中国科学院大学人工智能学院《文献阅读》课程—文献综述报告

课程编号：081104M05018H-01

在线优化与博弈论

撰写人：蔡怀广

学号：4325012000007287035

培养单位：自动化所

2023 年 6 月 20 日

在线优化与博弈论

摘要：当前，在线优化和博弈论都是机器学习中关于决策方向的热点研究内容。具体来说，在线优化研究的是单个个体在未来信息不可知情况下的序列决策方法，可以应用的场景为：改变的环境，信息不可知，求解近似最优，或者要求鲁棒性高的情况。博弈论研究的是许多个体之间策略相关的互动；很多种类的博弈（重复博弈、不完全信息博弈、不完美信息博弈）都可以和在线优化结合。本文首先回顾在线优化和博弈论的基本设定和经典方法，然后着重介绍最近几年在线优化和博弈论结合的研究：求解更宽松的的均衡，求解不完全信息的静态重复博弈，在线优化与公平性，研究重复博弈的最终状态(last iterate)，求解不完美信息的动态博弈，机制设计，复杂性与可学性理论。接着指出在线优化和博弈论实际上是个体在信息受限的情况下决策的两种主要方法，最后提出对在线优化和博弈论结合未来发展趋势的设想。

关键词：决策、不完全信息、在线优化、博弈论、遗憾、均衡

1. 在线优化与博弈论概述

1.1. 研究背景

这一部分我们简要介绍在线优化和博弈论这两个研究领域的基本设定、建模分类、方法分类，以及将在线优化和博弈论结合的原因和意义。

1.1.1 博弈论

金融市场、劳动力市场、社会团体之间存在大量个体之间的交互，最终的收益取决于所有个体的决策；因此单个个体的收益不仅仅取决于个体的决策，也取决于其他个体的决策。博弈论研究的就是这类情境下的决策方法。

具体来说，博弈论研究的是许多个体之间的互动，其中个体的策略是相互依赖的。值得注意的是，博弈不仅仅存在于对立或者部分对立的情景，也存在于合作的情景。对于合作的情景，个体的最优策略取决于对其他个体策略的判断。博弈的分类有很多种，这里只介绍和本文相关的：

1. 按照行动的先后次序：静态博弈，动态博弈（动作有先后顺序）
2. 按照博弈的次数：单次博弈，有限多次博弈，无限重复博弈
3. 按照参与人之间状态信息的知晓程度：完全信息博弈，不完全信息博弈（参与

者不完全知道其他人的收益矩阵）；完美信息博弈，不完美信息博弈（不是所有的博弈参与者都完全知晓他行动前的博弈过程和历史）。

博弈的结果往往是一个稳态——这个稳态中没有人愿意主动改变，这种稳态被称作均衡。纳什均衡指的就是博弈的各方陷入一个局面：没有任何一个个体可以通过仅仅改变自身行为获得额外收益。纳什均衡是博弈论中的基础概念，每一个有限博弈至少存在一个纳什均衡。但是求解纳什均衡仍然是很有挑战性的难题。目前已知的是求解一般的两、三、四人博弈纳什均衡均是 PPAD Hard 问题（不可解，除非 P=NP），但求解多人（五人及以上）纳什均衡是否仍是 PPAD Hard 问题还不确定。因此，求解复杂博弈的纳什均衡是不现实的。很多研究转而求解近似纳什均衡，但相关进展缓慢。

纳什均衡只是博弈中的稳态，它并不意味着就具备了我们希望的所有性质——比如社会福利最大化、个体诚信。因此机制设计（比如第一价格拍卖、第二价格拍卖）是博弈论的另一大块拼图——设计博弈的规则，使得各个参与者在达到纳什均衡的同时能够达到设计者所设定的目标。

以上所描述的内容均可和在线优化结合。接下来简要介绍一下在线优化。

1.1.2 在线优化

在线优化研究的是在有限信息下做出决策，希望与获得全部信息的情况下做出的决策差距不那么大。主要衡量指标是遗憾（Regret），通常定义如下[19]，表示 T 的时间长度内，环境不断变化（函数$f_t()$可以随着 t 的改变而改变），我们使用 A 算法付出的代价，和理论上最小代价的差值的上界：

$$\mathrm{Regret}_T(\mathcal{A}) = \sup_{\{f_1,\ldots,f_T\}\subseteq\mathcal{F}} \left\{ \sum_{t=1}^{T} f_t(\mathbf{x}_t^{\mathcal{A}}) - \min_{\mathbf{x}\in\mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}) \right\}$$

Regret 是一种 worst-case 指标，关注的是最坏情况下算法的性能。常见的在线优化算法有[19]：Hedge（或者说 Multiplicative Weights，零阶算法），在线梯度下降（一阶算法），在线镜像梯度下降（Online Mirror Descent，一阶算法），跟随领导者（Regularized Follow The Leader，RFTL，一阶算法）， 在线牛顿法（二阶算法），Regret-Matching[22]；这里备注的阶数是说反馈信息中包含的函数$f_t()$的几阶导数。上面的在线优化算法默认是单次反馈信息是完全的（也就是我们知道函数$f_t()$在所有合法输入下的值），但要是反馈信息是不完全的（也就是我们知道函数$f_t()$在某些输入下的值），就催生了另一大流派的算法——bandit 算法，比如 EXP3、UCB（强化学

习中常用）。这两类方法是近二十年机器学习理论中较为火热的研究内容。

在线优化算法可以在很多实际场景中得到应用，比如金融市场中股票购买方案，专家系统决策方法。若将适用范围限制在机器学习，在线优化算法主要是可以在机器学习中承担一个基础组件——优化算法。不仅仅是神经网络中直接使用在线优化算法训练神经网络，在线优化可以应用于更广泛的机器学习领域：比如一类著名的算法——Boosting（例如 AdaBoost）。这类算法刻画的是如何将一个弱学习器提升为（或者说优化为）强学习器的过程，而这个“提升过程”可以由任意一种在线凸优化算法完成[19]，对某些 Boosting 算法的改进也可以从其使用的在线凸优化算法入手。

机器学习领域中，在线优化受到关注的原因有两个：一个是“在线”的这种设定和符合大量现实情况（比如信息在交互中逐渐被揭晓），另一个是其具备的强大的理论保证。每一个在线算法都对应着一个 Regret，这个 Regret 就告诉了决策者在线优化算法的性能最差到什么程度。这种理论保证是神经网络这种目前主流范式严重缺乏的。虽然目前在线优化算法基本上只能处理凸函数的情况，但是在一些设定比较清晰和简单的情景，在线优化算法可以取得还不错的效果。如何将在线优化算法推广到非凸的情况、乃至进一步更加深入地和机器学习应用结合是研究者应该考虑的。

在以往的在线优化的研究[19]一般是针对单个参与者和环境的交互来进行的，很少有研究多个参与者的情况。实际上，这正是博弈论研究的问题。

1.2. 在线优化和博弈论结合的原因和意义（引言）

我们强调，在线优化与其说是一个研究领域，不如说是一个工具，可以应用于不同领域。这些领域具备以下一个或几个特质：改变的环境，信息不可知，近似最优，或者要求鲁棒性高的情况。这些特质恰好在不同类型的博弈中有所体现，比如“不完全信息的静态重复博弈”就是收益矩阵在改变，未来的收益矩阵信息不知道；比如“不完美信息的动态博弈”就是过去的动作信息不知道；在这两种博弈中，我们都希望采用的决策方案可以达到近似最优，而在线优化就能解决这一点。而对于一般的博弈的均衡求解，在线优化则可以求得近似纳什均衡或者更加宽松的均衡。

因此理论方面，在线优化可以很好地和博弈论中重复博弈、不完全信息博弈、甚至动态博弈、不完美信息博弈结合。此外，对于实际应用方面，2018 年[18]将 OMD 算法来训练 WGAN 可以有效解决 GAN 训练过程中循环的现象，17 年基于 CFR 算法，文章[21]解决了两人有限注德州扑克的问题，都是在线优化在博弈问题中的具体应用。

上面描述的是将在线优化运用到博弈场景中，但实际上在线优化也可以看做是算法和环境（或者说到来的数据、信息）两者之间的博弈，因此博弈也可以运用到在线优化中。这方面的主要成就是姚期智先生提出的 Yao's principle，它刻画了随机算法的性能下界。随机算法的意思是当数据到来时，以一定的概率选择一堆确定型算法的一种，相当于只有在数据到来时我们才确定最终采用的算法，因此随机算法本身就包含了在线的设定。Yao's principle 将随机算法看做是算法选择者和数据之间的博弈，用博弈论中经典的、由冯诺依曼提出的极大极小值原理证明了随机算法在最坏的数据分布的情况下期望性能差于在任意数据分布上期望性能最好的确定型算法。除此之外，也存在一些研究者研究涉及到多个个体的在线优化算法，此时博弈论中的纳什社会福利便作为指标衡量不同在线优化算法的性能。

总的来说，一方面，在线优化主要关注单个个体在信息受限情况下的决策，而博弈论的设置通常为多个个体和信息受限。因此在线优化为博弈论提供了研究工具，使得信息受限的博弈场景能够被研究和解决，沿用的在线优化算法不仅为单个个体的决策提供了解决方案，也确保了在所有个体都采用在线优化算法的情境下能达到更加宽松的均衡；另一方面，沿用在线优化的设置，新的类型的博弈也在不断出现。在线优化也可以看做是算法和环境（或者说到来的数据、信息）两者之间的博弈，博弈论中的一些指标比如纳什社会福利也反馈了在线优化的研究。但是将在线优化运用到博弈论的工作远远比将博弈论运用到在线优化的工作多，因此一方面本文的介绍将侧重于将在线优化运用到博弈论中，另一方面也指明将博弈论的方法应用到在线优化是一个还未充分发掘的方向。

2. 研究现状

接下来，介绍 2021~2023 年理论计算机顶会 STOC、FOCS、SODA，机器学习顶会 NIPS、ICML、ICLR，学习理论顶会 COLT，运筹学顶会 OR 还有 WWW、AAAI、EC 等顶会绝大部分关于在线优化和博弈论结合的研究，希望能给读者提供一个领域概览。

2.1. 求解更宽松的的均衡

在博弈中，纯策略纳什均衡（Pure Nash Equilibria，PNE）不一定存在；但是混合策略纳什均衡（Mixed Nash Equilibria，MNE）一定存在。但可惜的是，MNE 通常难以求解——对两人非零和博弈来说，求解纳什均衡已经是 PPAD 完全问题了，目前不存在多项式的解法；多人博弈的纳什均衡求解更为困难。因此，越来越多的研究转向可高效求解的均衡概念，比如相关均衡（Correlated Equilibria，CE）、粗相关均衡（Coarse Correlated Equilibria，CCE），或者近似求解纳什均衡。

而在线优化（或者说 No-Regret Algorithm）可以用来计算 CE、CCE。No-Regret Dynamics 收敛性定理讲的就是从在线优化得到 CCE 过程：每个智能体都采用 No-Regret 算法（比如 1.1.2 中介绍的 Hedge，在线镜像梯度下降等），那么历史平均策略收敛到 CCE。其中 No-Regret 算法指的是 Regret 为 o(T)，对于每个智能体意味着算法的长期平均结果收敛于最优解。类似的，对于相关均衡（CE），每个智能体使用 No-Swap-Regret Algorithm（NSRA）算法，那么历史平均策略收敛到 CE。而 NSRA 可以由 n 个 No-External-Regret 算法（Hedge）和一个 master 算法得到。

现在我们通过在博弈论场景中引入在线优化的技术，求得了比纳什均衡更加宽松的均衡——近似纳什均衡、CE、CCE。但在上述求解过程中，每个智能体付出的代价是长时间运行一个在线优化算法直至 regret 足够小，自然而然接下来的一个目标是，我们能不能加快这个收敛的过程？拿 CCE 举个例子：如果每个参与者使用 Hedge，那么对于参与者来说，Regret 是 $O(\sqrt{T})$；此时博弈以 $O(\sqrt{1/T})$的速率收敛于 CCE。我们可以让参与者使用特殊的算法，进而在特定场景下收敛速度更快。NIPS 21 的文章[2]便证明了在多参与者的 general-sum 博弈中，假设每个参与者使用 Optimistic Hedge 算法（更关注临近时间信息的 Hedge 变体），那么每个参与者的 Regret 是 Poly(log T)；此时博弈以 O(1/T)加一个 log 量的速率收敛于 CCE。类似的研究还有 NIPS 15 的文章[3]。

类似于上面对 CCE 收敛速率的提升，在 STOC 22 的文章[7]中作者们开发了算法使得博弈以 O(1/T)加一个 log 量的速率收敛于 CE。理论上这些研究都可以被改进，因为还没有研究指出在线优化算法求解均衡的最佳速率是多少。

除了 CCE 和 CE 外，还有针对近似 NE 的研究：ICML 22 的文章[17]作者们发现除了零和博弈外的几类博弈问题比如 constant-sum polymatrix games 和 strategically

zero-sum games，当所有参与者采用 optimistic mirror descent (OMD)算法，博弈将收敛于一个近似纳什均衡。其中近似纳什均衡的定义如下，直观上说就是任意一个用户逃离当前策略能获得最大新增效用不超过 ε 。

Definition 2.1 (Approximate Nash equilibrium)**.** A joint strategy profile $(\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*) \in \prod_{i\in[n]} \Delta(\mathcal{A}_i)$ is an ϵ-*approximate (mixed) Nash equilibrium* if for any player $i \in [n]$ and any unilateral deviation $\boldsymbol{x}_i \in \Delta(\mathcal{A}_i)$,

$$u_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i}^*) \leq u_i(\boldsymbol{x}^*) + \epsilon.$$

上面的例子展示了使用在博弈论中使用在线优化算法可能可以得到更紧的纳什均衡的近似系数，但其实在线优化并不是必要的。比如在 AAAI 22 的文章[10]作者们则研究了一个特定的博弈问题叫做 multi-leader single-follower congestion game （多个用户(领导者)每个从一组资源中选择一个资源；在观察到已实现的负载后，敌手(单追随者)攻击具有最大负载的资源，从而给领导者们造成额外的成本），这里面就没有用到在线优化算法，但却得到更紧的纳什均衡的近似系数。作者们首先证明了在这个博弈中纯策略纳什均衡可能不存在（通过设计一个特定的博弈例子，枚举纯策略纳什均衡点可能的每个位置，说明在每个位置上用户都有动机逃离这个点进而这个点不稳定）。因此就有必要研究近似纳什均衡。其中近似纳什均衡的定义如下，直观上说就是任意一个用户逃离当前策略所新付出的代价至少是原代价的 1/ α 倍。

Definition 2. *A strategy profile $x \in X$ is an α-approximate pure Nash equilibrium (α-PNE) of G for some $\alpha \geq 1$, if for all $i \in N$:*

$$\pi_i(x) \leq \alpha \cdot \pi_i(y_i, x_{-i}) \text{ for all } y_i \in X_i.$$

值得注意的是，这里的近似是“乘性”的，而上面截图中对 CCE 的近似都是“加性”的。而作者给出的理由是在本文的博弈中，由于各个量未曾归一化，所以使用“加性”近似是不合理的。进一步，作者给出了此博弈的紧的近似比：α =1.1974，这个系数是$-x^3 + x^2/2 + 1 = 0.$ 的唯一解。更进一步，作者给出了得到这个近似纯策略纳什均衡的一个算法，且多项式时间的。无论在理论（系数是紧的）和应用（算法多项式时间）作者都给出了令人满意的结论。

OR 22 的文章[10]作者们研究了重复折扣博弈的情况下，每个智能体的损失是向量的情况，提出了一种近似动态规划技术来使得智能体近似最优策略且此时损失得到理论保证，此算法在实际环境中优于 Hedge。

以上的研究都可以归类于“加速”，是一种典型的在线优化和博弈做结合的研究思路——改进以往的理论结果，使得 CE、CCE 的收敛更加快速或者近似 NE 的系数更小。这种工作主要依靠在线优化技术的发展，从而能在博弈的场景下证明更好的结果。

另一类工作是将在线优化中的典型场景，比如环境变化迁移到博弈中，形成新的博弈问题，见 2.2。

2.2. 求解不完全信息的静态重复博弈

上面一节求解 CCE 过程中，我们设定收益矩阵对于所有时刻是不变的。实际上这一类的研究大都遵循一下思路——设计一个能让所有参与者的 individual regret 较小的在线算法，同时让最终得到博弈结果尽可能接近纳什均衡。

但如果我们类似于在线优化算法中，假定收益矩阵是随时间变化的呢？此时同样地最小化所有参与者的 individual regret 会带来类似 CCE 的结果么？ICML 19 的文章[1]便证明了在二人零和博弈中，最小化两个人的 individual regret 和博弈收敛到纳什均衡点不可能同时存在！具体定理如下：

Theorem

Consider any algorithm that selects a sequence of x_t, y_t pairs given the past payoff matrices $A_1, \ldots, A_{t-1}$. Consider the following three objectives:

$$\left| \sum_{t=1}^{T} x_t^\top A_t y_t - \min_{x \in \Delta^{d_1}} \max_{y_t \in \Delta^{d_2}} \sum_{t=1}^{T} x^\top A_t y \right| = o(T), \quad (1)$$

$$\sum_{t=1}^{T} x_t^\top A_t y_t - \min_{x \in \Delta_X} \sum_{t=1}^{T} x^\top A_t y_t = o(T), \quad (2)$$

$$\max_{y \in \Delta_Y} \sum_{t=1}^{T} x_t^\top A_t y - \sum_{t=1}^{T} x_t^\top A_t y_t = o(T). \quad (3)$$

Then there exists an (adversarially-chosen) sequence $A_1, A_2, \ldots$ such that not all of (1), (2), and (3), are true.

从上面的定理来看，在不完全信息的静态重复博弈中，每个参与者使用在线优化算法来最小化 individual regret 似乎不是一个好的选择。但周志华等人在 ICML22 的文章[4]指出上面文章用于衡量最终结果和纳什均衡之间的差距的指标——NE-Regret 存在缺陷。他们构造了一个反例指出在此种输下，两个参与者每轮都达到单轮的纳什均衡，最终得到的 NE-Regret 线性正比于 T。接着他们提出一个更合理的指标 dynamic NE-regret，表示如下：

$$\text{DynNE-Reg}_T \triangleq \left| \sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} \min_{x \in \Delta_m} \max_{y \in \Delta_n} x^\top A_t y \right|.$$

和上面图中的差别仅在于将 min_max 移到了求和内部。最后作者类似于 2.1 中求 CE 的算法构造了叠加在一组基础算法之上的一个元学习算法，使得两个参与者的 individual regret 和 dynamic NE-regret 都达到了 o(T)。

上面的例子是一种典型的在线优化和博弈做结合的研究——对于新的博弈场景，提出新的、更合理的 Regret 指标，再去针对此 Regret 指标设计在线优化算法。

上面的研究将引入在线优化到博弈论中的原因是未来收益矩阵不可知，这是一种不完全信息，还有存在另一种不完全信息的情况是每个智能体收到的反馈信息存在噪声，此时我们就能使用在线优化算法的高鲁棒性来对抗这种噪声了。ICML 20 文章[12]在一类特殊的博弈——λ -cocoercive 博弈中，研究了反馈信息（效用梯度）存在噪声的情况，给出这种情况下 last-iterate 的定性和定量的结果。文章[13]也进行了类似的研究：反馈信息（这里指效用梯度）存在噪声的情况下（包括“乘性”和“加性”噪声），通过让每个个体采用特定的 no-regret 算法（optimistic gradient scheme with learning rate separation）的博弈更快收敛、且不需要先验知识。

介于 2.1 中提到的收益矩阵对于所有时刻是不变的和 2.2 中提到的收益矩阵对于所有时刻是变化的中间环境——零和重复博弈，但是收益矩阵是周期性变化的，NIPS 21 的文章[15]便证明了在这种设置下，每个智能体使用在线优化算法在 time-average 意义上可能不收敛！其中使用的分析技术涉及到了动力系统、常微分方程的使用。

2.3. 在线优化与公平性

在大多数在线优化与博弈论的研究中，在线优化被看做是一种技术。但实际上，博弈论中的一些概念特别是公平性也反哺了在线优化的研究。比如在 SODA 22 的文章[5]中，纳什社会福利（Nash social welfare，NSW）被当做优化指标。作者们希望在在线分配物品给 N 个参与者的过程中，最大化关于 NSW 的竞争比。作者首先证明了没有在线优化的算法可以达到 O(N)的竞争比。接着作者结合预测的思路，证明了如果告诉参与者它将获得的总物品价值的预测，那么运行提出的算法后，当在这种预测是准确时，NSW 将达到 O(log N)和 O(log T)的竞争比并且是紧的。且算法对预测误差鲁棒。

上面的例子是典型的在线优化的研究——指出针对某类问题的某个指标理论上界，证明提出的算法到达了此上界（于是此算法是 tight 的，这个问题也被 solved 了）。

值得指出的是，NSW 是在博弈论的研究中提出的，可以兼顾效率和公平性。这里被用作优化的指标，同时 NSW 也在一些计算机网络的流量控制、拥塞控制的文章中被使用，这里就不赘述了。

2.4. 研究重复博弈的最终状态(last iterate)

前面 2.1 中我们提到的 No-Regret Dynamics 收敛性定理讲的是从在线优化得到 CCE 过程：每个智能体都采用 No-Regret 算法，那么历史平均策略（这叫 averaged iterate）收敛到 CCE。实际上，这些智能体采用的最终策略可能不是任何一个运行过的策略，而是所有历史策略的平均，也就是说，最终策略可能是一个从来没运行过的、存在于虚空的策略。在理想情况下，智能体的策略会随时间慢慢收敛，那么智能体后期采用的策略就类似于最终的平均策略。但是在所有情况下，当博弈进行到后期，参与者真的会采用近似于历史平均的策略么？答案是否定的。

ICML 22 的文章[17]指出一类 No-Regret 算法（比如镜像下降）当被所有参与者采用时，最终会出现循环甚至混乱的行为。其原因是某个参与者的历史平均策略不代表他最终状态采用的策略！

实际上，从在线优化推导到 CE、CCE 的过程中，我们还有很多问题没有考虑：1. 假设不同参与者使用不同更新策略的方法呢？2. 不再是静态重复博弈，而不完美信息动态博弈？3. 到达均衡的效率？和最大社会福利之间的差距？

ICML 22 的文章[17]作者们主要将 OMD（optimistic mirror descent）算法应用于不同情景的博弈，部分给出了以上几个问题的答案。

On Last-Iterate Convergence Beyond Zero-Sum Games

Setting	Results
Games with nonnegative regrets (Sections 3.1 and 3.2)	• Bounded second-order path lengths (Theorem 3.1) • Optimal regret (Corollary 3.3) • $O(1/\sqrt{T})$ rates to Nash equilibria (NE) (Theorem 3.4)
Smooth games (Section 3.3)	• Outperforming the (robust) price of anarchy (Theorem 3.8)
Potential and near-potential games (Section 4)	• Optimal regret (Theorem 4.6) • $O(1/\sqrt{T})$ rates to NE (Theorems 4.7 and 4.10) • Convergence in Fisher markets (Appendix B.2)
Unconstrained general-sum games (Section 5)	• Convergence of OGD (Theorems 5.1 and 5.4) • Inefficiency of OGD (Proposition 5.2) • Instability of first-order methods (Theorem C.11)

Table 1. Overview of our main results.

其中对于 general-sum game 中的循环的行为，作者们用均衡的效率来考虑。而作者证明了每个参与者使用 OMD 时，要么博弈收敛到 ε -approximate NE，要么以 ε 的平方的量超出 PoA 。因此我们使用 OMD 的话就能有效解决博弈最终状态中循环的行为。

COLT 21 文章[11]研究了类似的问题，在所有稳定博弈（包括所有 monotone games 和 convex-concave zero-sum games）的情景下，所有参与者采用作者提出的一类基于 OMD 的 adaptive、no-regret 策略的话，将得到 O(1)的 social regret 和 individual regret，且在 last iterate 的意义下收敛于纳什均衡。特别地，ICLR18 文章[18]指出将 OMD 算法来训练 WGAN 可以有效解决 GAN 训练过程中循环的现象，而这是实际中机器学习的主流优化算法——梯度下降法无法做到的。这说明了在线优化算法不仅仅只是对传统优化算法的增量式改进，而的确在一些方面能够解决传统优化算法无法处理的情况。

但 ICML 20 文章[12]指出在一类特殊的博弈——λ -cocoercive 博弈中，使用在线梯度下降也不会出现 last iterate 不收敛的情况。具体来说，作者们修改了标准的在线梯度下降，采用了 adaptive 的思想使得算法可以在不知道博弈的先验知识（前面的 λ ）的情况下运行。除此之外，作者还研究了反馈信息中的梯度存在噪声的情况，并给出这种情况下 last-iterate 的定性和定量的结果。

上述研究可以分为两类：研究博弈场景中采用在线优化算法的智能体在 averaged iterate 和 last iterate 意义下的算法的收敛性和收敛速度，结论通常是“averaged iterate 意义下算法性能表现极好”或是“averaged iterate 处理不了的情况中，采用 last iterate 意义下的算法性能表现极好”。看上去似乎从 last iterate 要比 averaged iterate 角度设计的在线最优算法好，但是鲜有人真正严格比较 averaged iterate 和 last iterate 意义算法在同一个问题上的收敛性和收敛速度。在 COLT 20 的文章[28]作者们则在一类叫做 Smooth Convex-Concave Saddle Point Problems 的问题上说明了采用 Extragradient (EG) algorithm 时，last iterate 和 averaged iterate 意义下都能收敛，但是 averaged iterate 至少是 last iterate 收敛速度的平方量级那么快！具体来说，Convex-Concave Saddle Point Problems 指的是一类基础的博弈问题（形式化：$\min_x \max_y f(x,y)$），其中 f(x,y)对于 x 是凸函数，对于 y 来说是凹函数。这个设定是很通用的，比如 GAN 的目标函数就满足这一点。而 Smooth 指的是 f(x,y)的一阶导数和二阶导数都不是无穷值。Extragradient (EG)则是普通的梯度下降法的改进版（GD），采用 EG 而不是 GD 的原因是已经在理

论和实际中验证了当 x 和 y 两方都采用 GD 时，last iterate 意义下算法无法收敛。而 EG 则可以在 last iterate 意义下以及 averaged iterate 意义下都能收敛。EG（Extragradient）算法如下图所示：

- **Extragradient** algorithm *[Korpelevich, '76];* extra gradient at each time t:

$$x_{t+1/2} = x_t - \eta\nabla_x f(x_t, y_t), \qquad y_{t+1/2} = y_t + \eta\nabla_y f(x_t, y_t)$$
$$x_{t+1} = x_t - \eta\nabla_x f\left(x_{t+1/2}, y_{t+1/2}\right), \quad y_{t+1} = y_t + \eta\nabla_y f(x_{t+1/2}, y_{t+1/2})$$

COLT 20 的文章[28]作者们首先证明了 EG 在 last iterate 意义下的收敛速度就是 $O(\sqrt{1/T})$，这是通过首先证明 EG 的收敛速度快于 $O(\sqrt{1/T})$，然后再证明 EG 的收敛速度慢于 $O(\sqrt{1/T})$。接着证明 EG 的 averaged iterate 意义下收敛速度小于 O(1/T)，于是便说明了采用 EG 的历史平均策略（averaged iterate）的收敛速度是 EG 的实际策略（last iterate）平方量级。这个反直觉的结论作者也没能说明背后真正的机理。其实除此之外，我认为这份工作并没有标题写的那么厉害，因为它只在 EG 这种算法上验证了历史平均策略比实际策略快，对于哪些算法满足这个结论、哪些不满足、算法的共性之类地都还留待探究。作者也指明了未来的一些研究方向：对于目标函数 Nonconvex-Nonconcave 已经有实验（包括在大规模 GAN 中）验证 averaged iterate 是对收敛性有帮助的，但是这是一个很困难的理论研究问题。

2.5. 求解不完美信息的动态博弈

NIPS 07 文章[20]将遗憾定义在信息集上，定义了反事实遗憾，展示了如何最小化反事实遗憾从而最小化整体遗憾，因此在自我博弈中可以用来计算纳什均衡。从而给出了对扩展式博弈、不完美信息博弈的一个解决方案。基于上面给出的 CFR 算法，Science 17 文章[21]解决了两人有限注德州扑克的问题。

尽管 CFR 算法取得了极大的成功，但是 CFR 的收敛性仅在 time-average 的意义下成立，而非 last-iterate。NIPS 21 的文章[9]研究了扩展式博弈（不完美信息的动态博弈）中 last-iterate 的情况，具体来说，对于零和扩展式博弈，使用一些 optimistic regret-minimization 算法，将可以在 last-iterate 意义上收敛，并且某些情况下速度速度为指数级。

2.6. 机制设计

WWW 22 文章[8] 研究了在一价拍卖中，玩家通过在线优化的方法来出价。具体来说，一个物品，N 个人对物品的估值不变，每个人运行一个在线优化算法（满足 mean-based 性质的算法，例如 greedy(follow the leader)、epsilon-greedy、MWU），拍卖进行无穷轮。那从 time-average 和 last-iterate 的角度看，最终会达到一个均衡么？

作者的答案为取决于 N 个人中对物品的估值最高的人数：当最高估值人数大于等于 3，拍卖将在 time-average 和 last-iterate 收敛于纳什均衡。当最高估值人数等于 2，拍卖将在 time-average 的意义下收敛于纳什均衡，而 last-iterate 的意义下则不一定。当最高估值人数等于 1，拍卖将在 time-average 和 last-iterate 的意义下都不收敛于纳什均衡。

上面的研究的重点实际上是多次重复拍卖的环境下的均衡。只不过将常见的在线优化算法迁移到拍卖的情景下研究，并未涉及到最大社会福利、诚实出价等。

SODA 21 的文章[23]研究了在线环境下的组合拍卖的设定下，如何最大化社会福利。此处在线的设定是每天都有一些新的物品到达，且每个物品都有一个保质期，我们必须要在保质期前卖出这些物品。组合拍卖指的是竞价人可以对多种商品的组合进行竞价的拍卖方式。这篇文章就研究了 submodular 和 XOS 两种组合拍卖的情况下（定义如下）社会福利最大化的竞争比。submodular 的情况作者设计了在线拍卖算法，给出了 O(log m)的竞争比，对于其中贝叶斯的设定给出了 O(1)的竞争比；XOS 的情况作者给出了 $o((m/\log m)^{1/3})$的竞争比下界，m 是物品的总数。

Definition 2.2 (Additive, XOS, and Submodular Valuation)**.** *A valuation v is additive if for every bundle $S \subseteq U$, we have $v(S) = \sum_{j \in S} v(\{j\})$. A valuation v is XOS if there exist additive valuations $a_1, \cdots, a_q$ such that for every bundle $S \subseteq U$, we have $v(S) = \max_r a_r(S)$. A valuation v is submodular if for every bundle S and S', we have $v(S) + v(S') \geq v(S \cup S') + v(S \cap S')$.*

2.7. 复杂性与可学性理论

正如我们前面提到的：“如果每个参与者使用 Hedge，那么对于参与者来说，Regret 是 $O(\sqrt{T})$；此时博弈以 $O(\sqrt{1/T})$的速率收敛于 CCE。”Hedge（也叫 randomized weighted majority (RWM)、Multiplicative Weights Update）是最经典的在线优化算法。与开发新的算法不同，SODA 23 的文章[24]作者们研究了 Hedge 算法的效率——因为 Hedge 要对所有动作维持一个分数，在经济领域，动作空间是指数级的，这便限制了 Hedge 的实用性。但是作者们发现，对于一大类被称作结构化的博弈（structured game）中，

其实可以通过对动作空间采样得到一个高效率的 no-regret 算法——采样的时间为 polylogarithmic（多项式加一个 log 因子），从而解决一大类问题——比如 (discrete) Colonel Blotto game，matroid congestion， matroid security，和 basic dueling game。

COLT 21 的文章[25]作者们说明了在矩阵式博弈中进行学习（包括在线学习）可能是任意复杂度的。复制因子动力学（replicator dynamics），即对 hedge 算法的连续时间模拟，即使应用于非常有限的一类博弈（有限矩阵博弈），也能够近似任意的动力系统。他们的结论展示了机器学习中几乎无限的动态建模能力，但也有消极的意味，意味着这些能力可能以可解释性为代价。

COLT 21 的文章[26]作者们研究了在线学习能学到什么样概念类、学到什么程度，且推广了冯诺依曼的极大极小定理。

3. 分析总结

3.1. 已有方案分析

因为求解一般的两、三、四人博弈纳什均衡均是 PPAD Hard 问题，越来越多的研究转向可高效求解的均衡概念，比如 CE、CCE、或者近似纳什均衡。这类工作属于增量型，主要关注提出更好的在线优化技术，使得 CE、CCE 的收敛更加快速或者近似 NE 的系数更小。但是如果提出的算法到达了此问题的理论上下界（于是此算法是 tight 的），也意味着这个问题也被 solved 了，那这就是一个很深刻的结果——因为提出的在线优化算法可能反映出了问题的本质，启发性更大。

吸收了在线优化的设定，很多人转而研究在新的博弈场景下（比如收益矩阵是随时间变化的、收益矩阵是周期性变化的、不完美信息的动态博弈），提出新的、更合理的 Regret 指标（比如 Counterfactual Regret），再去针对此 Regret 指标设计在线优化算法，最终使得每个智能体最小化自身遗憾的同时，博弈也达到均衡。这类工作属于开创型，对在线优化算法的理论性能要求不那么高，研究的重点主要在指出研究方向（最小化提出的指标）。

上面研究的问题在于，虽然历史平均策略收敛到某个均衡，但博弈进行到后期，参与者不一定会采用近似于历史平均的策略。博弈可能会出现循环甚至混乱的行为。这里一大类研究就是提出可以让博弈在 last iterate 意义下收敛的在线优化算法。同样的更快的收敛速度也是研究者追求的。此外，还有研究关注相同算法算法在 averaged

iterate 和 last iterate 意义在同一个问题上的收敛性和收敛速度，理论和实验中都观察到在一些情况中 average 操作有利于加速收敛。

博弈中有一些很好的概念可能是在线优化的研究者未曾关注到的，比如兼顾了效率和公平的纳什社会福利（Nash social welfare，NSW）。研究者就要在各种在线问题中最大化关于 NSW 的竞争比。这类工作属于在线优化的传统思路——最大化或者最小化某个指标的竞争比。

吸收了在线优化的设定，机制设计也会出现新变化。比如无限轮拍卖会收敛么？在线的设定下，社会福利的竞争比最大为多少？用户能否诚实出价？这些机制设计的经典问题也会在在线的场景出现，这方面研究结果还不那么丰富。

此外，因为在线优化属于机器学习理论领域，自然有着针对在线优化算法的改进——采样、可学性、复杂度等，这类研究属于将机器学习理论的传统范式迁移到博弈场景中的在线优化算法中来，需要艰深广博的机器学习理论基础知识和高深的数学技巧，在研究难度上比前面的主题高一个层次，且抽象程度更大而应用可能性更低，属于真正的原理方面的研究。

3.2. 总结

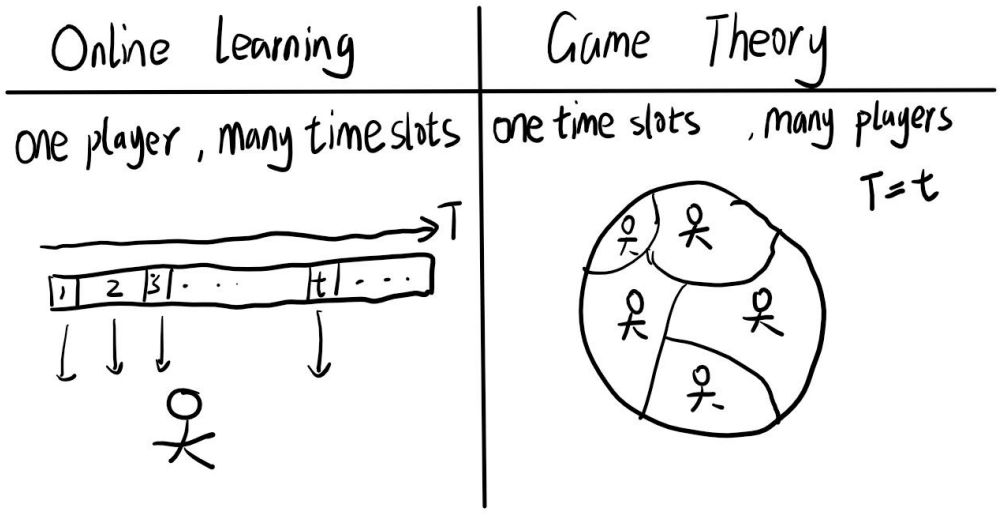
因为在线优化可以归类于机器学习理论，博弈论可以归类于理论计算机或者运筹学或者经济学。因此运筹学、经济学、数学、机器学习应用的研究者和机器学习理论的研究者都会或多或少关注在线优化和博弈论。不同领域的学者关注的问题也各有偏重（比如经济学家注重博弈的现实场景和定性结果，机器学习应用研究者关注博弈场景中在线优化的收敛速度和实际数据集上的表现，机器学习理论研究者关注博弈论和在线优化的统一模型的泛化性，运筹学者关注博弈论和在线优化的实际解和最优解之间的距离），研究思路也有区别，使用到的数学工具种类繁多。

总的来说，在线优化和博弈论的结合可以是非常紧密的，近年来也不断涌现出新的成果，是一个理论艰深但飞速发展的领域。一方面，在线优化为博弈论提供了研究工具，使得更加宽松的博弈能被达到；另一方面，沿用在线优化的设置，新的类型的博弈也在不断出现；而博弈论中的一些指标比如纳什社会福利也反馈了在线优化的研究。除了之外，还有研究者从传统机器学习理论的角度研究博弈论和在线优化的结合，在这部分研究中，博弈论和在线优化被统一到一种模型中，从而可以从采样、可学性、复杂度、泛化性等角度研究这个模型。

4. 展望

接下来我将从信息分发角度试图将在线优化和博弈论统一到同一个决策框架下。在线优化中每个时间段单个个体会收到部分信息，个体直到最后一刻或者永远也不会知道全部信息，也就是信息在时间上存在断裂。博弈论中每个个体会收到部分信息，大部分现实情况中（不完全信息静态博弈），每个个体收到的信息没有重合且没有个体知道全部信息。如果将每个个体掌握的信息看做拼块，那么大部分现实情况中每个个体的拼块都是不同的，各个个体的拼块组合起来才是完整的拼图，也就是信息在空间上存在断裂。在线算法和博弈论这两种方法就是分别研究信息在时间和空间存在断裂时如何决策的方法。值得说明的是，当信息全部已知的情况下进行决策，就是传统的“优化”的研究内容，这个领域虽仍有一些问题没能解决（比如说非凸问题），但已存在一些令人满意的解决方案（梯度下降法、遗传算法等）。

下面是我画的示意图，左侧是最基础的在线优化的设定（单个人，序列决策，存在未知信息），右侧是博弈论中的主流范式（多个人，同时做决策，可以进行多轮但是每个人掌握的信息不同）：



在如今以模式识别技术为代表的的感知方向已经充分发展的前提下，机器学习未来的研究热点当属于决策方向，在线优化和博弈论就对应着决策中两种核心范式（不确定性决策、多智能体决策）的解决方案，更加重要的是，在线优化和博弈论的 regret 和 price of anarchy 都描绘了算法的解和理论的最优解之间的距离，因此这两种方案给出的决策结论是有性能保证的。这种性能保证类似于一种对决策的有根据的解释，而

这是深度学习无法给出的。因此，我认为，为了发展可信的决策技术，在线优化和博弈论是绝对无法绕开的研究领域。

在线优化和博弈论两者都是 worst-case 的考虑观点（在线优化中假设环境会和算法对着干、博弈论中假设对手足够聪明），因此在线优化和博弈论给出的决策都偏“保守”，虽有性能保证，但往往实际效果远低于理想情况。如何平衡性能保证和实际效果的比重是在线优化和博弈论两个研究领域以及其交叉领域的未来研究趋势。

除此之外，尽管在线优化和博弈论给出的理论结果似乎很强大，但是这是建立在对现实情况进行精细的建模的基础上的，而且建模的结果应该比较简洁清晰，否则研究者便无法进行研究。因此在线优化和博弈论的融合通常发生在简单情景，无法直接进一步应用于复杂的现实情况。但深度学习恰恰相反，深度学习的核心在于将对复杂的现实情况的建模交由神经网络完成，但是建模的结果人类还无法理解。因此将在线优化和博弈论结合进而改进深度学习是一个研究方向，如何使获得的深度学习方法具备在线优化和博弈论类似的的性能保证是研究的难点，也是未来的研究方向。

总结来说，决策是未来机器学习的研究热点和主流研究方向。为了获得更加可信的决策方案，在线优化和博弈论是研究的必经路线。如何克服在线优化和博弈论偏保守的决策结果是未来这两个领域的研究趋势。而将深度学习和在线优化和博弈论相结合从而得到性能更佳且有理论保证的深度学习技术是未来的一个研究方向。

参考文献

1. A. R. Cardoso, J. Abernethy, H. Wang, and H. Xu, “Competing Against Nash Equilibria in Adversarially Changing Zero-Sum Games,” in Proceedings of the 36th International Conference on Machine Learning, Jun. 2019, vol. 97, pp. 921–930.
2. C. Daskalakis, M. Fishelson, and N. Golowich, “Near-Optimal No-Regret Learning in General Games,” in Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 27604–27616.
3. V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire, “Fast Convergence of Regularized Learning in Games,” in Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2015, pp. 2989–2997.
4. M. Zhang, P. Zhao, H. Luo, and Z.-H. Zhou, “No-Regret Learning in Time-Varying Zero-Sum Games,” in International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, 2022, vol. 162, pp. 26772–26808.

5. S. Banerjee, V. Gkatzelis, A. Gorokh, and B. Jin, “Online Nash Social Welfare Maximization with Predictions,” in Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022, 2022, pp. 1–19.
6. Z. Zhou, P. Mertikopoulos, A. L. Moustakas, N. Bambos, and P. W. Glynn, “Robust Power Management via Learning and Game Design,” Oper. Res., vol. 69, no. 1, pp. 331–345, 2021.
7. I. Anagnostides, C. Daskalakis, G. Farina, M. Fishelson, N. Golowich, and T. Sandholm, “Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games,” in STOC ’22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022, 2022, pp. 736–749.
8. X. Deng, X. Hu, T. Lin, and W. Zheng, “Nash Convergence of Mean-Based Learning Algorithms in First Price Auctions,” in WWW ’22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022, 2022, pp. 141–150.
9. C.-W. Lee, C. Kroer, and H. Luo, “Last-iterate Convergence in Extensive-Form Games,” in Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 14293–14305.
10. T. Harks, M. Henle, M. Klimm, J. Matuschke, A. Schedel, Multi-Leader Congestion Games with an Adversary, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022 , AAAI Press, 2022: pp. 5068–5075.
11. Y.-G. Hsieh, K. Antonakopoulos, and P. Mertikopoulos, “Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Nash Equilibrium,” in Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA, 2021, vol. 134, pp. 2388–2422.
12. T. Lin, Z. Zhou, P. Mertikopoulos, and M. I. Jordan, “Finite-Time Last-Iterate Convergence for Multi-Agent Learning in Games,” in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, 2020, vol. 119, pp. 6161–6171.
13. Y.-G. Hsieh, K. Antonakopoulos, V. Cevher, and P. Mertikopoulos, “No-Regret Learning in Games with Noisy Feedback: Faster Rates and Adaptivity via Learning Rate Separation,” 2022.hal-03694134
14. D. Q. Vu, K. Antonakopoulos, and P. Mertikopoulos, “Fast Routing under Uncertainty: Adaptive Learning in Congestion Games via Exponential Weights,” in Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 14708–14720.
15. T. Fiez, R. Sim, S. Skoulakis, G. Piliouras, and L. J. Ratliff, “Online Learning in Periodic Zero-Sum Games,” in Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 10313–10325.
16. A. Giannou, E.-V. Vlatakis-Gkaragkounis, and P. Mertikopoulos, “On the Rate of Convergence of Regularized Learning in Games: From Bandits and Uncertainty to Optimism and Beyond,” in Advances in Neural Information Processing Systems 34:

Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 22655 – 22666.
17. I. Anagnostides, I. Panageas, G. Farina, and T. Sandholm, "On Last-Iterate Convergence Beyond Zero-Sum Games," in International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, 2022, vol. 162, pp. 536 – 581.
18. C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training GANs with Optimism," ICLR, 2018.
19. E. Hazan, "Introduction to Online Convex Optimization," CoRR, vol. abs/1909.05207, 2019, [Online]. Available: http://arxiv.org/abs/1909.05207
20. M. Zinkevich, M. Johanson, M. H. Bowling, and C. Piccione, "Regret Minimization in Games with Incomplete Information," in Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007, 2007, pp. 1729 – 1736.
21. M. Bowling, N. Burch, M. Johanson, and O. Tammelin, "Heads-up limit hold'em poker is solved," Commun. ACM, vol. 60, no. 11, pp. 81 – 88, 2017.
22. A. Greenwald, Z. Li, and C. Marks, "Bounds for Regret-Matching Algorithms," 2006.
23. Y. Deng, D. Panigrahi, and H. Zhang, "Online Combinatorial Auctions," in Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021, 2021, pp. 1131 – 1149.
24. D. Beaglehole, M. Hopkins, D. Kane, S. Liu, and Shachar Lovett"Sampling Equilibria: Fast No-Regret Learning in Structured Games," in Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023.
25. G. P. Andrade, R. M. Frongillo, and G. Piliouras, "Learning in Matrix Games can be Arbitrarily Complex," in Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA, 2021, vol. 134, pp. 159–185.
26. S. Hanneke, R. Livni, and S. Moran, "Online Learning with Simple Predictors and a Combinatorial Characterization of Minimax in 0/1 Games," in Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA, 2021, vol. 134, pp. 2289–2314.
27. V. Kamble, P. Loiseau, J. Walrand ,"An Approximate Dynamic Programming Approach to Repeated Games with Vector Losses." Operations Research 2022.
28. Noah Golowich, Sarath Pattathil, Constantinos Daskalakis,"Last Iterate is Slower than Averaged Iterate in Smooth Convex-Concave Saddle Point Problems. "In Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria] (pp. 1758–1784). PMLR.