2021年9 月　　运 筹 学 学 报　　第25卷 第3期
Sep., 2021　　Operations Research Transactions　　Vol.25 No.3

DOI: 10.15960/j.cnki.issn.1007-6093.2021.03.004

非凸极小极大问题的优化算法与复杂度分析*

徐 姿[1,†]　张慧灵[1]

摘要 非凸极小极大问题是近期国际上优化与机器学习、信号处理等交叉领域的一个重要研究前沿和热点，包括对抗学习、强化学习、分布式非凸优化等前沿研究方向的一些关键科学问题都归结为该类问题。国际上凸-凹极小极大问题的研究已取得很好的成果，但非凸极小极大问题不同于凸-凹极小极大问题，是有其自身结构的非凸非光滑优化问题，理论研究和求解难度都更具挑战性，一般都是NP-难的。重点介绍非凸极小极大问题的优化算法和复杂度分析方面的最新进展。

关键词 极小极大优化问题，复杂度分析，一阶算法，(随机)梯度下降上升算法，交替梯度投影算法，非凸优化，机器学习

中图分类号 O221.2

2010 数学分类号 90C47, 90C26, 90C30

Optimization algorithms and their complexity analysis for non-convex minimax problems*

XU Zi[1,†]　ZHANG Huiling[1]

Abstract The non-convex minimax problem is an important research front and hot spot in the cross-fields of optimization, machine learning, signal processing, etc. Some key scientific issues in frontier research directions such as adversarial learning, reinforcement learning, and distributed non-convex optimization, all belong to this type of problem. Internationally, the research on convex-concave minimax problems has achieved good results. However, the non-convex minimax problem is different from the convex-concave minimax problem, and it is a non-convex and non-smooth optimization problem with its own structure, for which, the theoretical analysis and the algorithm design are more challenging than that of the convex-concave minimax problem, and it is generally NP-hard. This paper focuses on the latest developments in optimization algorithms and complexity analysis for non-convex minimax problems.

Keywords minimax optimization problem, complexity analysis, first order method, (stochastic) gradient descent ascent algorithm, alternating gradient projection algorithm, nonconvex optimization, machine learning

Chinese Library Classification O221.2

2010 Mathematics Subject Classification 90C47, 90C26, 90C30

收稿日期: 2021-03-24

* 基金项目: 国家自然科学基金(Nos. 12071279, 11771208), 上海市自然科学基金(No. 20ZR1420600)

1. 上海大学理学院数学系, 上海 200444; Department of Mathematics, College of Sciences, Shanghai University, Shanghai 200444, China

† 通信作者　E-mail: xuzi@shu.edu.cn

极小极大优化问题，又称鞍点问题，模型如下：

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y), \tag{1}$$

其中$x\in\mathbb{R}^n$, $y\in\mathbb{R}^m$, $\mathcal{X}$, $\mathcal{Y}$分别是$\mathbb{R}^n$和$\mathbb{R}^m$欧式空间中的闭凸集，函数$f(x,y)$是定义在$\mathcal{X}\times\mathcal{Y}$上的光滑或者非光滑的实值函数。当$f(x,y)$关于$x$是(非)凸函数关于$y$是(非)凹函数时，这类问题被称为(非)凸-(非)凹极小极大问题。凸-凹极小极大问题因为对应的变分不等式是单调算子，且与拉格朗日对偶问题有密切的联系，已被一大批国际国内著名优化专家学者以及数学、统计、经济、计算机科学等很多领域的专家学者广泛研究[1−11]。而非凸极小极大问题(包括非凸-凹、凸-非凹、非凸-非凹极小极大问题)是有其自身特殊结构的非凸非光滑优化问题，一般都是NP-难的。近期，非凸极小极大问题的优化算法及复杂度分析，成为优化和机器学习、人工智能等交叉领域国际研究的前沿和热点问题。

比如，机器学习中的生成对抗网络(GANs)问题[12]:

$$\min_G\max_D V(D,G)=\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))], \tag{2}$$

其中$D(\cdot)$是估计数据点是否真实的判别函数，$G(\cdot)$是生成函数。自2014年该问题被“GANs之父”Ian J. Goodfellow提出以来，引起机器学习领域的轰动，被图灵奖得主、深度学习领域三巨头之一的Yann LeCun教授描述为“过去10年机器学习领域最有趣的问题之一”。该问题是非凸极小极大问题的一个特例。

又比如多域上的鲁棒学习问题[13]：

$$\min_x\max_{y\in\Delta} y^{\mathrm{T}}F(x)-\frac{\lambda}{2}D(y||q), \tag{3}$$

其中y描述不同域上的对抗分布，$F(x)=[f_1(x);f_2(x);\cdots;f_M(x)]\in\mathbb{R}^{M\times 1}$, 其中$f_m(x)=\frac{1}{S_m}\sum_{i=1}^{|S_m|}l\left(s_i^m,t_i^m,x\right)$ $(m=1,\cdots,M)$是第m个域上可能的经验风险函数，是非凸函数，其中 $S_m=\{(s_i^m,t_i^m)\}$, $1\leqslant m\leqslant M$是 M 个来自不同领域的训练集，用来训练机器学习模型，$s_i^m\in\mathbb{R}^N$,$t_i^m\in\mathbb{R}$, x 是需要学习的模型的参数, $l(\cdot)$ 是非负损失函数，$D(\cdot)$是不同概率分布之间的距离，q是某种先验概率分布，Δ 是标准单纯形。该模型考虑 M 个不同域中可能出现的最坏分布的同时还需要使得经验风险最小, 也是非凸极小极大优化问题的一个特例。

此外，国际上热门研究领域包括分布式非凸优化[14]，信号处理中的功率控制和收发器设计问题[14]、对抗性学习中的鲁棒性[15], 强化学习[16], 统计学习[17]等很多方向中的关键性科学问题和新兴热点问题，都可以建模为非凸极小极大问题。

理论上，很多著名的数学和优化难题可以看作是非凸极小极大问题的特例。比如，著名数学家S. Smale在1998年提出的18个数学公开问题的第7个：球面点散布问题[18](也是由文[19]提出的9个公开问题的第5个)在三维空间单位球面上的情形就是非凸极小极大问题的一个特例。又比如，著名的有限多个二次函数最大值的极小化问题，该问题是2008年由教育部、科技部、中科院、国家自然科学基金委员会联合编写的《10000个科学难题：数学卷》[20]中的12个运筹学相关的难题之一，也是2012年由中国科学院胡晓东研究员、袁亚湘院士、章祥荪研究员共同撰写的“运筹学发展的回顾和展望”[21]中的12个运筹学未解难题的第2 个，也是非凸极小极大问题的一个特例。

本文着重介绍非凸极小极大问题的优化算法和复杂度分析方面的最新研究进展。第一章主要介绍非凸极小极大问题的最优性条件等理论方面的研究现状。第2节、第3节分别

介绍非凸-凹和(非)凸-非凹极小极大问题的一阶优化算法和复杂度最新研究进展。第4节主要介绍非凸极小极大问题的零阶算法及复杂度分析方面研究进展。本文的主要侧重点有两个方面：一是求解这类问题的有效优化算法；另一个是这些算法的复杂度分析。2019年，袁亚湘院士、文再文教授、蓝光辉教授等合著的综述论文[22]中就指出，优化算法的收敛性分析是优化中很重要的一个领域，然而收敛性并不足以作为比较不同算法效率的标准，因此我们需要另外一套衡量优化问题难易程度以及优化算法效率高低的理论，这套理论被称为优化算法的复杂度分析理论。非凸极小极大问题的优化算法及复杂度分析，是近期优化和机器学习、信号处理等交叉领域国际研究的前沿和热点。

符号约定：对于向量，我们使用$\|\cdot\|$表示l_2范数。对于函数$f(x,y):\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$，我们使用$\nabla_x f(x,y)$ (或$\nabla_y f(x,y)$)表示f在点(x,y)处相对于第一个变量(或第二个变量)的局部梯度。$\mathcal{P}_{\mathcal{X}}$和$\mathcal{P}_{\mathcal{Y}}$表示对集合$\mathcal{X}$和$\mathcal{Y}$的投影。最后，我们使用符号$\mathcal{O}(\cdot)$来隐藏不依赖于任何问题参数的绝对常量，并使用$\tilde{\mathcal{O}}(\cdot)$来隐藏不依赖于任何问题参数的绝对常量和对数因子。

1 最优性条件

非凸极小极大问题的最优性条件是理论上重点关注的一个研究方向。对极小极大问题的一个著名的最优解的概念就是纳什均衡点(鞍点)，即$\forall x\in\mathcal{X},\forall y\in\mathcal{Y}$，满足

$$f(x^*,y)\leqslant f(x^*,y^*)\leqslant f(x,y^*) \tag{4}$$

的点。对于凸-凹极小极大问题，可以证明梯度下降上升法就可以很有效地找到一阶ε-纳什均衡点[23]。对于非凸极小极大问题，一般情形下求解纳什均衡点是NP-难的。退而求其次，一些最新的研究结果是尝试设计算法去寻找局部纳什均衡点[24,25]。2019年，美国加州大学伯克利分校、著名机器学习专家M. I. Jordan 团队的最新研究结果[26]表明，即使对$f(x,y)=\sin(x+y)$这种性态很好的函数，非凸极小极大问题的纳什均衡点，甚至是局部纳什均衡点，都有可能不存在。他们提出了全局和局部极小极大点的概念，定义如下:

定义 1 若存在 δ_0 和函数 h, 使得 $\lim\limits_{\delta\to 0}h(\delta)=0$, 其中 $\delta\in(0,\delta_0]$, 且对任意 (x,y), 当 $\|x-x^*\|\leqslant\delta,\|y-y^*\|\leqslant\delta$ 时, 有

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

则 (x^*,y^*) 为局部极小极大值点。

但对有些问题全局极小极大点可能都不是稳定点，并且局部极小极大点可能不存在。文[26] 将提出的局部极小极大值点与梯度下降上升方法(GDA)建立了一定的联系, 即在某些条件下, 梯度下降上升算法的所有稳定极限点在退化之前都是局部极小极大值点。2020年，戴彧虹研究员和张立卫教授[27]将局部和全局极小极大点的概念推广到了有约束的非凸-非凹极小极大问题的情形。目前，从理论上对该类问题最优性条件的刻画的已有研究结果非常少，仍是值得进一步研究的公开问题。已有的研究结果大多是退而求其次去寻找这类问题的ε-近似一阶稳定点[14,28]，即存在(x^*,y^*), 使得 $\|\nabla G(x^*,y^*)\|\leqslant\varepsilon$，其中

$$\nabla G\left(x^*,y^*\right):=\begin{pmatrix}\beta\left(x^*-\mathcal{P}_{\mathcal{X}}\left(x^*-\frac{1}{\beta}\nabla_x f\left(x^*,y^*\right)\right)\right)\\ \gamma\left(y^*-\mathcal{P}_{\mathcal{Y}}\left(y^*+\frac{1}{\gamma}\nabla_y f\left(x^*,y^*\right)\right)\right)\end{pmatrix}。 \tag{5}$$

或者当$f(x,y)$关于y是强凹函数时，求得满足$\|\nabla\Phi(x^*)\|\leqslant\varepsilon$的解，其中

$$\Phi(\cdot)=\max_{y\in\mathcal{Y}}f(\cdot,y)。$$

除此以外，还有文章中使用一阶ε-纳什均衡点[29], 即存在 (x^*,y^*), 使得 $\mathcal{X}(x^*,y^*)\leqslant\varepsilon$ 且 $\mathcal{Y}(x^*,y^*)\leqslant\varepsilon$, 其中

$$\mathcal{X}(x^*,y^*):=-\min_x\langle\nabla_xf(x^*,y^*),x-x^*\rangle\quad\text{s.t. }x\in\mathcal{X},\|x-x^*\|\leqslant1,$$

$$\mathcal{Y}(x^*,y^*):=\max_y\langle\nabla_yf(x^*,y^*),y-y^*\rangle\quad\text{s.t. }y\in\mathcal{Y},\|y-y^*\|\leqslant1。$$

2 非凸-凹极小极大优化问题的优化算法与复杂度

尽管凸-凹极小极大优化问题已有很多有效的算法可以快速求解，比如A. Nemirovski等提出的mirror-prox算法[30], 在 $\mathcal{X}$ 和 $\mathcal{Y}$ 都是有界集时, 可以以 $\mathcal{O}(1/\varepsilon)$ 的复杂度得到目标函数的近似鞍点；随后A. Auslender和M. Teboulle[31]将该算法推广到一类距离生成函数；R. Monteiro和B. Svaiter[32]采用了具有不同误差准则的混合近端外梯度算法, 并将复杂度结果推广到无界集合和复合目标；Y. Nesterov 等[1]提出的对偶外推算法以及P. Tseng等[33]的加速近端梯度算法, 都被证明了迭代复杂度为$\mathcal{O}(1/\varepsilon)$，文献[34] 已经证明了该速率是利用一阶算法求解这类光滑的凸-凹极小极大问题的最优迭代复杂度。但这些算法能否类似地推广到求解非凸极小极大问题，相关的理论分析是否可以类似推广，都在进一步研究中。目前，已有研究结果大多考虑的是非凸-凹极小极大问题这种情形。求解非凸-凹情形的极小极大优化问题的算法通常有两类：多循环算法和单循环算法。

2.1 多循环算法

多循环算法是目前正被广泛研究的一类算法。2018年，Mertikopoulos等[35]提出一种镜像下降算法，在很强的假设条件下，可以证明算法的收敛性。2018年，Rafique等[36]提出一种近端引导的随机镜像下降方法(PG-SMD/PGSVRG)求解非凸-线性极小极大问题，可以依概率收敛到$\Phi(\cdot)=\max\limits_{y\in\mathcal{Y}}f(\cdot,y)$的近似稳定点。2018年，M. Sanjabi等[37]提出一种基于非凸随机梯度下降算法求解GANs问题，其中极大化子问题是非精确求解的，并证明了对这种特殊的非凸-强凹极小极大问题，算法找到ε-近似一阶稳定点的迭代复杂度是$\mathcal{O}(1/\varepsilon^2)$。此后，他们还提出一种原始对偶算法来求解一类GANs 问题，可以证明该算法能收敛到ε-近似一阶稳定点[38]。2019年，Nouiehed等[39]提出一种多步上升下降梯度方法求解非凸-凹极小极大问题，见算法1。该算法在外层循环的每一步迭代中, 对于 y 的更新, 先利用正则化将原函数转换成非凸-强凹函数 $f_\lambda(x,y)=f(x,y)-\frac{\lambda}{2}\|y-\bar{y}\|^2$，$\bar{y}\in\mathcal{Y}$ 是某个给定的点, 再利用多步加速投影梯度上升算法求解内部正则化后的子问题 $\max_y f_\lambda(x_t,y)$ 的近似解得到 y_{t+1}, x_t的更新是直接利用投影梯度下降法。文中还证明了该算法内层循环的迭代复杂度为 $\mathcal{O}\left(\varepsilon^{-0.5}\right)$, 外层循环的迭代复杂度为 $\mathcal{O}\left(\varepsilon^{-3}\right)$, 因此算法得到目标函数的一阶 ε-纳什均衡点的复杂度是 $\mathcal{O}\left(\varepsilon^{-3.5}\right)$。

算法 1 (交替多步梯度算法[29])

输入: $x_0,y_0,\eta_1,\eta_2,K,T,N,\gamma_1=1$

for $t=0,\cdots,T$

for $k=0,\cdots,\lfloor K/N\rfloor$

if k=0, $z_1=y_t$; else $z_1=m_N$

for $i=1,\cdots,N$

$m_t=\mathcal{P}_{\mathcal{Y}}\left(z_t+\eta_1\nabla_y f_\lambda(x_t,z_t)\right)$,

$\gamma_{t+1}=\frac{1+\sqrt{1+4\gamma_t^2}}{2}$,

$z_{t+1}=m_t+\frac{\gamma_t-1}{\gamma_{t+1}}(m_t-m_{t-1})$

end for

end for

$y_{t+1}=m_N$,

$x_{t+1}=\mathcal{P}_{\mathcal{X}}\left(x_t-\eta_2\nabla_x f(x_t,y_{t+1})\right)$

end for

2019年， Thekumparampil等[39]提出了一种近端对偶隐式加速梯度算法求解非凸-凹极小极大问题，并证明了当$\mathcal{X}$是无约束, $\mathcal{Y}$是凸紧集时，算法求得$\Phi(\cdot)=\max\limits_{y\in\mathcal{Y}} f(\cdot,y)$的近似稳定点的迭代复杂度是$\tilde{\mathcal{O}}\left(\varepsilon^{-3}\right)$。Kong 和Monteiro[40]也提出了一种加速非精确临近点光滑化方法求解非凸-凹极小极大问题，能得到同样的迭代复杂度，但该方法在每一步外层迭代需要求解内层极大化问题的扰动光滑逼近子问题，而子问题的求解复杂度没有计算在内。2020年，Lin等[41]提出了一类加速算法求解光滑非凸-凹极小极大问题，该算法通过加入正则项把原函数$f(x,y)$ 转换成强凸-强凹函数, 再基于 Nesterov 加速梯度法求解, 该算法证明了在非凸-凹情形下可以达到 $\mathcal{O}\left(\varepsilon^{-2.5}\right)$ 的迭代复杂度, 在非凸-强凹情形下可以达到 $O(\sqrt{\kappa_y}\varepsilon^{-2})$ 的迭代复杂度, 其中κ_y是强凹条件数。迭代格式详见算法2，其中MAXIMIN-AG2子程序是利用加速梯度上升和下降分别求解内层和外层子问题，AGD就是加速梯度法，详见文[41]。这是目前迭代复杂度最好的多循环算法。

算法 2 (加速梯度算法[41])

输入：初始点x_0, y_0; 临近项系数l; 强凸系数μ，最大迭代步数T; 函数关于y的条件数κ_y，$\mathcal{Y}$的直径D_y; 精度ε

初始化：令$\delta=\frac{\varepsilon^2}{(10\kappa_y)^4 l}\cdot\left(\frac{\varepsilon}{lD_y}\right)^2$

for $t=1,\cdots,T$ **do**

$g_t(x,y):=f(x,y)+l\|x-x_{t-1}\|^2$,

$x_t\leftarrow$MAXIMIN-AG2$(g_t,x_0,y_0,3l,l,\mu,\delta)$

end for

在$\{1,2,\cdots,T\}$中依均匀分布产生下标s,

$y_s\leftarrow$AGD$(-f(x_s,\cdot),\,y_0,\,l,\,\mu,\,\delta)$

输出：(x_s,y_s)

已有的多循环算法的一个共同点是，这些算法在每一次更新内层变量y的时候都需要精确或者在ε-精度内非精确地求解内层的极大化子问题或者正则化的内层极大化子问题，这使得算法实现起来会比较复杂，且需要调的参数较多，相对比较难以推广到求解其他类型的或更为一般化的非凸极小极大问题。多循环算法的另一个缺点是难以求解具有多分块结构的极小极大问题, 因为它们的内层循环使用的加速步骤不容易扩展到有多块结构的情况, 而具有多分块结构的极小极大问题在机器学习和信号处理的分布式训练中是很常见的。

2.2 单循环算法

单循环算法因实现简单，且更易于推广到求解更一般化的非凸极小极大问题，倍受工程界青睐。目前非凸极小极大问题的单循环算法方面的已有研究成果很少。最为简单的单循环算法就是梯度下降上升算法(GDA)，类似于非线性优化中的经典梯度下降法，该算法采用交替梯度下降和梯度上升步，是一类一开始就被机器学习、统计等领域广泛应用于求解该类问题的优化算法，见算法3。

算法 3 (梯度下降上升算法(GDA))

步骤1 输入 x_0, y_0, η_x, η_y, 令 $t = 1$;

步骤2 更新 x 和y：

$$\begin{aligned} x_{t+1} &= \mathcal{P}_{\mathcal{X}}(x_t - \eta_x \nabla_x f(x_t, y_t)), \\ y_{t+1} &= \mathcal{P}_{\mathcal{Y}}(y_t + \eta_y \nabla_y f(x_t, y_t)); \end{aligned}$$

步骤3 当算法满足终止准则时, 停止; 否则, 令 $t := t + 1$, 回到步骤2。

2019年最新研究结果[42]表明，即使是对于双线性的极小极大问题这类最简单的凸-凹极小极大问题，当$\mathcal{X} = \mathbb{R}^n$且$\mathcal{Y} = \mathbb{R}^m$ 时，GDA方法也不能保证收敛。此后，一些改进的GDA算法被提出。2019年，Lin等[43]证明了，在$\mathcal{X}$ 是无约束集, $\mathcal{Y}$是凸紧集的情形下，GDA方法求解非凸-凹极小极大问题时，通过将x的迭代步长设置为ε^4量级，算法得到$\Phi(\cdot) = \max\limits_{y\in\mathcal{Y}} f(\cdot, y)$ 的ε-稳定点的迭代复杂度是$\mathcal{O}(\varepsilon^{-6})$。但对于GDA算法的理论研究结果目前非常有限，这也从一个侧面说明了极小极大问题与传统非线性优化的不同。2019年，Jin等[44]提出了一种GDmax算法求解非凸- 凹极小极大问题，迭代复杂度也是$\tilde{\mathcal{O}}\left(\varepsilon^{-6}\right)$，但内层极大化子问题的计算复杂度没有考虑在内。

2019年，美国明尼苏达大学M. Hong的研究小组[14]，类似于块上界梯度下降上升法的思想，提出杂交分块序列近似算法(HiBSA) 求解非凸-凹极小极大问题，可以证明该算法找到ε-近似一阶稳定点的迭代复杂度是$\mathcal{O}(1/\varepsilon^4)$。该算法在求解非凸-强凹极小极大问题是一个单循环算法，算法的迭代格式如下：

$$x_{k+1} = \arg\min_{x\in\mathcal{X}} \langle \nabla_x f\left(x_k, y_k\right), x - x_k\rangle + \tfrac{\beta}{2}\|x - x_k\|^2, \tag{6}$$

$$y_{k+1} = \arg\max_{y\in\mathcal{Y}} \langle \nabla_y f\left(x_{k+1}, y_k\right), y - y_k\rangle - \tfrac{\rho}{2}\|y - y_k\|^2。 \tag{7}$$

当$\mathcal{X}$ 和 $\mathcal{Y}$ 是凸紧集时, HiBSA算法求解非凸-强凹极小极大问题得到$f(x, y)$的一阶ε-稳定点的迭代复杂度是$\mathcal{O}\left(\varepsilon^{-2}\right)$。与非凸-强凹情形不同的是, HiBSA算法在求解一般非凸-凹极小极大问题时，迭代格式有所不同。对于y的更新, 需要求解原函数$f(x, y)$加上正则项的一个最大化子问题, 见算法4。

算法 4 (非凸-凹情形下的HiBSA算法)

步骤1 输入 x_1, y_1, 令 $k=1$;

步骤2 计算 β_k 并更新 x_k :

$$x_{k+1}=\arg\min_{x\in\mathcal{X}}\langle\nabla_x f\left(x_k,y_k\right),x-x_k\rangle+\tfrac{\beta_k}{2}\|x-x_k\|^2;$$

步骤3 计算 c_k 并更新 y_k :

$$y_{k+1}=\arg\max_{y\in\mathcal{Y}} f(x_{k+1},y)-\tfrac{\rho}{2}\|y-y_k\|^2-\tfrac{c_k}{2}\|y\|^2;$$

步骤4 当算法满足终止准则时, 终止; 否则, 令 $k:=k+1$, 回到步骤2。

在$\mathcal{X}$和$\mathcal{Y}$是凸紧集时, HiBSA算法求解非凸-凹极小极大问题得到$f(x,y)$的一阶ε-稳定点的迭代复杂度是$\mathcal{O}\left(\varepsilon^{-4}\right)$。但该算法每步迭代需要求解一个正则化的内层子问题，并未考虑求解内层子问题的计算复杂度，此时算法就不再是单循环算法了。

2020年6月，我们提出了一种一致的单循环交替梯度投影算法(AGP)[28]求解非凸-凹和凸-非凹极小极大问题，见算法5。

算法 5 (交替梯度投影算法(AGP))

步骤1 输入$x_1,y_1,\beta_1,\gamma_1,b_1,c_1$, 令$k=1$;

步骤2 计算β_k和b_k, 更新x_k:

$$x_{k+1}=\mathcal{P}_{\mathcal{X}}\left(x_k-\frac{1}{\beta_k}\nabla_x f(x_k,y_k)-\frac{1}{\beta_k}b_k x_k\right);$$

步骤3 计算γ_k和c_k, 更新y_k:

$$y_{k+1}=\mathcal{P}_{\mathcal{Y}}\left(y_k+\frac{1}{\gamma_k}\nabla_y f(x_{k+1},y_k)-\frac{1}{\gamma_k}c_k y_k\right);$$

步骤4 当算法满足终止准则时, 终止; 否则, 令 $k:=k+1$, 回到步骤2。

在算法的每一步迭代中, 仅使用简单的投影梯度步交替更新x和y。算法的主要思想是考虑目标函数$f(x,y)$的如下正则化函数，

$$\tilde{f}(x,y)=f(x,y)+\frac{b_k}{2}\|x\|^2-\frac{c_k}{2}\|y\|^2,$$

其中 $b_k\geqslant 0$ 和 $c_k\geqslant 0$ 是正则化参数。对于给定的 $(x_k,y_k)\in\mathcal{X}\times\mathcal{Y}$, AGP算法通过极小化 $\tilde{f}(x,y_k)$ 的线性近似和一个正则项的和来更新 x_k, 这等价于一个梯度投影步，即

$$\begin{aligned}x_{k+1} &= \arg\min_{x\in\mathcal{X}}\langle\nabla_x\tilde{f}\left(x_k,y_k\right),x-x_k\rangle+\frac{\beta_k}{2}\|x-x_k\|^2\\ &= \mathcal{P}_{\mathcal{X}}\left(x_k-\frac{1}{\beta_k}\nabla_x f(x_k,y_k)-\frac{1}{\beta_k}b_k x_k\right),\end{aligned}$$

而对于 y_k的更新, 类似地极大化 $\tilde{f}\left(x_{k+1}, y\right)$ 的线性近似和一个正则项的差，也等价于一个梯度投影步，即

$$\begin{aligned} y_{k+1} &= \arg\max_{y\in\mathcal{Y}} \langle \nabla_y \tilde{f}\left(x_{k+1}, y_k\right), y-y_k\rangle - \frac{\gamma_k}{2}\|y-y_k\|^2 \\ &= \mathcal{P}_{\mathcal{Y}}\left(y_k + \frac{1}{\gamma_k}\nabla_y f(x_{k+1}, y_k) - \frac{1}{\gamma_k}c_k y_k\right), \end{aligned}$$

其中 $\mathcal{P}_{\mathcal{X}}$ 和 $\mathcal{P}_{\mathcal{Y}}$ 分别是在 $\mathcal{X}$ 和 $\mathcal{Y}$ 的投影, $\beta_k > 0$ 和 $\gamma_k > 0$ 是步长参数。需要指出的是，AGP算法在求解非凸-强凹极小极大问题时，无需正则项，也即$b_k = c_k = 0$，算法简化到如下的迭代格式：

$$x_{k+1} = \mathcal{P}_{\mathcal{X}}\left(x_k - \frac{1}{\eta}\nabla_x f(x_k, y_k)\right), \tag{8}$$

$$y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(y_k + \rho\nabla_y f(x_{k+1}, y_k)\right)。 \tag{9}$$

请注意，在将β_k和γ_k固定为常数之后，AGP算法与GDA算法[43]密切相关。主要区别在于，GDA使用$\nabla_x f x_k, y_k)$和$\nabla_y f(x_k, y_k)$ 同步更新x_k和y_k，而AGP在x_k上执行梯度下降步之后，通过使用最新的梯度$\nabla_y f(x_{k+1}, y_k)$，而不是 $\nabla_y f(x_k, y_k)$，在y_k处执行梯度上升步。这是交替的梯度下降上升算法。我们也证明了此时AGP方法求解非凸-强凹极小极大问题得到$f(x, y)$的一阶ε-稳定点的迭代复杂度是$\mathcal{O}\left(\varepsilon^{-2}\right)$[28]。

与HiBSA算法相比，AGP算法更新y时不再需要求解内层子问题，而仅仅需要一次梯度投影计算即可。我们可以证明，AGP算法求解非凸-凹极小极大问题得到$f(x, y)$的一阶ε-稳定点的迭代复杂度是$\mathcal{O}\left(\varepsilon^{-4}\right)$。据我们所知，这是求解非凸-凹极小极大问题的第一个达到–4阶迭代复杂度的单循环算法。我们还证明了对非凸-线性极小极大问题[45]，交替梯度投影上升下降(AGP)算法找到ε-近似一阶稳定点的迭代复杂度是$\mathcal{O}(1/\varepsilon^3)$，并且将该算法应用于极大极小散度问题这类NP-难问题的求解，数值表现优于该问题的已有算法。

此外, 我们将AGP算法的思想用于求解具有如下形式的分块非光滑的极小极大优化问题[46]:

$$\min_{x_i\in\mathcal{X}_i}\max_{y_j\in\mathcal{Y}_j} f(x_1, x_2, \cdots, x_{K_1}, y_1, y_2, \cdots, y_{K_2}) + \sum_{i=1}^{K_1} h_i(x_i) - \sum_{j=1}^{K_2} g_j(y_j), \tag{10}$$

其中 $\mathcal{X}_i$ 和 $\mathcal{Y}_j$ 是凸紧集, $x := [x_1; \cdots; x_{K_1}]$ 和 $y := [y_1; \cdots; y_{K_2}]$ 是块变量, $f(x, y)$ 是连续可微的函数; h_i 和 g_j 是简单的凸函数但不一定光滑。我们提出分块交替近端梯度算法来求解分块非光滑的极小极大优化问题，见算法6。

算法 6 (分块交替近端梯度算法(APGP))

步骤1 输入 $x^1, y^1, \beta_1, c_1, \mu, \rho$, 令 $r = 1$;

步骤2 计算 c_r 并更新 x^r:

$$\begin{aligned} x^{r+1} &= \arg\min_{x\in\mathcal{X}} \langle \nabla_x f(x^r, y^r), x-x^r\rangle + h(x) + \frac{1}{2\rho}\|x-(1-c_r\rho)x^r\|^2 \\ &= \mathcal{P}_{\mathcal{X}}\left(x^r - \rho\nabla_x \tilde{f}\left(x^r, y^r\right)\right); \end{aligned} \tag{11}$$

步骤3 对于 $i=1,\cdots,K$，计算 β_r 并更新 y_i^r:

$$\begin{aligned} y_i^{r+1} &= \arg\max_{y_i\in\mathcal{Y}_i}\langle\nabla_{y_i}f(x^{r+1},w_i^{r+1}),y_i-y_i^r\rangle-g_i(y_i)-\frac{\mu_i+\beta_r}{2}\|y_i-y_i^r\|^2 \\ &= \mathcal{P}_{\mathcal{Y}}\left(y_i^r+\frac{1}{(\beta_r+\mu_i)}\nabla_{y_i}f\left(x^{r+1},w_i^{r+1}\right)\right); \end{aligned} \tag{12}$$

步骤4 当算法满足某种收敛准则时，终止；否则，令 $r:=r+1$，回到步骤2。

我们可以证明，APGP算法求解问题(10)的迭代复杂度也是 $\mathcal{O}\left(\varepsilon^{-4}\right)$。

2020年10月，Zhang等[47]提出了一种单循环光滑梯度下降上升算法(smoothGDA)，见算法7。对于一般化的非凸-凹极小极大问题，解到一阶ε-稳定点的迭代复杂度是$\mathcal{O}\left(\varepsilon^{-4}\right)$，对于非凸-线性情形，迭代复杂度是$\mathcal{O}\left(\varepsilon^{-2}\right)$。

算法 7 (光滑梯度下降上升算法(smoothGDA))

步骤1 输入 x^0,y^0, 令 $t=1$;

步骤2 更新 x^t,y^t,z^t

$$\begin{aligned} x^{t+1} &= \mathcal{P}_{\mathcal{X}}\left(x^t-c\nabla_x K\left(x^t,z^t;y^t\right)\right), \\ y^{t+1} &= \mathcal{P}_{\mathcal{Y}}\left(y^t+\alpha\nabla_y K\left(x^{t+1},z^t;y^t\right)\right), \\ z^{t+1} &= z^t+\beta\left(x^{t+1}-z^t\right); \end{aligned}$$

步骤3 当算法满足终止准则时, 终止; 否则, 令 $t:=t+1$, 回到步骤2。

对于一般化的非凸-凹极小极大问题，是否存在比-4阶的迭代复杂度更好的单循环算法，目前仍是一个公开问题。

3 (非)凸-非凹极小极大优化问题

对于(非)凸-非凹极小极大问题，由于对任意给定的 x, 当 $f(x,y)$ 关于 y 是非凹函数时, 求解内层最大化子问题 $\max\limits_{y\in\mathcal{Y}}f(x,y)$ 将会是一个NP-难问题, 因而已有的多循环算法，以及HiBSA和GDmax等单循环算法均无法直接用于求解这类问题。目前这类问题的优化算法方面的结果非常少。2020年，我们提出的一致的单循环交替梯度投影算法(AGP)[28]可用于求解凸-非凹极小极大问题，我们也证明了，AGP算法求解非凸-(强)凹情形得到$f(x,y)$的一阶ε-稳定点的迭代复杂度是$(\mathcal{O}\left(\varepsilon^{-2}\right))\,\mathcal{O}\left(\varepsilon^{-4}\right)$。目前所知，这是求解凸-非凹极小极大问题的第一个具有复杂度保证的算法。已有的求解非凸-凹和凸-非凹极小极大问题的单循环算法的复杂度结果见表1，注意到其中非凸-凹(非强凹)的情形下，HiBSA和GDmax算法不能算严格意义上的单循环算法，因为没有考虑内层子问题的计算复杂度。

表1 非凸-凹和凸-非凹极小极大问题的单循环算法

单循环算法	ε-稳定点	非凸-凹极小极大问题		凸-非凹极小极大问题	
		强凹	凹	强凸	凸
GDA[43]	$\|\nabla\Phi(x^*)\|\leqslant\varepsilon$	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\varepsilon^{-6}\right)$	未知	未知
GDmax[44]	$\|\nabla\Phi(x^*)\|\leqslant\varepsilon$	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\varepsilon^{-6}\right)$	无法求解	无法求解
HiBSA[14]	$\|\nabla G(x^*,y^*)\|\leqslant\varepsilon$	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\tilde{\mathcal{O}}\left(\varepsilon^{-4}\right)$	无法求解	无法求解
AGP[28]	$\|\nabla G(x^*,y^*)\|\leqslant\varepsilon$	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\mathcal{O}\left(\varepsilon^{-4}\right)$	$\mathcal{O}\left(\varepsilon^{-2}\right)$	$\mathcal{O}\left(\varepsilon^{-4}\right)$
smoothGDA[47]	$\|\nabla_x f(x^*,y^*)+\partial 1_{\mathcal{X}}(x^*)\|\leqslant\varepsilon$ $\|\nabla_y f(x^*,y^*)-\partial 1_{\mathcal{Y}}(y^*)\|\leqslant\varepsilon$	$\mathcal{O}(\varepsilon^{-2})$	$\mathcal{O}(\varepsilon^{-4})$	未知	未知

对于更一般化的非凸-非凹极小极大问题，2018年，Q. Lin等[48]提出了一种基于邻近点法的算法框架来解决弱凸-弱凹极小极大问题，证明了非渐近收敛性，但需要很强的、不可验证的假设条件，即函数 $f(x,y)$ 需要满足广义单调变分不等式条件[49]。更多的研究进展参见文献[29, 41, 50]。是否存在更好复杂度的一阶算法求解一般化的非凸-非凹极小极大问题，目前已有的优化算法，特别是复杂度分析方面的研究结果很少，值得进一步研究。

4 非凸极小极大优化问题的零阶算法

如果极小极大问题中函数$f(x,y)$是一个随机函数或者随机函数的近似函数, 比如有限和函数时(即$f(x,y)=\frac{1}{n}\sum_{i=1}^{n}f_i(x,y)$)，我们往往无法直接获得$f(x,y)$的精确梯度，或者计算精确梯度的代价非常大，此时求解该类问题的优化方法称为零阶方法。2018年，Menickelly等[51]提出了一个无导数方法求解极小极大问题，但没有非渐近的复杂度分析结果。此外，针对鲁棒优化和鲁棒学习里的应用背景，贝叶斯优化算法和一些启发式算法也被提出来求解极小极大问题[52]，均没有复杂度分析结果。2019年，Roy等[53]研究了强凸-强凹极小极大问题的零阶Frank-Wolfe算法，并给出了非渐近的复杂度分析。2019 年, Liu 等[54]提出了非凸-强凹极小极大问题的ZO-MIN-MAX零阶算法，证明了迭代复杂度是$\mathcal{O}(1/\varepsilon^6)$。2020年，Wang等[55]提出了ZO-SGDA 和ZO-SGDMSA零阶算法求解非凸-强凹极小极大问题，迭代复杂度是$\mathcal{O}(1/\varepsilon^4)$。Xu等[56]提出了加速版零阶算法求解非凸-强凹极小极大问题，迭代复杂度是$\mathcal{O}(1/\varepsilon^3)$。Huang等[57]利用基于动量的方差缩减技术，提出了加速零阶动量梯度下降上升算法，迭代复杂度是$\mathcal{O}(1/\varepsilon^3)$。2020年，Luo等[58]在更强的均方Lipschitz梯度的假设条件下，提出了非凸-强凹极小极大问题的零阶算法，迭代复杂度算法是$\mathcal{O}(1/\varepsilon^3)$。2020年，Yang等[50]中提出了方差减少的交替梯度下降上升算法, 且是有限和形式 $f(x,y)=\frac{1}{n}\sum_{i=1}^{n}f_i(x,y)$, $\mathcal{X}$ 和 $\mathcal{Y}$ 是无约束的情形，证明了交替梯度下降上升算法能全局线性收敛到鞍点, 算法的复杂度为 $\mathcal{O}\left(n\kappa^3\log\varepsilon^{-1}\right)$, 其中 $\kappa=\frac{l}{\min\{\mu_1,\mu_2\}}$, l 为 f_i 的 Lipschitz 常数， $\mu_1,\mu_2>0$ 为使得$\forall x,y$,

$$\|\nabla_x f(x,y)\|^2\geqslant 2\mu_1[f(x,y)-\min_x f(x,y)],\quad \|\nabla_y f(x,y)\|^2\geqslant 2\mu_2[\max_y f(x,y)-f(x,y)]$$

同时满足的常数。目前非凸极小极大问题的零阶算法方面的研究结果不多，大多集中在非凸-强凹情形。

5 结束语

目前国际上非凸极小极大问题的研究，无论是理论还是算法方面，都还有许多重要而又难度大的问题需要解决。比如，极小极大问题所特有的复杂度下界，即便是凸-凹的情形，目前紧的复杂度下界也仍是一个公开问题。在一般的假设条件下，是否存在基于梯度的算法，找到ε-近似一阶稳定点的迭代复杂度达到$\mathcal{O}(1/\varepsilon^2)$，这也仍是一个公开问题[59]。如何从机器学习和信号处理中实际的非凸极小极大问题的具体结构出发，设计相应的优化算法求解大规模问题，进行复杂度分析，对机器学习和信号处理等与优化的交叉领域的研究与发展有着重要的指导意义。这些方面本文都未涉及，有兴趣的读者可以查阅相关文献。

参 考 文 献

[1] Nesterov Y. Dual extrapolation and its applications to solving variational inequalities and related problems [J]. *Mathematical Programming*, 2007, **109**(2/3): 319-344.

[2] Monteiro R, Svaiter B. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean [J]. *SIAM Journal on Optimization*, 2010, **20**(6): 2755-2787.

[3] Juditsky A, Nemirovski A. Solving variational inequalities with monotone operators on domains given by linear minimization oracles [J]. *Mathematical Programming*, 2016, **156**(1/2): 221-256.

[4] Facchinei F, Pang J. *Finite-Dimensional Variational Inequalities and Complementarity Problems* [M]. Berlin: Springer Science & Business Media, 2007.

[5] Von Neumann J, Morgenstern O. *Theory of Games and Economic Behavior* [M]. Princeton: Princeton University Press, 2007.

[6] Gidel G, Berard H, Vignoud G, et al. A variational inequality perspective on generative adversarial networks [C]//*International Conference on Learning Representations*, 2019.

[7] Chen X. Global and superlinear convergence of inexact uzawa methods for saddle point problems with nondifferentiable mappings [J]. *SIAM Journal on Numerical Analysis*, 1998, **35**(3): 1130-1148.

[8] 袁亚湘. 非线性优化计算方法 [M]. 北京：科学出版社，2008.

[9] He B, Yuan X. Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective [J]. *SIAM Journal on Imaging Sciences*, 2012, **5**(1): 119-149.

[10] Chen Y, Lan G, Ouyang Y. Optimal primal-dual methods for a class of saddle point problems [J]. *SIAM Journal on Optimization*, 2014, **24**(4): 1779-1814.

[11] Chen Y, Lan G, Ouyang Y. Accelerated schemes for a class of variational inequalities [J]. *Mathematical Programming*, 2017, **165**(1): 113-149.

[12] Goodfellow I, Pouget-Abadie J, Mirza M. *Generative Adversarial Nets* [M]. Cambridge: MIT Press, 2014.

[13] Qian Q, Zhu S, Tang J, et al. Robust optimization over multiple domains [C]//*Proceedings of the AAAI Conference on Artificial Intelligence*, 2019: 4739-4746.

[14] Lu S, Tsaknakis I, Hong M, et al. Hybrid block successive approximation for one-sided nonconvex min-max problems: algorithms and applications [J]. *IEEE Transactions on Signal Processing*, 2020, **68**: 3676-3691.

[15] Sinha A, Namkoong H, Duchi J. Certifiable distributional robustness with principled adversarial training [C]//*International Conference on Learning Representations*, 2018.

[16] Dai B, Shaw A, Li L. Sbeed: Convergent reinforcement learning with nonlinear function approximation [C]//*International Conference on Machine Learning*, PMLR, 2018: 1125-1134.

[17] Shafieezadeh-Abadeh S, Esfahani P, Kuhn D. Distributionally robust logistic regression [C]//*Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015: 1576-1584.

[18] Smale S. Mathematical problems for the next century [J]. *Mathematical Intelligencer*, 1998, **20**: 7-15.

[19] Hiriart-Urruty J. A new series of conjectures and open questions in optimization and matrix analysis [J]. *ESAIM: Control, Optimisation and Calculus of Variations*, 2009, **15**: 454-470.

[20] 10000个科学难题数学编委会. 10000个科学难题：数学卷 [M]. 北京: 科学出版社，2008.

[21] 胡晓东，袁亚湘，章祥荪. 运筹学发展的回顾与展望 [J]. 中国科学院院刊, 2012, **27**(2): 145-160.

[22] 王奇超，文再文，蓝光辉，等. 优化算法的复杂度分析 [J]. 中国科学：数学，2020, **50**(9): 144-209.

[23] Bubeck S. Convex optimization: Algorithms and complexity [J]. *Foundations and Trends in Machine Learning*, 2015, **8**(3/4): 231-357.

[24] Daskalakis C, Panageas I. The limit points of (optimistic) gradient descent in min-max optimization [C]//*Advances in Neural Information Processing Systems*, 2018: 9256-9266.

[25] Mazumdar E, Jordan M, Sastry S. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games [J]. 2019, arXiv: 1901.00838.

[26] Jin C, Netrapalli P, Jordan M. What is local optimality in nonconvex-nonconcave minimax optimization? [C]//*International Conference on Machine Learning*, PMLR, 2020: 4880-4889.

[27] Dai Y, Zhang L. Optimality conditions for constrained minimax optimization [J]. *CSIAM Transactions on Applied Mathematics*, 2020, **1**: 296-315.

[28] Xu Z, Zhang H, Xu Y, et al. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems [J]. 2020, arXiv: 2006.02032.

[29] Nouiehed M, Sanjabi M, Huang T, et al. Solving a class of nonconvex min-max games using iterative first order methods [J]. *Advances in Neural Information Processing Systems*, 2019: 14905-14916.

[30] Nemirovski A. A prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems [J]. *SIAM Journal on Optimization*, 2004, **15**(1): 229-251.

[31] Auslender A, Teboulle M. Interior projection-like methods for monotone variational inequalities [J]. *Mathematical programming*, 2005, **104**(1): 39-68.

[32] Monteiro R, Svaiter B. Complexity of variants of Tseng' s modified FB splitting and Korpelevich' s methods for generalized variational inequalities with applications to saddle point and convex optimization problems [J]. *SIAM Journal on Optimization*, 2010, **21**(4): 1688-1720.

[33] Tseng P. On accelerated proximal gradient methods for convex-concave optimization [J/OL]. *SIAM Journal on Optimization*, (2008-01-23)[2021-03-02]. https://www.csie.ntu.edu.tw/b97058/tseng/papers/apgm.pdf.

[34] Ouyang Y, Xu Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems [J]. *Mathematical Programming*, 2019, **185**: 1-35.

[35] Mertikopoulos P, Zenati H, Lecouat B, et al. Mirror descent in saddle-point problems: Going the extra (gradient) mile [J]. 2018, arXiv: 1807.02629.

[36] Rafique H, Liu M, Lin Q, et al. Non-convex min-max optimization: Provable algorithms and applications in machine learning [J]. 2018, arXiv: 1810.02060.

[37] Sanjabi M, Ba J, Razaviyayn M, et al. On the convergence and robustness of training gans with regularized optimal transport [C]//*Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018: 7091-7101.

[38] Sanjabi M, Razaviyayn M, Lee J. Solving non-convex non-concave min-max games under polyak-lojasiewicz condition [J]. 2018, arXiv: 1812.02878.

[39] Thekumparampil K, Jain P, Netrapalli P et al. Efficient algorithms for smooth minimax optimization [J]. 2019, arXiv: 1907.01543.

[40] Kong W, Monteiro R. An accelerated inexact proximal point method for solving nonconvex concave min-max problems [J]. 2019, arXiv: 1905.13433.

[41] Lin T, Jin C Jordan M. Near-optimal algorithms for minimax optimization [J]. 2020, arXiv: 2002.02417.

[42] Letcher A, Balduzzi D, Racaniere S, et al. Differentiable game mechanics [J]. *Journal of Machine Learning Research*, 2019, **20**(84): 1-40.

[43] Lin T, Jin C, Jordan M. On gradient descent ascent for nonconvex-concave minimax problems [J]. 2019, arXiv: 1906.00331.

[44] Jin C, Netrapalli P, Jordan M. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal [J]. 2019, arXiv: 1902.00618.

[45] Pan W, Shen J, Xu Z. An efficient algorithm for nonconvex-linear minimax optimization problem and its application in solving weighted maximin dispersion problem [J]. *Computational Optimization and Applications*, 2021, **78**(1): 287-306.

[46] 张慧灵, 徐洋, 徐姿. 分块凸-非凹极小极大问题的交替近端梯度算法 [J/OL]. 运筹学学报, (2021-03-15)[2021-02-26]. https://www.ort.shu.edu.cn/CN/abstract/abstract18248.shtml.

[47] Zhang J, Xiao P, Sun R, et al. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems [C]//*Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[48] Lin Q, Liu M, Rafique H, et al. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality [J]. 2018, arXiv: 1810.10207.

[49] Dang C, Lan G. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators [J]. *Computational Optimization and Applications*, 2015, **60**(2): 277-310.

[50] Yang J, Kiyavash N, He N. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems [C]//*Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[51] Menickelly M, Wild S. Derivative-free robust optimization by outer approximations [J]. *Mathematical Programming*, 2018, **179**(1): 1-37.

[52] Picheny V, Binois M, Habbal A. A bayesian optimization approach to find nash equilibria [J]. *Journal of Global Optimization*, 2019, **73**(1): 171-192.

[53] Roy A, Chen Y, Balasubramanian K, et al. Online and bandit algorithms for nonstationary stochastic saddle-point optimization [J]. 2019, arXiv: 1912.01698.

[54] Liu S, Lu S, Chen X, et al. Min-max optimization without gradients: Convergence and applications to adversarialml [C]//*International Conference on Machine Learning*, 2020, 6282-6293.

[55] Wang Z, Balasubramanian K, Ma S, et al. Zeroth-order algorithms for nonconvex minimax problems with improved complexities [J]. 2020, arXiv: 2001.07819.

[56] Xu T, Wang Z, Liang Y, et al. Enhanced first and zeroth order variance reduced algorithms for min-max optimization [J]. 2020, arXiv: 2006.09361.

[57] Huang F, Gao S, Pei J, et al. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization [J]. 2020, arXiv: 2008.08170.

[58] Luo L, Ye H, Zhang T. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems [C]//*Proceedings of the 33th International Conference on Neural Information Processing Systems*, 2020.

[59] Carmon Y, Duchi J, Hinder O, et al. Lower bounds for finding stationary points II: first-order methods [J]. *Mathematical Programming*, 2021, **185**: 315-355.