arXiv:2301.06619v3 [math.OC] 9 Jun 2023

弱凸损失下的分布鲁棒学习：收敛速率与有限样本保证

Landi Zhu LZ401@SCARLETMAIL.RUTGERS.EDU

Mert G¨urb ¨uzbalaban MG1366@RUTGERS.EDU

安杰伊·鲁什琴斯基 RUSZ@RUTGERS.EDU

管理科学与信息系统系，罗格斯商学院，美国皮斯卡塔韦。

摘要

我们研究一个分布鲁棒随机优化问题，并利用均值–半偏差风险度量将其表述为一个随机两层复合优化问题。在此背景下，我们考虑一种单时间尺度算法，其中包含两种内层函数值跟踪方法：对连续可微损失函数进行线性化跟踪，以及对弱凸损失函数进行SPIDER跟踪。我们采用 Moreau 包络的梯度范数作为驻点性的度量，并证明在这两种情况下，$\mathcal{O}(\varepsilon^{-3})$的样本复杂度都是可实现的，只是第二种情况下的常数更大。最后，我们通过一个鲁棒学习示例和一个弱凸非光滑回归示例，展示了该算法的性能。

1. 引言

我们研究如下形式的分布鲁棒学习问题

$$\min_{x\in X} \max_{\mathbb{Q}\in\mathscr{M}(\mathbb{P})} \mathbb{E}_{D\sim\mathbb{Q}}\left[\ell(x,D)\right], \tag{1}$$

其中，$\ell:\mathbb{R}_n\times\mathbb{R}^d\to\mathbb{R}$ 是预测器 x 在随机数据 D 上的损失函数，其分布经扰动后服从概率律 $\mathbb{Q}$；$\mathscr{M}(\mathbb{P})$ 是由概率测度构成的闭凸集（即歧义集），用于刻画对参考概率律 $\mathbb{P}$ 的扰动；而 $X\subset\mathbb{R}^n$ 是可行集。这类建模方式通过考虑输入分布在集合 $\mathscr{M}(\mathbb{P})$ 内变化时的最坏情形，使得我们能够利用数据训练预测模型，使其对输入数据分布 $\mathbb{P}$ 的扰动具有鲁棒性。这种最坏情形方法属于随机优化；近年来，它也逐渐应用于机器学习。这些应用包括但不限于逻辑回归的凸与非凸建模、深度学习，以及更广泛地，以风险最小化的方式进行数据鲁棒的预测模型监督学习（G¨urb¨uzbalaban 等，2022年；Zhang 等，2021年、2022年；Laguel 等，2022年；Mehrotra 和 Zhang，2014年；Kuhn 等，2019年）。这些应用面临的挑战在于，模型维度 n 和数据点数量可能很大，而且损失函数可能既非光滑又非凸。

歧义集 $\mathscr{M}(\mathbb{P})$ 用于刻画基准数据分布 $\mathbb{P}$ 的不确定性，其选取以及损失函数的结构会影响所得问题形式的计算可处理性，以及求解 (1) 的优化算法设计。$\mathscr{M}(\mathbb{P})$ 的各种可能选取包括以 $\mathbb{P}$ 为中心的瓦瑟斯坦球（Shafieezadeh Abadeh 等人，2015；Esfahani 和 Kuhn，2018；Mehrotra 和 Zhang，2014；Kuhn 等人，2019；Gao 等人，2017；Sinha 等人，2018）、

©2023 Landi Zhu、Mert G¨urb¨uzbalaban 和 Andrzej Ruszczy´nski。许可协议：CC-BY 4.0，详见 https://creativecommons.org/licenses/by/4.0/。

基于 f-散度的不确定集（Bagnell，2005年；Duchi 和 Namkoong，2021年；Namkoong 和 Duchi，2016年；Zhang 等人，2021年），以及基于风险度量的方法，例如条件风险价值（Takeda 和 Kanamori，2009年；Laguel 等人，2022年）和均值–半偏差风险（G¨urb¨uzbalaban 等人，2022年）。在这一背景下，一阶优化算法的性能可根据其收敛速率保证和样本复杂度来评估；样本复杂度即为找到近似一阶驻点解所需采样的数据点数量。

我们首先介绍不确定集 $\mathscr{M}(\mathbb{P})$ 的基于均值–半偏差的建模，然后总结我们的贡献。对于每个模型参数 $x \in X$，我们考虑定义在配备 σ-代数 $\mathscr{F}$ 的样本空间 Ω 上的随机损失 $Z_x = \ell(x, D)$。我们假设期望值 $\mathbb{E}(Z_x) = \int_\Omega Z_x(\omega)\ \mathbb{P}(d\omega)$ 有限；即 $Z_x \in \mathscr{L}_1(\Omega, \mathscr{F}, \mathbb{P})$。我们可以通过如下定义的均值–半偏差风险度量来评估其质量：

$$\rho[Z_x] = \mathbb{E}[Z_x] + \varkappa \mathbb{E}\Big[\max\big(0, Z_x - \mathbb{E}[Z_x]\big)\Big], \quad \varkappa \in [0,1], \tag{2}$$

该度量会对随机损失中超出期望值的大额损失进行惩罚。这种风险度量是一种相干风险度量，具有若干理想性质，并广泛应用于统计学和随机优化的许多领域（Rockafellar 等，2006年；G¨urb¨uzbalaban 等，2022年）。众所周知，随机变量Z_x 的均值–半偏差风险 $\rho[Z_x]$ 具有如下对偶表示（Ruszczy´nski 和 Shapiro，2006年）：

$$\begin{aligned} \rho[Z_x] &= \max_{\mu \in \mathscr{A}} \int_\Omega Z_x(\omega)\mu(\omega)\ \mathbb{P}(d\omega) \\ &= \max_{\mathbb{Q}\,:\,\frac{d\mathbb{Q}}{d\mathbb{P}} \in \mathscr{A}} \int_\Omega Z_x(\omega)\ \mathbb{Q}(d\omega) = \max_{\mathbb{Q}\,:\,\frac{d\mathbb{Q}}{d\mathbb{P}} \in \mathscr{A}} \mathbb{E}_{\mathbb{Q}}[Z_x], \end{aligned} \tag{3}$$

其中

$$\mathscr{A} = \big\{\mu = \mathbb{1} + \xi - \mathbb{E}[\xi] :\ \xi \in \mathscr{L}_\infty(\Omega, \mathscr{F}, \mathbb{P}),\ \|\xi\|_\infty \le \varkappa,\ \xi \ge 0\big\}$$

是一个凸且闭集。因此，由（2）和（3）可知，具有歧义集的极小极大形式（1） et

$$\mathscr{M}(\mathbb{P}) = \Big\{\mathbb{Q} : \frac{d\mathbb{Q}}{d\mathbb{P}} \in \mathscr{A}\Big\} \tag{4}$$

等价于

$$\min_{x \in X} F(x) = \min_{x \in X} \rho[\ell(x, D)] = \min_{x \in X}\ f(x, h(x)) \tag{5}$$

其中函数为

$$\begin{aligned} f(x,u) &= \mathbb{E}\Big[u + \varkappa \max\big(0, \ell(x,D) - u\big)\Big], \\ h(x) &= \mathbb{E}[\ell(x,D)]. \end{aligned} \tag{6}$$

这样，利用均值-半偏差风险度量，可以将极小极大问题转化为一个两层随机优化问题，从而实现隐式鲁棒形式，其鲁棒性水平由参数 $\varkappa$ 控制：当 $\varkappa = 0$ 时，不确定集 (4) 仅包含原始概率测度 $\mathbb{P}$；而当 $\varkappa > 0$ 时，测度 $\mathbb{Q} \in \mathscr{M}(\mathbb{P})$ 是 $\mathbb{P}$ 的扭曲。允许的相对扭曲范围$\frac{d\mathbb{Q}}{d\mathbb{P}} - \mathbb{1}$由 $\varkappa$ 控制。挑战在于，这种形式是非光滑的；当损失函数为非凸时，它通常也是非凸的。此外，

在非线性函数$f(x\ \cdot)$内部使用期望值，会导致随机次梯度估计产生偏差。

贡献。本文基于重构形式（5），提出了一种名为随机复合次梯度（SCS）的优化算法，用于求解（1），并建立了有限时间收敛分析结果和样本复杂度结果。我们的算法是 G¨urb¨uzbalaban 等人（2022年）提出的单时间尺度方法的一种变体和简化版本，其中次梯度不再取平均，也不再需要次梯度跟踪。在第2节和第3节中，我们假设损失函数连续可微，并以投影次梯度下降框架为基础，采用线性化跟踪，并将 Moreau 包络的梯度用作收敛度量。我们证明，在这种情况下，SCS 方法的样本复杂度为 $\mathcal{O}(\varepsilon^{-3})$ 量级。在第4节中，我们假设损失函数是 δ-弱凸的（见定义（7）），并使用 Fang 等人（2018年）提出的 SPIDER 估计器来估计损失函数的期望。值得强调的是，在这种情况下，（5）中的 $f(\cdot,\cdot)$ 和 $h(\cdot)$ 都是非光滑的，且 $h(\cdot)$ 是非凸的。我们证明，尽管常数较大，SCS 方法的样本复杂度仍具有相同的 $\mathcal{O}(\varepsilon^{-3})$ 量级。

相关工作 （双层）随机复合优化问题 (5) 的凸性与非光滑结构由损失函数的选择决定。当损失函数为凸且可能非光滑时，复合目标函数 $F(x)$ 关于 x 也将是凸且非光滑的。在这种情况下，Wang 等人（2017a）提出的多层凸随机优化算法等方法均可适用；当损失函数为凸时，样本复杂度为 $\mathcal{O}(\varepsilon^{-4})$，当损失函数为强凸时，样本复杂度为 $\mathcal{O}(\varepsilon^{-1.5})$。当损失函数为非凸时，无论其是否光滑，复合目标函数 $F(x)$ 关于 x 都将是非凸且非光滑的。目前尚无这一一般设定下的收敛速率结果，但若仅考虑光滑问题，则已有一些复杂度结果。Wang 等人（2017a）分析了在目标函数满足不同假设时的随机梯度算法，并分别证明了光滑凸问题和光滑强凸问题的样本复杂度为 $\mathcal{O}(\varepsilon^{-3.5})$ 和 $\mathcal{O}(\varepsilon^{-1.25})$。通过适当的正则化，这些速率还可进一步提升（Wang 等人，2017b）。Ghadimi 等人（2020）针对光滑非凸复合优化问题提出了一种单时间尺度的嵌套平均随机逼近（NASA）方法，并证明其样本复杂度为 $\mathcal{O}(\varepsilon^{-2})$。对于更高层级（多于两层）的问题，Ruszczynski（2021）通过分析一个微分包含系统，证明了随机次梯度法的渐近收敛性；在假设光滑性的情况下，其样本复杂度为 $\mathcal{O}(\varepsilon^{-2})$。Balasubramanian 等人（2022）在不要求有界性假设的情况下，为光滑多层问题得到了另一种与层级无关的速率 $\mathcal{O}(\varepsilon^{-2})$。

此外，还有一些方法处理（1）中并非基于复合随机优化的情形。特别是，Ho-Nguyen 和 Wright（2022年）研究了受 Wasserstein歧义集约束、针对“零一损失函数”的线性分类问题；该损失函数是非凸且非光滑的。作者证明，该问题等价于最小化正则化斜坡损失目标，并提出了一类斜坡损失的光滑近似形式，可利用标准连续优化算法求解相应的光滑问题（近似求解亦可）。此外，还有其他方法能够在损失函数为光滑或凸时给出复杂度结果。

光滑损失函数 Sinha等人（2018）将 $\mathscr{M}(\mathbb{P})$ 定义为概率律 $\mathbb{P}$ 在 Wasserstein 度量下的 ρ 邻域。他们证明，对于光滑损失函数且鲁棒性水平 ρ 足够小时，随机梯度下降（SGD）方法能够达到与标准光滑非凸优化相同的收敛速率。在 Jin等人（2021）的研究中，作者考虑了光滑

和 Lipschitz 非凸损失函数，并使用基于f-散度的软惩罚项来建模分布偏移。他们分析了带动量的小批量归一化随机梯度下降，并证明其样本复杂度为 $\mathscr{O}(\varepsilon^{-4})$（使损失函数梯度的范数不超过 ε），这也与标准光滑非凸优化中可达到的速率相匹配。对于 CVaR 的平滑版本，作者针对 Lipschitz 光滑非凸损失函数获得了类似的收敛保证。在 Soma 和 Yoshida（2020）中，作者提出了条件风险价值（CVaR）公式。他们证明，对于凸、Lipschitz 且光滑的损失函数，其基于随机梯度下降的算法复杂度为 $\mathscr{O}(1/\varepsilon^2)$；而对于非凸、光滑且 Lipschitz 的损失函数，作者得到的复杂度为 $\mathscr{O}(1/\varepsilon^6)$。在 Curi 等（2020）的研究中，作者提出了一种自适应采样算法，用于随机优化损失函数经验分布的 CVaR，并基于 CVaR 的对偶表示将该优化问题重新表述为双人博弈。对于凸问题，他们在 $\mathscr{O}(T)$ 次T 迭代中得到遗憾界；对于非凸问题，在能够使用不精确的经验风险最小化（ERM）预言机的假设下，他们得到遗憾界 $\mathscr{O}(T)$。然而，对于非凸问题，实现该预言机需要在每次迭代中求解加权经验损失最小化问题，而这通常是 NP 困难的（Curi 等，2020，第 4.2 节）。

当损失函数在原始空间 X 上光滑且利普希茨连续时，对式 (1) 进行基于样本的近似，即用数据点上的有限平均来近似期望，会得到光滑的非凸/仅凹的极小极大优化问题；若对偶空间$\mathscr{M}(P)$由 f-散度确定，则该空间是有限维的（Zhang et al., 2022）。在这种情况下，可以使用原始-对偶算法，而（Zhang et al., 2022）中的 SAPD+ 算法具有 $\mathscr{O}(1/\varepsilon^6)$ 复杂度。此外，还有原始方法。具体而言，Qi 等（2021）提出了一种原始方法，用于求解一类具有光滑非凸目标函数、采用基于样本近似形式的分布鲁棒优化（DRO）问题。他们考虑对偶变量上的 KL 散度正则化，并将 DRO 问题转化为光滑的两层复合问题。对于光滑损失函数，作者证明，计算近似解 w_ε（即满足 $\|\nabla F_P(w_\varepsilon)\|^2 \le \varepsilon$ 的解）具有 $\tilde{\mathscr{O}}(1/\varepsilon^{3/2})$ 复杂度，其中 F_P 是原始函数。在附加满足 PL 条件时，若以原始函数的次优性作为性能指标，作者还针对光滑损失函数推导出了改进的 $(\mathscr{O})(\frac{1}{\varepsilon})$ 复杂度。最近，Qi 等（2022a）研究了光滑损失函数下的 KL 约束 DRO，并提出了可扩展复合算法，该算法在每次迭代中都可使用固定的批量大小。基于样本的 DRO 形式在机器学习中的各类问题上也得到了研究，例如处理不平衡数据（Qi et al., 2022b）和部分 AUC 最大化（Zhu et al., 2022）。我们的方法不需要基于样本的近似，能够处理 $\mathscr{M}(\mathbb{P})$ 可能为无限维的一般情形；此外，我们还针对弱凸损失给出了复杂度结果。这类损失不仅是非凸的，同时也可能是非光滑的。

凸损失若将分布鲁棒问题 (1) 表述为有限维凸规划（Shafieezadeh Abadeh 等人，2015年；Esfahani 和 Kuhn，2018年；Mehrotra 和 Zhang，2014年；Kuhn 等人，2019年），则可在多项式时间内求解。当 $\mathscr{M}(\mathbb{P})$ 通过 f-散度定义，且损失函数为凸且光滑时，可使用赌博机镜像下降算法（Namkoong 和 Duchi，2016年）求解 (1) 的基于样本的近似，其迭代次数与随机梯度下降的迭代次数相当。对于相同表述下的凸损失，也可以使用锥内点求解器或采用回溯 Armijo 线搜索的梯度下降法（Duchi 和 Namkoong，2021年），但当维度或样本数量较大时，这些方法在某些应用中计算开销可能很大。

当不确定集 $\mathscr{M}(\mathbb{P})$ 基于数据的经验分布，并通过 χ^2-散度或条件风险价值定义，且损失函数为凸且 Lipschitz 连续时，Levy 等人（2020）提出了能够达到最优 $\mathscr{O}(\varepsilon^{-2})$ 速率的算法，该速率与训练数据集的大小和参数数量无关。

当 $\ell(\cdot, D)$ 非凸且不可微时，分布鲁棒随机优化问题会导出非凸非光滑极小极大优化问题。据我们所知，在这一一般情形下，现有算法均无法证明其能够收敛到 (1) 的一个驻点，也无法给出迭代复杂度界。我们的结果适用于这一情形，并为可能非光滑的弱凸损失函数提供迭代复杂度估计。

符号与预备知识。函数 $q:\mathbb{R}^n\to\mathbb{R}$ 称为 δ-弱凸函数，如果正则化函数 $x\mapsto q(x)+\frac{\delta}{2}\|x\|^2$ 是凸的（Nurminskii，1973）。这是一类范围广泛的函数，可能非光滑且非凸，其中包括所有凸函数，以及具有全局利普希茨连续梯度的光滑函数。δ-弱凸函数 $q(x)$ 还具有以下性质：在每个点 $x\in\mathbb{R}^n$，都存在一个向量 $g\in\mathbb{R}_n$，使得

$$q(y)\geq q(x)+\langle g,y-x\rangle-\frac{\delta}{2}\|y-x\|^2,\qquad \forall y\in\mathbb{R}^n, \tag{7}$$

（例如，参见 Davis 等人（2018b））。满足上述关系的向量集合 $\partial q(x)$ 是次微分 $q(\cdot)$ 在 x 处的次微分；它非空、凸且闭。事实上，对于这一类函数，它与 Clarke 次微分一致（Rockafellar 和 Wets，2009年）。若连续可微函数 $q:\mathbb{R}^n\to\mathbb{R}$ 在凸集 X 上满足 $\nabla q(x)$ 在 X 上利普希茨连续，即对所有 x、$y\in X$，都有 $\|\nabla q(x)-\nabla q(y)\|\leq L\|x-y\|$，则称其在该凸集上是$L$-光滑的。

2. 随机复合次梯度（SCS）方法

我们首先考虑损失函数连续可微的情形，非光滑情形将在第4节中讨论。

假设1 集合 $X\subset\mathbb{R}^n$ 是凸且紧的。

假设 2 对于集合 X 的某个邻域中的所有 x：
(i) 函数 $\ell(x,\cdot)$ 是可积的；(ii) 函数 $\ell(\cdot,D)$ 连续可微，并且存在可积常数 $\tilde{\Delta}_h(D)$ 和 $\tilde{\delta}(D)$，使得

$$\|\nabla\ell(x,D)\|\leq\tilde{\Delta}_h(D),\quad \forall D\in\mathbb{R}^d,$$

and

$$\|\nabla\ell(x,D)-\nabla\ell(y,D)\|\leq\tilde{\delta}(D)\|x-y\|,\quad \forall x,y\in X,\quad \forall D\in\mathbb{R}^d.$$

备注1 由于损失函数 $\ell(x, D)$ 是 $\delta(D)$-光滑的，且可行集 X 是紧的，因此 $\ell(x, D)$ 也是 $\delta(D)$-弱凸的（弱凸性的定义见第*1*节最后一段）。假设*1*和假设 *2* 保证期望值函数 $h(\cdot)$ 定义良好，且在 X 上是 $\tilde{\delta}$-光滑、δ-弱凸的，其中 $\delta=\mathbb{E}[\delta(D)]$。

对于统计学习中的许多问题，假设1和2均成立，包括各种分类和回归任务的非凸约束形式，例如深度学习、最小二乘和逻辑回归（G¨urb¨uzbalaban等，2022年；Negiar等，2020年）。

如果已知$f(\cdot)$的一个次梯度以及$h(\cdot)$的梯度，就可以通过应用链式法则（Rockafellar 和 Wets，2009年）计算复合函数$F(x) = f(x,h(x))$的一个次梯度；也就是说，若 $\begin{bmatrix} g_{fx} & g_{fu} \end{bmatrix} \in \partial f(x,u)$且$g_h = \nabla h(x)$，那么我们就有

$$g_{fx} + g_{fu} g_h \in \partial F(x). \tag{8}$$

遗憾的是，在我们的设定中，我们既无法获得式（8）中的次梯度，也无法获得 h(x) 的值；我们只能得到它们的随机估计。为了解决这一问题，我们提出的方法——随机复合次梯度（SCS）——基于投影随机次梯度更新规则生成近似解 $\{\mathrm{x}^k\}_{\mathrm{k}=1,2,\ldots} \in \mathrm{Rn}$。该规则通过利用 f 和 h 的随机次梯度来估计复合函数 F(x) 的次梯度，并将迭代点投影回约束集 X，以确保迭代点保持有界。算法 1 中描述的方法还会生成随机内层函数估计 $\{\mathrm{u}^k\}_{\mathrm{k}=1,2,\ldots} \in \mathrm{R}$；在此，我们假设可以获得 f 和 h 的次梯度以及 h 的值的无偏随机估计，且其方差有界。更准确地说，记$\mathscr{F}_k$为由 $\{\mathrm{x}^0, \mathrm{u}^0, \mathrm{x}^1, \mathrm{u}^1, \ldots, \mathrm{x}^k, \mathrm{u}^k\}$生成的 σ-代数，其中 $\mathrm{x}^0 \in \mathrm{X}$, $\mathrm{u}^0 \in \mathbb{R}$ 为初始值，我们作出如下假设。

假设 3 对于所有 k，我们都可以访问满足以下条件的随机向量 $\tilde{g}_f^k$、 $\tilde{g}_h^k$、 $\tilde{J}^k$ 和随机变量 $\tilde{h}^k$：

- $\tilde{g}_{kf} = g_{kf} + e_{kf}$，$g_{kf} \in \partial f(x_k, u_k)$, $\mathbb{E}\left[e_{kf} \,|\, \mathscr{F}_k\right] = 0$, $\mathbb{E}\left[\|e_{kf}\|^2 \,|\, \mathscr{F}_k\right] \le \sigma^2$;
- $\tilde{g}_{kh} = g_{kh} + e_{kh}$, $g_{kh} = \nabla h(x_k)$, $\mathbb{E}\left[e_{kh} \,|\, \mathscr{F}_k\right] = 0$, $\mathbb{E}\left[\|e_{kh}\|^2 \,|\, \mathscr{F}_k\right] \le \sigma^2$;
- $\tilde{h}_k = h(x_k) + e_{k\ell}$, $\mathbb{E}\left[e_{k\ell} \,|\, \mathscr{F}_k\right] = 0$, $\mathbb{E}\left[\|e_{k\ell}\|^2 \,|\, \mathscr{F}_k\right] \le \sigma^2$;
- $\tilde{J}_k = g_{kh} + E_k$, $\mathbb{E}\{E_k \,|\, \mathscr{F}_k\} = 0$, $\mathbb{E}\{\|E_k\|^2 \,|\, \mathscr{F}_k\} \le \sigma^2$;

其中 σ 为常数，并且在给定 $\mathscr{F}_k$的条件下，误差 e_f^k、e_h^k、e_ℓ^k 和 E^k 条件独立。

注 2 在假设 *1* 和 *2* 下，f 和 h 的所有取值及次梯度：g_f^k、g_h^k 和 $h(x^k)$，都是有界的。对所有 $x \in X$、$u \in \mathbb{R}$，记作

$$\|\partial_x f(x,u)\| \le \Delta_{fx}, \qquad \|\nabla h(x)\| \le \Delta_h,$$

其中 $\Delta_h = \mathbb{E}[\tilde{\Delta}_h(D)]$。在假设 *3* 下，将 (8) 中的真实（次）梯度替换为其随机估计后，得到的 $\partial F(x^k)$中一个元素的随机估计具有有界的平方范数期望：

$$\mathbb{E}\left[\|\tilde{g}_{fx}^k + \tilde{g}_{fu}^k \tilde{g}_h^k\|^2 \,\middle|\, \mathscr{F}_k\right] \le M^2 \quad with \quad M^2 = (\Delta_{fx} + \Delta_h)^2 + 2\sigma^2 + 2\sigma\Delta_h, \tag{9}$$

其中我们使用了随机误差的条件独立性、柯西-施瓦茨不等式，以及由(12)可推出 $\tilde{g}_{fu}^k \in [0,1]$这一事实。

统计学习和随机优化中的一种常见设定是，根据有放回随机抽取的数据点子集估计次梯度（Bottou，2010年）。在此设定中，当定义域 X 无界时，这类随机次梯度估计量的方差可能无界（G¨urb¨uzbalaban 等人，2021年；Jain 等人，2018年；G¨urb¨uzbalaban 等人，2022年）。然而，在我们的设定中，可行集X 是紧集，因此假设 3 自然成立（见第 5.3 节）。

算法 1SCS 方法

输入： 初始点 $x^0 \in X$，$u^0 \in \mathbb{R}$，恒定步长 $\tau \in (0,1]$。1：**对于** $k = 0,1,...,N-1$ **执行**2：在 $\mathscr{F}_k$ 条件下独立抽取D_1^{k+1}、D_2^{k+1}和D_3^{k+1}，并计算估计量

$$G^k \in \partial_x \ell(x^k, D_1^{k+1}), \tag{10}$$

$$\tilde{g}_{fx}^k = \begin{cases} 0 & \text{if } \ell(x^k, D_1^{k+1}) < u^k, \\ \varkappa G^k & \text{if } \ell(x^k, D_1^{k+1}) \geq u^k, \end{cases} \tag{11}$$

$$\tilde{g}_{fu}^k = \begin{cases} 1 & \text{if } \ell(x^k, D_1^{k+1}) < u^k, \\ 1-\varkappa & \text{if } \ell(x^k, D_1^{k+1}) \geq u^k, \end{cases} \tag{12}$$

$$\tilde{g}_h^k \in \partial_x \ell(x^k, D_2^{k+1}), \tag{13}$$

$$\vec{J}^k \in \partial_x \ell(x^k, D_3^{k+1}), \tag{14}$$

$$\tilde{h}^k = \frac{1}{3}(\ell(x^k, D_1^{k+1}) + \ell(x^k, D_2^{k+1}) + \ell(x^k, D_3^{k+1})). \tag{15}$$

3：更新解估计

$$x^{k+1} = \Pi_X \left(x^k - \tau \left(\tilde{g}_{fx}^k + \tilde{g}_{fu}^k \tilde{g}_h^k \right)^T \right),$$

4：更新内函数估计

$$u^{k+1} = u^k + \tau \left(\tilde{h}^k - u^k \right) + \vec{J}^k \left(x^{k+1} - x^k \right). \tag{16}$$

5：**结束循环**

输出： x^R其中 R从 $\{0,1,\ldots,N-1\}$中均匀采样。

3. 连续可微损失函数的收敛速率

由于均值-半偏差风险度量的表达式包含非光滑的 max(·) 项，即使损失函数 $\ell(x, D)$ 连续可微，外层函数 (6) 也可能是非光滑的。因此，问题 (5) 面临的挑战在于其非光滑且非凸。不过，我们将论证它是弱凸的。对于弱凸目标函数，衡量一阶驻点性的一种常用指标是 Moreau 包络梯度的范数（Moreau，1965），我们将在下文介绍这一指标。我们首先考虑主问题 (5) 的另一种表述：

$$\min_{x\in\mathbb{R}^n} \varphi(x) := F(x) + r(x),$$

其中 $F(x) = f(x,h(x))$，而 $r(x)$ 是凸紧可行集 $X \subset \mathbb{R}_n$ 的示性函数，即 $r(x) = 0$ 当 $x \in X$ 时成立，而 $r(x) = +\infty$ 在其他情况下成立。Moreau 包络和近端映射分别定义为

$$\varphi_\lambda(x) := \min_y\{\varphi(y) + \frac{1}{2\lambda}\|y-x\|^2\},$$
$$\mathrm{prox}_{\lambda\varphi}(x) := \operatorname*{argmin}_y\{\varphi(y) + \frac{1}{2\lambda}\|y-x\|^2\},$$

。由于内层函数 $h(x)$ 是 δ-弱凸的，且外层函数 $f(x,u)$ 关于 x 是弱凸的，关于 u 是凸且非递减的（见备注1），复合函数 $F(x)$ 也具有 ρ-弱凸性，其中 $\rho = (1+2\varkappa)\delta$。在这种情况下，$\varphi_\lambda(x)$ 在 $\lambda \in (0, \rho^{-1})$ 上连续可微（Moreau，1965年），其梯度为

$$\nabla\varphi_\lambda(x) = \lambda^{-1}(x - \mathrm{prox}_{\lambda\varphi}(x)). \tag{17}$$

还可以证明，量 $\|\nabla\varphi_\lambda(x)\|$ 是驻点性的度量，也就是说，当 $\|\nabla\varphi_\lambda(x)\|$ 很小时，x 会接近某个近似驻点 $\hat{x}$，而该点的次微分也会接近 0（Davis 和 Drusvyatskiy，2019），即 $\hat{x}$ 满足以下关系：

$$\begin{cases} \|\hat{x} - x\| = \lambda\|\nabla\varphi_\lambda(x)\|, \\ \varphi(\hat{x}) \le \varphi(x), \\ \mathrm{dist}(0; \partial\varphi(\hat{x})) \le \|\nabla\varphi_\lambda(x)\|. \end{cases}$$

这里，$\mathrm{dist}(0;\partial\varphi(\hat{x}))$ 表示原点到集合 $\partial\varphi(\hat{x})$ 的距离。因此，本文关于 Moreau 包络的梯度的收敛保证，可以转化为以次微分表示的保证。

现在，我们可以继续证明 SCS 方法在损失函数连续可微时的收敛速率。首先，我们量化内函数估计 $\{u^k\}$ 对序列 $\{h(x^k)\}$ 的跟踪效果。

引理3 若假设*1*、*2*和*3*成立，则由算法*1*生成的序列 $\{u^k\}$ 满足：

$$\mathbb{E}\left[|u^k - h(x^k)|\right] \le \sigma(1+M)\tau^{1/2} + \delta M\tau + (1-\tau)^k|u^0 - h(x^0)|,$$
$$k = 0, 1, \ldots, N-1. \tag{18}$$

证明 根据假设 1 和 2，内层函数 $h(x)$ 是 δ-光滑的。因此，

$$h(x^{k+1}) = h(x^k) + \left[g_h^k\right]^T (x^{k+1} - x^k) + A_k, \quad \|A_k\| \le \delta \|x^{k+1} - x^k\|^2.$$

由 $\{u^k\}$的更新规则 (16)，可得

$$u^{k+1} = u^k + \tau\big(h(x^k) - u^k\big) + \tau e_\ell^k + \left[g_h^k\right]^T \big(x^{k+1} - x^k\big) + E^k(x^{k+1} - x^k).$$

因此，

$$u^{k+1} - h(x^{k+1}) = (1-\tau)\big[u^k - h(x^k)\big] + \tau e_\ell^k + E^k(x^{k+1} - x^k) - A_k.$$

递归地使用此等式，我们得到

$$u^{k+1} - h(x^{k+1}) = \sum_{j=0}^{k} (1-\tau)^{k-j}\big(\tau e_\ell^j + E^j(x^{j+1} - x^j) - A_j\big) + (1-\tau)^{k+1}(u^0 - h(x^0)). \quad (19)$$

鞅项的范数很容易得到界：

$$\mathbb{E}\left[\Big(\sum_{j=0}^{k} (1-\tau)^{k-j}\tau e_\ell^j\Big)^2\right] \le \sum_{j=0}^{k} (1-\tau)^{2(k-j)}\sigma^2\tau^2 \le \frac{\sigma^2\tau^2}{1-(1-\tau)^2} \le \sigma^2\tau,$$

从而

$$\mathbb{E}\left[\Big|\sum_{j=0}^{k} (1-\tau)^{k-j}\tau e_\ell^j\Big|\right] \le \sigma\tau^{1/2}. \quad (20)$$

注意，由E^k与$x^{k+1}-x^k$的条件独立性可知，$\mathbb{E}\big[E^k(x^{k+1}-x^k)\big|\mathscr{F}_k\big]=0$。此外，根据柯西–施瓦茨不等式以及投影算子的非扩张性

$$\mathbb{E}\left[\big|E^k(x^{k+1} - x^k)\big|^2 \,\Big|\, \mathscr{F}_k\right] \le \sigma^2 M^2 \tau^2,$$

其中M 是式 (9) 中的常数。因此，仿照式 (20)，我们得到

$$\mathbb{E}\left[\Big|\sum_{j=0}^{k} (1-\tau)^{k-j} E^j(x^{j+1} - x^j)\Big|\right] \le \sigma M \tau^{1/2}. \quad (21)$$

第三个求和项可以直接进行界定：

$$\mathbb{E}\left[\Big|\sum_{j=0}^{k} (1-\tau)^{k-j} A_j\Big|\right] \le \delta M \tau^2 \mathbb{E}\left[\Big|\sum_{j=0}^{k} (1-\tau)^{k-j}\Big|\right] \le \delta M \tau. \quad (22)$$

将估计式 (20)–(22) 代入式 (19)，我们得出

$$\mathbb{E}\big|u^{k+1} - h(x^{k+1})\big| \le \sigma(1+M)\tau^{1/2} + \delta M\tau + (1-\tau)^{k+1}|u^0 - h(x^0)|,$$

正如所要求的。 ■

现在考虑 Moreau 包络 $\varphi_{1/\bar{\rho}}(x)$，其中 $\bar{\rho}=\rho+(1+\varkappa)\delta$，并得到其梯度在一个迭代点 x^R 处的期望平方范数上界；该迭代点从前 N 个迭代点中随机选取。无约束随机弱凸最小化的复杂度结果已见于文献（Davis 等人，2018年a），但这些结果并不能直接应用于我们的设定，因为我们面对的是关于x的双层非光滑弱凸问题。我们的证明利用$f(x,u)$关于u 的单调性来处理其非光滑性，同时利用损失函数关于x的弱凸性。

定理 4 假设假设 *1* 至 *3* 成立。对于任意给定的迭代预算 N，考虑算法 *1* 的轨迹 $\{x^k\}_{k=0}^{N-1}$。我们有

$$\mathbb{E}[\|\nabla\varphi_{1/\bar{\rho}}(x^R)\|^2]\le 2\frac{\varphi_{1/\bar{\rho}}(x^0)-\min_{x\in X}F(x)+2\bar{\rho}|u^0-h(x^0)|+NC_3\tau^{3/2}}{N\tau},$$

其中 $\bar{\rho}=\rho+(1+\varkappa)\delta$，$C_3=2\bar{\rho}\sigma(1+M)$，期望是相对于算法 *1* 生成的轨迹以及随机变量 R 计算的，R 从 $\{0,1,...,N-1\}$中独立于该轨迹均匀抽样。

证明 定义 $\hat{x}^k:=$近端算子$_{\varphi/\bar{\rho}}(x^k)$，我们有

$$\begin{aligned}f(\hat{x}^k,h(\hat{x}^k))-f(x^k,u^k)&=h(\hat{x}^k)-u^k\\&\quad+\varkappa\mathbb{E}\big[\max(0,\ell(\hat{x}^k,D_1^{k+1})-h(\hat{x}^k))-\max(0,\ell(x^k,D_1^{k+1})-u^k)\big|\mathscr{F}_k\big].\end{aligned}\tag{23}$$

另一方面，根据算法 1，若记

$$I^k=\begin{cases}1 & \text{if}\quad \ell(x^k,D_1^{k+1})\ge u^k,\\ 0 & \text{if}\quad \ell(x^k,D_1^{k+1})<u^k,\end{cases}$$

我们可以写成

$$\tilde{g}^k_{fx}=\varkappa I^kG^k\quad\text{and}\quad \tilde{g}^k_{fu}=1-\varkappa I^k,$$

其中我们用到了定义 (10)、(11) 和 (12)。我们还可以估计式 (23) 中“最大值”项的差为

$$\begin{aligned}&\max(0,\ell(\hat{x}^k,D_1^{k+1})-h(\hat{x}^k))-\max(0,\ell(x^k,D_1^{k+1})-u^k)\\&\quad\ge\begin{cases}\ell(\hat{x}^k,D_1^{k+1})-h(\hat{x}^k)-(\ell(x^k,D_1^{k+1})-u^k), & \text{if } \ell(x^k,D_1^{k+1})\ge u^k,\\ 0, & \text{if } \ell(x^k,D_1^{k+1})<u^k,\end{cases}\\&\quad=I^k\big(\ell(\hat{x}^k,D_1^{k+1})-\ell(x^k,D_1^{k+1})-(h(\hat{x}^k)-u^k)\big)\\&\quad\ge I^k\big(\langle G^k,\hat{x}^k-x^k\rangle-\frac{\tilde{\delta}(D_1^{k+1})}{2}\|\hat{x}^k-x^k\|^2-(h(\hat{x}^k)-u^k)\big),\end{aligned}\tag{24}$$

其中我们使用了G^k 的定义（10），以及由于 ℓ的弱凸性，使用了满足$q(x)=\ell(x,D_1^{k+1})$的不等式（7）。记

$$A^k=\varkappa\mathbb{E}\big[I^kG^k\big|\mathscr{F}_k\big],$$

并利用（23）和（24），我们得到以下下界

$$f(\hat{x}^k,h(\hat{x}^k))-f(x^k,u^k)\geq\langle\hat{x}^k-x^k,A^k\rangle-\frac{\varkappa\delta}{2}\|\hat{x}^k-x^k\|^2 \\ +(1-\varkappa\mathbb{E}\left[I^k\middle|\mathscr{F}_k\right])(h(\hat{x}^k)-u^k). \tag{25}$$

记 $B^k=(1-\varkappa\mathbb{E}\left[I^k\middle|\mathscr{F}_k\right])g_h^k$，我们可以将最后一项估计为

$$\begin{aligned}&(1-\varkappa\mathbb{E}\left[I^k\middle|\mathscr{F}_k\right])(h(\hat{x}^k)-u^k)\\&\geq(1-\varkappa\mathbb{E}\left[I^k\middle|\mathscr{F}_k\right])(h(x^k)-u^k+\langle\hat{x}^k-x^k,g_h^k\rangle-\frac{\delta}{2}\|\hat{x}^k-x^k\|^2)\\&\geq\langle\hat{x}^k-x^k,B^k\rangle-\frac{\delta}{2}\|\hat{x}^k-x^k\|^2+(1-\varkappa\mathbb{E}\left[I^k\middle|\mathscr{F}_k\right])(h(x^k)-u^k).\end{aligned} \tag{26}$$

结合式（25）和式（26），可得

$$f(\hat{x}^k,h(\hat{x}^k))-f(x^k,u^k)\geq\langle\hat{x}^k-x^k,A^k+B^k\rangle-\frac{(1+\varkappa)\delta}{2}\|\hat{x}^k-x^k\|^2 \\ +(1-\varkappa\mathbb{E}\left[I^k\middle|\mathscr{F}_k\right])(h(x^k)-u^k). \tag{27}$$

现在我们考虑 Moreau 包络的变化：

$$\begin{aligned}&\mathbb{E}\left[\varphi_{1/\bar{\rho}}(x^{k+1})\,\middle|\,\mathscr{F}_k\right]\leq\mathbb{E}\left[F(\hat{x}^k)+\frac{\bar{\rho}}{2}\|\hat{x}^k-x^{k+1}\|^2\,\middle|\,\mathscr{F}_k\right]\\&=F(\hat{x}^k)+\frac{\bar{\rho}}{2}\mathbb{E}\left[\left\|\Pi_X\left(x^k-\tau\left(\tilde{g}_{fx}^k+\tilde{g}_{fu}^k\tilde{g}_h^k\right)^T\right)-\Pi_X(\hat{x}^k)\right\|^2\,\middle|\,\mathscr{F}_k\right]\\&\leq F(\hat{x}^k)+\frac{\bar{\rho}}{2}\mathbb{E}\left[\left\|x^k-\hat{x}^k-\tau(\tilde{g}_{fx}^k+\tilde{g}_{fu}^k\tilde{g}_h^k)^T\right\|^2\,\middle|\,\mathscr{F}_k\right]\\&\leq F(\hat{x}^k)+\frac{\bar{\rho}}{2}\|x^k-\hat{x}^k\|^2+\bar{\rho}\tau\mathbb{E}\left[\langle\hat{x}^k-x^k,\tilde{g}_{fx}^k+\tilde{g}_{fu}^k\tilde{g}_h^k\rangle\,\middle|\,\mathscr{F}_k\right]+\frac{\bar{\rho}M^2}{2}\tau^2.\end{aligned}$$

注意到 $\mathbb{E}[\tilde{g}_{fx}^k+\tilde{g}_{fu}^k\tilde{g}_h^k|\mathscr{F}_k]=A^k+B^k$，并代入下界（27），可得：

$$\begin{aligned}\mathbb{E}[\varphi_{1/\bar{\rho}}(x^{k+1})|\mathscr{F}_k]&\leq\varphi_{1/\bar{\rho}}(x^k)+\frac{\bar{\rho}M^2}{2}\tau^2+\bar{\rho}\tau(f(\hat{x}^k,h(\hat{x}^k))-f(x^k,u^k)\\&\quad+\frac{(1+\varkappa)\delta}{2}\|\hat{x}^k-x^k\|^2+(1-\varkappa\mathbb{E}\left[I^k\middle|\mathscr{F}_k\right])(u^k-h(x^k)))\\&\leq\varphi_{1/\bar{\rho}}(x^k)+\frac{\bar{\rho}M^2}{2}\tau^2+\bar{\rho}\tau(F(\hat{x}^k)-F(x^k)\\&\quad+\frac{(1+\varkappa)\delta}{2}\|\hat{x}^k-x^k\|^2)+\bar{\rho}\tau\|u^k-h(x^k)\|\\&\quad+\bar{\rho}\tau(F(x^k)-f(x^k,u^k)).\end{aligned}$$

由于函数 $x\mapsto F(x)+\frac{\rho}{2}\|x^k-x\|^2$ 是参数为 $\bar{\rho}-\rho>0$ 的强凸函数，因此

$$F(x^k)-F(\hat{x}^k)=(F(x^k)+\frac{\bar{\rho}}{2}\|x^k-x^k\|^2)-(F(\hat{x}^k)+\frac{\bar{\rho}}{2}\|x^k-\hat{x}^k\|^2) \\ +\frac{\bar{\rho}}{2}\|x^k-\hat{x}^k\|^2\geq(\bar{\rho}-\frac{\rho}{2})\|x^k-\hat{x}^k\|^2.$$

回想 $\bar{\rho}=\rho+(1+\varkappa)\delta$，并利用式 (17)，我们得到

$$\begin{aligned}\mathbb{E}[\varphi_{1/\bar{\rho}}(x^{k+1})|\mathscr{F}_k] &\le \varphi_{1/\bar{\rho}}(x^k)+\frac{\bar{\rho}M^2}{2}\tau^2+\bar{\rho}\tau(-(\bar{\rho}-\frac{\rho}{2})\|x^k-\hat{x}^k\|^2\\ &\quad+\frac{(1+\varkappa)\delta}{2}\|\hat{x}^k-x^k\|^2)+\bar{\rho}\tau\|u^k-h(x^k)\|\\ &\quad+\bar{\rho}\tau(F(x^k)-f(x^k,u^k))\\ &=\varphi_{1/\bar{\rho}}(x^k)+\frac{\bar{\rho}M^2}{2}\tau^2-\frac{1}{2}\tau\|\nabla\varphi_{1/\bar{\rho}}(x^k)\|^2\\ &\quad+\bar{\rho}\tau\|u^k-h(x^k)\|+\bar{\rho}\tau(F(x^k)-f(x^k,u^k)).\end{aligned}\tag{28}$$

利用$f(x,u)$关于u是 1-Lipschitz 这一事实，并结合跟踪误差界 (18)，我们有

$$\mathbb{E}[F(x^k)-f(x^k,u^k)]\le\mathbb{E}[|h(x^k)-u^k|]\le C_2\tau^{1/2}+(1-\tau)^k|u^0-h(x^0)|,$$

其中$C_2=\sigma+\sigma M$。在 (28) 中，对两边取期望可得

$$\begin{aligned}\mathbb{E}[\varphi_{1/\bar{\rho}}(x^{k+1})]\le\mathbb{E}[\varphi_{1/\bar{\rho}}(x^k)]&-\frac{1}{2}\tau\mathbb{E}[\|\nabla\varphi_{1/\bar{\rho}}(x^k)\|^2]+\frac{\bar{\rho}M^2}{2}\tau^2+2\bar{\rho}C_2\tau^{3/2}\\ &+2\bar{\rho}\tau(1-\tau)^k|u^0-h(x^0)|.\end{aligned}$$

将k从 0 到$N-1$求和，得到

$$\begin{aligned}\mathbb{E}[\varphi_{1/\bar{\rho}}(x^N)]\le\varphi_{1/\bar{\rho}}(x^0)&-\frac{1}{2}\tau\sum_{k=0}^{N-1}\mathbb{E}[\|\nabla\varphi_{1/\bar{\rho}}(x^k)\|^2]\\ &+NC_3\tau^{3/2}+2\bar{\rho}|u^0-h(x^0)|,\end{aligned}$$

其中$C_3=2\bar{\rho}C_2$。将左侧以下界 $\min_{x\in X}F$ 进行估计并重新整理，得到如下界

$$\begin{aligned}\mathbb{E}[\|\nabla\varphi_{1/\bar{\rho}}(x^R)\|^2]&=\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}[\|\nabla\varphi_{1/\bar{\rho}}(x^k)\|^2\\ &\le 2\frac{\varphi_{1/\bar{\rho}}(x^0)-\min_{x\in X}F(x)+2\bar{\rho}|u^0-h(x^0)|+NC_3\tau^{3/2}}{N\tau},\end{aligned}$$

这正是我们要证明的结论。 ■

注 5 若我们选择 $\tau=cN^{-2/3}$，其中 $c>0$为某个常数，则定理 4 的一个推论是 $E[\|\nabla\varphi_{1/\bar{\rho}}(x^R)\|^2]=\mathscr{O}(\ ^1N^{1/3})$，因此，为了获得一个 ε-最优点，即满足 $\mathbb{E}[\|\nabla\varphi_{1/\bar{\rho}}(x^R)\|^2]\le\varepsilon$，我们需要 $\mathscr{O}(\varepsilon^{-3})$ 次迭代。由于算法 1 的每一步使用三个数据样本，我们得到总样本复杂度为 $S_\varepsilon=24(\varphi_{1/\bar{\rho}}(x^0)\ -min_{x\in X}F(x)+2\bar{\rho}|u^0-h(x^0)|+C_3)^3\varepsilon^{-3}=\mathscr{O}(\varepsilon^{-3})$。

4. 非光滑弱凸损失函数的收敛速率

SCS 方法使用内层函数 $h(x)$ 值的有偏估计量。当损失函数非光滑时，由此产生的误差更难界定。这是非光滑非凸复合随机优化中固有的难题，阻碍了收敛速率的推导。因此，我们不再通过线性跟踪滤波器更新内层函数估计 $\{u^k\}$，而是构造 SPIDER 估计器（Fang 等人，2018年）：

$$\begin{aligned} u^0 &= \ell_{\mathscr{B}^0}(x^0), \\ u^k &= u^{k-1} + \ell_{\mathscr{B}^k}(x^k) - \ell_{\mathscr{B}^k}(x^{k-1}), \end{aligned} \tag{29}$$

其中，$\mathscr{B}^k$ 是在第k次迭代时随机抽取的数据小批量，且 $\ell_{\mathscr{B}^k}(x) \equiv \frac{1}{|\mathscr{B}^k|}\sum_{D\in\mathscr{B}^k}\ell(x, D)$。该估计器按周期运行，包含三个参数：周期长度$T$、标准批量大小$b$，以及更大的批量大小 $B > b$。它在每个周期开始时重新启动，第一步使用大批量，后续步骤使用标准批量。此外，SPIDER 估计器是无偏估计量：

$$\mathbb{E}[u^k - h(x^k)] = 0,$$

其中，期望是相对于截至第k次迭代的所有随机观测取的。采用 SPIDER 的 SCS 方法如算法 2 所示。

为处理非光滑损失函数，我们不采用假设 2，而是假设损失函数仅为弱凸：

假设 4 对于集合 X 邻域内的所有 x：

(i) 函数 $\ell(x\ \cdot)$ 可积；~(ii) 函数 $\ell(\cdot,D)$ 是弱凸的，且具有可积常数 $\delta(D)$。

~**注 6** 根据假设 *4*，内层函数 $h(x)$ 是 δ-弱凸的，其中 $\delta := \mathbb{E}[\delta(D)]$。由于外层函数 $f(x,u)$ 关于 x 弱凸，且关于 u 单调不减且为凸函数，因此复合函数 $F(x)$也是弱凸的。

根据假设 1 和假设 4，对于任意给定的 D，损失函数 $\ell(\cdot,D)$（作为一个凸函数与一个二次函数之差）在可行集 X 上利普希茨连续，其利普希茨常数为 ~$L(D)$。我们对该常数作出一个额外假设。

假设 5 损失函数 $\ell(x, D)$ 关于 x 的利普希茨常数 ~$L(D)$ 是平方可积的：

$$L^2 \equiv \mathbb{E}[\tilde{L}^2(D)] < +\infty.$$

注 7 假设 *5* 蕴含 *SPIDER* 估计器所要求的均方利普希茨（*MSL*）性质（*Nguyen* 等人，*2022*年；*Pham* 等人，*2020*年）：

$$\mathbb{E}[|\ell(x,D) - \ell(y,D)|^2] \le L^2\|x-y\|^2.$$

在这种情况下，复合次梯度界(9)可自动得出。

算法 2SCS 方法与 SPIDER

输入： 初始点 $x^0 \in X$，恒定步长 $\tau \in (0,\ 1]$；SPIDER周期长度 T，大批量大小 B和小批量大小b。

1: **对于** $k = 0, 1, \ldots, N-1$ **执行**

2: **若** $k \bmod T == 0$ **则**

3: 随机抽取一个数据批次 $\mathscr{B}^k$，大小为 $|\mathscr{B}^k| == B$

4: 重新初始化内函数估计：

$$u^k = \ell_{\mathscr{B}^k}(x^k);$$

5: **else**

6: 随机抽取一个数据批次 $\mathscr{B}^k$，大小为 $|\mathscr{B}^k| == b$

7: 更新内函数估计：

$$u^k = u^{k-1} + \ell_{\mathscr{B}^k}(x^k) - \ell_{\mathscr{B}^k}(x^{k-1}).$$

8: **结束 if**

9: 随机抽取D_1^{k+1}和D_2^{k+1}，并得到估计值

$$\begin{aligned}
&G^k \in \partial_x \ell(x^k, D_1^{k+1}),\\
&\tilde{g}_{fx}^k = \begin{cases} 0 & \text{if } \ell(x^k, D_1^{k+1}) < u^k,\\ \varkappa G^k & \text{if } \ell(x^k, D_1^{k+1}) \geq u^k, \end{cases}\\
&\tilde{g}_{fu}^k = \begin{cases} 1 & \text{if } \ell(x^k, D_1^{k+1}) < u^k,\\ 1-\varkappa & \text{if } \ell(x^k, D_1^{k+1}) \geq u^k, \end{cases}\\
&\tilde{g}_h^k \in \partial_x \ell(x^k, D_2^{k+1}).
\end{aligned}$$

10：更新解估计

$$x^{k+1} = \Pi_X\left(x^k - \tau\left(\tilde{g}_{fx}^k + \tilde{g}_{fu}^k \tilde{g}_h^k\right)^T\right), \tag{30}$$

11: **结束循环**

输出： x^R，其中 R从 $\{0, 1, \ldots, N-1\}$中均匀采样。

~现在，损失函数 $\ell(x,\ D)$ 关于 x 在任意 D处为 $\delta(D)$-弱凸（而非光滑）。注意，除跟踪误差界 (18) 外，第 3 节中的速率收敛分析仍然全部成立。因此，我们需要估计 SPIDER 估计器的跟踪误差；下一条结果中所述的文献中已有此类估计。

引理 8 *(Fang* 等人，*2018*年，引理 *1)* 假设损失函数 $\ell(x, D)$ 满足均方 *Lipschitz* 条件，常数为 L。那么(29) 中的估计量的均方误差可界定为

$$\mathbb{E}[|u^k - h(x^k)|^2] \le \mathbb{E}[|u^0 - h(x^0)|^2] + \sum_{r=1}^{k} \frac{L^2}{|\mathscr{B}^r|}\mathbb{E}[\|x^r - x^{r-1}\|^2], \qquad k = 1, ..., T-1.$$

如果我们能够控制步长 $\|x^k - x^{k-1}\| \le s$，其中 $k = 1,..., T-1$，那么

$$\mathbb{E}[|u^0 - h(x^0)|^2] \le \cdots \le \mathbb{E}[|u^{T-1} - h(x^{T-1})|^2] \le \frac{\sigma^2}{B} + \frac{TL^2s^2}{b}.$$

基于此引理，我们接下来针对 SPIDER参数 T、b 和 B 的特定选择，推导序列 $\{u^k\}$ 的跟踪误差界。

引理 9 假设假设*1*和假设 *3* 至 *5*成立。则当 $B = \frac{2\sigma^2}{\tau^2}$、$b = 2LM\sigma/\tau$、$T = \frac{\sigma}{LM\tau}$时，由算法~~*2*~~生成的序列 $\{u^k\}$ 满足：

$$\mathbb{E}\left[|u^k - h(x^k)|\right] \le \tau, \qquad k = 0, 1, ..., N-1. \tag{31}$$

证明 根据更新函数 (30) 和复合次梯度界 (9)，步长有界：

$$\|x^k - x^{k-1}\| \le M\tau, \qquad k = 0, 1, ..., N-1.$$

根据假设 5，$\ell(x,D)$ 是均方Lipschitz（MSL）的，其常数为 L。此外，根据引理 8，$\{u^k\}$ 的跟踪误差满足：

$$\mathbb{E}[|u^k - h(x^k)|^2] \le \frac{\sigma^2}{B} + \frac{TL^2M^2\tau^2}{b}, \qquad k = 0, 1, ..., N-1.$$

因此，如果我们选择

$$B = \frac{2\sigma^2}{\tau^2}, \qquad b = 2LM\sigma/\tau, \qquad T = \frac{\sigma}{LM\tau},$$

我们可以得到跟踪误差的上界：

$$\mathbb{E}[|u^k - h(x^k)|^2] \le \tau^2.$$

利用詹森不等式，我们立即得到（31）。 ■

其余的速率分析与连续可微情形相同，由此得到以下结果。

定理 10 若假设*1*和假设 *3* 至 *5*成立，则对于任意 $N \geq 1$，由算法 *2*生成的随机点 x^R 满足：

$$\mathbb{E}[\|\nabla \varphi_{1/\bar{\rho}}(x^R)\|^2] \leq 2\frac{\varphi_{1/\bar{\rho}}(x^0) - \min_{x\in X} F(x) + NC_4\tau^2}{N\tau},$$

其中 $C_4 = 2\bar{\rho}(\frac{1}{4}M^2+1)$，期望是针对算法 *2* 生成的轨迹取的，随机变量 R 从 $\{0,1,\ldots,T-1\}$ 中均匀采样，且独立于该轨迹。

证明 证明过程与定理 4 相同，唯一的区别是将跟踪界 (18) 替换为界 (31)。 ■

注 11 若选择 $\tau = N^{-1/2}$，为了得到一个 ε-最优点，我们需要 $\mathcal{O}(\varepsilon^{-2})$ 次迭代。采用引理 *9* 中的超参数选择时，每次迭代的平均批量大小为

$$b + B/T = 4LM\sigma/\tau,$$

因此，我们得到的总样本复杂度为

$$S_\varepsilon = 32(\varphi_{1/\bar{\rho}}(x^0) - \min_{x\in X} F(x) + C_4)^3 LM\sigma\varepsilon^{-3} = \mathcal{O}(\varepsilon^{-3}).$$

我们得出结论：对于光滑和非光滑情形，样本量为 $O(\varepsilon^{-3})$ 就足够了，非光滑情形下仅常数更大（见注 *5*）。

5. 数值实验

本节报告数值实验的结果，以展示我们的 SCS 方法在深度学习中的一项对抗学习任务，以及若干带有非光滑非凸正则项的逻辑回归问题上的收敛性和鲁棒性。数值结果使用 Python（版本 3.7）在一台配备 3.60 GHz CPU（i7-2677M）和 16GB 内存的 Alienware Aurora R8 台式机上获得。

5.1 深度学习

我们考虑一个应用于 MNIST 数据集的卷积神经网络（LeCun 等人，2010年）。该网络由三个卷积层和一个全连接层组成。所有隐藏层均采用 ELU 激活函数，输出层采用 softmax 激活函数。三个卷积层分别包含 16、32 和 32 个大小为 8、6 和 5 的卷积核。我们以这种特定方式构建网络，以便生成可与 Sinha等人（2018年）的结果进行比较的结果。

MNIST数据集由带有 0 到 9 整数标签的手写图像组成，目标是对这些图像进行分类。训练期间，我们使用交叉熵损失（参见此处的参考文献）。由此得到的损失函数 $\ell(x, D)$ 是卷积神经网络与交叉熵损失的复合函数；在我们的设定中，它是连续可微的，因为 ELU 激活函数是连续可微的。我们还对所有隐藏单元的权重设置了界限，这相当于选择可行集 $X = \{x : \|x\|_\infty \leq 10\}$。此类盒约束在实践中常用于正则化（Srivastava 等人，2014年）。

我们使用不同的优化器训练卷积神经网络，具体包括随机梯度下降、SCS（在算法 1 中使用不同的 $\varkappa$ in 值），以及另一种最先进的方法——Wasserstein鲁棒方法（WRM）（Sinha等人，2018年）。为研究训练所得网络的鲁棒性，我们考虑对测试数据集施加两类（对抗攻击）扰动：PGM 攻击（Sinha等人，2018年；Madry等人，2017年）和半偏差攻击，接下来将对后者进行介绍。

PGM攻击。 给定模型参数 x 和数据点 $D = (a,b)$，其输入为 a 、输出为 b，核心思路是在数据点周围的一个球内，对损失函数应用多步投影梯度上升，从而生成对抗性输入数据。具体而言，对于每个数据点 (a_i,b_i)，我们进行迭代

$$\nabla a_i^t(x) := \underset{\|\eta\|_2 \le \varepsilon_{adv}}{\operatorname{argmax}} \{\nabla_a \ell(x; a_i^t, b_i)^T \eta\},$$
$$a_i^{t+1} := \Pi_{B(a_i^t)}\{a_i^t + \tau_{adv} \nabla a_i^t(x)\},$$

对于 $t = 1,\ldots,T_{adv}$，其中 $B(a_i^t) := \{a\colon \|a - a_i^t\|_2 \le \varepsilon_{adv}\}$ 是以 a_i^t为中心的一个球， ε_{adv}控制对抗扰动的程度，T_{adv} 是迭代次数。更多细节请参阅（Madry等人，2017年）。

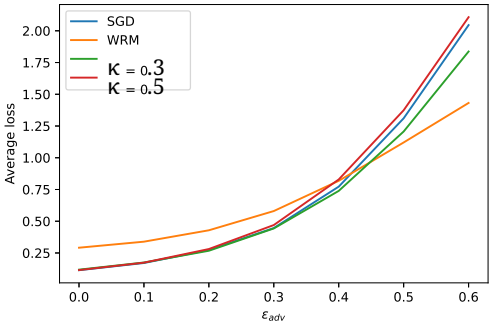
半偏差攻击。 虽然PGM 攻击会构造强大的最坏情况对手，但如果对数据集的扰动具有随机性而非最坏情况特性，这种做法可能过于保守。我们引入半偏差攻击，通过用“平均”样本替换“优质”样本，构造一个不那么保守的对手。更具体地说，对于每个数据点D_i，我们将相应的损失函数 $\ell(x,D_i)$ 替换为 $\ell_{sd}(x,D_i)$：

$$\bar{\ell}(x) := \frac{1}{|\mathscr{B}_{test}|} \sum_{D_j \in \mathscr{B}_{test}} \ell(x, D_j),$$
$$\ell_{sd}(x, D_i) := \bar{\ell}(x) + \varkappa_{adv} \max(0, \ell(x, D_i) - \bar{\ell}(x)),$$

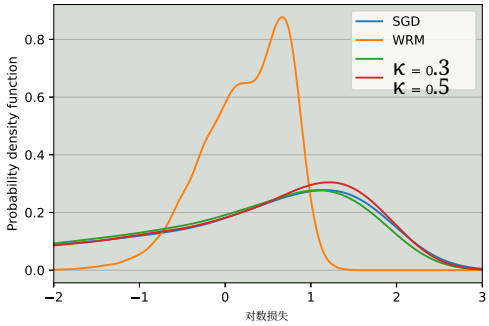
其中， $\mathscr{B}_{test}$ 表示测试数据集， $\varkappa_{adv}$ 控制对抗扰动级别。

我们控制训练过程中的梯度计算次数，使不同方法的计算次数相同。 [1] 训练数据为原始（未受污染的）MNIST数据，而模型则使用受到PGM攻击和半偏差攻击污染的数据进行测试：对于测试集中的每个数据点，我们分别在不同扰动级别下应用PGM攻击和半偏差攻击，并在图1的图(a)和(c)中绘制不同模型在各个扰动级别下的平均测试损失。当 $\varepsilon_{adv} = 0.6$ 针对PGM攻击，或 $\varkappa_{adv} = 1$ 针对半偏差攻击时，我们还计算测试集损失的（自然）对数，并在图1的图(b)和(d)中展示不同模型损失分布的对数。我们发现，采用合适的$\varkappa$ 值时，SCS在两种攻击下都能获得比随机梯度下降更好的解。它在PGM攻击下不如最坏情况风险最小化，这符合预期，因为最坏情况风险最小化的训练所针对的攻击与此类攻击非常相似。[2]出于同样的原因，SCS在半偏差攻击下优于最坏情况风险最小化。

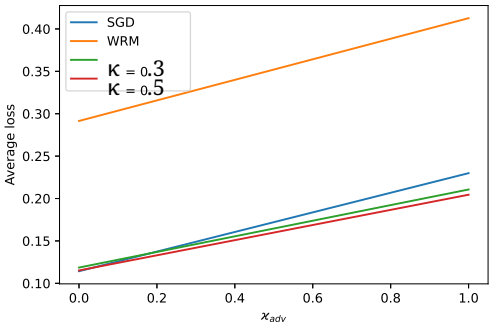
1. 我们使用 SCS 训练模型 20 轮次，使用 WRM 训练 4 轮次，因为在每一步中，SCS 计算 3 个梯度，而 WRM 计算 16 个梯度。2. 在第k步，WRM 对数据点$D^k = (a^k,b^k)$ 应用梯度上升来扰动代价函数$a \mapsto \ell(x;a$, $-$ γ_{adv} , ^{k} γ_{adv} , $^k = \| - {}^k\|^2$ $b^k)c(aa)$，其中，是控制鲁棒性的参数，$c(aa)$:$a\,a$ 是正则化项。在该实验中，我们在每一步的梯度上升过程中迭代 15 次。



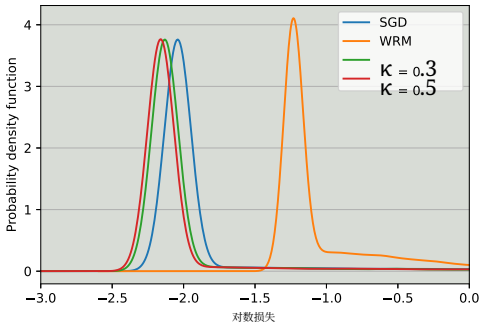
（a）PGM 攻击在不同扰动水平下的平均损失函数



(b) PGM攻击下损失函数对数的概率密度函数（PDF）



(c) 不同扰动
水平下的半偏差攻击平均损失



(d) 半偏差攻击下损失函数对数的概率密度函数

图1：PGM攻击和半偏差攻击下的测试损失。训练数据为原始（未受污染）的MNIST数据，而模型则使用受污染的数据进行测试。

5.2 非凸惩罚项

我们研究了 Blog Feedback 数据集上的一项回归任务（Buza，2014年），该数据集包含从博客文章中提取的281个变量。任务是根据其他280个变量（特征）预测评论数量。2010年和2011年的样本纳入训练集（共52396个）， 02/01/2012 的样本纳入验证集（共114个）， 02/02/2012 至 03/31/2012 之间的样本纳入测试集（共7511个）。测试集分为60个子集，每个子集包含在2月或3月某一天生成的样本。我们采用带有平均绝对差（MAD）损失函数的线性回归模型，并加入一个正则化项。损失函数的形式为 $\ell(x, D) = |a^T x - b| + r(x)$，其中$D = (a, b)$是输入数据，而$r(x)$是正则化项。选择不同的正则化项时，损失函数可以是凸的，也可以是非凸的。这里我们研究 Lasso 惩罚项（Frank 和 Friedman，1993年）以及两种非凸惩罚项：平滑剪切绝对偏差惩罚项（Fan 和 Li，2001年）和 MCP 惩罚项（Zhang，2010年）。对应的正则化器如下：

- Lasso：

$$r(x) = \lambda |x|, \tag{32}$$

• 平滑剪切绝对偏差:

$$r(x)=\begin{cases}\lambda|x| & \text{if } |x|\le\lambda,\\ \frac{\gamma\lambda|x|-0.5(x^2+\lambda^2)}{\gamma-1} & \text{if } \lambda<|x|\le\lambda\gamma,\\ \frac{\lambda^2(\gamma+1)}{2} & \text{if } |x|>\lambda\gamma,\end{cases}\tag{33}$$

• 极小凹罚函数:

$$r(x)=\begin{cases}\lambda|x|-\frac{x^2}{2\gamma} & \text{if } |x|\le\lambda\gamma,\\ \frac{\lambda^2\gamma}{2} & \text{if } |x|>\lambda\gamma,\end{cases}\tag{34}$$

其中 $\lambda>0$ 和 $\gamma>0$ 是参数。采用平滑剪切绝对偏差或极小凹罚函数惩罚项时，损失函数将变为非光滑且弱凸。我们将约束集设为 $X=\{x:\|x\|_\infty\le 10\}$。

在图 2 中，我们比较了这三种不同惩罚项对应的训练损失变化历程、测试损失对数的分布，以及训练阶段目标函数的衰减曲线。该方法在（Lasso）凸情形和（平滑剪切绝对偏差与极小凹罚函数）非凸情形下，表现出相近的收敛速度和测试性能。

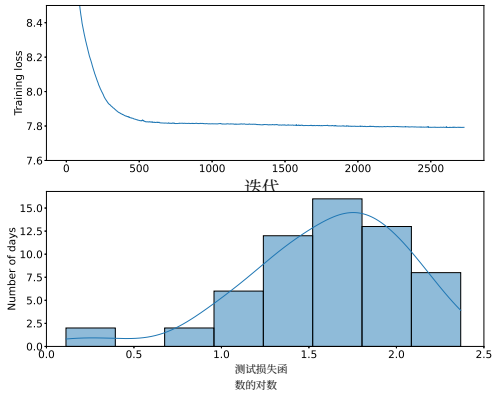
5.3 关于假设的备注

对于我们的深度学习实验，可行集 $\mathrm{X}=\{\mathrm{x}:\|\mathrm{x}\|_\infty\le 10\}$ 显然是凸且紧的，并且假设1成立。输入数据经过归一化且有界，迭代点始终位于紧集 X 中；在该集合上，连续可微损失 ℓ(x, D) 的梯度连续且有界，因此对于每个固定的 $\mathrm{x}\in\mathrm{X}$，损失函数 ℓ(x, D) 关于 D 可积。因此，假设 2 成立。此外，从 (10) 到 (15) 可以看出，序列 $\tilde{\mathrm{J}}^k$、$\tilde{\mathrm{g}}^k$ 和 $\tilde{\mathrm{h}}^k$ 均一致有界。因此，它们在给定自然滤过 $\mathscr{F}_k$的条件下方差有界，次微分 ∂F(x^k) 中元素的随机估计也有界。对这些估计取期望时，由于次微分是有界集，我们可以交换次微分与期望算子（Mikhalevich 等人，1987 年，定理 23.1）。由于 D_1^k、D_2^k 和 D_3^k 是来自经验数据分布的独立同分布样本，因此可以推知 $\tilde{\mathrm{J}}^k$、$\tilde{\mathrm{g}}^k$ 和 $\tilde{\mathrm{h}}^k$ 是无偏估计。根据这些观察，我们得出假设 3 也满足。

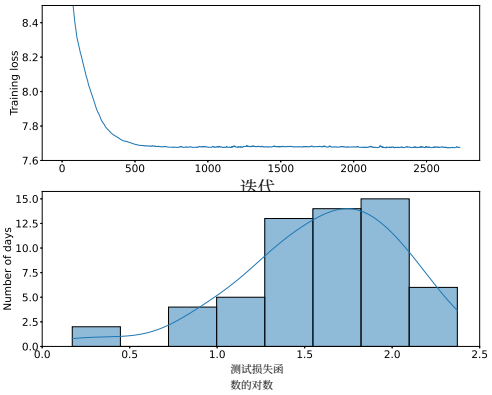
对于正则化逻辑回归示例，与 (Mei 等人，2018年) 类似，约束集X 是一个 ℓ_∞ 球，因此它是凸且紧的，并且假设1成立。我们在 (32)–(34) 中考虑的所有惩罚函数都是弱凸的，因此损失函数 $\ell(x,D)$ 关于x也是弱凸的。通过与深度学习情形类似的论证，对于每个固定的$x\in X$，损失函数 $\ell(x,D)$ 关于D 可积，且假设 3 和 4 成立。输入数据经过归一化且有界，因此随机 Lipschitz 常数 $\tilde{L}(D)$ 也有界，假设 5 成立。因此，我们的假设适用于本研究开展的数值实验。

6. 结论

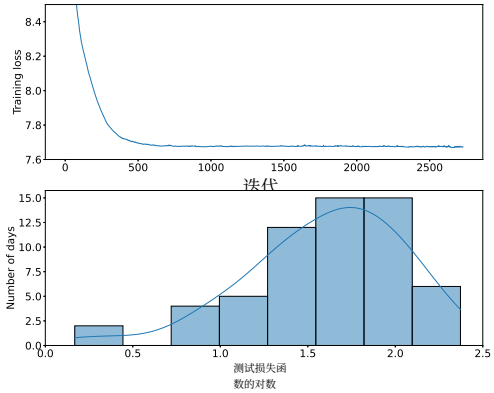
本文研究了一个分布鲁棒随机优化问题，其中歧义集采用均值–半偏差风险度量定义。我们将该问题重新表述为随机双层非光滑优化问题，并提出了一种名为随机复合次梯度（SCS）的单时间尺度方法。我们的方法支持两种不同的内层值跟踪方式：（i）对连续可微损失



(a) Lasso 惩罚项



(b) 平滑剪切绝对偏差惩罚项



(c) MCP 惩罚项

图 2：对于 Lasso、SCAD 和 MCP 惩罚项，每个图中的上图展示了训练损失随迭代的变化，下图展示了测试数据上损失分布的对数。

函数进行线性化跟踪；（ii）通过 SPIDER 估计器跟踪弱凸损失函数。我们证明，在这两种情况下，$\mathcal{O}(\varepsilon^{-3})$ 均可实现相应的样本复杂度，只是第二种情况下的常数较大。据我们所知，这是关于非凸非光滑损失的分布鲁棒学习的首个样本复杂度结果。最后，我们在一个鲁棒深度学习问题以及一个采用弱凸非光滑正则化项的逻辑回归问题上展示了该算法的性能。

7. 致谢

本研究部分由以下资助支持：美国海军研究办公室奖项编号 N00014- 21-1-2244、美国国家科学基金会（NSF）CCF-1814888、NSF DMS-2053485、NSF DMS-

1723085、美国国家科学基金会奖项 DMS-1907522，以及美国海军研究办公室奖项 N00014-21-1-2161。

参考文献

J Andrew Bagnell。鲁棒监督学习。载于*AAAI*，第 714–719 页，2005年。

Krishnakumar Balasubramanian、Saeed Ghadimi 和 Anthony Nguyen。《具有层级无关收敛速率的随机多层组合优化算法》。*SIAM* 优化期刊，32(2)：519–544，2022年。L´eon Bottou。《使用随机梯度下降的大规模机器学习》。载于*COMPSTAT'2010* 论文集，第177–186页。Springer，2010年。Krisztian Buza。《博客的反馈预测》。载于数据分析、机器学习与知识发现，第145–152页。Springer，2014年。Sebastian Curi、Kfir Y Levy、Stefanie Jegelka 和 Andreas Krause。《面向随机风险厌恶学习的自适应采样》。神经信息处理系统进展，33：1036–1047，2020年。D. Davis 和 D. Drusvyatskiy。《弱凸函数的随机模型法最小化》。*SIAM* 优化期刊，29(1)：207–239，2019年。D. Davis、D. Drusvyatskiy、S. Kakade 和 J. D. Lee。《随机次梯度法在温和函数上收敛》。计算数学基础，第1–36页，2018a年。Damek Davis、Dmitriy Drusvyatskiy、Kellie J MacPhee 和 Courtney Paquette。《尖锐弱凸函数的次梯度方法》。优化理论与应用期刊，179 (3)：962–982，2018b年。John C. Duchi 和 Hongseok Namkoong。《通过分布鲁棒优化学习具有一致性能的模型》。统计年刊，49(3)：1378–1406，2021年。doi：10.1214/20-AOS2004。URL https://doi.org/10.1214/20-AOS2004。Peyman Mohajerin Esfahani 和 Daniel Kuhn。《利用Wasserstein度量进行数据驱动的分布鲁棒优化：性能保证与可处理的重新表述》。数学规划，171(1-2)：115–166，2018年。Jianqing Fan 和 Runze Li。《通过非凹惩罚似然进行变量选择及其预言机性质》。美国统计学会会刊，96(456)：1348–1360，2001年。Cong Fang、Chris Junchi Li、Zhouchen Lin 和 Tong Zhang。《Spider：通过随机路径积分差分估计器实现近最优非凸优化》。神经信息处理系统进展，31，2018年。LLdiko E Frank 和 Jerome H Friedman。《若干化学计量学回归工具的统计学视角》。*Technometrics*，35(2)：109–135，1993年。

Rui Gao、Xi Chen 和 Anton J Kleywegt。《统计学习中的Wasserstein分布鲁棒性与正则化》。*arXiv*预印本 *arXiv:1712.06050*，2017年。Saeed Ghadimi、Andrzej Ruszczynski 和 Mengdi Wang。《用于嵌套随机优化的单时间尺度随机逼近方法》。*SIAM* 优化期刊，30(1)：960–979，2020年。Mert G¨urb¨uzbalaban、Umut Si msekli 和 Lingjiong Zhu。《随机梯度下降中的重尾现象》。载于国际机器学习会议，第3964–3975页。PMLR，2021年。Mert G¨urb¨uzbalaban、袁汉 Hu、Umut Simsekli、袁坤和 Lingjiong Zhu。《去中心化随机梯度下降中的重尾现象》。*arXiv*电子预印本，文献编号 arXiv:2205.06689，2022年5月。Mert G¨urb¨uzbalaban、Andrzej Ruszczy´nski 和 Landi Zhu。《用于分布鲁棒非凸非光滑学习的随机次梯度法》。优化理论与应用期刊，194(3)：1014–1041，2022年。Nam Ho-Nguyen 和 Stephen J Wright。《通过具有Wasserstein歧义集的分布鲁棒性实现对抗分类》。数学规划，第1–37页，2022年。Prateek Jain、Sham M Kakade、Rahul Kidambi、Praneeth Netrapalli 和 Aaron Sidford。《加速最小二乘回归的随机梯度下降》。载于学习理论会议，第545–604页，2018年。金继凯、张博航、王海洋和王立伟。《非凸分布鲁棒优化：非渐近分析》。神经信息处理系统进展，34：2771–2782，2021年。Daniel Kuhn、Peyman Mohajerin Esfahani、Viet Anh Nguyen 和 Soroosh Shafieezadeh-Abadeh。《Wasserstein分布鲁棒优化：理论及其在机器学习中的应用》。载于分析时代的运筹学与管理科学，第130–166页。INFORMS，2019年。Yassine Laguel、J´erˆome Malick 和 Zaid Harchaoui。《基于超分位数的学习：一种直接采用基于梯度的优化的方法》。信号处理系统期刊，94(2)：161–177，2022年。Yann LeCun、Corinna Cortes 和 CJ Burges。《MNIST手写数字数据库》。*ATT*实验室 *[*在线*]*。可访问：*http://yann.lecun.com/exdb/mnist*，2，2010年。Daniel Levy、Yair Carmon、John C Duchi 和 Aaron Sidford。《分布鲁棒优化的大规模方法》。神经信息处理系统进展，33：8847–8860，2020年。Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。《迈向能够抵御对抗攻击的深度学习模型》。统计，1050：9，2017年。Sa njay Mehrotra 和 He Zhang。《分布鲁棒最小二乘问题的模型与算法》。数学规划，146(1)：123–141，2014年。doi：10.1007/s10107-013-0681-9。URL https://doi.org/10.1007/s10107-013-0681-9。

Song Mei、Yu Bai 和 Andrea Montanari。非凸损失下经验风险的景观。统计年刊，46(6A):2747–2774，2018年。V. S. Mikhalevich、A. M. Gupal 和 V. I. Norkin。非凸优化方法。Nauka，莫斯科，1987 年。Jean Jacques Moreau。希尔伯特空间中的邻近性与对偶性。法国数学会´公´会´通报，93:273–299，1965年。Hongseok Namkoong 和 John C Duchi。基于 f-散度的分布鲁棒优化随机梯度方法。载于神经信息处理系统进展，第2208–2216页，2016年。Geoffrey Negiar、Gideon Dresdner、Alicia Tsai、Laurent El Ghaoui、Francesco Locatello、Robert Freund 和 Fabian Pedregosa。用于约束有限和最小化的随机 Frank-Wolfe 方法。载于 Hal Daum´e III 和 Aarti Singh（编），第*37*届国际机器学习会议论文集，第119卷，机器学习研究论文集，第7253–7262页。PMLR，2020年7月13日至18日。网址：https://proceedings.mlr.press/v119/ negiar20a.html。Lam M Nguyen、Marten van Dijk、Dzung T Phan、Phuong Ha Nguyen、Tsui-Wei Weng 和 Jayant R Kalagnanam。基于 SARAH 的有限和光滑优化。计算优化与应用，第1–33页，2022年。Evgeni Alekseevich Nurminskii。求解非线性规划问题的拟梯度法。控制论，9(1):145–150，1973年。Nhan H Pham、Lam M Nguyen、Dzung T Phan 和 Quoc Tran-Dinh。ProxSARAH：一种用于随机复合非凸优化的高效算法框架。《机器学习研究杂志》，21(110):1–48，2020年。Qi Qi、Zhishuai Guo、Yi Xu、Rong Jin 和 Tianbao Yang。一种适用于一类具有非凸目标函数的分布鲁棒优化问题的在线方法。载于 M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan（编），神经信息处理系统进展，第34卷，第10067–10080页。Curran Associates, Inc.，2021年。网址：https://proceedings.neurips.cc/paper_files/paper/2021/file/533fa796b43291fc61a9e812a50c3fb6-Paper.pdf。Qi Qi、Jiameng Lyu、Kung sik Chan、Er Wei Bai 和 Tianbao Yang。与样本量无关复杂度的随机约束分布鲁棒优化，2022a年。Qi Qi、Yi Xu、Rong Jin、Wotao Yin 和 Tianbao Yang。注意力偏置随机梯度下降，2022b年。R Tyrrell Rockafellar 和 Roger J-B Wets。变分分析，第317卷。Springer Science & Business Media，2009年。R Tyrrell Rockafellar、Stan Uryasev 和 Michael Zabarankin。风险分析中的广义偏差。金融与随机，10(1):51–74，2006年。

A. Ruszczy´nski 和 A. Shapiro。凸风险函数优化。《运筹研究数学》，31:433–452，2006年。Andrzej Ruszczynski。一种用于非光滑非凸多层复合优化的随机次梯度法。《*SIAM* 控制与优化杂志》，59(3):2301–2320，2021年。Soroosh Shafieezadeh Abadeh、Peyman Mohajerin Mohajerin Esfahani 和 Daniel Kuhn。分布鲁棒逻辑回归。载于 C. Cortes、N. D. Lawrence、D. D. Lee、M. Sugiyama 和 R. Garnett（编），《神经信息处理系统进展》第*28*卷，第1576–1584页。Curran Associates, Inc.，2015年。网址：http://papers.nips.cc/paper/5745-distributionally-robust-logistic-regression.pdf。Aman Sinha、Hongseok Namkoong 和 John Duchi。通过有原则的对抗训练认证某些分布鲁棒性。载于国际学习表征会议，2018年。Tasuku Soma 和 Yuichi Yoshida。基于条件风险价值的统计学习。*arXiv* 预印本 *arXiv:2002.05826*，2020年。Nitish Srivastava、Geoffrey Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。Dropout：一种防止神经网络过拟合的简单方法。《机器学习研究杂志》，15(1):1929–1958，2014年。Akiko Takeda 和 Takafumi Kanamori。一种基于条件风险价值度量的统计学习问题鲁棒方法。《欧洲运筹学杂志》，198(1):287–296，2009年。ISSN 0377-2217。doi：https://doi.org/10.1016/j.ejor.2008.07.027。网址：http: // www.sciencedirect.com/science/article/pii/S0377221708005614。M. Wang、E. X. Fang 和 B. Liu。随机复合梯度下降：最小化期望值函数复合的算法。数学规划，161(1-2):419–449，2017a年。M. Wang、J. Liu 和 E. X. Fang。加速随机复合优化。《机器学习研究杂志》，18:1–23，2017b年。Cun-Hui Zhang。在极小极大凹惩罚下近乎无偏的变量选择。统计年刊，38(2):894–942，2010年。Xuan Zhang、Necdet Serhat Aybat 和 Mert G¨urb¨uzbalaban。用于求解鞍点的鲁棒加速原始-对偶方法。*arXiv* 预印本 *arXiv:2111.12743*，2021年。Xuan Zhang、Necdet Serhat Aybat 和 Mert G¨urb¨uzbalaban。Sapd+：一种用于非凸-凹极小极大问题的加速随机方法。已被神经信息处理系统（*NeurIPS*）会议论文集接收，2022年。Dixian Zhu、Gang Li、Bokun Wang、Xiaodong Wu 和 Tianbao Yang。当曲线下面积遇上分布鲁棒优化：以非凸收敛保证为深度学习优化部分曲线下面积。载于国际机器学习会议，第27548–27573页。PMLR，2022年。