilon_0>0$，且$\left(y(\cdot),\eta(\cdot),\lambda(\cdot)\right)$由引理 9 给出，则值函数$\varphi(x)=f\left(x,y(x)\right)$在$B_{\delta_0}\left(x^*\right)$为局部 Lipschitz 连续，对$x\in B_{\delta_0}\left(x^*\right)$，满足:

a) φ在 x 处的方向导数满足

$$\begin{aligned}\varphi'(x;d_x)\in\ & \nabla_x\mathcal{L}(x;y(x),\mu(x),\lambda(x))d_x\\ & -\left\{\nabla_{y,\mu,\lambda}\mathcal{L}(x;y(x),\mu(x),\lambda(x))^T H(x,W)d_x:\right.\\ & \qquad\left.W\in\partial_B\Pi_{\Re^{m_2}_-}(\lambda(x)+g(x,y(x)))\right\}\end{aligned}$$

b) φ在 x 处的 B 次微分满足

$$\begin{aligned}\partial_B\varphi(x)\subset\ & \nabla_x\mathcal{L}(x;y(x),\mu(x),\lambda(x))d_x\\ & -\left\{H(x,W)^T\nabla_{y,\mu,\lambda}\mathcal{L}(x;y(x),\mu(x),\lambda(x)):\right.\\ & \qquad\left.W\in\partial_B\Pi_{\Re^{m_2}_-}(\lambda(x)+g(x,y(x)))\right\}\end{aligned}$$

c) φ在 x 处的 Clarke generalized Jacobian 满足

$$\begin{aligned}\partial\varphi(x)\subset\ & \nabla_x\mathcal{L}(x;y(x),\mu(x),\lambda(x))d_x\\ & -\left\{H(x,W)^T\nabla_{y,\mu,\lambda}\mathcal{L}(x;y(x),\mu(x),\lambda(x)):\right.\\ & \qquad\left.W\in\partial\Pi_{\Re^{m_2}_-}(\lambda(x)+g(x,y(x)))\right\}\end{aligned}$$

三、优化条件

3.1 一阶问题

假设$\varphi(x)$由公式 (5) 定义，将约束极小化极大问题 (1) 局部化为

$$\begin{aligned}\min\quad & \varphi(x)=f(x,y(x))\\ \text{s.t.}\quad & x\in\Phi\cap\mathbf{B}_{\delta_0}(x^*),\end{aligned}\tag{7}$$

其中$y(x)$是y^*附近的(P_x)的局部极小化值，Φ由公式 (2) 定义。

3.2 约束集 Φ 和临界锥的 MFCQ

对于$x^*\in\Phi$，Mangasarian–Fromovitz 约束条件 (Mangasarian–Fromovitz constraint qualification) 在x^*时保持约束集Φ，满足

(1) 向量组$\nabla H_j\left(x^*\right),j=1,\ldots,n_1$是线性独立的;

(2) 存在一个向量$\overline{d}\in\Re^n$使得

$$\nabla H_j(x^*)^T \bar{d} = 0, j = 1, \ldots, n_1, \nabla G_i(x^*)^T \bar{d} < 0, i \in I(x^*),$$

where $I(x^*) = \{i : G_i(x^*) = 0, i = 1, \ldots, n_2\}$.

在x^*处，问题（7）的临界锥$C\left(x^*\right)$被定义为

$$\mathcal{C}(x^*) = \left\{ d_x \in \Re^n : \begin{array}{l} \nabla G_i(x^*)^T d_x \leq 0, i \in I(x^*); \\ \mathcal{J}H(x^*) d_x = 0; \varphi'(x^*; d_x) \leq 0 \end{array} \right\} \tag{8}$$

在雅可比唯一性条件下，临界锥$C\left(x^*\right)$可以表示为

$$\mathcal{C}(x^*) = \left\{ d_x : \begin{array}{l} \mathcal{J}H(x^*) d_x = 0; \nabla G_i(x^*)^T d_x \leq 0, i \in I(x^*); \\ \nabla_x \mathcal{L}(x^*; y^*, \mu^*, \lambda^*)^T d_x \leq 0 \end{array} \right\} \tag{9}$$

3.3 雅可比唯一性条件下的必要最优性

定理 11: 令$\left(x^*, y^*\right) \in \Re^n \times \Re^m$是$f, g, h$在$\left(x^*, y^*\right)$附近的两次连续可微的点，H，G 为$x^*$附近的连续两次微分。令$\left(x^*, y^*\right)$为问题（1）的局部极小化极大点，假设约束集合$Y\left(x^*\right)$在$y^*$处的线性独立约束条件成立，则存在唯一向量$\left(\mu^*,\ \lambda^*\right) \in \Re^{m_1} \times \Re^{m_2}$使得

$$\begin{aligned} &\nabla_y \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) = 0, \\ &h(x^*, y^*) = 0, \\ &0 \geq \lambda^* \perp g(x^*, y^*) \leq 0. \end{aligned}$$

对于任意$d_y \in C_{x^*}\left(y^*\right)$，满足

$$\langle \nabla^2_{yy} \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) d_y, d_y \rangle \leq 0.$$

假设问题$\left(P_{x^*}\right)$在$\left(y^*, \mu^*, \lambda^*\right)$处满足雅可比唯一条件，且 Mangasarian–Fromovitz 约束条件在x^*处满足约束集Φ，则存在$\left(u^*,\ v^*\right) \in \Re^{n_1} \times \Re^{n_2}$使得

$$\begin{aligned} &\nabla_x \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) + \mathcal{J}H(x^*)^T u^* + \mathcal{J}G(x^*)^T v^* = 0, \\ &H(x^*) = 0, \\ &0 \leq v^* \perp G(x^*) \leq 0. \end{aligned} \tag{10}$$

满足公式（10）的所有$\left(u^*,\ v^*\right)$的集合由$\nabla\left(x^*\right)$表示，该集合是一个非空紧致凸集。此外，每个$d_x \in C\left(x^*\right)$（$C\left(x^*\right)$由公式（9）定义）满足

$$\begin{aligned} \max_{(u,v) \in \Lambda(x^*)} &\left\{ \left\langle \left[\sum_{j=1}^{n_1} u_i \nabla^2_{xx} H_j(x^*) + \sum_{i=1}^{n_2} v_i \nabla^2_{xx} G_i(x^*) \right] d_x, d_x \right\rangle \right\} \\ &+ \left\langle \left[\nabla^2_{xx} \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) - N(x^*)^T K(x^*)^{-1} N(x^*) \right] d_x, d_x \right\rangle \geq 0, \end{aligned}$$

其中 $K(x)$ 和 $N(x)$ 由公式 (6) 定义。

定理 12 (二阶充分最优条件)： 假设 $x^* \in \Phi$ 且 $y^* \in Y(x^*)$。令 $(\mu^*, \lambda^*) \in \Re^{m_1} \times \Re^{m_2}$，假设问题 (P_{x^*}) 在 $(y^*, \mu^*, \lambda^*), \wedge(x^*) \neq \varnothing$ 处满足雅可比唯一条件，且对每个 $d_x \in C(x^*) \setminus \varnothing$ （$C(x^*)$ 由公式 (9) 定义）满足

$$\sup_{(u,v)\in\Lambda(x^*)} \left\{ \left\langle \left[\sum_{j=1}^{n_1} u_i \nabla_{xx}^2 H_j(x^*) + \sum_{i=1}^{n_2} v_i \nabla_{xx}^2 G_i(x^*) \right] d_x, d_x \right\rangle \right\} + \left\langle \left[\nabla_{xx}^2 \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) - N(x^*)^T K(x^*)^{-1} N(x^*) \right] d_x, d_x \right\rangle > 0.$$

存在 $\delta_1 \in (0, \delta_0)$，$\varepsilon_1 \in (0, \varepsilon_0)$ 且 $\gamma_1 > 0, \gamma_2 > 0$，使得对 $x \in B_{\delta_1}(x^*) \cap \Phi$，$y \in B_{\varepsilon_1}(y^*) \cap Y(x^*)$ 满足

$$f(x^*, y) + \gamma_1 \|y - y^*\|^2/2 \leq f(x^*, y^*) \leq \sup_{z \in Y(x) \cap \mathbf{B}_{\varepsilon_0}(y^*)} f(x, z) - \gamma_2 \|x - x^*\|^2/2,$$

则意味着 (x^*, y^*) 是问题 (1) 的局部极小化极大点。

3.4 假设 A 下的一阶必要最优性

定理 13： 令 $(x^*, y^*) \in \Re^n \times \Re^m$ 是 f, g, h 在 (x^*, y^*) 附近的两次连续可微的点，H，G 为 x^* 附近的连续两次微分。令 (x^*, y^*) 为问题 (1) 的局部极小化极大点。 假设约束集合 $Y(x^*)$ 在 y^* 处的线性独立约束条件成立，则存在唯一向量 $(\mu^*, \lambda^*) \in \Re^{m_1} \times \Re^{m_2}$ 使得

$$\begin{aligned} &\nabla_y \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) = 0, \\ &h(x^*, y^*) = 0, \\ &0 \geq \lambda^* \perp g(x^*, y^*) \leq 0. \end{aligned}$$

对于任意 $d_y \in C_{x^*}(y^*)$，满足

$$\langle \nabla_{yy}^2 \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) d_y, d_y \rangle \leq 0.$$

假设问题 (P_{x^*}) 在 (y^*, μ^*, λ^*) 处满足假设 A，且约束集 Φ 在 x^* 处满足 Mangasarian–Fromovitz 约束条件，则存在 $(u^*, v^*) \in \Re^{n_1} \times \Re^{n_2}$，$W \in \partial \prod_{R^{m_2}} (\lambda^* + g(x^*, y^*))$ 使得

$$\begin{aligned} &\nabla_x \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) - H(x^*, W)^T \nabla_{y,\mu,\lambda} \mathcal{L}(x^*; y^*, \mu^*, \lambda^*) \\ &\qquad + \mathcal{J} H(x^*)^T u^* + \mathcal{J} G(x^*)^T v^* = 0, \\ &H(x^*) = 0, \\ &0 \leq v^* \perp G(x^*) \leq 0. \end{aligned} \tag{11}$$

满足公式（11）的所有$\left(u^*,\ v^*\right)$的集合由$\wedge\left(x^*\right)$表示，该集合是一个非空紧致凸集。

3.5 如何设计有效算法来约束极小化极大优化?

在下一步的研究中，戴彧虹表示将考虑复合约束的极小化极大优化问题

$$\min_{x\in X}\max_{y\in Y} L(x,y) := f(x) + K(x,y) - g(y),$$

其中 X 和 Y 是两个有限维 Hilbert 空间，$K : X \times Y \to R$ 是连续可微函数，$f : X \to R$ 和 $g : Y \to R$ 是适当的下半连续凸函数。他们提出的交替坐标法能够在共同假设下具有一定的收敛性。

参考文献

[1] Roger B Myerson. Game Theory. Harvard University Press, 2013.

[2] Oskar Morgenstern and John Von Neumann. Theory of Games and Economic Behavior. Princeton University Press, 1953.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages26722680, 2014

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017

[5] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. arXiv preprint arXiv:1703.06182, 2017.

[6] Lillian J. Ratliff, Samuel Burden, and S. Shankar Sastry. Characterization and computation of local nash equilibria in continuous games. In 51st Annual Allerton Conference on Communication, Control, and Computing, Allerton 2013, Allerton Park & Retreat Center, Monticello, IL, USA, October 2-4, 2013, pages 917924, 2013.

[7] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In Advances in Neural Information Processing Systems, pp. 1825-1835, 2017

[8] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In International Conference on Machine Learning, pp. 363—372, 2018

[9] Nedi, Ozdaglar A. Subgradient Methods for Saddle-Point Problems[J]. Journal of Optimization Theory & Applications, 2009, 142(1):p.205-228

[10] Mertikopoulos, Panayotis, et al. "Mirror descent in saddle-point problems: Going the extra (gradient) mile." arXiv preprint arXiv:1807.02629 (2018)

[11] B. S. He, X. M. Yuan and W.X. Zhang, A customized proximal point algorithm for convex minimization with linear constraints, Comput. Optim. Appl., 56: 559-572, 2013.

[12] Gidel, Gauthier, et al. "A variational inequality perspective on generative adversarial networks." arXiv preprint arXiv:1802.10551 (2018)

[13] Wei Peng, Yu–Hong Dai, et al. Training GANs with centripetal acceleration. OMS, accepted
[14] Jin C., Netrapalli P. and Jordan M. I., What is local optimality in nonconvex–nonconcave minimax optimization? arXiv:1902.00618v2 [cs.LG] 3 Jun 2019.
[15] Dempe S., Foundations of Bilevel Programming, Kluwer, Dordrecht, 2002.
[16] Dempe S., A necessary and a sufficient optimality condition for bilevel programming problems, Optimization, 1992, Vol. 25, pp. 341–354.
[17] Falk J. E. and Liu J., On bilevel programming, Part I: nonlinear cases, Mathematical Programming, 70(1995), pp. 47–72.
[18] Ye J. J. and Zhu D. L., Optimality conditions for bilevel programming problems, Optimization, 33(1995), pp. 9–27. with correction in Optimization, 39(1997), pp. 361–366.
[19] Dempe S., Dutta J. and Mordukhovich B. S., New necessary optimality conditions in optimistic bilevel programming, Optimization, 56:5–6(2007), pp. 577–604.
[20] Dempe S. and Zemkoho A. B., The bilevel programming problem: reformulations, constraint qualifications and optimality conditions, Math. Program., Ser. A, 138(2013), pp. 447–473.
[21] Dempe S., Mordukhovich B. S. and Zemkoho A. B., Necessary optimality conditions in pessimistic bilevel programming, Optimization, 63:4(2014), pp. 505–533.
[22] Mehlitz P. and Zemkoho A. B., Sufficient optimality conditions in bilevel programming, arXiv:1911.01647v1 [math.OC] 5 Nov 2019.
[23] Jin C., Netrapalli P. and Jordan M. I., What is local optimality in nonconvex–nonconcave minimax optimization? arXiv:1902.00618v2 [cs.LG] 3 Jun 2019.

北大研究员林伟：多因果推理的工具变量——新与旧

整理：智源社区 王惠远

林伟，北京大学研究员，智源学者。
林伟本次演讲的主题是《Instrumental Variables for Multiple Causal Inference: Old and New》。

在报告中，林伟强调了因果推断在实现人类水平的人工智能中的巨大作用。如何利用机器学习领域中比较流行的方法去学习蕴含在数据中的因果结构是一个十分重要的问题。现代的机器学习任务通常需要同时考虑多个甚至高维的潜在原因。收集到的数据集包含多个潜在原因和结果时，如何对于任意一个原因去估计平均潜在结果的问题，是一个多重因果推断问题。因此考虑多重因果推断问题十分贴合实际需要。

这类问题的主要困难在于，如果原因和结果同时受一个未观测到的混杂变量 (confounding) 影响，在估计潜在原因对结果的影响时就会出现偏差。为了控制未被观测的混杂因素，在因果推断领域一个很常见的方法是引入工具变量 (instrumental variable)。一般而言，工具变量满足一定的条件之后才能有效减少混杂变量带来的偏差。工具变量的有效性假设包括:

- 工具变量与混杂变量独立;
- 工具变量与潜在原因不独立;
- 给定潜在原因和混杂变量，结果和工具变量独立。

对于多重因果推断问题，在有效性假设下，我们可以用两阶段最小二乘法来估计多个潜在原因对结果的影响。我们可以先用收集到的潜在原因和工具变量估计工具变量对潜在原因的影响，而后用工具变量预测得到的潜在原因作为自变量，结果作为因变量，利用线性回归得到潜在原因对结果影响的估计。

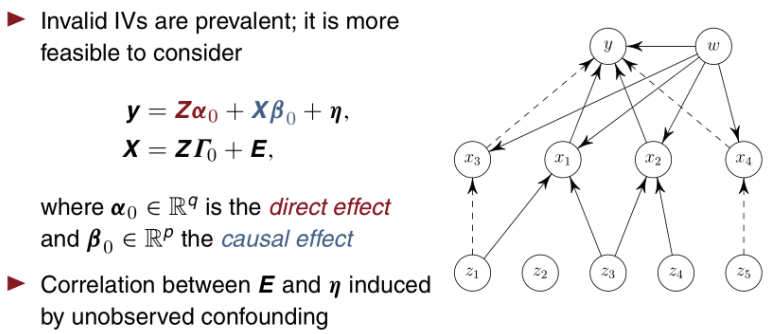
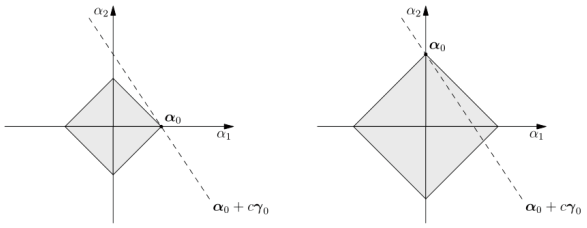


图 1：有效性条件不满足时的高维工具变量模型

但是当有效性假设第三条不满足时，也就是工具变量可能会对结果产生直接影响时 (如图 1)，直接采用两阶段最小二乘法估计存在着可识别性的问题。由于只有潜在原因对结果的影响需要估计，工具变量对结果的影响和潜在原因对结果的影响必须被区分开来。而一般情况下，工具变量对结果的影响和潜在原因对结果的影响是无

法被区分开的。林伟老师借用了带结构的压缩感知方面的数学工具来处理这个问题，并给出了十分优雅的可识别性的结果。



- Identifiability results
  - *Definition.* The *spark* of a vector space Φ is spark$(\Phi) = \min\{\|\phi\|_0 : 0 \neq \phi \in \Phi\}$. The *cospark* of a matrix is the spark of its column space
  - *Local identifiability in ℓ_0 balls:* Assume $\boldsymbol{\Gamma}_{0S}$. is of full column rank. The ℓ_0 problem with constraint $\|\boldsymbol{\alpha}_{0S}\| \leq k$ has at most one solution for any $\boldsymbol{\delta}_{0S}$ if and only if $k <$ cospark$(\boldsymbol{\Gamma}_{0S\cdot})/2$, where $S = \{1 \leq j \leq q : \boldsymbol{\gamma}_{0j} \neq \mathbf{0}\}$
  - Equivalence of ℓ_0 and ℓ_r solutions via *column space property*

图 2：可识别性

可以看到，尽管工具变量的维数可以很高，但是只要工具变量对结果能产生影响的个数比工具变量对潜在原因影响系数的某个称为 cospark 的维数的一半还要少，工具变量和潜在原因对结果的影响就可以被区分开来，可识别性由此可以得到保证。

- The *2SR-II* methodology
  - *Stage 1.* Estimate $\boldsymbol{\Gamma}_0$ from the cause model by

$$\hat{\boldsymbol{\Gamma}} = \arg\min_{\boldsymbol{\Gamma}} \left\{ \frac{1}{2n} \|\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{\Gamma}\|_F^2 + \sum_{j=1}^{p} \lambda_j \|\boldsymbol{\gamma}_j\|_1 \right\}$$

    and obtain the predicted causes $\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\boldsymbol{\Gamma}}$
  - *Stage 2.* Estimate $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ from the outcome model by

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\alpha} - \hat{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2 + \mu \|\boldsymbol{\alpha}\|_1 \right\}$$

  - Use the validity-restricted debiased estimator

$$\begin{pmatrix} \tilde{\boldsymbol{\alpha}}_{\hat{\mathcal{V}}^c} \\ \tilde{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}_{\hat{\mathcal{V}}^c} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} + \frac{1}{n} \hat{\boldsymbol{M}} (\boldsymbol{Z}_{\cdot \hat{\mathcal{V}}^c}, \hat{\boldsymbol{X}})^T (\boldsymbol{y} - \boldsymbol{Z}_{\cdot \hat{\mathcal{V}}^c} \hat{\boldsymbol{\alpha}}_{\hat{\mathcal{V}}^c} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

    for statistical inference

图 3：两阶段正则化方法

为了高效地把潜在原因对结果的影响估计出来，林伟提出了两阶段正则化方法。这个方法分为两步。第一步，利用 Lasso 方法估计工具变量对潜在原因的影响，并据此得到预测的潜在原因。第二步，给工具变量对结果的影响系数施加 L1 正则项，且以此分别求得工具变量和潜在原因对结果的影响。最后可选择对得到的结果进行去偏差化以方便统计推断。去偏差的估计可以证明是渐近正态分布。与经典的两步线性回归不同，额外的 L1 正则项经过适当的调参可以保证结果是稀疏的，进而确保估计的值满足可识别性。具体的理论保证按照协变量是否有随机性（确定性设计和随机设计）如下所示：

- *Deterministic design* conditions
  - $\kappa(\boldsymbol{Z}, \mathcal{M}_1(s_1), 3) \geq \kappa_1 > 0$, where $\mathcal{M}_1(s_1) = \{J \subset \{1, \ldots, q\} : |J| \leq s_1\}$
  - $\kappa((\boldsymbol{Z}, \boldsymbol{Z}\boldsymbol{\Gamma}_0), \mathcal{M}_2(s_2), 3) \geq \kappa_2 > 0$, where $\mathcal{M}_2(s_2) = \{J \cup \{q+1, \ldots, q+p\} : J \subset \{1, \ldots, q\}, |J| \leq s_2\}$
- Nonasymptotic error bounds
  - First-stage estimation
  - Second-stage estimation:

$$\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|_1 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq \frac{64C_0}{\kappa_1\kappa_2^2}(s_2 + p)\sqrt{\frac{s_1 \log q}{n}},$$

$$\|\boldsymbol{Z}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + \widehat{\boldsymbol{X}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 \leq \frac{64C_0^2}{\kappa_1^2\kappa_2^2} s_1(s_2 + p) \log q$$

图 4：两阶段正则法估计值的理论保证（协变量没有随机性）

可以看到，如果协变量是确定性设计，则两阶段正则化法的估计值在 L1 度量下是相合的，而且相应的预测误差（标准化后）也是趋于 0 的。

- *Random design* conditions
  - $0 < \tau_1 \leq \lambda_{\min}(\boldsymbol{\Theta}) \leq \lambda_{\max}(\boldsymbol{\Theta}) \leq \tau_2 < \infty$
  - $\|\boldsymbol{D}_{B^cB}(\boldsymbol{D}_{BB})^{-1}\|_\infty \leq 1 - a$ for $0 < a \leq 1$, where $B = A \cup \{q+1, \ldots, q+p\}$
- Distributional properties
  - *Valid set recovery*: $P(\widehat{V} = V) \geq 1 - c_1 q^{-c_2}$
  - *Asymptotic distribution:*

$$\sqrt{n}\begin{pmatrix} \widetilde{\boldsymbol{\alpha}}_{V^c} - \boldsymbol{\alpha}_{0V^c} \\ \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} = \boldsymbol{U} + \boldsymbol{\Delta},$$

where $\boldsymbol{U} \mid \boldsymbol{Z} \sim N(\boldsymbol{0}, \sigma_\eta^2 \boldsymbol{M}\boldsymbol{K}\boldsymbol{M}^T)$ and

$$P\left(\|\boldsymbol{\Delta}\|_\infty \geq (C_1 s_2\sqrt{s_1} + C_2 s_1)\frac{\log q}{\sqrt{n}} + (C_3\xi s_2\sqrt{s_1} + C_4 r_n)\sqrt{\log q}\right) \leq q^{-1}$$

图 5：两阶段正则法估计值的理论保证（协变量有随机性）

如果协变量是随机设计，则两阶段正则化法的解在去偏差化后具有渐近正态性。

工具变量可能是最早用来控制混杂变量的方式，但是它的思想至今仍然有用。演奏笛子时，用手指按住很少的音孔就可以吹出动人的声音，高维的工具变量也是如此。尽管工具变量的个数像笛子总的音孔数一样可能会很多，但只要对结果和潜在原因真正产生影响的个数较少，工具变量一样可以展现出强大的效用。这种思想对其他统计和机器学习方法也有深刻的启发和影响。

圆桌论坛：人工智能基础理论研究的回顾与展望

整理：智源社区　贾伟

在人工智能发展的今天，数学家被赋予了新的使命。

作为自然科学的基石，在任何一门科学发展成熟的时候，对其进行抽象、定义以及严格证明，都是数学发挥功力的时刻。人工智能进入以深度学习为代表的第三波爆发期后，迄今为止，大多数工作都还主要是凭借计算机科学家们的经验、灵感，以工程的思维来推动。近几年来有不少数学家已经开始认识到，对人工智能数理基础的研究或许将带来数学的又一春天。传统的数学（特别是统计）主要是从线性模型做起，直接分析优化，不用考虑学习；而深度学习在数学上本质上则是非凸的，学习策略影响学习结果。如何刻画这种学习？如何为以深度学习为代表的机器学习技术建立坚实的数理基础？对深度学习的研究，让原来局限在一个小圈子里的数学家们也有了很多机会，与计算机学家、人工智能学家、物理学家、脑科学家、计算神经科学家等坐在一起，共商人工智能的科学之本。与人工智能的交叉，将给数学界带来新的灵感。

当前，已经有不少数学家开始研究机器学习问题，例如 GAN 的数学描述即优化问题，双下降问题，极大极小优化问题，因果推断等。这些研究已取得或大或小的进展，但，正如北京大学张平文院士所言：“人工智能的数理基础，还不是一个成熟的、被明确定义的领域，人工智能数理基础研究的领导者还没有产生；正是因为这样，广大的青年学者现在还有很大的机会。”

在第二届智源大会“人工智能的数理基础”专题论坛中，8 位数学家共同回顾并探讨了人工智能基础理论在近几年取得的重要进展、当前最核心的挑战以及未来潜在的新思路和方向。

论坛嘉宾

张平文

- BAAI数理基础方向首席科学家
- 北京大学长聘教授
- 中国科学院院士

张志华

- BAAI数理基础方向研究员
- 北京大学长聘教授

史作强

- BAAI数理基础方向研究员
- 清华大学长聘副教授

董　彬

- BAAI数理基础方向研究员
- 北京大学长聘副教授

朱占星

- BAAI数理基础方向青年科学家
- 北京大学助理教授

朱宏图

- 滴滴出行首席统计学家
- 北卡大学教堂山分校终身教授

季春霖

- 光启高等研究院副院长
- 浸会大学兼职教授

邓　柯

- BAAI数理基础方向研究员
- 清华大学长聘副教授

一、回顾：人工智能基础理论研究近年来取得了哪些重要进展?

机器学习视角——

朱占星：在基础理论方面，我觉得近几年进展比较多的包括以下四个方面:

1、神经网络学习内在的工作机制是什么。近一两年大家研究比较多的是宽网络，即网络很宽时，神经网络的行为将是什么。大家发现这种学习很类似 kernellearning。但也有人提出质疑，因为宽网络有很强的限制，和现在大家普遍用的深度网络并不一样，根据 kernel 做出来的结果和我们深度学习做出来的结果仍然有很大差距。

2、用新的视角来看神经网络。有人尝试利用物理中的平均场理论，把每个神经元视作一个粒子，根据中心极限定理，来分析神经网络整体的行为。

3、DoubleDescent 问题。这种现象表明我们对传统的统计机器学习模型理解并不够透彻，例如前面 Johannes 对 bias–variancetrade–off 的新理解。

4、把机器学习看做一个动力系统。例如在训练的时候，把输入当做初始点，输出当做终点，训练时把步长不断缩小，这个过程可以看做一个连续 ODE，因此我们可以用已有的数学方法来解决一些问题。

其他方面进展，我觉得都还并不很顺利，例如神经网络性能与 data 之间的关系如何更好地进行数学上的刻画，如何 de–couple 训练策略和训练模型之间的关系等。

张志华：我对这个问题的理解有两个方面。

首先，机器学习（特别是深度学习）现在发现了很多现象，对这些现象，数学上能够提供什么样的刻画？这方面确实发现了一些现象，例如双下降等，针对这些现象确实已经有一些工作，但这些工作都包含了太多的假设，这些假设与真实的机器学习过程有很大的差距。从数学上解决比较好的是 GAN，原因在于：1）GAN 本身数学的定义就比较清楚；2）我们对 GAN 做分析时，已经把“深度”（最难的一块儿）去掉了。因此对 GAN 的分析就比较漂亮。此外便是对 min–max 的研究，现在也是研究比较清楚的，这里也没有考虑“深度”。把“深度”加进去的研究，还处于起步阶段。

其次，人工智能的数学基础，并不一定是对人工智能的数学刻画，也可以是用数学的工具提出一些新的方法。例如无监督，如何从数学的角度，给我们一些启示，提出一些新的方法。这方面还是取得了一定的进展，包括统计的鲁棒性以及林伟讲的因果推理。我觉得这方面的进展还是比较清楚明晰的。

统计学习视角——

季春霖：从统计的角度，我关注的有几个方面，

1、近似贝叶斯推断方面。近似贝叶斯推断最早是为了贝叶斯模型做后验分布计算，是一种替代蒙特卡

洛计算的手段。最近近似贝叶斯推断与 ML 结合的比较多，也受到了很大的关注，特别是随机变分推断 (StochasticVariationalInference) 能够处理复杂的、大规模的数据。其中 VE 应该是近似推断比较成功的例子。最近的一些突破主要是，近似推断尝试打破一些传统的基于模型假设的方法，提出了很多 model-free 的设想，在变分推断中会用到 proposal，这是一个分布，现在人们提出了很多不需要标准模型的 proposal，使得 proposal 更加灵活，把原来用模型来算的 likelihood 和 prior 变成用统计量直接度量，或用 GAN 直接替代 likelihood 和 prior 等。

2、另外关注比较多的是生成模型。其实 GAN 网络、VE 都属于生成模型，能生成很多复杂的数据，这对传统的统计来说是一个技术的节约。VE 本身还有一些基于统计模型的假设，它的重构损失等效于一个 likelihood，这就限制了模型的灵活性。而 GAN 用统计量直接生成数据和真实数据的距离，比较像统计学里的 ABC 计算。人们用不同的统计量去做 GAN 网络中的损失函数，尝试生成更好的效果。除了这些，还有尝试改变它的结构，引入条件或其他领域的先验知识，从而让生成模型更加逼真。

3、统计量除了在 GAN 中用的比较多之外，它还被用在跨域的度量，例如用在 TransferLearning、特征解纠缠。但统计量不是 on-line 的学习，因此我们应当关注利用 on-line 的方法去解这种统计量，这样会更有助于在机器学习里面使用。

4、数据生成。我们知道生成模型可以生成很多复杂数据，例如 GAN，在最初生成图片等，可以满足大家的好奇心，但实际上现在更多的关注是用生成的数据提高监督学习或强化学习的性能，包括 few-shot 或 zero-shotlearning 里面，利用生成数据提高监督学习的性能，都能达到很好的效果。这里值得关注的是，如何利用这个模型对数据里面的先验知识进行提取，并把这些先验知识转换成数据来喂给监督学习的模型，这还有很多创新的地方值得关注。

朱宏图：首先大家从理论上对 bias-variance 的研究还是不错的，也有很多人试图从逼近论的角度做深度学习的理论，不过还没有看到非常激动人心的东西，大家都还在尝试去做。

其次，大家尝试把统计模型和深度学习融合在一起去解决一些问题，因为本质上来说统计模型在解释性方面比较好。

另外，就是在强化学习中进行因果推断的研究。最近有越来越多的 IT 公司开始做因果推断。我们最近也有一些研究，结果已经出来了，效果还不错。

应用数学视角——

史作强：前面几位老师已经说的很全面了，把我想说的基本已经说完了。我再补充一点，现在有些研究会把物理中的一些约束放到 RNN 或 reenforcementlearning 中，构建一些网络。例如在 RNN 中，加入某种能量或其他一些物理量，就可以利用数学上的一些理论来处理，例如常识的依赖性等克服梯度消失 / 爆炸现象。这可能也是现在应用那个数学研究的一个趋势，即：考虑传统上的一些物理模型，看是否对 deeplearning 有一些启发。

董彬：近年来我们看到的一个趋势是，机理与数据的融合。不管你是 modeldriven，还是 datadriven，我们在做的就是把基于数据和基于我们已知的机理与知识进行结合。

我们原来做模型一般都是凭经验、直觉或基于非常强的假设做的设计，这些模型和算法普适性很好，可以在很大的空间中得到问题较好的解，但对于更具体的任务，特别是我们很多时候关心的具体问题的解是在一个较小的空间中的，普世的模型和算法就未必是最优的方法，不能充分挖掘这个小空间的结构，而深度学习方法却可以很好的刻画这些空间，这也是为什么深度学习方法在很多具体问题中都比传统方法要好。但是理论上，我们一直不知道怎么去描述这个小的空间，也就没法很好的解释为什么深度学习有如此好的性能，这也是理论上需要进一步探索的方向。在建模方面，我们需要把传统建模思想和深度学习思想融合，其关键是甄别哪些环节我们应该用传统的方法，哪些环节我们又需要利用机器学习的工具？这个目前已有很多成功的例子，但是整体规律和原则并不是很清楚，很多时候只能是 casebycase，需要有一个系统的指导。

这些年，我认为进展是大家意识到了我们需要把机理和数据融合。但我们还不是很清楚，到底是否存在一些系统的指导性原则，来指导 AI 更好的解决实际问题。

张平文：感谢以上六位专家分别从机器学习、统计学和应用数学三个角度来讲述数理基础的研究进展。

但什么是“人工智能的数理基础”呢？我觉得这个目前我们还说不太清楚。首先，它还不是一个成熟的领域，还不是一个被明确定义的领域。也正是因为这样，广大的青年学者就有很大的机会；因为在全球范围内人工智能数理基础研究方面的领导者还没有产生，所以大家都有机会。

第一，当前人工智能的数理基础研究最多的还是深度学习的数学理论，主要是因为第三轮人工智能的浪潮主要是因为深度学习到了，深度学习虽然在一些领域效果很好，但是人们不理解，可解释性成问题，所以这是当前最热的领域，但还有很多其它方面的研究。在我看来，应用数学，特别是计算，过去就没有可解释性的问题，因为我们都是从知识开始、从机理开始，所以没有可解释性的问题。传统的统计在我看来主要是怎么从数据到知识，就是用统计的手段，特别是在社会科学领域用得特别多，其实真正简洁与美的知识（像量子力学），并不是通过统计来的，主要是靠天才的努力。但是这样的东西毕竟有限，大量的还是社会科学、复杂科学，这里没有那么高的精度，但它也是知识，过去统计在里面起着极大的作用。我们来看机器学习想要干什么？实际上过去我们在数学圈子里面，阵地是划得很清楚的，从数据到知识是统计人的领域，从知识到决策或者到预测是计算人的领域。机器学习要一下子从数据到决策到预测，这就是机器学习想要干的，要把两个群体干的事情一把手接过去了，所以出现很多新的问题。当然这样挺好，但这些问题也不是短时间能够解决得了的。也就是说，可解释其实有两个层次：一个层次就是从算法的角度来说怎么可解释，另一个层次就是从模型和知识的角度来说可解释。这两个还是有区别。

第二，人工智能的数理基础真的是给了我们广义的应用数学，包含做应用数学、做统计、做机器学习甚至一些做工程的人，还有做脑科学、计算神经科学的人，真的是把我们团结在了一起，否则的话我就很难有机会来听统计学家的报告，很难有机会来听机器学习专家的报告，所以我觉得这是一个可能重构广义应用数学的机会，是一个非常重要的方向。

今天有非常多的年轻人在这里。所以我想强调：这个领域还不成熟，没有领导者，但它确实具有活力，所以希

望大家投身到我们人工智能的数理基础这样一个研究领域，不断地在大家的努力下让它变得更加成熟，然后做出一些原创性的成果。你们有很多的机遇，但挑战也不小。

二、挑战：人工智能基础理论研究当前最核心的挑战还有哪些?

朱宏图：统计学整个的理论基础都是基于线性模型做起来的，但我们现在想要处理的系统和问题太复杂，旧的一套框架完全不能适应这个发展，所以这个理论已经不能满足需求。现在大家做机器学习的人，一上来就说我有一个具体的问题，然后搜集一组数据，如果能够标注的话，我就是用这个标注的数据去做后面所有模型的开发算法，跳过整个理论研究去做。对于更复杂的系统我们基本上就开始做模拟器，尽量去模拟这个物理系统里面的粒子之间的交互或者人与人之间的交互，那么从模拟器的角度去做后面所有我认为重要的模块，所以这些东西是现有的数学以及所有的理论科学目前为止不知道怎么去刻画这个系统，就会造成理论和实践是有一个很大的间隙。因为能够证明出来的都是一些比较简单的情形，但我的情形比你的更复杂，所以基本就搞不定。因此我们面临的最大的挑战就是，我们对我们的目标没有一个很深刻的数学或其它的理论框架去刻画，以前那些简单的、比较容易处理的数学工具还是无效，所以造成了我们面临的挑战非常之大。

张志华：现在最大的挑战肯定还是深度学习的挑战。原来我们大部分的统计模型都是基于浅层的，一般都是一个凸问题，我们研究这个问题就相对比较容易，比如原来我做计算数学，计算数学本身原来可能就是一个连续方程，然后怎样去解它，这些数学的问题相对比较明确。对于深度学习只有两个问题，第一，数学刻画不明确，我们用一个什么样的数学定义去证明什么东西； 第二，用什么样的工具能够解决这个问题，现在也不是那么清楚。这是我认为的核心挑战。

季春霖：我想和大家探讨一下网络中不确定性的分析。现在我们在做经典的人脸识别的时候，都会把图像嵌入到空间里面，大家研究的时候更多关注的就是嵌入空间怎么设计比较合理，怎么度量嵌入比较合理，这样的话能够得到一个比较好的泛化能力。其实嵌入空间中是填不满，还有很多空余的区域。针对这些空余的区域，经典的统计方法中，偏离主要关注区域的话概率比较小，它至少有一个概率的描述，但深度学习却没有专门描述这些区域，这里可能和安全性比较高的 AI 领域有非常大的关系，描述一些风险事件就需要这种刻画。现在深度学习当中没有这种不确定性的刻画，或者做的相对比较少，人们不能像经典的贝叶斯模型可以把后验概率全部取样出来，网络这么大也不可能对参数所有的样本取样。现在人们在尝试着去做，但是这方面的工作还是做得不足，没有一个完整的方法去把这个不确定性和网络结合在一起。

朱占星：其实说到机器学习，一个很重要的问题就是表示学习。从 2006 年 Hinton 发《Science》的时候就有了，DNN 或者受限玻尔兹曼机，能够学到一个好的表示。到现在已经有十多年了，但也没有搞清楚什么是深度学习学到的表示。这个难点在于深度学习是由很多小的 buildingblock 堆起来的，每个 block 都可能对想要关心的表示都有关系。我们之前考虑的统计都是从线性模型做起来，这些模型求解起来比较困难，但根本不用关心学习，直接分析最优化就完事。但深度学习是非凸的，这个事情就和怎么选数据、怎么学习有很大的关系，不同的学习策略就会有不同的结果。所以现在大家的数据复杂了，模型也是非凸的，学习策略也很多样化，相互之间非常依赖，我们没有办法解耦，拿出一个单个的东西去研究，说白了就是刚才张老师说的，我们没有什么好的数学刻画。我觉得这是目前最困难的问题，现在大家做的很多事情其实相当于把这三者之间做最大化的假设，然后放到已有的数学分析的区域去做，但我们真正关心的那些问题还离得很远。

三、展望：人工智能基础理论研究的下一个阶段有哪些潜在的新思路、新方向?

朱宏图：我一般做什么东西必须要有应用场景，这个东西又要足够复杂。

首先，我们的衣食住行现在已经有了 IoT，所有东西都整合在一个平台上面之后，服务商和用户通过这个平台进行交易，我们叫做双边市场。双边市场某种程度上就是用 IT 所有的可能性，对老百姓的衣食住行各个方面进行改善。这里产生的问题比现在深度学习所做的三个主要方向（CV、NLP、语音）要更广。其实本质上来说很简单，就是你能不能给用户创造价值。

我认为最重要的点就是因果推断。有了一个 Action 之后，用户的满意度就提升了，效率就提高了，我要知道这个抓手是什么； 从数据的层面，我要知道怎么搜集数据，找到原因，然后再提高用户的满意度。在这里面，深度学习只是一个工具而已。

另外增强学习会变得越来越重要。因为现在收集数据的频次越来越频繁，这些数据会带来一个机遇，就是我能够不断地调整模型策略，比如慢性病和高血压，不断收集数据，实时调整治疗方案。

再就是匹配的问题，就是针对用户如何做最好的服务。这个问题很早就有了，但是未来配准问题会变成非常基础的数学问题，相比以前更重要，就是我给用户提供什么样的服务是最优的，或者在什么样的环境下是最优的。这些问题和深度学习融合在一起。

未来机器学习人工智能基础理论当中这是需要考虑的几个方向。

董彬：我们知道 Regularization is the key，但有没有一种统一的视角，写出一类正则化去做分析？这是我自己比较好奇的一点。

史作强：刚才几位老师的总结当中都有提到可解释性是深度学习的一个非常关键的东西，张老师的总结我特别认同，那个可解释性实际上是分为不同的层次，我们可以考虑模型的可解释性，也可以考虑结果的可解释性，某种意义上这个结果的可解释性是更复杂的。最传统的物理模型，流体力学的方程，我们认为那些方程是可解释的，因为我们是从物理的规律推出来的； 但由流体的现象，比如湍流，并不太好解释。因此，我觉得在未来可能更应该关注模型的可解释性； 结果过于复杂，依赖不同的场景、不同的应用，可解释性都是不一样的。关于模型的可解释性，至少我个人考虑的一件事情，就是模仿物理问题当中的过程，我们先建立一些规律，比如动量守恒和能量守恒，利用这些规律我们把模型推出来。六七年前我看到过 Stephen 有一些工作，就是从图像不变性把模型推出来。我们也可以尽量把模型限制在更小的范围内，减少模型中我们需要拟合的参数数目，尽量把模型的类型定下来，这也是我最近在思考的一些问题。

四、总结

张平文：首先特别感谢各位嘉宾在人工智能数理基础的进展、挑战和展望上发表自己的看法，大家说的都很有深度，对我们团队的研究也非常有指导意义，代表我们的团队感谢大家。

同时特别感谢北京智源人工智能研究院。我们这个方向还没有成型，还没有被明确定义，北京智源人工智能研究院就把我们这个方向作为第一个重大研究方向成立，其眼光真的是深远。这个平台能够使得我们不同领域的

人聚在一起，我们今天有这样的机会非常重要。

北京智源人工智能研究院这四天的会议有很多的报告其实跟我们密切相关，特别是后天专门有一个机器学习的论坛。机器学习专家们会有不同的视角，他们的讨论跟我们也非常的近，所以我希望大家多去听一听他们的报告，很多思想性的报告都会对我们有启发。

我真的是认为人工智能数理基础这个方向对数学的发展会非常的重要。当年统计学的基础是概率论，现在已经成为了数学最核心的方向。早几年我在北大数院当常务副院长的时候有一个改革，就是要把大学数学最基础的教育从“三高”(分析、几何、代数) 变成“四高”(加入概率论和随机分析)。现在概率论和随机分析已经渗透到了数学的方方面面，已经是最核心的数学。相信「学习理论」真的有可能在基础数学方面有新的突破，现在有人开始研究「离散拓扑」,「组合论」也焕发新的青春，能不能产生类似概率论这样新的数学是我们期待的。尽管短时间内不太可能产生的，但也还是有这种可能性。现在我们人工智能基础更多的是关注怎么去理解机器学习（特别是深度学习）的一些算法，我们也希望有一些新的算法产生； 但更高层次的研究是，我们希望人工智能能够回馈数学，产生类似于概率论和随机分析这样核心的数学。这需要一个过程，因为概率论有几百年的历史，真正成熟不到一百年，成为数学的核心也是最近的一二十年的事情，这是一个漫长的过程。但对人工智能的数理基础这个方向，我是充满期待。

今天有很多来听报告的年轻人，如果是学数学的，不管你是学基础数学还是学应用数学、计算数学、统计学、信息科学或者计算机科学，人工智能的数理基础是一个非常有活力、非常有前景的方向，欢迎大家加入。