arXiv:2304.00459v1 [cs.LG] 2 Apr 2023

过参数化与 Polyak-ojasiewicz 条件下随机重排的快速收敛

陈帆[1],Christos Thrampoulidis[2], 和 Mark Schmidt[1]

[1] 计算机科学系[2] 不列颠哥伦比亚大学电气与计算机工程系 加拿大不列颠哥伦比亚省温哥华{fanchen2,schmidtm}@cs.ubc.ca,{cthrampo@ece.ubc.ca}

摘要。 现代机器学习模型通常是过参数化的，因此能够插值训练数据。在这种情形下，我们研究一种无放回抽样的随机梯度下降（SGD）变体——随机重排（RR）的收敛性质。与每次迭代都进行有放回抽样的 SGD 不同，RR 在每个周期开始时随机排列数据，并在每次迭代时从该排列中选取下一个样本。对于欠参数化模型，已有研究表明，在某些假设下，RR 的收敛速度可以快于 SGD。然而，除了一些限制性很强的情形外，以往研究并未表明 RR 在过参数化情形下优于 SGD。对于 Polyak-Łojasiewicz（PL）函数类，我们证明，当满足以下任一条件时，RR 在过参数化情形下可以优于 SGD：（i）样本数（n）小于条件数（κ）与弱增长条件（WGC）参数（α）的乘积；或（ii） n 小于强增长条件（SGC）的参数（ρ）。

1 引言

我们考虑如下形式的有限和最小化问题

$$\min\left\{f(x)=\frac{1}{n}\sum_{i=1}^{n}f(x;i)\right\}. \tag{1}$$

随机梯度下降（SGD）是解决这类机器学习问题的一种常用算法。人们已投入大量精力研究其理论和实证性质 [5]。SGD 的更新规则很简单：每次迭代时，从样本中有放回地随机选取一个样本 i_k ，并计算 $x^{k+1}=x^k-\eta^k\nabla f(x^k;i_k)$。这比每次迭代都使用全梯度的成本更低。然而，众所周知，SGD 的收敛速率可能远逊于全梯度下降的收敛速率。例如，对于强凸函数，SGD 的收敛速率是次线性的，而全梯度下降则具有线性收敛速率。

鉴于现代学习模型日益复杂，一个具有实际意义的问题是：在过参数化情形下，SGD 的表现如何？在这种情形下，模型能够插值（精确拟合）数据。此前已有研究表明，对于强凸函数，在多种插值条件下，SGD 能像全梯度下降一样实现线性收敛速率 [24,35,26,37]，例如满足强增长条件（SGC）或弱增长条件（WGC）。Polyak-Łojasiewicz（PL）条件 [29] 是一种弱于强凸性的假设，在该假设下，全梯度下降仍能实现线性收敛速率。近年来，PL 条件在机器学习领域日益受到关注 [14]，且已有研究表明，若干能够插值数据的过参数化模型满足 PL 条件 [2]。与强凸情形类似，在插值条件和 PL 条件下，SGD 能够实现与全梯度下降相近的线性收敛速率 [2,37]。

随机梯度下降（SGD）的一种常见变体是随机重排（RR）。在每个周期 $t \in \{1, 2, ..., T\}$，算法 RR 都会随机选择一个排列 π^t。也就是说，从集合 $\{1, 2, ..., n\}$ 中为 $j \in \{0, 1, ..., n-1\}$ 不放回地抽取 π^t_{j+1}。然后，它按照 π^t 指定的顺序遍历数据集，并执行以下更新。

$$x^t_{j+1} = x^t_j - \eta^t_j \nabla f(x^t_j; \pi^t_{j+1}). \tag{2}$$

注意，$x_0 \triangleq x^1_0$ 是初始化，$x^{t+1}_0 = x^t_n \ \forall t \geq 1$。我们在算法1中总结了该方法。RR 的一种变体是增量梯度（IG）方法，其中 π^t 在所有周期中都是确定且固定的。

算法1 随机重排（RR）

输入： x_0, T, 步长 $\{\eta^t_j\}$

对于 $t = 1, 2, \ldots, T$ **do**
  从所有排列的集合中选择一个排列 π^t；若 $t = 1$，则设置 $x^t_0 = x_0$；若 $t > 1$，则设置 $x^t_0 = x^{t-1}_n$。
  对于 $j = 1$、2、.....、$n-1$ **执行**
    $x^t_{j+1} = x^t_j - \eta^t_j \nabla f(x^t_j; \pi^t_{j+1})$
  结束循环
结束循环

众所周知，对于某些问题，随机重排（RR）在经验上比随机梯度下降（SGD）收敛得更快 [3,4]。然而，随机重排的分析比随机梯度下降更困难，因为在给定过去迭代点的条件下，每个单独的梯度都不再是全梯度的无偏估计。因此，关于随机重排的分析直到近期一系列研究中才开始出现 [13,25,33,12]。先前的研究表明，在各种欠参数化设定中，当训练轮次（T）足够多时，对于强凸目标函数，随机重排的表现优于随机梯度下降。然而，随机重排的这些次线性收敛速率比过参数化设定下随机梯度下降所达到的线性收敛率更慢。此外，目前关于过参数化设定下随机重排的收敛速率分析

要么基于不切实际的假设（见下一节），要么除非作出非常强的假设，否则其收敛速度也比随机梯度下降慢。

在这项工作中，我们研究了在 SGC 和 WGC 过参数化假设下，PL 函数的随机重排和增量梯度。我们给出了随机重排的显式收敛速率；在过参数化最小二乘等实际情形所满足的现实假设下，该速率可以快于目前已知的随机梯度下降最佳速率。我们的结果还表明，在某些情况下，增量梯度的收敛速度可以快于随机梯度下降（但不及随机重排）。我们还研究了 SGC 和 WGC 的放宽形式，其中插值仅近似满足；结果表明，随机重排能够更快地收敛到解的邻域。

2 相关工作

PL 条件下的优化：PL不等式最早由 Polyak [29]和 -Lojasiewicz [19]研究。它适用于许多重要的机器学习问题，例如在紧集上的最小二乘和逻辑回归[14]。更一般地，任何具有如下形式的函数 $f(x) = g(Ax)$，其中矩阵 A 具有 μ-强凸函数 g，都满足 μ-PL 条件 [14]。近期多项研究指出，围绕极小点的局部 PL 条件可用于建模分析随机梯度下降在神经网络训练中的有效性 [9,17,28,36]。

Polyak [29] 证明了，在 PL 条件下，对光滑目标函数进行全梯度下降可以实现线性收敛速率。但近期研究指出，PL 条件可用于证明多种方法具有线性收敛速率 [14]。通常，PL 条件带来的收敛速率与更强的强凸性条件下得到的相似。对于插值情形下的 SGD，已有研究表明，在 PL 条件下 SGD 的收敛速率是线性的 [2,37]。然而，对于 μ-PL 函数，插值情形下 SGD 的收敛速率慢于强凸函数的收敛速率（见表 1）。

随机重排在强凸欠参数化问题中的应用：本段将重点讨论强凸函数的情形，这是该主题大多数文献研究的对象。Bottou 推测，随机重排的收敛速率为 $O(\frac{1}{n^2T^2})$ [3]，，其中 T为训练轮次。但经过数年，研究才在证明这一点上取得进展。随机重排分析的难点源于梯度条件期望存在偏差，

$$\mathbb{E}\left[\nabla f(x_i^t; \pi_{i+1}^t) \mid x_i^t\right] \neq \nabla f(x_i^t). \tag{3}$$

早期分析随机重排（RR） [32] 的尝试并未成功，因为他们关于非交换算术—几何平均不等式的猜想被证明是错误的[15]。G¨urb¨uzbalaban 等人最早渐近地证明了一个 $\widetilde{\mathcal{O}}(\frac{1}{n^2T^2})$ 收敛速率[12]。Haochen 和 Sra [13] 在强凸常数为 μ 的条件下，在较强的假设下首次给出了 $\tilde{\mathcal{O}}(\frac{1}{\mu^4}(\frac{1}{n^2T^2} + \frac{1}{T^3}))$ 的非渐近收敛结果。对于较弱的假设，当 $T \gtrsim \frac{1}{\mu^2}$ 时，已有研究证明了 $\tilde{\mathcal{O}}(\frac{1}{\mu^3 nT^2})$ 的收敛速率

其依据是每个 $f(\cdot; i)$ [25], 都具有逐分量凸性，这与 $\Omega(\frac{1}{nT^2})$ 的下界相匹配 [31]。[3] 请注意，当满足较大的轮次要求时，这一收敛速率快于随机梯度下降（SGD）的 $\tilde{\mathcal{O}}(\frac{1}{nT})$ 收敛速率。Mischenko 等人 [22] 得到了相同的 $\tilde{\mathcal{O}}(\frac{1}{\mu^3 nT^2})$ 收敛速率，但只要求 $T \gtrsim \frac{1}{\mu}$。他们的分析依赖于底层的逐分量凸结构。Ahn 等人 [1]消除了这种依赖，并在 $T \gtrsim \frac{1}{\mu}$ 条件下得到了相同的收敛速率。然而，他们的分析要求每个 $f(\cdot; i)$ 都是 G-Lipschitz 连续的（对所有 i, $\|\nabla f(\cdot; i)\| \leq G$），这可能需要对问题（1）施加约束，并且需要进行投影操作以确保迭代点有界 [1,27]。Nguyen 等人 [27] 对随机重排之外的洗牌方案给出了统一分析；而 Safran 和 Shamir [33]在 f 是 n 个二次函数之和时，证明了一个 $\Omega(\frac{1}{\mu n^2 T^2} + \frac{1}{\mu n T^3})$ 下界。在一项较新的研究中，他们表明，在最坏情况下，除非 T 大于 $\Theta(\frac{1}{\mu})$，否则随机重排并不会显著优于随机梯度下降 [34]。我们还可以考虑其他洗牌方案；Lu 等人 [20] 设计了 GraB 算法，通过将先前的梯度信息用于确定当前排列，实现了更快的 $\tilde{\mathcal{O}}(\frac{1}{\mu^3 n^2 T^2})$ 收敛速率。此外，Cha 等人 [7]给出了该算法的匹配下界。需要指出的是，所有这些次线性收敛速率都慢于过参数化问题中随机梯度下降可能达到的线性收敛率。

强凸过参数化问题中的随机重排： 尽管随机重排（RR）广泛用于过参数化模型的训练，但分析这种设定的文献相对较少。Haochen 和 Sra [13] 证明，即使没有任何轮次要求，在插值条件下，随机重排的收敛速率也至少与随机梯度下降（SGD）一样快。然而，他们的结果并未表明两者的收敛速率可能存在差距，而且只适用于每个函数都强凸的退化情形。[4] Ma 和 Zhou [21]证明，在过参数化条件和“弱强凸性”假设下，随机重排的收敛速率快于随机梯度下降；但他们的结果也有些退化，因为其假设排除了过参数化最小二乘等标准问题。[5] 我们可以将 Mischenko 等人 [22]和 Nguyen 等人 [27]的结果作为特例，在更现实的假设下获得插值结果。然而，在插值设定下，这些关于凸函数和强凸函数的研究所得到的收敛速率，慢于 Vaswani 等人 [37] 为随机梯度下降证明的速率（我们在表 1 中总结了过参数化条件下已知的线性收敛率）。

[3] 按照文献中的惯例 [25,34,7],，我们采用 $\kappa = \Theta(\frac{1}{\mu})$进行比较。[4] 在这种情况下,– 我们只需对任意一个单独的函数应用梯度下降，并忽略所有其他训练样本，即可解决该问题。[5] 通常的“弱强凸性”假设是，对于每个 f_i，相对于 f_i 的极小值集合进行投影时，强凸性不等式成立；最小二乘问题满足这一条件。但 Ma 和 Zhou 改为投影到该集合与相对于 f 的极小值集合的交集上，而这对于最小二乘问题通常并不成立。

Polyak-Lojasiewicz 问题中的随机重排：已有多项研究给出了随机重排在 PL 问题上的收敛速率 [13,1,22,27]。Haochen 和 Sra 首次为欠参数化情形给出了非渐近收敛速率 $\tilde{\mathcal{O}}(\frac{1}{\mu^4}(\frac{1}{n^2T^2}+\frac{1}{T^3}))$，但要求每个单独的函数都具有 Lipschitz 连续海森矩阵 [13]。Ahn 等人将速率提升至 $\tilde{\mathcal{O}}(\frac{1}{\mu^3 nT^2})$，条件是 $T \gtrsim \frac{1}{\mu}$ 假设梯度有界 [1]，但这些适用于欠参数化情形的次线性收敛速率，慢于随机梯度下降在过参数化问题中的线性收敛率。Mischenko 等人关于 μ-PL 函数 [22] 的结果可用于在过参数化条件下获得线性收敛率，[6] 但仅适用于所有函数具有相同梯度的退化情形。Nguyen 等人关于 PL 函数 [27, 的定理 1]讨论的是欠参数化情形。但在附录 E 中，我们展示了如何修改 Nguyen 等人 [27] 定理 1 的证明，从而得到随机重排在过参数化设定下的线性收敛速率。随机重排的这一收敛速率可能快于随机梯度下降，但适用情形比本文结果的限制更多（具体细节见第 4 节）。

3 假设

本节介绍我们在分析中采用的假设。我们的第一个假设是，单个函数和整体函数均为光滑且有下界。

假设 1（光滑性和有界性） 目标函数 f 是 L-光滑的，且每个单个损失函数 $f(\cdot;\ i)$ 都是 L_i-光滑的，使得 $\forall x, x' \in \mathrm{dom}(f)$,

$$\|\nabla f(x) - \nabla f(x')\| \le L\|x - x'\|, \tag{4}$$

$$\|\nabla f(x;i) - \nabla f(x';i)\| \le L_i\|x - x'\|, \forall i. \tag{5}$$

记 $L_{\max} \triangleq \max_i L_i$。我们假设 f 有下界 f^*，且该下界在某个 x^*处取得，因此 $f^* = f(x^*)$。我们还假设每个 $f(\cdot;\ i)$ 都有某个下界 f_i^*。

接下来，我们给出 Polyak-Lojasiewicz（PL）不等式的定义，该不等式以下界的形式约束梯度的大小，使其随 f 的值增大而增大。

假设 2（μ-PL） 存在某个 $\mu > 0$ 使得

$$f(x) - f^* \le \frac{1}{2\mu}\|\nabla f(x)\|^2 \quad \forall x \in \mathrm{dom}(f), \tag{6}$$

其中，f^* 是 f的最优值。

所有强凸函数都满足PL不等式。然而，许多重要函数也满足PL不等式，但它们不一定是强凸函数；线性预测函数 [14]对应的平方预测误差就是一个经典例子。PL函数可以是非凸的，但该条件意味着

[6] 取 Mischenko 等人定理 4 中的 $B = 0$ 。

凸类性，这是凸性的一种推广。对于光滑函数，凸类性等价于 f 的每个驻点都是全局最小值（不过，与强凸函数不同，驻点可能不止一个） [8]。近期文献表明，PL条件及其变体有助于分析各种复杂模型 [2,9,17,28,36]。

接下来，我们正式定义插值，这是过参数化所蕴含的优化问题的关键性质。

假设3（插值）我们处于插值情形，这意味着

$$\nabla f(x^*) = 0 \quad \Longrightarrow \quad \nabla f(x^*; i) = 0. \tag{7}$$

因此，我们所说的插值是指：相对于函数 f 的驻点，也是相对于各个函数 $f(\cdot; i)$ 的驻点。不过，我们不假设 x^* 是唯一的。我们的结果还考虑强增长条件或弱增长条件 [35,37]，描述各个梯度在 x 远离 x^* 时如何变化。

假设 4（强增长条件）存在一个常数 $\rho \geq 1$，使得以下条件成立

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x; i)\|^2 \leq \rho\|\nabla f(x)\|^2, \quad \forall x \in \mathrm{dom}(f). \tag{8}$$

假设5（弱增长条件）存在一个常数 $\alpha \geq 0$，使得以下条件成立

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x; i)\|^2 \leq 2\alpha L(f(x) - f(x^*)), \quad \forall x \in \mathrm{dom}(f), \tag{9}$$

其中，f 有下界 f^*，且 $f^* = f(x^*)$。

插值、强增长条件（SGC）和弱增长条件（WGC）之间存在密切关系。SGC 蕴含插值；对于光滑函数，SGC 还蕴含 WGC [23,引理 5]。此外，对于满足光滑性和 μ-PL（假设 2 和 1）的函数，若 f_i 属于凸类，则这三个条件等价（SGC 蕴含插值，插值蕴含 WGC，WGC 又蕴含 SGC）。对于具有凸类 f_i的光滑 μ-PL 函数，WGC 中的常数 α 上界为 $\frac{L_{\max}}{L}$，并且根据 ρ [23, 引理 5-6],[7]，SGC 中的常数 ρ 上界为 $\frac{\alpha L}{\mu}$ [37, 命题 1]，因此也有上界 $\frac{L_{\max}}{\mu}$。SGC 中的 ρ 下界为 1；附录 A 中我们证明，WGC 中的 α 下界为 $\frac{\mu}{L}$。

一个经典情形是：对于数据线性可分的二分类问题，采用线性分类器时，我们预期这些插值性质成立。在这种情形下，每个样本的梯度都可以

[7] 我们将该结果列于附录 A 中，因为它是在更强的假设下证明的， [37] 或者说， [22] 先前工作中已有不同表述。

因线性可分性而收敛到零（对于标准损失函数）。在附录 A 中，我们推广了已有的一项结果 [37, 引理 1]，证明在该线性可分情形下，平方合页损失和逻辑回归等一类函数满足强增长条件。

在我们结果的最一般形式中，我们考虑对强增长条件和弱增长条件的放宽版本，这些版本不要求数据被精确拟合 [30,6]。

假设 6 存在常数 $\rho \geq 0$ 和 $\sigma \geq 0$，使得以下条件成立：$\forall x \in \mathrm{dom}(f)$,

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;i)\|^2 \leq \rho\|\nabla f(x)\|^2 + \sigma^2. \tag{10}$$

假设 7 存在常数 $\alpha \geq 0$ 和 $\sigma \geq 0$，使得以下条件成立：$\forall x \in \mathrm{dom}(f)$,

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;i)\|^2 \leq 2\alpha L(f(x) - f(x^*)) + \sigma^2. \tag{11}$$

当 $\sigma = 0$ 时，这些假设可化为强增长条件和弱增长条件。我们注意到，松弛 WGC 与 Gower 等人提出的期望光滑性条件有关 [11]。我们还注意到，当 $\rho = 1$ 时，假设 6 可化为有界方差假设 $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;\, i) - \nabla f(x)\|^2 \leq \sigma^2$，该假设常用于随机梯度下降的分析 [5]。

4 贡献

表 1 总结了在插值假设（假设 3、4 和 5）下，随机梯度下降在强凸函数和 PL 函数上的收敛已知结果与新结果。从表中可以看到，对于强凸函数，现有的随机重排收敛率比随机梯度下降更慢。相比之下，对于 μ-PL 目标函数，本文展示的随机重排收敛率可能快于随机梯度下降的收敛率。我们的主要贡献总结如下：

WGC 下的随机重排：对于满足 WGC 的 μ-PL 函数，我们推导出随机重排的样本复杂度为 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}}{\mu}\sqrt{n\frac{\alpha L}{\mu}}\})$。相比之下，在这种情况下，随机梯度下降的样本复杂度为 $\tilde{\mathcal{O}}(\alpha\frac{L^2}{\mu^2})$（我们在附录 B 中证明了这一点）。因此，当 $\boldsymbol{\alpha\frac{L}{\mu} > n}$ 且 $L_{\max} \sim L$ 时，随机重排优于随机梯度下降，无需大量训练轮次。因此，对于满足 WGC 的 μ-PL 目标函数，当 Lipschitz 常数相近且问题病态程度足够高时，随机重排比随机梯度下降收敛更快。

表 1：各算法为获得一个 ϵ精度的解所需的梯度评估次数，其定义为 $\|x - x^*\|^2 \leq \epsilon$ 和 $f(x) - f(x^*) \leq \epsilon$分别对应于 μ-强凸和 μ-PL 目标函数。本工作的贡献是蓝色标出的收敛速率。(1)(2)(3)

	插值	WGC	SGC	插值+凸类
随机梯度下降/强凸	-	$\alpha\frac{L}{\mu}$ [37]	$\rho\frac{L}{\mu}$ [35]	$\frac{L_{\max}}{\mu}$ [26]
随机重排/强凸	-	-	-	$\frac{L_{\max}}{\mu}n$ [22]
SGD/PL	$\frac{L^2_{\max}}{\mu^2}$ [2]	$\alpha\frac{L^2}{\mu^2}$	$\rho\frac{L}{\mu}$ [37]	$\frac{L^2_{\max}}{\mu^2}$ [2]
IG/PL	$\frac{L^2_{\max}}{\mu^2}n$ [27]	$\frac{L_{\max}}{\mu}n\sqrt{\alpha\frac{L}{\mu}}$	$\frac{L_{\max}}{\mu}n(\sqrt{n}\vee\rho)$	$(\frac{L_{\max}}{\mu})^{3/2}n$
RR/PL	$\frac{L^2_{\max}}{\mu^2}n$ [27]	$\frac{L_{\max}}{\mu}\sqrt{n}(\sqrt{n}\vee\sqrt{\frac{\alpha L}{\mu}})$	$\frac{L_{\max}}{\mu}\sqrt{n}(\sqrt{n}\vee\sqrt{\rho})$	$\frac{L_{\max}}{\mu}\sqrt{n}(\sqrt{n}\vee\sqrt{\frac{L_{\max}}{\mu}})$

(1) 我们忽略数值常数和对数因子。注意，μ-SC 表示 μ-强凸。Interp、SGC 和 WGC 分别指假设3、4 和 5。注意，对于 WGC 常数 α，我们有 $\frac{\mu}{L} \leq \alpha \leq \rho$ [23, 附录 A]，；而对于 SGC 常数 ρ，我们有 $\max\{1, \alpha\} \leq \rho \leq \frac{\alpha L}{\mu}$ [37]；如果所有函数都是凸类，还可以得到 $\alpha \leq L_{\max}L$ 以及 $\rho \leq \frac{L_{\max}}{\mu}$ [23]。

(2)“插值+凸类”的结果假设每个函数都是凸类函数，而强凸结果 [26,22] 还假设每个函数都是凸函数。(

3)符号“$\vee$”表示取两个值中的最大值。

***RR* 在强增长条件下：** 在强增长条件下，我们给出了 RR 的样本复杂度$\tilde{\mathcal{O}}$(最大值$\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}}{\mu}\sqrt{n\rho}\}$)，优于随机梯度下降的 $\tilde{\mathcal{O}}(\rho\frac{L}{\mu})$，前提是$\boldsymbol{\rho > n}$且 $L_{\max} \sim L$。当梯度之间存在大量分歧时，$\boldsymbol{\rho > n}$这种情况就会发生。请注意，当 $\rho \leq \alpha\frac{L}{\mu}$时，强增长条件下的收敛速率可能快于弱增长条件下的收敛速率。例如，当$\alpha \sim 1$且 $L_{\max} \sim L$时，便可满足这一条件。

***WGC/SGC* 下的增量梯度：** 对于 μ-PL 目标函数，当$\boldsymbol{\rho > n^2}$ 且 $L_{\max} \sim L$ 满足 SGC 时，或当 $\boldsymbol{\alpha\frac{L}{\mu} > n^2}$ 且 $L_{\max} \sim L$满足 WGC 时，增量梯度也可能优于随机梯度下降。对于 μ-PL 目标函数，随机重排的收敛速率可能比增量梯度快 $\sqrt{n}$ 倍；而且不会更差，因为 $\rho \geq 1$ 且 $\alpha \geq \frac{\mu}{L}$（见附录中的命题 3）。

实验： 我们针对满足假设的目标函数进行了实验。实验结果表明，类似 $\boldsymbol{\rho > n}$ 的条件对于随机重排实现比随机梯度下降更快的收敛速率至关重要。对于过参数化问题，我们证明在实践中，即使在训练初期，随机重排也能比随机梯度下降更快收敛；但随着 n 增大，随机重排的优势会减小。

与现有速率的比较，适用于 μ-*PL* 函数： 附录 E 讨论了如何利用 Nguyen 等人 [27, 的定理 1] 来获得随机重排在过参数化 μ-PL 目标函数上的收敛速率。在插值（假设 3）条件下，Nguyen 等人的分析推出随机重排的速率为 $\tilde{\mathcal{O}}((\frac{L_{\max}}{\mu})^2 n)$，该速率

始终劣于随机梯度下降的速率 $\tilde{\mathcal{O}}((\frac{L_{\max}}{\mu})^2)$，也劣于我们在此假设下随机重排的速率 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, (\frac{L_{\max}}{\mu})^{3/2}\sqrt{n}\})$（不过，该速率要求假设每个函数均为凸类）。另一方面，在强增长条件（假设 4）下，Nguyen 等人的分析推出的速率为 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}}{\mu^2}\sqrt{n}\sqrt{n+\rho-1}\})$，但这劣于我们在强增长条件下得到的结果 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}}{\mu}\sqrt{n}\sqrt{\rho}\})$，而且只有在问题常数取值范围更窄时才快于随机梯度下降。

5 收敛结果

我们给出了在 SGC 和 WGC 条件下，针对 μ-PL 目标函数的 RR 和 IG 的收敛结果。下文采用常数步长 η。定理 1 和定理 2 的证明见附录 D。

定理 1（RR + μ-PL）。假设 f 满足假设 *1* 和假设 *2*，则当学习率 η 满足

$$\eta \leq \min\{\frac{1}{2nL_{\max}}, \frac{1}{2\sqrt{2}L_{\max}\sqrt{n\rho}}\} \quad under\ Assumption\ 6,$$

or

$$\eta \leq \min\{\frac{1}{2nL_{\max}}, \frac{1}{2\sqrt{2}L_{\max}\sqrt{n}\sqrt{\frac{\alpha L}{\mu}}}\} \quad under\ Assumption\ 7, \tag{12}$$

时，对于 μ-*PL* 目标函数，可达到以下收敛速率：

$$\mathbb{E}[f(x_n^T) - f(x^*)] \leq (1 - \frac{1}{4}n\mu\eta)^T(f(x^0) - f(x^*)) + \frac{4L_{\max}^2\eta^2 n\sigma^2}{\mu}. \tag{13}$$

注释 *1*。在 (13) 中令 $\sigma = 0$，并利用附录 C 中的引理 5 求解递推关系，则 SGC 条件下的样本复杂度为 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}}{\mu}\sqrt{n}\sqrt{\rho}\})$，WGC 条件下的样本复杂度为 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}}{\mu}\sqrt{n}\sqrt{\frac{\alpha L}{\mu}}\})$。当 $n < \rho$ 或 $n < \frac{\alpha L}{\mu}$ 成立且 $L_{\max} \sim L$时，对于 μ-PL 目标函数，这些速率比相应假设下的 SGD 更快。当 $L_{\max}$ 与 L不相似时，若 SGC 条件下满足 $n < \frac{\rho L^2}{L_{\max}^2}$，或 WGC 条件下满足 $n < \frac{\alpha L^3}{\mu L_{\max}^2}$，RR 仍可能优于 SGD。

定理 2（增量梯度 + μ-PL）。假设 f 满足假设 *1* 和 *2*，则增量梯度采用满足以下条件的学习率 η

$$\eta \leq \min\{\frac{1}{\sqrt{2}nL_{\max}}, \frac{1}{2L_{\max}n\sqrt{\rho}}\} \quad under\ Assumption\ 6,$$

or

$$\eta \leq \min\{\frac{1}{\sqrt{2}nL_{\max}}, \frac{1}{2L_{\max}n\sqrt{\frac{\alpha L}{\mu}}}\} \quad under\ Assumption\ 7, \tag{14}$$

对于 μ-PL 目标函数，可达到以下收敛速率：

$$f(x_n^T) - f(x^*) \leq (1 - \frac{1}{2}n\mu\eta)^T (f(x^0) - f(x^*)) + \frac{2L_{\max}^2\eta^2 n^2\sigma^2}{\mu}. \quad (15)$$

注释 *2*。与注释 1 类似，我们可以采用相同的方法，得到在强增长条件下，增量梯度的样本复杂度为 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}}{\mu}n\sqrt{\rho}\}) = \tilde{\mathcal{O}}(\frac{L_{\max}}{\mu}n\sqrt{\rho})$，在弱增长条件下为 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}}{\mu}n\sqrt{\frac{\alpha L}{\mu}}\}) = \tilde{\mathcal{O}}(\frac{L_{\max}}{\mu}n\sqrt{\frac{\alpha L}{\mu}})$。对于 μ-PL 目标函数，由于强增长条件中的参数 ρ 满足 $\rho \geq 1$，而弱增长条件中的参数 α 满足 $\alpha \geq \frac{\mu}{L}$，因此这些收敛率比随机重排更差。

5.1 证明概要

本节概要介绍定理1和定理2在强增长条件下的证明。对于弱增长条件，我们采用类似的方法。首先，我们用训练轮次来表示目标函数 f 的下降量。对于步长 $\eta \leq \frac{1}{nL}$ [27]，有如下结论

$$f(x_0^{t+1}) \leq f(x_0^t) - \frac{n\eta}{2}\|\nabla f(x_0^t)\|^2 + \frac{L_{\max}^2\eta}{2}\sum_{i=0}^{n}\|x_i^t - x_0^t\|^2. \quad (16)$$

关键在于对一个周期内每一步相对于其起点的偏差和进行上界估计，即 $\sum_{i=0}^{n}\|x_i^t - x_0^t\|^2$。为了获得快速收敛速率，我们希望这一项足够小。为此，在 SGC 下，我们得到 IG 和 RR 的以下结果（见附录 C 中的引理3和引理4，其中令 $\sigma = 0$）。

$$\text{IG:} \quad \sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq 2\eta^2 n^2(n\rho)\|\nabla f(x_0^t)\|^2,$$

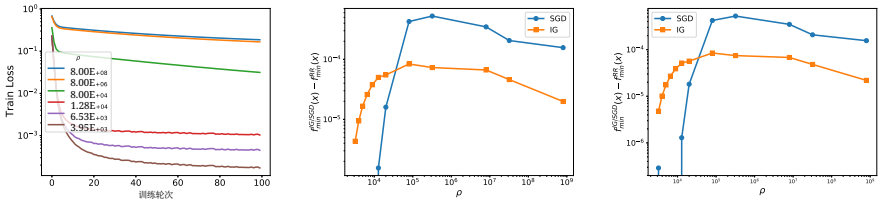
$$\text{RR:} \quad \mathbb{E}[\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2] \leq 2\eta^2 n^2(\rho + n)\mathbb{E}[\|\nabla f(x_0^t)\|^2].$$

忽略取期望的影响，RR 与 IG 的偏差界差异体现在因子 $\rho + n$ 和 $n\rho$ 上。将这些结果代入 (16) 后，我们得到

$$\text{IG:} \quad f(x_0^{t+1}) \overset{(a)}{\leq} f(x_0^t) - \frac{n\eta}{2}(1 - 2L_{\max}^2\eta^2 n^2\rho)\|\nabla f(x_0^t)\|^2, \quad (17)$$

$$\text{RR:} \quad \mathbb{E}[f(x_0^{t+1})] \overset{(b)}{\leq} \mathbb{E}[f(x_0^t)] - \frac{n\eta}{4}(1 - 4L_{\max}^2\eta^2 n\rho)\mathbb{E}[\|\nabla f(x_0^t)\|^2], \quad (18)$$

其中，(a) 和 (b) 都要求 $\eta \sim \tilde{\mathcal{O}}(\frac{1}{nL_{\max}})$。请注意，为了使 IG 在降低 f 方面取得进展，我们还要求 $\eta \sim \tilde{\mathcal{O}}(\min\{\frac{1}{nL_{\max}}, \frac{1}{L_{\max}n\sqrt{\rho}}\}) = \frac{1}{L_{\max}n\sqrt{\rho}}$），因为 $\rho \geq 1$；而对于 RR，我们要求 $\eta \sim \tilde{\mathcal{O}}(\min\{\frac{1}{nL_{\max}}, \frac{1}{L_{\max}\sqrt{n\rho}}\})$。因此，在 SGC 下，RR 的学习率可以高于 IG。



(a) 训练损失与训练轮次的关系　(b) 平方合页损失　(c) 平方合页 L_2 损失

图 1：在具有 $n = 800$的线性可分数据集上的二分类。(a) 我们绘制了使用平方合页损失时，RR 的训练损失随训练轮次变化的曲线，并比较了不同的 ρ取值。我们观察到，当 $\rho > \mathcal{O}(10^4)$时，训练速度非常慢；当 ρ 接近 n 的量级或更小时，训练速度会加快。(b) 我们绘制了 RR 与 SGD 或 IG 之间的最小训练损失差异随 ρ变化的曲线。对于平方合页损失这类凸目标函数，我们观察到，RR 相较于 SGD 的性能优势主要出现在 $\rho \gtrsim \mathcal{O}(10^4)$时。对于 $\rho < \mathcal{O}(10^4)$，SGD 的训练损失低于 RR。我们还观察到，对于所有 ρ 取值，RR 的表现始终优于 IG。(c) 对于 μ-PL 目标函数，也有与 (b) 类似的观察结果：采用 L_2 正则化的平方合页损失。

6 实验结果

本节旨在通过实验回答以下问题：① RR、SGD 和 IG 的经验收敛率如何依赖于 n和 ρ? ② 在插值条件下，RR 是否需要经过大量训练轮次才能超越 SGD? 除非另有说明，我们选择范围为 $[10^{-3}, 10^{-1}]$ 且能使训练损失最小化的恒定学习率。

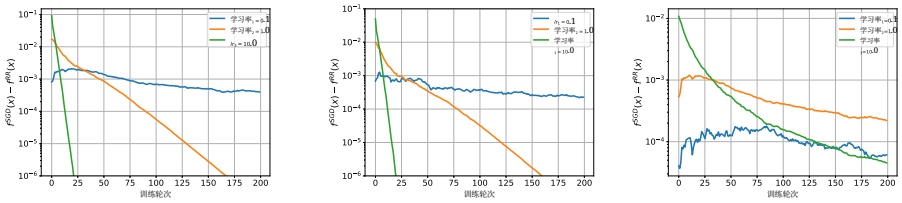
6.1 合成实验

我们首先在一个二分类合成数据集上进行实验。具体而言，我们使用 Loizou 等人提供的代码 [18]生成间隔为 $\tau > 0$ 的线性可分数据。训练时，我们采用带或不带 L_2 正则化的平方合页损失，分别模拟凸情形和 μ-PL 情形。注意，平方合页损失满足 SGC，且 $\rho \leq \frac{n}{\tau^2}$（见附录 A 中的命题 1）。因此，通过改变 τ，我们可以研究随机梯度下降、随机重排和增量梯度的收敛性如何随 SGC常数 ρ变化。结果如图 1 所示。与我们的界一致，我们观察到：(i) 当$\rho \gg n$时，随机重排优于随机梯度下降；(ii) 对于所有 ρ，随机重排都优于增量梯度；(iii) 随着 ρ 增大，收敛速度（自然地）变慢（这对应于间隔τ减小）。

6.2 使用径向基函数核的二分类

除了了解 ρ 的影响之外，我们还通过实验研究在插值条件下最小化凸损失时随机重排的收敛速率——

下的情况。为此，我们使用 LIBSVM 中的蘑菇数据集和径向基函数核 [18]，采用平方合页损失、平方损失和逻辑损失。我们按照 Vaswani 等人 [38] 的方法选择径向基函数核的带宽。正如我们的界所示，随机重排无需经过很多训练轮次就能优于随机梯度下降。如图 2 所示，随机重排在训练的前几个周期内便优于随机梯度下降；而且，训练早期的性能差距明显大于后期。



(a) 平方合页损失　　(b) 平方损失　　(c) 逻辑损失

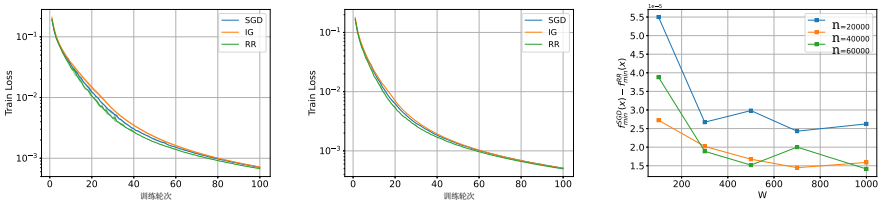
图 2：使用径向基函数核在蘑菇数据集上进行二分类。(a) 我们绘制了 SGD 与 RR 的平方损失差异随训练轮次的变化情况，对应的步长为 (lr) 0.1，1.0 和 10.0。(b)(c) 与 (a) 类似，分别展示平方损失和逻辑损失的情况。我们观察到，对于所有步长，RR 都比 SGD 收敛得更快，而且训练损失差异在训练初期更加显著。对于逻辑损失和 $lr = 10.0$，训练损失差距在一个周期后为 $\tilde{\mathcal{O}}(10^{-2})$，最终降至 $\tilde{\mathcal{O}}(10^{-4})$。同样，当使用平方合页损失或平方损失时，我们观察到训练损失差异在训练的前 25 个周期内下降得更快。

6.3 使用深度网络进行多分类

我们开展了实验，以研究插值条件下非凸设置中RR的收敛速率。为此，我们在MNIST图像分类数据集上，使用Softmax损失和具有不同宽度的单隐藏层多层感知机（1-MLP） [16]。图3中的结果表明，当 T较小时，RR的收敛速度可能快于SGD；较小的 n 有利于RR，而当 n 较大时，使用RR的优势则较小。

我们从实验结果中得到的关键结论是，对于过参数化问题：

1. RR 始终优于 IG。2. 条件 $n \ll \rho$ 对于 RR 优于 SGD 至关重要。3. 在过参数化条件下，RR 在训练早期就可以优于 SGD（不需要像此前针对欠参数化模型的理论所暗示的那样经过较大的周期）。



(a) 训练损失（$W = 100$） (b) 训练损失（$W = 1000$） (c) 不同 ns

图3：使用 Softmax 损失进行 MNIST 数字分类。(a) (b) 我们绘制了宽度分别为（W）100 和 1000 的单层多层感知机在不同训练轮次下的训练损失。在这两种情况下，我们观察到，训练初期随机重排的收敛速度都快于随机梯度下降。(c) 我们绘制了随机重排与随机梯度下降之间的最小训练损失差异，对比 n =20000、40000 和 60000 时不同 W 值的结果。尽管随机重排在所有 Ws 和 ns 下都优于随机梯度下降，但当 $n = 20000$ 时，性能差距更为显著。

7 结论

本文推导了插值条件下，在 SGC 或 WGC 对 μ-PL 目标函数所蕴含条件下 RR 的收敛速率。在此设置下，如果 SGC 下的关键条件 $\rho > n$ 或 WGC 下的关键条件 $\alpha \frac{L}{\mu} > n$ 成立（假设各个 Lipschitz 常数与整体Lipschitz常数相差不大），则 RR 的收敛速度快于 SGD。此外，我们证明，在 SGC 和 WGC 下，RR 分别对所有 ρ 和 α 的取值都优于 IG。除此之外，我们进一步证明，当 SGC 下满足 $\rho > n^2$ 或 WGC 下满足 $\alpha \frac{L}{\mu} > n^2$ 时，IG 可以优于 SGD。需要指出的是，在强凸性假设下的既有分析无法得出上述任何结论。我们的实验结果支持这些理论发现。

伦理声明 本文的贡献在于对现有算法进行理论分析，因此不会直接产生社会或伦理影响。

参考文献

1. Ahn, K., Yun, C., Sra, S.：随机打乱 SGD：无需分量凸性和较高轮次要求的最优速率。神经信息处理系统进展 **33**，17526–17535（2020） 2. Bassily, R., Belkin, M., Ma, S.：关于非凸过参数化学习中 SGD 的指数收敛。arXiv 预印本 arXiv:1811.02564（2018） 3. Bottou, L.：某些随机梯度下降算法为何能够实现异常快速的收敛。载于：学习与数据科学研讨会论文集，巴黎。第 8 卷，第 2624–2633 页（2009） 4. Bottou, L.：随机梯度下降技巧。载于：神经网络：行业诀窍，第 421–436 页。Springer（2012）

5. Bottou, L., Curtis, F.E., Nocedal, J.：大规模机器学习优化方法。Siam Review **60**(2)，223–311（2018） 6. Cevher, V., V˜u, B.C.：关于恒定步长随机梯度法的线性收敛。Optimization Letters **13**(5)，1177–1187（2019） 7. Cha, J., Lee, J., Yun, C.：随机打乱 SGD 的更紧下界：随机排列及其他方法。arXiv 预印本 arXiv:2303.07160（2023） 8. Craven, B.D., Glover, B.M.：凸类函数与对偶性。澳大利亚数学学会杂志 **39**(1)，1–20（1985） 9. Du, S.S., Zhai, X., Poczos, B., Singh, A.：梯度下降能够理论保证地优化过参数化神经网络。arXiv 预印本 arXiv:1810.02054（2018） 10. Garrigos, G., Gower, R.M.：随机梯度方法收敛定理手册。arXiv 预印本 arXiv:2301.11235（2023） 11. Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., Richt´arik, P.：SGD：一般分析与改进速率。载于：国际机器学习会议论文集。第 5200–5209 页。PMLR（2019） 12. G¨urb¨uzbalaban, M., Ozdaglar, A., Parrilo, P.A.：为什么随机重排优于随机梯度下降。Mathematical Programming **186**(1)，49–84（2021） 13. Haochen, J., Sra, S.：有限轮次后随机重排优于 SGD。载于：国际机器学习会议论文集。第 2624–2633 页。PMLR（2019） 14. Karimi, H., Nutini, J., Schmidt, M.：在 Polyak-Łojasiewicz 条件下梯度方法与近端梯度方法的线性收敛。载于：机器学习与数据库知识发现联合欧洲会议论文集。第 795–811 页。Springer（2016） 15. Lai, Z., Lim, L.H.：Recht-r´e 非交换算术–几何平均猜想是错误的。载于：国际机器学习会议论文集。第 5608–5617 页。PMLR（2020） 16. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.：应用于文档识别的基于梯度的学习。IEEE 会议录 **86**(11)，2278–2324（1998） 17. Liu, C., Zhu, L., Belkin, M.：过参数化非线性系统和神经网络中的损失景观与优化。Applied and Computational Harmonic Analysis **59**，85–116（2022） 18. Loizou, N., Vaswani, S., Laradji, I.H., Lacoste-Julien, S.：用于 SGD 的随机 Polyak 步长：实现快速收敛的自适应学习率。载于：人工智能与统计国际会议论文集。第 1306–1314 页。PMLR（2021） 19. Lojasiewicz, S.：实解析子集的一个拓扑性质。法国国家科学研究中心文集，《偏微分方程》**117**(87-89)，2（1963） 20. Lu, Y., Guo, W., Sa, C.D.：GRAB：寻找理论上优于随机重排的数据排列（2023） 21. Ma, S., Zhou, Y.：理解模型不相干性对随机重排增量随机梯度下降收敛的影响。载于：国际机器学习会议论文集。第 6565–6574 页。PMLR（2020） 22. Mishchenko, K., Khaled, A., Richt´arik, P.：随机重排：简单分析与大幅改进。神经信息处理系统进展 **33**，17309–17320（2020） 23. Mishkin, A.：插值、增长条件与随机梯度下降。博士论文，不列颠哥伦比亚大学（2020） 24. Moulines, E., Bach, F.：面向机器学习的随机近似算法非渐近分析。神经信息处理系统进展**24** （2011）

25. Nagaraj, D., Jain, P., Netrapalli, P.：无放回随机梯度下降：一般光滑凸函数的更优收敛速率。载于：国际机器学习会议。第 4703–4711 页。PMLR（2019） 26. Needell, D., Ward, R., Srebro, N.：随机梯度下降、加权抽样与随机化 Kaczmarz 算法。载于：神经信息处理系统进展 **27** （2014） 27. Nguyen, L.M., Tran-Dinh, Q., Phan, D.T., Nguyen, P.H., Van Dijk, M.：随机重排型梯度方法的统一收敛分析。机器学习研究杂志 **22**（1），9397–9440（2021） 28. Oymak, S., Soltanolkotabi, M.：过参数化非线性学习：梯度下降是否走最短路径？载于：国际机器学习会议。第 4951–4960 页。PMLR（2019） 29. Polyak, B.T.：泛函最小化的梯度方法。《苏联计算数学与数学物理》**3**（4），864–878（1963） 30. Polyak, B., Tsypkin, Y.Z.：伪梯度自适应与训练算法。《自动化与遥控》**34**，45–67（1973） 31. Rajput, S., Gupta, A., Papailiopoulos, D.：缩小无放回随机梯度下降的收敛差距。载于：国际机器学习会议。第 7964–7973 页。PMLR（2020） 32. Recht, B., R´e, C.：迈向非交换算术—几何平均不等式：猜想、案例研究与推论。载于：学习理论会议。第 11–1 页。JMLR 研讨会与会议论文集（2012） 33. Safran, I., Shamir, O.：随机重排下的随机梯度下降效果如何？载于：学习理论会议。第 3250–3284 页。PMLR（2020） 34. Safran, I., Shamir, O.：在病态问题上，随机重排只有经过许多训练轮次后才优于随机梯度下降。神经信息处理系统进展 **34**，15151–15161（2021） 35. Schmidt, M., Roux, N.L.：强增长条件下随机梯度下降的快速收敛。arXiv 预印本 arXiv:1308.6370（2013） 36. Soltanolkotabi, M., Javanmard, A., Lee, J.D.：过参数化浅层神经网络优化景观的理论见解。《IEEE 信息理论汇刊》**65**（2），742–769（2018） 37. Vaswani, S., Bach, F., Schmidt, M.：过参数化模型中随机梯度下降的快速与更快收敛，以及一种加速感知机。载于：第 22 届国际人工智能与统计会议。第 1195–1204 页。PMLR（2019） 38. Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G., Lacoste-Julien, S.：轻松随机梯度：插值、线搜索与收敛速率。神经信息处理系统进展 **32** （2019）

关于强增长条件和弱增长条件的更多结果

下文的命题将 Vaswani 等人 [37, 引理 1]推广到一个函数类，该函数类包含平方 hinge 损失或逻辑损失等特殊情形。

命题 1. 假设二元线性可分数据集 $(a_i, y_i)\ i \in [n]$ 的大小为 n $\tau = \max_{\|x\|_2=1} \min_{i \in [n]} y_i a_i^T x > 0$ $\|a_i\|_2 \leq$，间隔为 $\arg\max\min$，具有归一化特征和 $y_i \in \{\pm 1\}$。令 $f(;)\,(y_i a_i^T x)$ 为一个光滑单调函数。则强增长条件 (9) 在 $\rho \leq n/\tau^2$ 下成立。

证明。 我们考虑形式为 $f_i(x) = l(y_i a_i^T x) = l(z_i^T x)$ 的光滑单调非增函数，其中 $z_i = y_i a_i$，且 a_i 是第 i 个样本的特征向量。我们假设 z_i 的取值经过适当归一化，使得 $\max_i \|z_i\| \leq 1$。

定义 $x^* = \arg\max_{\|x\|_2=1} \min_i z_i^t x$ 和 $\tau = \max_{\|x\|_2=1} \min_i z_i^T x$。那么我们有

$$\|\frac{1}{n}\sum_{i=1}^n \nabla f(x;i)\|_2^2 = \|\frac{1}{n}\sum_{i=1}^n l^{'}(z_i^T x) z_i\|_2^2. \tag{19}$$

$$\tag{20}$$

注意，对于向量 u，其二范数为 $\|u\|_2 = \max_{\|v\|_2=1} v^T u$。因此，我们有如下

$$\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^n \nabla f(x;i)\|_2^2 &\geq \big(\frac{1}{n}\sum_{i=1}^n l^{'}(z_i^T x) z_i^T x^*\big)^2 \\
&= \tau^2\big(\frac{1}{n}\sum_{i=1}^n l^{'}(z_i^T x)\big)^2 \\
&= \tau^2\Big\{\frac{1}{n^2}\sum_{i=1}^n l^{'}(z_i^T x)^2 + \frac{1}{n^2}\sum_{i\neq j} l^{'}(z_i^T x) l^{'}(z_j^T x)\Big\} \\
&\geq \frac{\tau^2}{n}\Big(\frac{1}{n}\sum_{i=1}^n l^{'}(z_i^T x)^2\Big) \qquad (21) \\
&\geq \frac{\tau^2}{n}\Big(\frac{1}{n}\sum_{i=1}^n l^{'}(z_i^T x)^2 \|z_i\|^2\Big) \\
&= \frac{\tau^2}{n}\frac{1}{n}\sum_{i=1}^n \|\nabla f(x;i)\|^2, \qquad (22)
\end{aligned}$$

其中，(21) 中的不等式成立是因为 l 是单调的，也就是说 $l^{'}(t_1) l^{'}(t_2) \geq 0 \forall t_1,\ t_2 \in \mathbb{R}$。因此，我们有如下

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f(x;i)\|^2 \leq \frac{n}{\tau^2}\|\frac{1}{n}\sum_{i=1}^n \nabla f(x;i)\|_2^2,$$

因此我们得到 $\rho \leq \frac{n}{\tau^2}$。

下一个命题已在先前工作 [23] 中得到证明，但表述方式有所不同。

命题 2. 假设 f 满足假设 *1* 和假设 *4*，且常数为 $\rho \geq 1$，那么它也满足假设*5*，其中 $\alpha \leq \rho$。

证明. $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;\, i)\|^2 \overset{(a)}{\leq} \rho\|\nabla f(x)\|^2 \overset{(b)}{\leq} 2L\rho(f(x) - f(x^*))$，其中 (a) 基于假设 4，(b) 基于 L-光滑函数的一个标准结果（证明见 Garrigos 和 Gower [10,引理 2.28] ）。因此，当参数为 $\alpha \leq \rho$ 时，弱增长条件成立。

下面给出一个结果，为满足弱增长条件的光滑μ-PL 目标函数的参数 α 给出下界。

命题 3. 假设f满足假设 *2*、*1* 和 *5*，那么弱增长条件中的参数α 满足 $\alpha \geq \frac{\mu}{L}$。

证明。

$$2\mu(f(x) - f(x^*)) \overset{(a)}{\leq} \|\nabla f(x)\|^2 \overset{(b)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\|\nabla f(x;i)\|^2 \overset{(c)}{\leq} 2\alpha L(f(x) - f(x^*)), \tag{23}$$

其中，(a) 基于假设 2；(b) 基于詹森不等式和平方式范数的凸性；(c) 基于假设5。

B 插值下的随机梯度下降证明

定理 3（弱增长条件下的随机梯度下降 μ-PL）。假设 f 满足假设 *1* 和 *5*，则采用恒定学习率 $\eta = \frac{\mu}{L^2\alpha}$ 的随机梯度下降，对于 μ-*PL* 目标函数具有以下收敛速率：

$$\mathbb{E}[f(x^{K+1}) - f(x^*)] \leq (1 - \frac{\mu^2}{L^2\alpha})^K (f(x^0) - f(x^*)), \tag{24}$$

其中，K 是迭代总次数。

证明。 对 x^k取条件期望

$$\begin{aligned}
\mathbb{E}_k[f(x^{k+1})] &\overset{(a)}{\leq} f(x^k) - \langle \nabla f(x^k), \mathbb{E}_k[x^{k+1} - x^k]\rangle + \frac{L}{2}\mathbb{E}_k[\|x^{k+1} - x^k\|^2] \\
&\overset{(b)}{=} f(x^k) - \eta_k\|\nabla f(x^k)\|^2 + \frac{L\eta_k^2}{2}\mathbb{E}_k[\|\nabla f(x^k; i_k)\|^2] \\
&\overset{(c)}{\leq} f(x^k) - \eta_k 2\mu(f(x^k) - f(x^*)) + L^2\eta_k^2\alpha(f(x^k) - f(x^*)),
\end{aligned} \tag{25}$$

其中，(a) 基于假设 1；(b) 由 $\mathbb{E}[\nabla f(x^k; i_k)] = \nabla f(x^k)$ 得出；(c) 基于假设5和假设2。两边取全期望，并减去 $f(x^*)$

$$\begin{aligned}\mathbb{E}[f(x^{k+1}) - f(x^*)] &\leq [1 - 2\mu\eta_k(1 - \frac{L^2\eta_k\alpha}{2\mu})]\mathbb{E}[f(x^k) - f(x^*)] \\ &\overset{(d)}{=} (1 - \frac{\mu^2}{L^2\alpha})\mathbb{E}[f(x^k) - f(x^*)], \end{aligned} \tag{26}$$

其中，(d) 是通过代入 $\eta_k = \eta = \frac{\mu}{L^2\alpha}$ 得到的。然后递归求解该关系。

C 有用引理

我们借鉴 Nguyen 等人 [27] 和 Mishchenko 等人 [22] 的思路来构造证明。总体思路是先界定目标函数值的下降幅度（见引理2），然后分别界定增量梯度的进展项 $\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2$ 和随机重排的进展项 $\mathbb{E}[\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2]$（见引理3和引理4）。本节将介绍理论证明中要用到的这些引理。引理1是对 Mishchenk 等人 [22, 引理 1] 的重述。

引理1. 设 $X_1, ..., X_n$ 是 $\mathbb{R}^d$ 中给定的一组向量，其平均值记为 $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$，总体方差记为 $\sigma^2 = \frac{1}{n}\sum_{i=1}^n\|X_i - \bar{X}\|^2$。固定 $k \in \{1, ..., n\}$，设 $X_{\pi_1}, ..., X_{\pi_k}$ 是从 $\{X_1, ..., X_n\}$ 中不放回均匀抽样得到的样本，其平均值为 X_π。则以下结论成立

$$\mathbb{E}[\bar{X}_\pi] = \bar{X} \qquad \mathbb{E}[\|\bar{X}_\pi - \bar{X}\|^2] = \frac{n-k}{k(n-1)}\sigma^2. \tag{27}$$

引理2在 Nguyen 等人的 [27, 引理 8]中得到了证明。这里我们将 L 替换为 $L_{\max}$，并令步长 η 在所有迭代次数和训练轮次中保持恒定。

引理2. 假设假设 *1* 成立。给定洗牌方案 $\{\pi^t\}_t$ 和常数步长 η，满足$\eta \leq \frac{1}{nL}$，则有：

$$f(x_0^{t+1}) \leq f(x_0^t) - \frac{n\eta}{2}\|\nabla f(x_0^t)\|^2 + \frac{L_{\max}^2\eta}{2}\sum_{i=0}^n\|x_i^t - x_0^t\|^2. \tag{28}$$

对随机性取全期望

$$\mathbb{E}[f(x_0^{t+1})] \leq \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2}\mathbb{E}[\|\nabla f(x_0^t)\|^2] + \frac{L_{\max}^2\eta}{2}\sum_{i=0}^n\mathbb{E}[\|x_i^t - x_0^t\|^2]. \tag{29}$$

剩余的引理3-4讨论式（2）中的更新规则。回想一下，$x_0^{t+1} = x_n^t \ \forall t \geq 0$。还要回想一下，对于随机重排，置换向量 π^t 会在每个周期 t 随机选取，而对于增量梯度，它是固定的。引理3将与引理2结合，用于增量梯度的证明。引理4将与引理2结合，用于随机重排的证明。

引理3。 假设假设 *1* 成立。对于步长 $\eta \leq \frac{1}{\sqrt{2n}L_{\max}}$，在假设 *6* 下，~~，~~增量梯度满足

$$\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq 2\eta^2 n^3 \rho \|\nabla f(x_0^t)\|^2 + 2\eta^2 n^3 \sigma^2, \tag{30}$$

并且在假设 *7* 下，

$$\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq 4\eta^2 n^3 \alpha L (f(x_0^t) - f(x^*)) + 2\eta^2 n^3 \sigma^2. \tag{31}$$

证明。 根据更新规则

$$\begin{aligned}
\|x_i^t - x_0^t\|^2 &= \eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_j^t; j+1)\|^2 \\
&= \eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_j^t; j+1) - \nabla f(x_0^t; j+1) + \nabla f(x_0^t; j+1)\|^2 \\
&\overset{(a)}{\leq} 2\eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_j^t; j+1) - \nabla f(x_0^t; j+1)\|^2 + 2\eta^2 \|\sum_{j=0}^{i-1} \nabla f(x_0^t; j+1)\|^2 \\
&\overset{(b)}{\leq} 2\eta^2 i \sum_{j=0}^{i-1} \|\nabla f(x_j^t; j+1) - \nabla f(x_0^t; j+1)\|^2 + 2\eta^2 i \sum_{j=0}^{i-1} \|\nabla f(x_0^t; j+1)\|^2 \\
&\overset{(c)}{\leq} 2\eta^2 L_{\max}^2 i \sum_{j=0}^{i-1} \|x_j^t - x_0^t\|^2 + 2\eta^2 i \sum_{j=0}^{i-1} \|\nabla f(x_0^t; j+1)\|^2,
\end{aligned} \tag{32}$$

其中，(a) 和 (b) 基于詹森不等式和平方式范数的凸性，(c) 基于假设 1。对 $i = 0, ..., n-1$ 求和，得到

$$\begin{aligned}
\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 &\leq 2\eta^2 L_{\max}^2 \sum_{i=0}^{n-1} i \sum_{j=0}^{i-1} \|x_j^t - x_0^t\|^2 + 2\eta^2 \sum_{i=0}^{n-1} i \sum_{j=0}^{i-1} \|\nabla f(x_0^t; j+1)\|^2 \\
&\leq 2\eta^2 L_{\max}^2 \sum_{i=0}^{n-1} i \sum_{j=0}^{n-1} \|x_j^t - x_0^t\|^2 + 2\eta^2 \sum_{i=0}^{n-1} i \sum_{j=0}^{n-1} \|\nabla f(x_0^t; j+1)\|^2 \\
&\overset{(d)}{\leq} \eta^2 L_{\max}^2 n^2 \sum_{i=0}^{n-1} \|x_i^t - x_0^t\|^2 + \eta^2 n^3 [\frac{1}{n} \sum_{i=0}^{n-1} \|\nabla f(x_0^t; i+1)\|^2],
\end{aligned} \tag{33}$$

其中在 (d) 中我们使用了 $\sum_{i=0}^{n-1} i \leq \frac{n^2}{2}$。重新整理后得到：

$$\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq \frac{\eta^2 n^3}{1 - \eta^2 L_{\max}^2 n^2} [\frac{1}{n} \sum_{i=0}^{n-1} \|\nabla f(x_0^t; i+1)\|^2]. \tag{34}$$

选择学习率 $\eta \leq \frac{1}{\sqrt{2nL_{\max}}}$ 可得——

$$\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq 2\eta^2 n^3 [\frac{1}{n}\sum_{i=0}^{n-1}\|\nabla f(x_0^t; i+1)\|^2]. \tag{35}$$

若假设 6 成立，则可得

$$\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq 2\eta^2 n^3 \rho \|\nabla f(x_0^t)\|^2 + 2\eta^2 n^3 \sigma^2. \tag{36}$$

若假设 7 成立，则可得

$$\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2 \leq 4\eta^2 n^3 \alpha L(f(x_0^t) - f(x^*)) + 2\eta^2 n^3 \sigma^2. \tag{37}$$

引理4. 设假设 *1* 成立。对于步长 $\eta \leq \frac{1}{\sqrt{3nL_{\max}}}$，在假设 *6* 下，随机重排满足

$$\mathbb{E}\Big[\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2\Big] \leq 2n^2\eta^2(\rho+n)\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^2 \sigma^2, \tag{38}$$

在假设 *7* 下，

$$\mathbb{E}[\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2] \leq 4\eta^2 n^2 \alpha L \mathbb{E}[f(x_0^t) - f(x^*)] + 2\eta^2 n^3 \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^2 \sigma^2. \tag{39}$$

证明。 根据更新规则

$$\begin{aligned}
\|x_i^t - x_0^t\|^2 &= \eta^2 \|\sum_{j=0}^{i-1}\nabla f(x_j^t; \pi_{j+1}^t)\|^2 \\
&= \eta^2 i^2 \|\frac{1}{i}\sum_{j=0}^{i-1}[\nabla f(x_j^t; \pi_{j+1}^t) - \nabla f(x_0^t; \pi_{j+1}^t) + \nabla f(x_0^t; \pi_{j+1}^t) - \nabla f(x_0^t) + \nabla f(x_0^t)]\|^2 \\
&\overset{(a)}{\leq} 3\eta^2 i^2 \|\frac{1}{i}\sum_{j=0}^{i-1}[\nabla f(x_j^t; \pi_{j+1}^t) - \nabla f(x_0^t; \pi_{j+1}^t)]\|^2 + 3\eta^2 i^2 \|\frac{1}{i}\sum_{j=0}^{i-1}[\nabla f(x_0^t; \pi_{j+1}^t) \\
&\quad - \nabla f(x_0^t)]\|^2 + 3\eta^2 i^2 \|\nabla f(x_0^t)\|^2 \\
&\overset{(b)}{\leq} 3\eta^2 i \sum_{j=0}^{i-1}\|\nabla f(x_j^t; \pi_{j+1}^t) - \nabla f(x_0^t; \pi_{j+1}^t)\|^2 + 3\eta^2 i^2 \|\frac{1}{i}\sum_{j=0}^{i-1}[\nabla f(x_0^t; \pi_{j+1}^t) \\
&\quad - \nabla f(x_0^t)]\|^2 + 3\eta^2 i^2 \|\nabla f(x_0^t)\|^2 \\
&\overset{(c)}{\leq} 3\eta^2 i L_{\max}^2 \sum_{j=0}^{i-1}\|x_j^t - x_0^t\|^2 + 3\eta^2 i^2 \|\frac{1}{i}\sum_{j=0}^{i-1}\nabla f(x_0^t; \pi_{j+1}^t) - \nabla f(x_0^t)\|^2 \\
&\quad + 3\eta^2 i^2 \|\nabla f(x_0^t)\|^2,
\end{aligned} \tag{40}$$

其中 (a) 和 (b) 基于詹森不等式和平方范数的凸性；(c) 基于假设 1。令 σ^t 为迭代点 $\{x_0^1, ..., x_0^t\}$ 上的一个 σ-代数。我们对 σ^t 取条件期望，并应用引理1，得到

$$\begin{aligned}\mathbb{E}\Big[\|\frac{1}{i}\sum_{j=0}^{i-1}\nabla f(x_0^t;\pi_{j+1}^t)-\nabla f(x_0^t)\|^2|\sigma^t\Big] &= \frac{n-i}{i(n-1)}\frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(x_0^t;j+1)-\nabla f(x_0^t)\|^2\\ &= \frac{n-i}{i(n-1)}\big[\frac{1}{n}\sum_{j=0}^{n-1}[\|\nabla f(x_0^t;j+1)\|^2+\|\nabla f(x_0^t)\|^2-2\langle\nabla f(x_0^t;j+1),\nabla f(x_0^t)\rangle]\big]\\ &= \frac{n-i}{i(n-1)}\big[\frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(x_0^t;j+1)\|^2-\|\nabla f(x_0^t)\|^2\big]\\ &\le \frac{n-i}{i(n-1)}\frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(x_0^t;j+1)\|^2. \end{aligned}\tag{41}$$

对（40）取条件期望，并将（41）代回

$$\begin{aligned}\mathbb{E}\Big[\|x_i^t-x_0^t\||\sigma^t\Big] \le 3\eta^2 iL_{\max}^2\sum_{j=0}^{n-1}\mathbb{E}[\|x_j^t-x_0^t\||\sigma^t]+3\eta^2\frac{i(n-i)}{(n-1)}\frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(x_0^t;j+1)\|^2+\\ +3\eta^2 i^2\|\nabla f(x_0^t)\|^2.\end{aligned}\tag{42}$$

接下来，取全期望并对 $i=0,...,n-1$ 求和：

$$\begin{aligned}\mathbb{E}\Big[\sum_{i=0}^{n-1}\|x_i^t-x_0^t\|^2\Big] &\le 3\eta^2L_{\max}^2\mathbb{E}[\sum_{j=0}^{n-1}\|x_j^t-x_0^t\|^2]\sum_{i=0}^{n-1}i+3\eta^2\frac{1}{n-1}\mathbb{E}[\frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(x_0^t;j+1)\|^2]\sum_{i=0}^{n-1}i(n-i)+\\ &\quad+3\eta^2\mathbb{E}[\|\nabla f(x_0^t)\|^2]\sum_{i=0}^{n-1}i^2\\ &\overset{(d)}{\le}\frac{3}{2}\eta^2n^2L_{\max}^2\mathbb{E}[\sum_{j=0}^{n-1}\|x_j^t-x_0^t\|^2]+\eta^2n^2\mathbb{E}[\frac{1}{n}\sum_{j=0}^{n-1}\|\nabla f(x_0^t;j+1)\|^2]+\\ &\quad+\eta^2n^3\mathbb{E}[\|\nabla f(x_0^t)\|^2]\end{aligned}\tag{43}$$

在（d）中，我们用到了 $\sum_{i=0}^{n-1}i\le\frac{n^2}{2}$、$\sum_{i=0}^{n-1}i^2\le\frac{n^3}{3}$ 和 $\sum_{i=0}^{n-1}i(n-i)\le\frac{n^2}{(n-1)3}$。取 $\eta\le\frac{1}{\sqrt{3}L_{\max}n}$，可得

$$\mathbb{E}[\sum_{i=0}^{n-1}\|x_i^t-x_0^t\|^2]\le 2\eta^2n^2\mathbb{E}[\frac{1}{n}\sum_{i=0}^{n-1}\|\nabla f(x_0^t;j+1)\|^2]+2\eta^2n^3\mathbb{E}[\|\nabla f(x_0^t)\|^2]\tag{44}$$

若假设 6 成立，则可得

$$\mathbb{E}[\sum_{i=0}^{n-1}\|x_i^t-x_0^t\|^2]\le 2\eta^2n^2(\rho+n)\mathbb{E}[\|\nabla f(x_0^t)\|^2]+2\eta^2n^2\sigma^2.\tag{45}$$

若假设 7 成立，则可得

$$\mathbb{E}[\sum_{i=0}^{n-1}\|x_i^t - x_0^t\|^2] \leq 4\eta^2 n^2 \alpha L \mathbb{E}[f(x_0^t) - f(x^*)] + 2\eta^2 n^3 \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^2 \sigma^2. \tag{46}$$

引理 5。 考虑以下满足条件的递推关系，其中某个 $\eta \leq$ 最小值$\{\frac{1}{\eta_1}, \frac{1}{\eta_2}\}$ — —

$$\delta_T \leq (1 - \eta\mu n)^T \delta_0, \tag{47}$$

其中 T 是总周期数，因此样本复杂度为

$$\mathcal{O}(\max\{\eta_1, \eta_2\}\frac{1}{\mu} log(\frac{\delta_0}{\epsilon})) \tag{48}$$

证明。我们希望

$$\delta_T \leq exp(-\eta\mu n T)\delta_0 \overset{(a)}{=} exp(-\eta\mu K)\delta_0 < \epsilon. \tag{49}$$

注意，(a) 成立是因为 $T = \frac{K}{n}$，其中 K 是迭代总次数。因此，我们有

$$K \geq \frac{1}{\eta}\frac{1}{\mu} log(\frac{\delta_0}{\epsilon}) \geq \max\{\eta_1, \eta_2\}\frac{1}{\mu} log(\frac{\delta_0}{\epsilon}). \tag{50}$$

D 主要定理的证明

D.1 定理 1 的证明

证明。 我们首先证明假设6下的结论。将引理4中的（38）代入引理2中的（29）

$$\begin{aligned}
\mathbb{E}[f(x_0^{t+1})] &\leq E[f(x_0^t)] - \frac{n\eta}{2}\mathbb{E}[\|\nabla f(x_0^t)\|^2] + \frac{L_{\max}^2\eta}{2}(2n^2\eta^2(\rho+n)\mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^2\sigma^2) \\
&= E[f(x_0^t)] - \frac{n\eta}{2}(1 - 2L_{\max}^2 n^2\eta^2)\mathbb{E}[\|\nabla f(x_0^t)\|^2] + L_{\max}^2\eta^3 n^2\rho\mathbb{E}[\|\nabla f(x_0^t)\|^2] + L_{\max}^2\eta^3 n^2\sigma^2 \\
&\overset{(a)}{\leq} E[f(x_0^t)] - \frac{n\eta}{4}\mathbb{E}[\|\nabla f(x_0^t)\|^2] + L_{\max}^2\eta^3 n^2\rho\mathbb{E}[\|\nabla f(x_0^t)\|^2] + L_{\max}^2\eta^3 n^2\sigma^2 \\
&= E[f(x_0^t)] - \frac{n\eta}{4}(1 - 4L_{\max}^2\eta^2 n\rho)\mathbb{E}[\|\nabla f(x_0^t)\|^2] + L_{\max}^2\eta^3 n^2\sigma^2 \\
&\overset{(b)}{\leq} E[f(x_0^t)] - \frac{n\eta}{8}\mathbb{E}[\|\nabla f(x_0^t)\|^2] + L_{\max}^2\eta^3 n^2\sigma^2 \\
&\overset{(c)}{\leq} E[f(x_0^t)] - \frac{n\mu\eta}{4}\mathbb{E}[f(x_0^t) - f(x^*)] + L_{\max}^2\eta^3 n^2\sigma^2,
\end{aligned} \tag{51}$$

其中，在 (a) 中我们选择 $\eta \leq \frac{1}{2L_{\max}n}$，在 (b) 中我们选择 $\eta \leq \frac{1}{2\sqrt{2}L_{\max}\sqrt{n\rho}}$，
而 (c) 根据假设 2 得出。两边同时减去 $f(x^*)$

$$\mathbb{E}[f(x_0^{t+1}) - f(x^*)] \leq (1 - \frac{n\mu\eta}{4})\mathbb{E}[f(x_0^t) - f(x^*)] + L_{\max}^2\eta^3 n^2\sigma^2 \tag{52}$$

递归地求解,

$$\mathbb{E}[f(x_0^{T+1}) - f(x^*)] \leq (1 - \frac{n\mu\eta}{4})^T (f(x_0) - f(x^*)) + L_{\max}^2 \eta^3 n^2 \sigma^2 \sum_{j=0}^{T} (1 - \frac{n\mu\eta}{4})^j. \tag{53}$$

这意味着（回顾 $x_0^{T+1} = x_n^T$）：

$$\begin{aligned}
\mathbb{E}[f(x_n^T) - f(x^*)] &\leq (1 - \frac{n\mu\eta}{4})^T (f(x_0) - f(x^*)) + L_{\max}^2 \eta^3 n^2 \sigma^2 \sum_{j=0}^{\infty} (1 - \frac{n\mu\eta}{4})^j \\
&\leq (1 - \frac{n\mu\eta}{4})^T (f(x_0) - f(x^*)) + L_{\max}^2 \eta^3 n^2 \sigma^2 (\frac{4}{n\mu\eta}) \\
&= (1 - \frac{n\mu\eta}{4})^T (f(x_0) - f(x^*)) + \frac{4 L_{\max}^2 \eta^2 n \sigma^2}{\mu}.
\end{aligned} \tag{54}$$

接下来，我们在假设 7 下给出证明。将引理4中的式（39）代入引理2中的式（29）

$$\begin{aligned}
\mathbb{E}[f(x_0^{t+1})] &\leq \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2} \mathbb{E}[\|\nabla f(x_0^t)\|^2] + \frac{L_{\max}^2 \eta}{2} (4\eta^2 n^2 \alpha L \mathbb{E}[f(x_0^t) - f(x^*)] \\
&\quad + 2\eta^2 n^3 \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2\eta^2 n^2 \sigma^2) \\
&= \mathbb{E}[f(x_0^t)] - \frac{n\eta}{2} (1 - 2L_{\max}^2 \eta^2 n^2) \mathbb{E}[\|\nabla f(x_0^t)\|^2] \\
&\quad + 2L_{\max}^2 \eta^3 n^2 \alpha L \mathbb{E}[f(x_0^t) - f(x^*)] + L_{\max}^2 \eta^3 n^2 \sigma^2 \\
&\overset{(d)}{\leq} \mathbb{E}[f(x_0^t)] - \frac{n\eta}{4} \mathbb{E}[\|\nabla f(x_0^t)\|^2] + 2L_{\max}^2 \eta^3 n^2 \alpha L \mathbb{E}[f(x_0^t) - f(x^*)] + L_{\max}^2 \eta^3 n^2 \sigma^2 \\
&\overset{(e)}{\leq} \mathbb{E}[f(x_0^t)] - \frac{n\mu\eta}{2} \mathbb{E}[f(x_0^t) - f(x^*)] + 2L_{\max}^2 \eta^3 n^2 \alpha L \mathbb{E}[f(x_0^t) - f(x^*)] \\
&\quad + L_{\max}^2 \eta^3 n^2 \sigma^2 \\
&= \mathbb{E}[f(x_0^t)] - \frac{n\mu\eta}{2} (1 - \frac{4L_{\max}^2 \eta^2 n \alpha L}{\mu}) \mathbb{E}[f(x_0^t) - f(x^*)] + L_{\max}^2 \eta^3 n^2 \sigma^2 \\
&\overset{(f)}{\leq} E[f(x_0^t)] - \frac{n\mu\eta}{4} \mathbb{E}[f(x_0^t) - f(x^*)] + L_{\max}^2 \eta^3 n^2 \sigma^2,
\end{aligned} \tag{55}$$

其中，在 (d) 中我们选择 $\eta \leq \frac{1}{2L_{\max} n}$；(e) 由假设 2 得出；在 (f) 中我们选择 $\eta \leq \frac{\mu}{2\sqrt{2} L_{\max} \sqrt{n\alpha L^{\sigma}}}$ 然后从等式两边减去 $f(x^*)$，并按照上述方法求解递推关系，以得到最终结果。

D.2 定理2证明

证明。 我们首先给出假设6下的证明。将引理3中的 (30) 代入引理2中的 (28)

$$\begin{aligned} f(x_0^{t+1}) &\leq f(x_0^t) - \frac{n\eta}{2}\|\nabla f(x_0^t)\|^2 + \frac{L_{\max}^2\eta}{2}(2\eta^2 n^3\rho\|\nabla f(x_0^t)\|^2 + 2\eta^2 n^3\sigma^2) \\ &= f(x_0^t) - \frac{n\eta}{2}(1 - 2L_{\max}^2 n^2\eta^2\rho)\|\nabla f(x_0^t)\|^2 + L_{\max}^2\eta^3 n^3\sigma^2 \\ &\overset{(a)}{\leq} f(x_0^t) - \frac{n\eta}{4}\|\nabla f(x_0^t)\|^2 + L_{\max}^2\eta^3 n^3\sigma^2 \\ &\overset{(b)}{\leq} f(x_0^t) - \frac{n\mu\eta}{2}(f(x_0^t) - f(x^*)) + L_{\max}^2\eta^3 n^3\sigma^2, \end{aligned} \tag{56}$$

其中，(a) 中我们选择 $\eta \leq \frac{1}{2L_{\max}n\sqrt{\rho}}$，而 (b) 基于假设 2。在两边减去 $f(x^*)$

$$f(x_0^{t+1}) - f(x^*) \leq (1 - \frac{n\mu\eta}{2})(f(x_0^t) - f(x^*)) + L_{\max}^2\eta^3 n^3\sigma^2, \tag{57}$$

然后按照定理 1 的证明类似地求解递推式，可得

$$f(x_n^T) - f(x^*) \leq (1 - \frac{n\mu\eta}{2})(f(x_0) - f(x^*))^T + \frac{2L_{\max}^2\eta^2 n^2\sigma^2}{\mu} \tag{58}$$

在引理3中，我们还需要 $\eta \leq \frac{1}{\sqrt{2n}L_{\max}}$。因此，总体要求为 $\eta \leq \min\{\frac{1}{\sqrt{2n}L_{\max}}, \frac{1}{2L_{\max}n\sqrt{\rho}}\}$。接下来，我们在假设 7 下给出证明。将引理3中的 (31) 代入引理2中的 (28)

$$\begin{aligned} f(x_0^{t+1}) &\leq f(x_0^t) - \frac{n\eta}{2}\|\nabla f(x_0^t)\|^2 + \frac{L_{\max}^2\eta}{2}(4\eta^2 n^3\alpha L(f(x_0^t) - f(x^*)) + 2\eta^2 n^3\sigma^2) \\ &\overset{(a)}{\leq} f(x_0^t) - n\mu\eta(f(x_0^t) - f(x^*)) + 2L_{\max}^2\eta^3 n^3\alpha L(f(x_0^t) - f(x^*)) + L_{\max}^2\eta^3 n^3\sigma^2, \end{aligned} \tag{59}$$

其中，(a) 基于假设 2。在两边减去 $f(x^*)$，并重新整理各项

$$\begin{aligned} f(x_0^{t+1}) - f(x^*) &\leq (f(x_0^t) - f(x^*))[1 - n\mu\eta(1 - \frac{2L_{\max}^2\eta^2 n^2\alpha L)}{\mu}] + L_{\max}^2\eta^3 n^3\sigma^2 \\ &\overset{(b)}{\leq} (1 - \frac{n\mu\eta}{2})(f(x_0^t) - f(x^*)) + L_{\max}^2\eta^3 n^3\sigma^2, \end{aligned} \tag{60}$$

其中，在 (b) 中我们选择 $\eta \leq \frac{\sqrt{\mu}}{2L_{\max}n\sqrt{\alpha L}}$。因此，在这种情况下，我们要求 $\eta \leq \min\{\frac{1}{\sqrt{2n}L_{\max}}, \frac{\sqrt{\mu}}{2L_{\max}n\sqrt{\alpha L}}\}$。然后，按照定理 1 的证明类似地求解递推关系。

E Nguyen 等人关于过参数化 μ-PL 函数随机重排的收敛速率

本节将详细说明如何从 Nguyen 等人 [27, 定理 1] 中得到线性收敛速率。在假设 3[8], 下，为了得到形如 (47) 的关系，他们要求学习率 $\eta \leq \min\{\frac{1}{2L_{\max}n}, \sqrt{\frac{3}{8}}\frac{\mu}{L_{\max}^2 n}\}$。将引理 5 应用于 $\frac{1}{\eta_1} = \frac{1}{2L_{\max}n}$ 和 $\frac{1}{\eta_2} = \sqrt{\frac{3}{8}}\frac{\mu}{L_{\max}^2 n}$，此时得到的收敛速率为 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}^2}{\mu^2}n\}) = \tilde{\mathcal{O}}(\frac{L_{\max}^2}{\mu^2}n)$。类似地，在假设 4[9], 下，Nguyen 等人 [27, 定理 1] 所要求的学习率为 $\eta \leq \min\{\frac{1}{\sqrt{3}L_{\max}n}, \frac{\mu}{2L_{\max}\sqrt{n}\sqrt{n+\rho-1}}\}$。将引理 5 应用于 $\frac{1}{\eta_1} = \frac{1}{\sqrt{3}L_{\max}n}$ 和 $\frac{1}{\eta_2} = \frac{\mu}{2L_{\max}\sqrt{n}\sqrt{n+\rho-1}}$，即可得到收敛速率 $\tilde{\mathcal{O}}(\max\{\frac{L_{\max}}{\mu}n, \frac{L_{\max}}{\mu^2}\sqrt{n}\sqrt{n+\rho-1}\})$。

[8] 取 Nguyen 等人定理 1 中的 $\sigma^* = 0$。[9] 取 Nguyen 等人定理 1 中的 $\Theta = \rho - 1$ 和 $\sigma = 0$。