摘 要

理想的机器学习范式需要从“完美”的数据中学习，它们通常被假设是合适数量的向量。然而，现实世界中的数据以海量、流式、在线、分布式和受污染数据等不受欢迎的形式出现。这些问题相互叠加，雪上加霜。

在本文中，我们研究了海量受污染数据的问题。我们首先考虑数据的大规模性。处理海量数据的一个自然想法是对其进行压缩。我们在这里采用的关键概念是 核心集，它是原始数据集的加权子集，对于任意模型参数都能近似目标函数的值。然而研究表明，保证全局近似的传统核心集需要相当大的采样复杂度——逻辑回归这种简单的线性模型也至少需要 $\Omega(n/\operatorname{polylog}(n))$ 的样本量 (Mai et al., 2021)。近年来研究者们提出了局部核心集 (Ding et al., 2020; Huang et al., 2021)，它仅对一个范数球内的模型参数逼近目标函数值。研究表明局部核心集的采样复杂度可以减小到 $\mathcal{O}(\operatorname{polylog}(n))$，所以本文也采用局部核心集概念。

然后我们考虑数据受到污染的问题。特别地，我们从概率分布的角度来看这个问题：受污染分布和真实分布之间的 Wasserstein 距离是一个有界正数，这种情况对应着当前备受关注的数据分布偏移现象。在数据分布偏移的情况下训练机器学习模型可以被形式化为 Wasserstein 分布鲁棒优化 (WDRO) 问题。对于大规模数据，解决 WDRO 问题的现有方法计算成本很高，这给了我们研究 WDRO 问题的核心集的动机。由于 WDRO 具有难解性，WDRO 的局部核心集构造是很困难的。为了解决这个问题，我们利用了 WDRO 的强对偶性并定义了“对偶核心集”，它比局部核心集要容易研究。为了构建对偶核心集，我们提出了一种特别适用于 WDRO 对偶形式的网格采样方法。为了给我们的方法提供理论依据，我们在一些温和的假设下证明了对偶核心集也是合格的局部核心集。我们实现了提出的核心集算法并验证了它的有效性。

关键词：核心集；Wasserstein 分布鲁棒优化；大规模优化；采样

目 录

第1章 绪 论

1.1 研究背景

近年来，基于机器学习的人工智能开始冲击各个行业，对人类社会的发展产生不可磨灭的推动作用。一般来说，机器学习模型的训练依赖于收集到的数据。如今，每天都有海量的数据在产生，数据产生和收集的形式也日渐多样化：在线数据，流数据，分布式数据，关系性数据，受污染数据 ... ... 在海量的各种形式的数据上有效训练高质量机器学习模型已成为人工智能技术落地过程中亟需解决的一个问题。

在理论计算机和计算几何领域，核心集 (Feldman, 2020) 是一种比较流行的用来处理大数据的方法。核心集是一种高度压缩的数据表示，可以被用来加速优化问题的求解。传统的核心集一般被用于加速几何优化问题求解，其中最具有代表性的是聚类问题：在一个小规模的核心集上计算得到的聚类中心也能将大规模数据集较好地划分成数个簇。核心集由于具有可合并性和可分解性，所以也能够用于处理流数据，在线数据和分布式数据的聚类问题。近年来，核心集方法还被扩展到含离群点数据，关系性数据等场景。如今的核心集方法几乎可以在各种场景下帮助求解聚类等几何优化问题。

近年来研究者们开始考虑一些简单机器学习模型（比如逻辑回归，ℓ_p 回归等）的传统核心集理论，并且给出了一些初步结果；关于一般机器学习模型的传统核心集理论的研究目前还比较稀缺。也有研究者从其他角度出发重新定义了机器学习模型的核心集，这种根据机器学习问题本身性质重新定义的核心集可能会比传统核心集更加适配模型训练。

鲁棒性是机器学习模型落地的关键。然而研究表明，对输入样本加上一个微小的对抗扰动，能够让一个训练好的机器学习模型的输出发生巨大变化 (Szegedy et al., 2014)。即使不对模型的输入样本刻意添加攻击，当模型面对更“难”的测试样本或者是测试样本随时间发生了演变，深度模型可能也会发生性能骤降 (Recht et al., 2019)。因此学者们提出了一些训练鲁棒机器学习模型的范式，比如对抗训练 (Madry et al., 2018)，重要性加权训练 (Sugiyama et al., 2007)，鲁棒优化 (Ben-Tal et al., 2009; Sinha et al., 2018) 等。然而这些鲁棒训练范式几乎都会让模型训练变得更加消耗计算资源，比如对抗训练通常比标准训练多花数倍乃至数十倍的时间。

本文从理论上研究了 Wasserstein 分布鲁棒优化问题的核心集。Wasserstein 分布鲁棒优化是鲁棒优化的一种，它对应了一种在数据分布受到扰动的场景下的模型训练。另外，我们还通过实验对提出的方法进行了验证。

1.2 研究现状

这一章我们介绍大规模鲁棒机器学习相关的研究现状。

1.2.1 核心集

核心集的概念最早由Agarwal et al. (2004) 在研究几何形状包含问题时提出(他们当时称之为 ε-核)。同时他们在Bentley et al. (1980) 的启发下，提出了一种支持在流数据场景下动态向核心集插入新数据的树状数据结构。可见核心集在诞生的时候，就已经显示了能在多种数据场景下处理大数据问题的潜力。

核心集的构建方法有非常多种，最简单的是基于均匀采样的核心集。在现实中，这种简单的方法对于有离群点的聚类等问题出乎意料地有效。这一点后来由Li et al. (2001); Ding et al. (2019) 给出了理论上的佐证，他们的分析依赖于问题的 "低伪维度" 性质。从方差缩减采样的角度，Feldman et al. (2011) 提出了用重要性采样来构建核心集的理论框架。重要性采样首先根据目标函数计算出每个数据的重要性分数，进而得到采样的概率分布。可以证明，重要性采样通常有着比均匀采样更低的理论采样复杂度，不过它也依赖于问题的低伪维度。Chen (2009) 提出了一种不依赖于伪维度的方差缩减采样方法：分层采样方法。这种方法一般先将数据按照某种指标分成数个环，然后在每个环内部进行均匀采样并调整权重。以上方法都是随机算法，一种确定性的构建核心集的方法是贪心选择 (Badoiu et al., 2003)，这种方法每一轮迭代都会贪心地并入对当前核心集提升最大的点，对于特定问题这种方法构造的核心集的规模很小，但是构造所需的计算开销通常很大。

近年来研究者开始把核心集技术推广至机器学习领域。由于机器学习模型的优化通常是基于梯度的，所以Mirzasoleiman et al. (2020) 提出了一种近似全梯度的核心集。类似地，Pooladzandi et al. (2022) 提出了动态近似全 Hessian 矩阵和梯度乘积的核心集，用来加速二阶优化方法。这些核心集的构建都可以转化为一个可以用贪心算法求解的子模优化问题，但是算法性能的分析依赖于目标函数的凸性。为了构建深度神经网络的核心集，有研究者提出了一种新的角度—将核心集构建问题看作双层优化的过程 (Borsos et al., 2020; Zhou et al., 2022)：内层优化求解在给定核心集上的最优模型，外层优化则是找到最优核心集权重系数使得对应最优模型在全数据集上损失最小。

1.2.2 鲁棒机器学习核心集

鲁棒性是机器学习模型落地的关键，也有一些学者研究鲁棒优化问题的核心集。Mirzasoleiman et al. (2020) 设计了一种核心集构建算法来低秩逼近神经网络的 Jacobian 矩阵，进而实现对抗标签噪声的鲁棒训练。离群点鲁棒聚类的核心集

是一个被广泛研究的经典问题 (Ding et al., 2019; Feldman et al., 2012; Huang et al., 2018)，对于一般的离群点鲁棒连续有界优化问题，Wang et al. (2021) 提出了一种动态核心集框架。对抗训练是神经网络获得对抗鲁棒性的主要方式 (Goodfellow et al., 2015)，但是对抗训练需要对每个训练数据都计算对应的对抗样本，所以可能需要花数倍于普通训练的时间。如何有选择性地对训练数据计算对抗样本，或者说，如何构建对抗鲁棒的核心集会是未来的一个比较有价值的研究问题。

1.2.3 其他数据压缩方法

矩阵稀疏化的 **Sketch** 方法 (Phillips, 2016) 是另一类数据压缩方法。与核心集不同，sketch 并不要求压缩得到的数据必须是原数据集的子集，可以是原数据的线性组合。在机器学习中，Sketch 通常被用于 (正则化或者无正则化) 线性回归中 (Avron et al., 2017; Chowdhury et al., 2018)。特别地，sketch 方法经常被用于设备存储容量受限的流数据以及在线数据场景 (Cohen et al., 2020; Shi et al., 2021)。

另外一种值得注意的数据压缩方法是**降维** (Dimensionality Reduction)。之前介绍的 sketch 和核心集都是减小数据的数量 n，而降维方法则是用来减小数据维度 d. 最著名的确定性降维方法应该是主成分分析 (Principal Component Analysis)(Hotelling, 1933)，它可以将数据集正交投影至一个低维子空间中，使得投影平方误差最小[①]。事实上，这个低维子空间由 $n \times d$ 的数据矩阵的右奇异向量张成，所以计算降维后的矩阵需要计算数据矩阵的奇异值分解 (Singular Value Decomposition)，其时间复杂度为 $O(n^3 + dn^2)$(Li et al., 2019)。基于随机投影 (Random Projection) 的不确定性降维方法更适合大数据场景，其中最有名的方法是 Johnson-Lindenstrauss (JL) 变换 (Dasgupta et al., 2003)。JL 变换的投影矩阵是随机高斯矩阵，因而可以比较方便地随机生成。

1.3 主要研究内容

核心集方法在几何优化问题中表现得很好，但是总体上在机器学习领域的理论研究和应用还相对较少。已有的研究表明，传统核心集很难扩展到机器学习领域。为了解决这个问题，近年来研究者们提出了一些 "改良" 版本的核心集的定义，特别地，我们采用了局部核心集。过去的鲁棒机器学习局部核心集研究聚焦于对离群点鲁棒，然而其他类型的数据污染则考虑得比较少。本文主要考虑了当前备受关注的数据污染场景：数据分布偏移 (Quinonero-Candela et al., 2008)。具体地，我们研究了 Wasserstein 分布鲁棒优化 (WDRO) 的局部核心集理论。WDRO 是一种对抗数据分布扰动的优化框架，这种框架假设数据分布扰动

[①]投影平方误差最小和投影点集的方差最大这两种叙述是等价的。

的 Wasserstein 距离不超过某个阈值 σ. 事实上，这种假设可以涵盖非常多的数据扰动情况，包括最常见的逐点欧式距离扰动。本文贡献如下:

(1) 提出对偶核心集

因为 WDRO 问题本身具有难解性，我们发现直接研究 WDRO 的局部核心集是很困难的。为了解决这个问题，我们先利用 WDRO 的强对偶性定义了对偶核心集，它比局部核心集要更容易研究。为了给我们的方法提供理论保障，我们在一些温和的假设下证明了对偶核心集也是合格的局部核心集。据我们所知，这是首个分布鲁棒优化的核心集方法。

(2) 给出构造对偶核心集的算法框架

为了构建对偶核心集，我们提出了一种特别适用于 WDRO 对偶形式的网格采样算法，并且分析了它的理论采样复杂度上界。这种网格采样算法是分层采样算法的推广，在保持和分层采样一样高效的同时，又能够适配更一般的机器学习模型的分布鲁棒训练。

1.4 论文组织结构

本文共分为六个章节，组织结构如下:

第一章是绪论，介绍了论文的研究背景，意义，现状和主要内容。

第二章详细介绍了核心集以及分布鲁棒优化的背景知识和主要相关研究工作。

第三章提出了 WDRO 问题的对偶核心集并从理论上证明了它的有效性。

第四章给出了一个构建对偶核心集的分层采样算法并从理论上分析了它的采样复杂度。

第五章给出了对偶核心集的具体应用，并在这些应用任务上做了实验验证对偶核心集算法的有效性。

第六章总结了全文。

第 2 章　背景知识和相关工作

这一章我们详细介绍核心集以及鲁棒优化相关的研究内容和已有工作。

2.1　核心集

假设输入数据集 $P_n = \{p_1, p_2, \dots, p_n\}$ 来自数据空间 $\mathbb{P} \subseteq \mathbb{R}^d$，我们要求解的优化问题如下:

$$\min_{S \in \mathbb{S}} \mathcal{L}(P_n, S), \tag{2.1}$$

其中 $\mathbb{S}$ 是可行域，$\mathcal{L}$ 是损失函数并且可以接受带权重数据集合输入。

如果输入数据集的规模 n 太大会导致优化问题（2.1）的求解变得低效，我们可以先构建一个 P_n 的核心集。

定义 2.1 (ε-核心集). *对于一个由数据集，损失函数以及可行域组成的三元组 $(P_n, \mathcal{L}, \mathbb{S})$，它的 ε-核心集是一个非负稀疏权重向量 $W = [w_1, w_2, \dots, w_n]$，使得对应的加权集合 $\tilde{P}$ 满足*

$$\mathcal{L}(\tilde{P}_n, S) \in (1 \pm \varepsilon) \cdot \mathcal{L}(P_n, S), \forall S \in \mathbb{S}. \tag{2.2}$$

注 2.2. *由于在给定 P_n 之后，权重向量 W 和对应的加权子集 $\tilde{P}_n$ 是等价的，所以在本文中我们有时候也称加权子集 $\tilde{P}_n$ 为核心集。*

在近似算法领域，人们通常用近似比衡量一个解 S 的质量。

定义 2.3 (α-近似解). *我们称 $S_0 \in \mathbb{S}$ 是优化问题 (2.1) 的一个 α-近似解，或者说 $\alpha \geq 1$ 是解 S_0 的近似比，如果*

$$\mathcal{L}(P_n, S_0) \leq \alpha \cdot \min_{S \in \mathbb{S}} \mathcal{L}(P_n, S). \tag{2.3}$$

不难发现，对核心集 $\tilde{P}_n$ 优化关于 S 的损失函数 $\mathcal{L}(\tilde{P}_n, S)$ 也能产生出一个高质量的解。具体地，我们有如下论断。

论断 2.4 (核心集的有效性). *如果 $\tilde{S}$ 是 ε-核心集 $\tilde{P}_n$ 上的优化问题*

$$\min_{S \in \mathbb{S}} \mathcal{L}(\tilde{P}_n, S) \tag{2.4}$$

的 α-近似解，那么 $\tilde{S}$ 也是原问题 (2.1) 的一个 $(\alpha \cdot \frac{1+\varepsilon}{1-\varepsilon})$-近似解。

研究表明，在机器学习领域中保证全局近似的传统核心集需要相当大的采样复杂度——逻辑回归这种简单的线性模型也至少需要 $\Omega(n/\operatorname{polylog}(n))$ 数量的样本 (Mai et al., 2021)。为了解决这个问题，研究者们提出了局部核心集 (Ding et al., 2020; Huang et al., 2021)。为了描述局部核心集的概念，我们假设可行域 $\mathbb{S}$ 被赋予度量，并记 $\mathbb{S}$ 中以 S 为球心，R 为半径的度量球为 $\mathcal{B}(S, R)$.

定义 2.5 (局部 ε-核心集). 对于一个由数据集，损失函数以及可行域组成的三元组 $(P_n, \mathcal{L}, \mathbb{S})$, 给定 $R > 0, S_{anc} \in \mathbb{S}$, 它的局部 ε-核心集是一个非负稀疏权重向量 $W = [w_1, w_2, \ldots, w_n]$，使得对应的加权集合 $\tilde{P}_n$ 满足

$$\mathcal{L}(\tilde{P}_n, S) \in (1 \pm \varepsilon) \cdot \mathcal{L}(P_n, S), \forall S \in \mathcal{B}(S_{anc}, R). \tag{2.5}$$

不难发现，局部核心集就是将核心集的全局近似要求放松到有界度量球内近似。研究显示局部核心集的采样复杂度可以缩减至 $\mathcal{O}(\text{polylog}(n))$(Ding et al., 2020; Huang et al., 2021)。

2.1.1　重要性采样

重要性采样 (Langberg et al., 2010) 最早是一种用来计算蒙特卡洛积分的方法，通过选取合适的重要性函数作为采样密度，这种方法能够有效减小计算的方差。Feldman et al. (2011) 用重要性采样来构建聚类问题的核心集，他们的方法被称为 *Feldman-Langberg* 框架，并不断被后来的学者研究改进和推广 (Braverman et al., 2016; Huang et al., 2020; Tukan et al., 2020)。在具体介绍 *Feldman-Langberg* 框架之前，我们给出有限和损失函数的定义。

定义 2.6 (有限和形式的损失函数). 我们称损失函数 $\mathcal{L}(P_n, S)$ 具有有限和形式，如果它可以分解为某个损失函数 $\ell(p, S)$ 的加权求和：

$$\mathcal{L}(P_n, S) = \sum_{p \in P_n} w(p) \cdot \ell(p, S), \tag{2.6}$$

这里 $w(p)$ 是 p 在 P_n 中的权重。

我们总结由Braverman et al. (2016) 改进的 *Feldman-Langberg* 框架如下：

1. 估计敏感度函数的一个上界 $\sigma : P_n \to \mathbb{R}$ 使得 $\sigma(p) \geq \sup_{S \in \mathbb{S}} \frac{\ell(p,S)}{\sum_{q \in P_n} \ell(q,S)}$;
2. 计算 $t := \sum_{p \in P_n} \sigma(p)$，令 p 被选中的概率为 $\sigma(p)/t$，进行独立同分布采样获得集合 D，并对每个 $p \in D$ 赋权重 $w(p) = \frac{t}{|D| \cdot \sigma(p)}$;
3. 输出 $\tilde{P}_n = D$ 为核心集。

为了给出这种方法的采样复杂度，我们介绍三元组 $(\mathbb{P}, \ell, \mathbb{S})$ 的 VC 维：

定义 2.7 (VC 维). 对于三元组 $(\mathbb{P}, \ell, \mathbb{S})$，$S \in \mathbb{S}$ 以及 $r \geq 0$, 我们定义

$$\text{ranges}(S, r) = \{p \in \mathbb{P} \mid \ell(p, S) \leq r\},$$

那么 $(\mathbb{P}, \ell, \mathbb{S})$ 的 *VC* 维 d_{VC} 是满足如下式的最大子集 $Q \subset P$ 的大小：

$$|\{Q \cap \text{ranges}(S, r) \mid S \in \mathbb{S}, r \geq 0\}| = 2^{|Q|}.$$

Braverman et al. (2016) 给出了一个 *Feldman-Langberg* 框架的采样复杂度上界如下：

定理 2.8 ((Braverman et al., 2016)). *令 $\varepsilon, \delta \in (0,1/2)$, $k \geq 1$ 且 $z \geq 1$，如果采样复杂度 $|D| = \mathcal{O}\left(\frac{t}{\varepsilon^2}(d_{\mathrm{VC}} \cdot \log t + \log(1/\delta))\right)$，其中 d_{VC} 为 $(\mathbb{P}, \ell, \mathbb{S})$ 的 VC 维，t 为总敏感度的一个上界，那么以至少 $1-\delta$ 的概率，上述 Feldman-Langberg 框架输出的 $\tilde{P}_n$ 为优化问题 (2.1) 的一个 ε-核心集。*

重要性采样方法的采样复杂度依赖于问题的 VC 维，而现实问题的 VC 维可能非常高，并且难以估计。比如Bartlett et al. (2019) 给出了 Relu 激活函数的神经网络的 VC 维的上下界为 $cWL\log(W/L) \leq d_{\mathrm{VC}} \leq C \cdot WL\log(WL)$，这里 W 是网络的总参数量，L 是网络的深度，c 和 C 是两个常数。

值得注意的是，重要性采样方法的采样复杂度不仅依赖于问题的 VC 维，还依赖于对敏感度上界的精确估计。敏感度估计要求的技术性比较强，不同的损失函数可能需要不同的理论工具。Munteanu et al. (2018) 研究了逻辑回归的敏感度函数并提出了一种数据依赖的复杂度度量：μ-复杂度。他们用 μ-复杂度给出了逻辑回归的敏感度函数上界估计，如果使用他们的估计，重要性采样构建的逻辑回归核心集的采样复杂度为 $\mathcal{O}\left(\frac{\mu^3}{\varepsilon^4}d^3\log^2(\mu nd)\log^2 n(\log\log n)^4\right)$，这个结果后来被Mai et al. (2021) 改进至 $\tilde{\mathcal{O}}\left(\frac{\mu^2 d}{\varepsilon^2}\right)$. 值得注意的是可以人为构造数据集的使得它的 μ-复杂度任意大，而目前还不确定现实数据集是否大多具有低 μ-复杂度。Tukan et al. (2020) 把 SVD 分解从矩阵推广至函数，得到了被称作 f-SVD 的工具，并且证明了对于一类 "近凸" 损失函数 f，可以通过 f-SVD 分解获得它的一个敏感度上界 $t = \frac{2c_2}{c_1} + \frac{cc_2}{c_1}\max\left\{n^{1-z}, 1\right\}\alpha^z d$，这里 c, c_1, c_2, α, z 是描述 "近凸" 损失函数需要的因子。特别地，对于 ℓ_2-正则化逻辑回归，$t = \mathcal{O}(d\sqrt{n})$，总采样复杂度为 $\mathcal{O}\left(\frac{d\sqrt{n}}{\varepsilon^2}(d\log(d\sqrt{n}) + \log(1/\delta))\right)$.

2.1.2　分层采样

Chen (2009) 给出了一种使用分层采样构建 k-median 和 k-means 聚类核心集的方法。在介绍他的方法之前，我们先给出一般度量空间中这两种 k-聚类问题的定义。

定义 2.9 (k-median/means 聚类). *假设 $\mathbb{P}$ 被赋予度量 $\mathrm{dist}(\cdot,\cdot)$ 成为一个完备度量空间，对有限集合 $S \subset \mathbb{P}$ 和 $p \in \mathbb{P}$，我们定义 $\mathrm{dist}(p,S)$ 为 p 到 S 中最近点的距离。那么我们可以定义 (k,z)-聚类的为损失函数*

$$\mathcal{L}_z(P_n, S) := \sum_{p \in P_n} w(p) \cdot \mathrm{dist}^z(p,S). \tag{2.7}$$

给定 (加权) 数据集合 P_n，(k,z)-聚类是找到一个包含 k 个点的聚类中心集合 $S \subset \mathbb{P}_{\mathbf{n}}$ 最小化 $\mathcal{L}_z(P_n,S)$ 的值。特别地，这个问题在 $z=1$ 时被称为 k-median 聚类，$z=2$ 时被称为 k-means 聚类。

高维欧式空间下的 k-median/means 聚类精确求解是非常困难的。事实上，对