arXiv:2601.01160v1 [math.OC] 3 Jan 2026

BRAIn Lab

无梯度方法是高效应对马尔可夫随机性的关键

BorisProkhorov[1], *, **Semyon Chebykin**[2], *, **Alexander Gasnikov**[3,4,5], **Aleksandr Beznosikov**[6]1洛桑联邦理工学院2多伦多大学3因诺波利斯大学4莫斯科人工智能独立研究院5俄罗斯科学院斯捷克洛夫数学研究所6人工智能基础研究实验室（BRAIn Lab）

本文研究带有零阶预言机和马尔可夫噪声的随机优化问题。我们提出并分析了一种新颖的无导数方法，用于求解强凸光滑和非光滑情形下的问题，并采用单点和双点反馈预言机。通过使用随机批处理方案，我们证明，当底层噪声序列的混合时间 τ 小于问题的维度 d时，该方法的收敛估计不依赖于 τ。这一发现为应对马尔可夫随机性提供了一种高效方式：与其调用代价高昂的一阶预言机，不如使用零阶预言机。最后，我们还给出了相应的上界和下界，从而证实了结果的最优性。

1 引言

S随机性是许多优化问题的基本特征，在机器
l学习 [48, 28]领域中自然存在。随机梯度下降（随机梯度下降） [45] 及其加速变体 [38, 25] 已成为现代大型模型训练的首选
f算法。随机梯度下降的理论性质已在多种统计框架下得到广泛研究
u， [36, 24, 10, 56], 这些框架通常假设噪声相互独立
a且服从相同分布（独立同分布）。然而，在许多现实应用中——包括强化学习
(强化学习） [6, 16], 分布式优化 [35, 31], 和赌博机问题 [3]——噪声并非独立同分布，而是呈现出
c相关性或马尔可夫型结构。
F例如，在上述不断发展的强化学习领域中，与环境的序贯交互会引发状态——
d
噪声的相关结构，因此需要能够应对非独立同分布噪声的算法。过去十年间，尽管已有研究探讨了若干适用于马尔可夫随机预言机的基于梯度的方法 [14, 18], 但强化学习中的策略优化完全依赖奖励反馈，因此传统方法并不适用，因为无法获取一阶信息 [46, 9, 19]。零阶优化 （ZOO）方法专为解决此类问题而开发，适用于无法获取梯度或计算梯度的成本高得令人难以承受的场景。除强化学习外，ZOO 技术还广泛应用于对抗攻击生成 [8], 超参数调优 [47, 57], 连续老虎机 [7, 49] 及其他应用领域 [54, 33]。尽管关于 ZOO 的文献十分丰富，据我们所知，本研究是首个针对同时具有零阶信息和马尔可夫噪声的优化问题的研究，旨在为这两个领域交叉形成的一大类问题开发最优算法。

电子邮箱：boris.prokhorov@epfl.ch, s.chebykin@mail.utoronto.ca, gasnikov@yandex.ru, anbeznosikov@gmail.com*共同贡献。作者撰写本文时均在 BRAIn Lab 工作。

1.1 相关工作

◇**零阶方法** 是优化领域中最重要、历史最悠久的研究方向之一。零阶方法种类繁多，这里简要介绍几种，例如一维方法 [32, 42] 及其高维类比方法 [41],椭球算法 [58] 以及沿随机方向搜索 [4]。目前，ZOO 方法中最受欢迎、研究最深入的机制，是 [43, 20, 40]中所述的梯度有限差分近似。其思路很简单：查询两个足够接近的点，实质上等同于求出函数的方向导数值：

$$\langle \nabla f(x), e\rangle \approx \frac{f(x+te)-f(x)}{t} \approx \frac{f(x+te)-f(x-te)}{2t}, \tag{1}$$

其中 e 是一个随机方向。它可以是随机坐标、欧几里得球面上的向量，或高斯分布的一个样本。近似（1）反过来会导出 Nesterov 的梯度法或坐标算法 [39]。不过，两者之间存在若干差异：

- 首先，为了获得完整的梯度信息，该算法需要进行 d 次查询，而不是调用一次梯度预言机（此处 d 是 x的维度）。

- 其次，如果零阶优化预言机并不精确，即只能获得带噪声的函数值，那么当噪声分量无法相互抵消时，有限差分方案可能会失效。

第二个问题的情形，即函数评估受到零均值加性扰动时，被称为随机零阶优化。如前所述，随机性在现代优化领域十分普遍。为了解决这一问题，需要对噪声结构作出额外假设。下面简要讨论文献中采用的两个主要思路，并请读者参阅第2节以了解精确定义。

对于双点反馈，我们假设，对于一个固定的噪声变量值，至少可以调用两次随机零阶预言机。这意味着我们可以计算如下形式的有限差分近似：

$$p(x,\xi,e)=\frac{f(x+te,\xi)-f(x-te,\xi)}{2t} \approx \langle \nabla_x f(x,\xi), e\rangle \tag{2}$$

这种近似给出了函数 $f(\cdot\ \xi)$ 的一个带噪实现f的方向导数估计。如前所述，近似式（2）可以在一阶方法中替代（随机）梯度。在独立随机性的情况下，许多研究都基于这一思路。针对非光滑和光滑凸问题，已有基于 Nesterov 和 Spokoiny 的经典梯度法和加速梯度方法的研究结果 [40]。本文关注 [17], 中关于光滑强凸问题的结果，即关于达到 ε-解所需零阶预言机调用次数的估计，其量级为 $\|x-x^*\|$:

$\mathcal{O}(\frac{d\sigma^2}{\mu^2\varepsilon^2})$。这里， σ_2表示梯度方差，即假设 $\mathbb{E}_\xi \nabla f(x,\xi) = \nabla f(x)$ 且$\mathbb{E}_\xi\|\nabla f(x,\xi) - \nabla f(x)\|^2 \le \sigma_2^2$.双点方法的主要局限在于，需要针对同一个噪声变量进行多次评估。这适用于经验风险优化等问题 [34], ，但对于强化学习或在线优化而言，可能构成重大障碍。

在 单点反馈 设定下，假设随机性更为一般。在这种情况下，每次调用零阶预言机都会产生新的随机变量。此时，近似（1）如下所示

$$p(x,\xi^{\pm},e)=\frac{f(x+te,\xi^{+})-f(x-te,\xi^{-})}{2t} \tag{3}$$

在（3）中使用不同的 ξ^+ 和 ξ^-，会使对 $\nabla f(\cdot\ \xi)$ 性质的任何条件都失去作用。相反，假设 $\mathbb{E}_\xi f(x,\xi) = f(x)$ 且 $\mathbb{E}_\xi|f(x,\xi) - f(x)|^2 \le \sigma_1^2$。对于单点反馈，主要问题在于为有限差分方案选择合适的位移 t 。位移过小会放大加性噪声，而 t 过大则会导致梯度估计不佳。由于这种方差权衡，单点近似方法的最优收敛速率劣于双点反馈方法。特别地，对于光滑强凸问题，零阶预言机调用次数 [23] 满足以下估计：$\mathrm{O}(\frac{d^2\sigma_1^2}{\mu^3\varepsilon^2})$。

尽管零阶梯度近似方案存在方差较高的问题，但它们有一个令人惊讶的特性，使其在非光滑优化中更具优势 [22, 44, 49]。这一思路可追溯到 20 世纪 70 年代，并利用了这样一个事实：

that

$$\mathbb{E}[e \cdot p(x, \xi^{(\pm)}, e)] = \frac{1}{d}\nabla f_t(x), \text{ where } f_t \text{ is a } \textit{smoothed} \text{ function, defined as}$$
$$f_t(x) = \mathbb{E}_r\left[f(x + tr)\right] \text{ with } r \sim RB_2^d$$

事实上，可以证明 f_t 是 $\frac{\sqrt{d}G}{t}$ -光滑的，条件是 f 是 G-利普希茨的。这使得零阶近似成为 f_t的随机梯度的一种合适候选。用一阶方法优化这个函数会得到一个解，但它未必是 f [22]的最优解。从这里开始，就需要在两者之间权衡：当 t 较小时，函数 f 和 f_t 更接近；而当 t 较大时，函数 f_t 会因变得更加光滑而更容易优化。

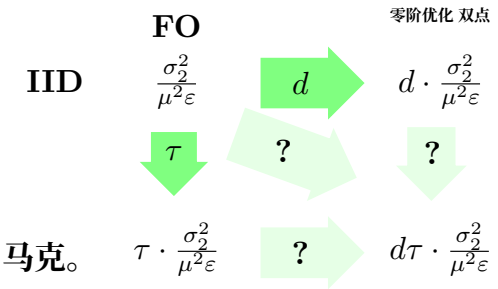
在较新的研究中，零阶优化方法的理论理解取得了许多进展。作者们研究了底层函数的高阶光滑性 [2], 处理非凸非光滑问题[44], 采用任意布雷格曼几何，以降低预言机复杂度 [49, 29],并提出了尖锐的信息论下界，以理解计算能力的极限 [15, 1]。但这些研究都没有考虑马尔可夫随机性。

◇ **马尔可夫一阶方法**。 尽管关于独立同分布噪声下随机优化的文献十分丰富，但针对马尔可夫情形的研究仍相对有限。在本文中，我们聚焦于最“友好”的一致几何遍历马尔可夫链（精确定义见第 2 节）。

Duchi 等人 [14] 率先研究了非独立同分布噪声，探讨了遍历镜像下降算法，并为非光滑凸问题确立了最优收敛速率。对于光滑问题，研究者曾尝试在一阶预言机 [12, 11, 59, 18]方面取得突破性估计。最终，Beznosikov 等人的研究为凸问题和非凸问题都得到了最优结果。 [5], Solodkin 等人[52]。特别是，对于马尔可夫噪声下的光滑强凸目标函数，作者给出了如下复杂度： $\mathcal{O}(\frac{\tau\sigma_2^2}{\mu^2\varepsilon})$，其中 τ定义为相应马尔可夫链的混合时间（见第 2节）。请注意，这些研究采用了多层蒙特卡洛（MLMC）批处理技术，有助于有效应对马尔可夫噪声。我们也需要采用这种方法。值得注意的是，Dorfman 和 Levy [13] 最初在马尔可夫梯度优化中提出了这种方法，用于自动适应未知的 τ。

◇ **假设**。 强凸一阶随机方法的复杂度估计为 $\mathcal{O}(\frac{\sigma_2^2}{\mu^2\varepsilon})$ [36, 37]。针对同一类问题和方法的下界表明，该结果无法进一步改进 [58]。如前所述，从独立同分布随机性转向马尔可夫型随机性，会使估计值增加 τ 倍。Beznosikov 等人证明了该结果也是最优的 [5]。与此同时，从梯度预言机转向零阶方法，会使双点反馈的复杂度增加 d 倍，单点情形下增加 d^2/ε 倍。而且这一估计同样无法进一步改进 [1, 15]。

由此产生这样的假设：转向零阶马尔可夫优化会同时增加两个乘数： $d\tau$ 和 $d^2\tau/\varepsilon$ ，分别对应双点和单点情形。下图展示了双点反馈的情况：



1.2 我们的贡献

我们的主要贡献是回答上述假设：令人惊讶的是，该假设并不成立。具体而言：

◇ **加速随机梯度下降**。我们首次分析了零阶加速随机梯度下降在马尔可夫噪声下的情形，同时考虑双点和单点反馈。与预期中收敛速率会随维数和混合时间共同按乘法关系缩放不同，我们的分析揭示了显著的加速效果，如表1所示。结果表明，如果 τ 小于 d，我们的结果与具有独立随机性的无梯度方法完全相同。第2.1节介绍了这一加速背后的关键技术。第3节还通过数值实验验证了该理论。

◇ **非光滑问题**。我们还研究了马尔可夫噪声下的非光滑问题。利用平滑技术，我们为这种情形推导出了相应的上界，如表1所示。这些界的详细内容见第B.2节。

◇ **计算效率**。 首先，如上所述，对于任意 $\tau \leq d$，我们的方法都具有相同的预言机复杂度。此外，如果我们假设调用零阶预言机比计算相应的

表 1: 上界汇总 的总结。关于符号，见 表 2

	光滑		非光滑	
	IID	**马尔可夫。**	**IID**	**马尔可夫。**
FO	$\frac{\sigma_2^2}{\mu^2\varepsilon}$ [45]	$\tau\frac{\sigma_2^2}{\mu^2\varepsilon}$ [5]	$\frac{G^2}{\mu^2\varepsilon}$ [50]	$\tau\frac{G^2}{\mu^2\varepsilon}$ [14][1]
零阶优化 2P	$d\frac{\sigma_2^2}{\mu^2\varepsilon}$ [30]	$(d+\tau)\frac{\sigma_2^2}{\mu^2\varepsilon}$	$d\frac{G^2}{\mu^2\varepsilon}$ [22]	$(d+\tau)\frac{G^2}{\mu^2\varepsilon}$
零阶优化 1P	$d^2\frac{\sigma_1^2}{\mu^3\varepsilon^2}$ [2][2]	$d(d+\tau)\frac{L\sigma_1^2}{\mu^3\varepsilon^2}$	$d^2\frac{\sigma_1^2G^2}{\mu^4\varepsilon^3}$ [23]	$d(d+\tau)\frac{\sigma_1^2G^2}{\mu^4\varepsilon^3}$

梯度便宜 倍，那么在马尔可夫噪声下，梯度法所需资源与 $d\cdot\tau$成正比——一次预言机调用的成本为 d，且表 τ 中的一阶方法的复杂度按 1 缩放。与此同时，我们的零阶方法的资源复杂度与 $d+\tau$ 成正比。

◇ **下界。** 在第 2.3节中，我们建立了针对单点和双点反馈的马尔可夫优化问题求解方法的首批信息论下界。我们的结果与算法的收敛保证仅相差对数因子，这表明分析是准确的，且不可能进一步改进。

表2: 符号与定义

Sym.	定义	Sym.	定义
$\|\cdot\|$, $\langle\cdot,\cdot\rangle$	范数、点积，默认假设为欧几里得	ε	$\|x-x^*\|^2$
Z, $\mathcal{Z}$	完备可分度量空间及其 Borel σ-代数	d	问题维数
Q	$\mathsf{Z}\times$上的马尔可夫核$\mathcal{Z}$	L	梯度的利普希茨常数
$\mathbb{P}_\xi$, $\mathbb{E}_\xi$	概率、初始分布下的期望 ξ[3]	μ	强凸常数
$\{Z_k\}$	核 Q 对应的典范过程	G	函数的利普希茨常数
RB_2^d, RS_2^d	单位 ℓ_2 球内的均匀分布，-球面	σ_1^2	$\lvert F(x,Z)-f(x)\rvert^2\le\sigma_1^2$
e	随机方向， $e\sim RS_2^d$	σ_2^2	$\|\nabla F(x,Z)-\nabla f(x)\|^2\le\sigma_2^2$
$a_n\lesssim b_n$	$\exists c\in\mathbb{R}$ （通用常数）： $a_n\le cb_n$ 对所有 n	τ	Z的混合时间
$a_n\simeq b_n$	$a_n\lesssim b_n$ 和 $b_n\lesssim a_n$	g, $\hat{g}$	梯度估计器
$T=\mathcal{O}(S)$	$T\le poly(\log S)\cdot S$ 为 $\varepsilon\to 0$	$f_t(x)$	$\mathbb{E}_r\left[f(x+tr)\right]$， $r\sim RB_2^d$

2 主要结果

现在我们已经准备好 进行更正式的阐述。在本文中，我们研究最小化问题

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{Z\sim\pi}\left[F(x,Z)\right], \tag{4}$$

其中 π 是一个未知分布（见假设3），并且可以通过随机单点或双点预言机 $F(x,\ Z)$访问函数 f （而非其梯度 ∇f）。

[1]作者考虑一般凸情形。使用标准的重启技术，我们得到强凸情形下相应的界。[2]假设噪声与点无关。[3]根据构造，对于任意 $A\in\mathcal{Z}$，都有 $\mathbb{P}_\xi(Z_k\in A\mid Z_{k-1})=\mathsf{Q}(Z_{k-1},A)$ $\mathbb{P}_\xi$-几乎必然成立。

在分析中，我们将对底层函数 f 及其预言机采用一组假设，首先是光滑性和凸性：

假设 1

函数 f 在 $\mathbb{R}^d$ 上是 L-光滑的，且满足 $L > 0$，即它是可微的，并且存在常数 $L > 0$，使得对所有 x，$y \in \mathbb{R}^d$，以下不等式成立：

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

在双点反馈设定下，我们需要以下推广：

假设 1′

对于所有 $Z \in \mathsf{Z}$，函数 $F(\cdot\ Z)$ 在 $\mathbb{R}^d$上是 L-光滑的。

注意，一致 1′ 意味着1。

假设 2

函数 f 连续可微，且在 $\mathbb{R}^d$ 上是 μ-强凸的，即存在常数 $\mu > 0$，使得对所有 x，$y \in \mathbb{R}^d$，以下不等式均成立：

$$\frac{\mu}{2}\|x - y\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y\rangle. \tag{5}$$

现在我们转而讨论噪声状态序列 $\{Z_i\}_{i=0}^{\infty}$ 的假设。具体而言，我们考虑$\{Z_i\}_{i=0}^{\infty}$ 构成齐次马尔可夫链的情形。令 Q 表示相应的马尔可夫核。为刻画其混合性质，我们对 Q 作出以下假设：

假设 3

$\{Z_i\}_{i=0}^{\infty}$ 是具有马尔可夫核 Q 和唯一不变分布 π 的平稳马尔可夫链，其状态空间为 $(\mathsf{Z},\ \mathcal{Z})$。此外，$\mathsf{Q}$ 是一致几何遍历的，混合时间为 $\tau \in \mathbb{N}$，即对每个 $k \in \mathbb{N}$，经过 k 步后，全变差按如下方式衰减：

$$\Delta(\mathsf{Q}^k) = \sup_{z,z' \in \mathsf{Z}} (1/2)\left\|\mathsf{Q}^k(z,\cdot) - \mathsf{Q}^k(z',\cdot)\right\|_{\mathsf{TV}} \leq (1/4)^{\lfloor k/\tau \rfloor}. \tag{6}$$

假设3 在马尔可夫随机性相关文献中很常见 [14, 12, 13, 5, 52]。例如，它涵盖不可约非周期有限马尔可夫链 [18]。混合时间 τ 反映了链的分布趋近平稳状态的速度，为数据中的时间依赖性提供了自然的度量。接下来，我们说明对预言机的假设。如第 1.1 节所述，这些假设因反馈类型而异。

假设 4（单点）

对于所有 $x \in \mathbb{R}^d$ it 满足 $\mathbb{E}_{\pi}[F(x, Z)] = f(x)$。此外，对于所有 $Z \in \mathsf{Z}$ 和 $x \in \mathbb{R}^d$，都有 t

$$|F(x, Z) - f(x)|^2 \leq \sigma_1^2\,.$$

假设 4′ (针对双点)

对于所有 $x \in \mathbb{R}^d$，都有 $\mathbb{E}_\pi[\nabla F(x, Z)] = \nabla f(x)$。此外，对于所有 $Z \in \mathsf{Z}$ 和 $x \in \mathbb{R}^d$，都有

$$\|\nabla F(x, Z) - \nabla f(x)\|^2 \leq \sigma_2^2 .$$

近期关于随机零阶优化方法的研究考虑了较弱的假设，例如有界方差（见第 1.1 节）。然而，在马尔可夫噪声分析中，假设4 和 4′ 中的一致有界性是标准做法 [14, 12, 13, 5, 52]。在更强的条件下，例如 $(\cdot\ Z)$ [18] 的一致凸性和光滑性，这些假设可以放宽。

假设 3 和 4 使我们能够通过批处理降低噪声方差，其方式与独立同分布情形类似。以下技术引理对此进行了说明：

引理 1. 设假设 *3* 和 *4(4′)* 成立。那么对于任意 $n \geq 1$ 和 $x \in \mathbb{R}^d$，以及任意定义在 ξ 上的初始分布$(\mathsf{Z}, \mathcal{Z})$，我们有

$$\mathbb{E}_\xi \left[\frac{1}{n}\sum_{i=1}^n F(x, Z_i) - f(x)\right]^2 \lesssim \frac{\tau}{n}\sigma_1^2, \quad \mathbb{E}_\xi \left\|\frac{1}{n}\sum_{i=1}^n \nabla F(x, Z_i) - \nabla f(x)\right\|^2 \lesssim \frac{\tau}{n}\sigma_2^2.$$

2.1 批处理技术

本节介绍用于建立误差率 $(d + \tau)$ 型缩放关系的主要工具。我们将重点介绍如何利用一种专门的批处理方法降低梯度估计量的方差和偏差。
我们首先固定梯度估计器在点 x 处的一个通用构件，该构件同时适用于单点反馈和双点反馈，正如第 1.1节所介绍的：

$$\hat{g}(x, Z^{(\pm)}, e) = d \cdot p(x, Z^{(\pm)}, e) \cdot e = e \cdot \begin{cases} d\dfrac{F(x + te, Z^+) - F(x - te, Z^-)}{2t} & \text{(one-point)}, \\ d\dfrac{F(x + te, Z) - F(x - te, Z)}{2t} & \text{(two-point)}. \end{cases}$$

这些估计器呈现出双重随机性，这会影响它们围绕真实梯度的集中速度，正如我们将在下文讨论的那样。

为便于说明，我们将讨论重点放在单点情形上，尽管我们的结论同样适用于双点情形。

一种广泛使用的方差缩减技术是小批量处理，即在一批噪声变量$\{Z_i\}_{i=1}^n$上计算 $F(x, Z_i)$。小批量梯度估计器定义为：

$$\hat{g}_{mb}(x) = \frac{1}{n}\sum_{i=1}^n \hat{g}(x, Z_i^\pm, e) = d\overbrace{\left(\frac{1}{n}\sum_{i=1}^n p(x, Z_i^\pm, e)\right)}^{p_{mb}} \cdot e.$$

让我们估计其方差 $\mathbb{E}_e\mathbb{E}_Z\|\hat{g}_{mb} - \nabla f\|^2$ 随噪声水平 σ_1^2 的缩放关系。当 $\mathbb{E}_Z \hat{p}_{mb} \approx \frac{f(x+te) - f(x-te)}{2t} \approx \langle \nabla f, e \rangle$ 时，我们希望对任意固定方向 e估计以下量：

$$\mathbb{E}_Z[p_{mb}(x) - \langle \nabla f, e\rangle]^2 \approx \frac{1}{t^2}\mathbb{E}_Z\Big[\frac{1}{n}\sum_{i=1}^n F(x + te, Z_i^+) - f(x + te)\Big]^2 \overset{(1)}{\approx} \frac{\tau}{n}\frac{\sigma_1^2}{t^2}. \tag{7}$$

据此，我们可以界定方差：

$$\mathbb{E}_e\mathbb{E}_Z\|\hat{g}_{mb} - \nabla f\|^2 \geq \mathbb{E}_e\mathbb{E}_Z\|\hat{g}_{mb} - \mathbb{E}_Z\hat{g}_{mb}\|^2 = \mathbb{E}_e\mathbb{E}_Z\|d \cdot [p_{mb} - \mathbb{E}_Z[p_{mb}]] \cdot e\|^2 = \tag{8}$$

$$d^2\mathbb{E}_e\mathbb{E}_Z|p_{mb}-\langle\nabla f,e\rangle|^2 \overset{(7)}{\approx} \frac{d^2\tau\sigma_1^2}{nt^2}.$$

能否改进小批量方案?

本小节探讨了一种出乎意料的改进来源，它与我们最初的假设相悖。具体而言，我们发现当前样本的使用方式存在低效之处 Z_i，这一点从两个角度都显而易见。公式（8）表明，方差随 $\frac{\tau}{n}$ 增长。如果我们能将 τ 缩小 k倍，就只需使用 k分之一的样本，便能维持相同的方差。这引出了稀疏化采样的想法。我们将马尔可夫噪声链$\{Z_i\}$划分为 k条子链$\{Z_{k\cdot i+r}\}$，其中 $r=0\dots k-1$。这意味着每条子链的混合时间为 $\lceil\frac{\tau}{k}\rceil$（见（3））_，从而有效降低时间相关性——这是每隔 k个元素从原始链中采样的自然结果。因此，从任意一条子链采样，都可能使所需样本数减少$\min(k,\tau)$倍（尽管该过程仍需进行所有中间预言机调用，因此无法带来计算加速）。

为具体说明这种低效性，考虑一条惰性马尔可夫链：它在均匀随机地转移之前，会在同一状态停留 τ 步（平均而言）。在这种情况下，对于固定的 x，所有预言机查询 $F(x,\ Z)$ 在连续的 τ 步中都会返回相同的值。因此，只保留每隔 τ 步的估计值 $\hat{}g$，就能得到质量相当的小批量。

总而言之，我们发现，原则上，小批量方案只需每隔 k 个样本保留一个并丢弃其余样本，就能同样有效地运行。这或许意味着，我们可以更充分地利用样本。然而，一阶方法无法利用这种冗余（正如 [5] 的下界所示），实际上只能等待 τ 步的混合窗口结束。相比之下，我们可以利用这一结构，沿不同方向查询有限差分，以更准确地估计梯度。具体而言，我们构造 d 条子链，其中第 r 条子链 $Z_{d\cdot i+r}$ 用于沿第 r 个坐标进行预言机调用：$\frac{F(x+te_r,\ Z)-F(x-te_r,\ Z)}{2t}$。这样就能逐坐标恢复完整梯度。

接下来估算由此带来的方差缩减。首先，通过重构所有 d个 d梯度坐标，我们实现了$\lceil\frac{\tau}{d}\rceil$倍的缩减。其次，现在每个坐标都在混合时间为$\lceil\frac{\tau}{d}\rceil$的链上运行，因此额外获得$\min(d,\tau)$这一缩减因子。不过，由于现在每个批次被分配到 d 个坐标上，因此每个批次的规模比之前小 d 倍，导致损失因子 d。因此，净方差缩减为$\min(d,\tau)$，最终的缩放变为 $d\cdot\frac{d\tau}{\min(d,\tau)}=d\cdot\max(d,\ \tau)\ \simeq d(d+\tau)$。

随机方向

这一思路可以扩展为一种更简单但同样有效的方法。我们不再确定性地分配方向，而是为每个样本关联一个随机方向 $e\in RS_2^d$，从而构造出如下估计量：

$$\hat{g}_{rd}[n](x,Z,e)=\frac{1}{n}\sum_{i=1}^{n}\hat{g}(x,Z_i,e_i).$$

上述讨论虽然直观，但现在我们将概述一种更形式化的方法（详见引理5 ）。由于惰性马尔可夫链实际上等价于随机独立同分布的 τ点反馈设定，我们采用 [15],的推论 2，将总方差分解为两项：

$$\mathbb{E}\|\hat{g}_{rd}-\nabla f(x)\|^2\le 2\mathbb{E}\|\hat{g}_{rd}-\mathbb{E}_e\hat{g}_{rd}\|^2+2\mathbb{E}\|\mathbb{E}_e\hat{g}_{rd}-\nabla f(x)\|^2.$$

这两个项中的每一项都单独消除了对 $d^2\tau$ 的一个因子依赖。第一项：

$$\mathbb{E}\|\hat{g}_{rd}-\mathbb{E}_e\hat{g}_{rd}\|^2=\mathbb{E}_Z\mathbb{E}_e\left\|\frac{1}{n}\sum_{i=1}^{n}\underbrace{[\hat{g}(x,Z_i,e_i)-E_{e_i}\hat{g}(x,Z_i,e_i)]}_{\mathbb{E}_e[\cdot]=0,\text{ independent w.r.t. }e}\right\|^2=\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\|\hat{g}(x,Z_i,e_i)-\mathbb{E}_{e_i}\hat{g}(x,Z_i,e_i)\|^2$$

与.....无关 τ 因为假设 4直接界定了每一项。

对于第二项，我们注意到 $\mathbb{E}_e\hat{}g_{rd}=\mathbb{E}_e\hat{}g_{mb}$，因此界中包含 $\mathbb{E}\|\mathbb{E}_e\hat{}g_{mb}-\nabla f(x)\|^2$。这与小批量情形中出现的 $d^2\tau$ 依赖性有着至关重要的区别；在小批量情形下，我们考虑的是 $\mathbb{E}\|\hat{}g_{mb}-\nabla f(x)\|^2$。

直观而言，对方向取期望有助于恢复完整梯度，而非仅恢复某个方向上的分量，从而降低相对于 d的方差。

多层蒙特卡罗

估计量 $\hat{}g_{rd}$ 并非我们最终的构造。虽然它能控制方差，但噪声中的时间相关性可能会引入显著偏差。缓解这一问题的一种成熟方法是多层蒙特卡罗，该方法在统计学文献中得到广泛应用 [27, 26], ，近年来也用于梯度优化 [13, 5]。下面是我们对此方法的解读。

采用表 J、l、M、B中的参数 3， $\{Z_i\}$ - $2^J l$ 个来自 Z的样本和 $\{e_i\}$ 个随机方向，我们引入多层蒙特卡罗估计量：

$$\hat{g}_{ml}(x) = \hat{g}_{rd}[l](x) + \begin{cases} 2^J \left[\hat{g}_{rd}\left[2^J l\right](x) - \hat{g}_{rd}\left[2^{J-1} l\right](x)\right], & \text{if } 2^J \leq M \\ 0, & \text{otherwise.} \end{cases}$$

为了便于理解上面的公式，考虑一个例子，其中 $l = 1$， $M = \infty$， $B = 1$ ，并将基础估计值依次记为 g_1， g_2，... ...。那么，多层蒙特卡罗估计值为 $\hat{}g_{ml} = g_1$，概率为 $1/2$； $g_1 + (g_3 - g_2)$，概率为 $1/4$；$g_1 + (g_4 + g_5 - g_2 - g_3)$，概率为 $1/8$，依此类推。参数 M是所用估计值数量的上界。参数 l 将基础估计量转换为一系列 l 个基础估计量，实际上将所有内容拉伸 l倍。最后， B 是一个超参数，可通过乘法增加 l。$\hat{}g_{ml}$ 是我们的最终梯度估计量，并具有以下保证：

引理 2 （单点）。 设假设 *1*、*3* 和*4* 成立。对于任意定义在[1] ξ 上的初始分布$(\mathsf{Z}, \mathcal{Z})$ ， 梯度估计$\hat{}g_{ml}$ 满足 $\mathbb{E}[\hat{}g_{ml}] = \mathbb{E}\left[\hat{}g_{rd}\left[2^{\lfloor \log_2 M \rfloor} l\right]\right]$。此外，

$$\mathbb{E}\|\nabla f_t(x) - \hat{g}_{ml}(x)\|^2 \lesssim \frac{d\|\nabla f(x)\|^2}{B} + \frac{d^2 L^2 t^2}{B} + \frac{d\,(d+\tau)\,\sigma_1^2}{B t^2}, \quad \|\nabla f_t(x) - \mathbb{E}[\hat{g}_{ml}(x)]\|^2 \lesssim \frac{d\tau\sigma_1^2}{t^2 B M}.$$

可以注意到，尽管 $\hat{}g_{ml}$ 平均需要 $l \log_2 M \simeq \log_2^2 M \cdot B$ 次预言机调用，但方差仅降低了 B倍。相比之下，偏差显著降低了 BM倍。

2.2 算法

表 3： 算法1的参数

超参数		动量		批次隐藏参数	
γ	步长， $\in (0; \frac{3}{4L}]$	β	$\sqrt{\frac{4p^2\mu\gamma}{3}}$	$2^J l$	批量大小。如果 $2^J > M$，则为 0
t	近似步骤	η	$\sqrt{\frac{3}{\mu\gamma}}$	J	随机变量， $J \sim$Geom(1/2)
B	批量大小乘数	θ	$\frac{p\eta^{-1}-1}{\beta p\eta^{-1}-1}$	M	批量大小上限， $M = \frac{1}{p} + \frac{2}{\beta}$
N	迭代次数	p	常数或[2] $\frac{B}{B+d}$	l	$(\lfloor \log_2 M \rfloor + 1) \cdot B$

现在给出算法 1 的完整版本。该算法纳入了上一节讨论的梯度估计量，并以略作修改的 Nesterov加速梯度下降 变体为核心。

[1]请注意，$\hat{g}_{ml}$ （具体而言， Z_1）间接取决于链的初始分布。由于我们的算法将反复调用$\hat{g}_{ml}$，下一次迭代的初始分布就是当前迭代的最终分布。这一事实会导致估计值之间存在相关性。我们通过假设任意初始分布来规避这一问题。[2]分别适用于单点或双点情形

尽管从技术上讲，我们证明了涵盖光滑和非光滑假设下单点和两点反馈的四个不同上界，但它们遵循相同的方案，下面我们将在单点光滑情形下进行说明。

引理 4确立了平滑目标函数的关键性质。引理5给出了基准估计量 $\hat{}g_{rd}$ 的偏差和方差界。引理 2随后量化了多层蒙特卡洛方案如何放大或降低这些统计量。最后，在第 D.4 节中，我们结合这些引理的结果，证明定理 1 的第一部分，给出算法 1 的误差界。通过适当调整参数，我们得到以下迭代复杂度界：

算法 1：随机加速零阶梯度下降

Initialization: $x_f^0 = x^0$; see Table 3.
for $k = 0, 1, 2, \ldots, N-1$ **do**
  $x_g^k = \theta x_f^k + (1-\theta)x^k$
  Sample $J, \{e_i\}, \left\{F(x_g^k \pm te_i, Z_i^{(\pm)})\right\}$
  Calculate $\hat{g}^k = \hat{g}_{ml}(x_g^k)$
  $x_f^{k+1} = x_g^k - p\gamma\hat{g}^k$
  $x^{k+1} = \eta x_f^{k+1} + (p-\eta)x_f^k +$
    $+(1-p)(1-\beta)x^k + (1-p)\beta x_g^k$
或d f

定理 1

令假设 1 至4 成立，并考虑由算法 1 求解的问题（4）。那么，对于任意目标精度ε和批量大小乘数 B（符号说明见表 2和 3），以及适当选择的 γ、t、p，确保 $\mathbb{E}\|x^N - x^*\|^2 \le \varepsilon$ 所需的预言机调用次数有界为

$$B \cdot \tilde{\mathcal{O}}\left(\max\left[1, \frac{d}{B}\right]\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon} + \frac{Ld\,(d+\tau)\,\sigma_1^2}{B\mu^3\varepsilon^2}\right) \quad \text{one-point oracle calls}.$$

定理 1′

假设 1′ 至 4′ 成立，并考虑由算法 4）求解的问题（1。那么，对于任意目标精度ε和批量大小乘数 B（符号见表 2和 3），若适当选择 γ, t, p, 确保 $\mathbb{E}\|x^N - x^*\|^2 \le \varepsilon$ 所需的预言机调用次数有以下界

$$B \cdot \tilde{\mathcal{O}}\left(\max\left[1, \frac{d}{B}\right]\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon} + \frac{(d+\tau)\sigma_2^2}{B\mu^2\varepsilon}\right) \quad \text{two-point oracle calls}.$$

注。 该算法的迭代复杂度，即生成的迭代点数量 x^k（等于预言机复杂度除以 B），有界于 $\tilde{}\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$，当批量大小乘数 B趋于无穷时。这与使用精确梯度进行优化的最优收敛速率相符 [38]。

2.3 下界

这里我们给出定理，证明在光滑、强凸情形下，无论采用单点还是双点反馈，都没有任何算法能够渐近地超越算法1 。

定理 2

（下界）对于任何求解问题（4）的算法（可能是随机化的），都存在一个函数f 满足假设 1 至 4 (1′ 至 4′)，为了达到 ε-近似期望解$\mathbb{E}\|x^N - x^*\|^2 \leq \varepsilon$，该算法至少需要

$$\Omega\left(\frac{d(d+\tau)\sigma_1^2}{\mu^2\varepsilon^2}\right) \quad \text{one-point or} \quad \Omega\left(\frac{(d+\tau)\sigma_2^2}{\mu^2\varepsilon}\right) \quad \text{two-point oracle calls.}$$

注。 这些结果假设二阶矩有界，而非噪声界一致。我们将在第F.2节说明如何将其调整到我们的设定中，且只需付出对数级开销。

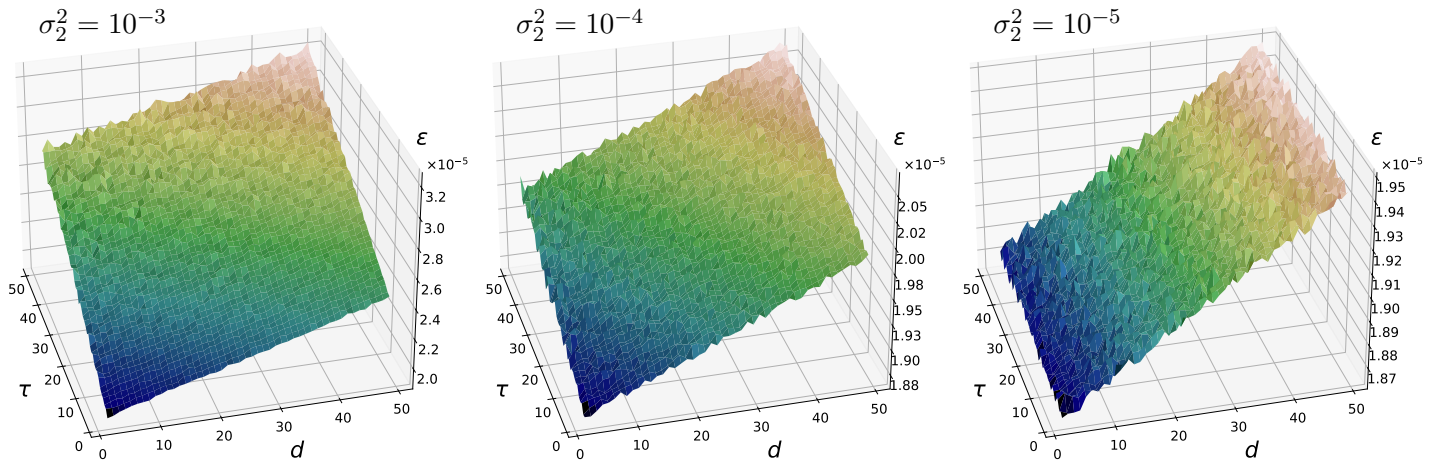
讨论。 现在我们将本文结果与已有工作进行比较。Akhavan 等人 [2] 分析了单点设定的一种特殊情形，其中噪声与查询点无关。这与我们的单点预言机模型一致，并允许将独立同分布采样视为具有固定混合时间的马尔可夫链 $\tau = 1$。他们未考虑的唯一因素是σ_1^2，不过，在使用缩放高斯噪声时，他们的证明中会出现额外的 μ^2 因子。我们将在第F节进一步讨论这一点。

Beznosikov 等人 [5], 的工作考虑了一阶马尔可夫预言机，但其困难实例问题是一维二次函数，因此一阶信息与零阶信息等价。因此，他们的结果对应于双点情形中的 $d = 1$ 情形。Duchi 等人 [15] 针对一般凸函数在双点反馈下给出了紧下界。通过在困难实例中加入一个共享的二次项，他们的方法可以扩展到强凸情形，详情见第 F节，定理10，由此得到我们为具有 $\tau = 1$ 的双点预言机陈述的界。

我们的主要贡献在于建立了一个下界，其尺度为 $d\tau$，适用于较大 τ时的单点情形；参见定理 8。虽然我们的分析依赖于多维假设检验等经典工具，但马尔可夫结构要求我们为联合分布之间的距离建立新的界，并使用截断。有关截断的详细证明、讨论和补充说明见第F节。

3 实验

本节通过实证结果支持我们的理论收敛速率和下界，尤其关注随机分量；我们声称该分量关于 $d + \tau$ 呈线性标度，而非关于 $d\tau$。



(a) (b) (c)

图1: 经过 $N = 10^3$ 次迭代后的优化误差 $\varepsilon = \|x^N - x^*\|^2$。起点误差 $\|x_0 - x^*\|^2 = 10^{-2}$。步长 $\gamma = 10^{-3}$, $t = 10^{-5}$。结果为 10^4 次运行的平均值。

设置。我们的设置沿用用于证明下界的问题（参见附录 F 节和 [51]）。我们考虑二次目标函数 $f(x) = \frac{1}{2}\|x\|^2$ 以及两点马尔可夫预言机 $F(x; Z) = f(x) + x, Z$。噪声序列 $\{Z_i\}$ 是方差为 σ_2^2 的惰性更新的标准高斯向量。图1展示了算法 1 的优化误差如何随混合时间、问题维数以及σ_2^2的不同取值而变化。

讨论。 结果证实，误差与问题维数 d 和混合时间 τ均呈线性关系。噪声参数 σ^2 控制随机部分的影响。在图 (a) 中， $\sigma_2^2 = 10^{-3}$，随机分量占主导；而在图 (c) 中， $\sigma_2^2 = 10^{-5}$，随机分量可以忽略。图 (b) 展示了一个在两者之间平滑过渡的中间状态，同时仍保持线性标度。确定性部分 (c) 不依赖混合时间，但随 d线性增长，这与我们的理论（定理 1′）一致。随机部分 (a) 按 $(d + \tau)$ 缩放，也符合定理 1′给出的界。

参考文献

[1] Arya Akhavan、Massimiliano Pontil 和 Alexandre Tsybakov。利用高阶光滑性进行无导数优化与连续老虎机研究。神经信息处理系统进展，33:9017–9027，2020年。[2] Arya Akhavan、Evgenii Chzhen、Massimiliano Pontil 和 Alexandre B Tsybakov。高度光滑函数的无梯度优化：改进的分析与一种新算法。机器学习研究期刊，25(370):1–50，2024年。[3] Peter Auer。多臂老虎机问题的有限时间分析。*Machine Learning*，47:235–256，2002。[4] El Houcine Bergou、Eduard Gorbunov 和 Peter Richtárik。用于无约束光滑最小化的随机三点法。*SIAM*优化期刊，30(4):2726–2749，2020年。[5] Aleksandr Beznosikov、Sergey Samsonov、Marina Sheshukova、Alexander Gasnikov、Alexey Naumov 和 Eric Moulines。马尔可夫噪声下的一阶方法：从加速到变分不等式。神经信息处理系统进展，36，2024年。[6] Jalaj Bhandari、Daniel Russo 和 Raghav Singal。线性函数逼近下时序差分学习的有限时间分析。载于学习理论会议，第1691–1692页。机器学习研究会论文集，2018年。[7] Sébastien Bubeck、Nicolo Cesa-Bianchi 等。随机与非随机多臂老虎机问题的遗憾分析。*Foundations and Trends® in Machine Learning*，5(1):1–122，2012年。[8] Pin-Yu Chen、Huan Zhang、Yash Sharma、Jinfeng Yi 和 Cho-Jui Hsieh。Zoo：无需训练替代模型、基于零阶优化的深度神经网络黑盒攻击。载于第*10*届*ACM*人工智能与安全研讨会论文集，第15–26页，2017年。[9] Krzysztof Choromanski、Mark Rowland、Vikas Sindhwani、Richard Turner 和 Adrian Weller。利用紧凑架构进行结构化演化，实现可扩展的策略优化。载于国际机器学习会议，第970–978页。机器学习研究会论文集，2018年。[10] Aymeric Dieuleveut、Nicolas Flammarion 和 Francis Bach。更难、更好、更快、更强：最小二乘回归的收敛速率。机器学习研究期刊，18(101):1–51，2017年。[11] Thinh T Doan。马尔可夫梯度下降的有限时间分析。*IEEE*自动控制汇刊，68 (4):2140–2153，2022年。[12] Thinh T Doan、Lam M Nguyen、Nhan H Pham 和 Justin Romberg。加速马尔可夫梯度下降的收敛速率及其在强化学习中的应用。*arXiv*预印本 *arXiv:2002.02873*，2020年。[13] Ron Dorfman 和 Kfir Yehuda Levy。在马尔可夫数据随机优化中适应混合时间。载于国际机器学习会议，第5429–5446页。机器学习研究会论文集，2022年。

[14] John C Duchi、Alekh Agarwal、Mikael Johansson 和 Michael I Jordan。遍历镜像下降。SIAM优化期刊，22(4):1549–1578，2012年。[15] John C Duchi、Michael I Jordan、Martin J Wainwright 和 Andre Wibisono。零阶凸优化的最优速率：两次函数评估的威力。IEEE 信息论汇刊，61 (5):2788–2806，2015年。[16] Alain Durmus、Eric Moulines、Alexey Naumov、Sergey Samsonov 和 Hoi-To Wai。马尔可夫噪声下随机矩阵乘积的稳定性：在线性随机逼近和 TD 学习中的应用。载于学习理论会议，第1711–1752页。机器学习研究会论文集，2021年。[17] Pavel Dvurechensky、Eduard Gorbunov 和 Alexander Gasnikov。用于光滑随机凸优化的加速方向导数方法。《欧洲运筹学杂志》，290(2):601–621，2021年。[18] Mathieu Even。马尔可夫采样方案下的随机梯度下降。载于国际机器学习会议，第9412–9439页。机器学习研究会论文集，2023年。[19] Maryam Fazel、Rong Ge、Sham Kakade 和 Mehran Mesbahi。线性二次型调节器中策略梯度方法的全局收敛性。载于国际机器学习会议，第1467–1476页。机器学习研究会论文集，2018年。[20] Abraham D. Flaxman、Adam Tauman Kalai 和 H. Brendan McMahan。多臂老虎机环境下的在线凸优化：无梯度的梯度下降。载于第十六届ACM-SIAM离散算法年会论文集，SODA '05，第385–394页，美国，2005年。工业与应用数学学会。ISBN 0898715857。[21] Alexander Gasnikov、Darina Dvinskikh、Pavel Dvurechensky、Eduard Gorbunov、Aleksandr Beznosikov 和 Alexander Lobanov。凸优化中的随机无梯度方法，第1–15页。Springer 国际出版集团，Cham，2020年。ISBN 978-3-030-54621-2。doi: 10.1007/978-3-030-54621-2_859-1。URLhttps://doi.org/10.1007/978-3-030-54621-2_859-1。[22] Alexander Gasnikov、Anton Novitskii、Vasilii Novitskii、Farshed Abdukhakimov、Dmitry Kamzolov、Aleksandr Beznosikov、Martin Takac、Pavel Dvurechensky 和 Bin Gu。一阶光滑优化在黑箱非光滑问题中的威力。载于Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato（编），第39届国际机器学习会议论文集，机器学习研究会论文集第162卷，第7241–7265页。机器学习研究会论文集，2022年7月17–23日。URLhttps://proceedings.mlr.press/v162/gasnikov22a.html。[23] Alexander V Gasnikov、Ekaterina A Krymova、Anastasia A Lagunovskaya、Ilnura N Usmanova 和 Fedor A Fedorenko。随机在线优化：单点与多点非线性多臂老虎机；凸与强凸情形。《自动化与远程控制》，78:224–234，2017年。[24] Saeed Ghadimi 和 Guanghui Lan。用于非凸随机规划的随机一阶与零阶方法。SIAM优化期刊，23(4):2341–2368，2013年。[25] Saeed Ghadimi 和 Guanghui Lan。用于非凸非线性与随机规划的加速梯度方法。《数学规划》，156(1):59–99，2016年。[26] Michael B. Giles。多层蒙特卡罗路径模拟。《运筹学》，56(3):607–617，2008年。doi: 10.1287/opre.1070.0496。URL https://doi.org/10.1287/opre.1070.0496。[27] Peter W. Glynn 和 Chang-Han Rhee。马尔可夫链平稳期望的精确估计。《应用概率杂志》，51A:377–389，2014年。ISSN 00219002。URL http://www.jstor.org/stable/43284129。

[28] Ian Goodfellow、Yoshua Bengio 和 Aaron Courville。《深度学习》。MIT出版社，2016年。http://www.deeplearningbook.org.[29] Eduard Gorbunov、Pavel Dvurechensky 和 Alexander Gasnikov。用于导数自由光滑随机凸优化的加速方法。《SIAM优化期刊》，32(2):1210–1238，2022年。doi: 10.1137/19M1259225。URL https://doi.org/10.1137/19M1259225.[30] Elad Hazan 和 Satyen Kale。超越遗憾最小化障碍：随机强凸优化的最优算法。《机器学习研究期刊》，15(71):2489–2512，2014年。URL http://jmlr.org/papers/v15/hazan14a.html.[31] Bjorn Johansson、Maben Rabi 和 Mikael Johansson。一种用于传感器网络分布式优化的简单对等算法。载于《2007年第46届IEEE决策与控制会议》，第4705–4710页，2007年。doi: 10.1109/CDC.2007.4434888.[32] J. Kiefer。最大值的序贯极小极大搜索。《美国数学会会刊》，4(3): 502–506，1953年。ISSN 00029939, 10886826。URL http://www.jstor.org/stable/2032161.[33] Xiangru Lian、Yijun Huang、Yuncheng Li 和 Ji Liu。用于非凸优化的异步并行随机梯度。《神经信息处理系统进展》，28，2015年。[34] Sijia Liu、Bhavya Kailkhura、Pin-Yu Chen、Paishun Ting、Shiyu Chang 和 Lisa Amini。用于非凸优化的零阶随机方差缩减。《神经信息处理系统进展》，31，2018年。[35] Cassio G. Lopes 和 Ali H. Sayed。分布式网络上的增量自适应策略。《IEEE信号处理汇刊》，55(8):4064–4077，2007年。doi: 10.1109/TSP.2007.896034.[36] Eric Moulines 和 Francis Bach。随机逼近算法在机器学习中的非渐近分析。载于 J. Shawe-Taylor、R. Zemel、P. Bartlett、F. Pereira 和 K.Q. Weinberger（编），《神经信息处理系统进展》，第24卷。Curran Associates, Inc.，2011年。URL https://proceedings.neurips.cc/paper_files/paper/2011/file/40008b9a5380fcacce3976bf7c08af5b-Paper.pdf.[37] Deanna Needell、Rachel Ward 和 Nati Srebro。随机梯度下降、加权抽样与随机化Kaczmarz算法。《神经信息处理系统进展》，27，2014年。[38] Yurii Nesterov。求解凸规划问题的一种方法，其收敛速度为 o (1/k2)。载于《苏联科学院通报》，第269卷，第543页，1983年。[39] Yurii Nesterov。坐标下降方法在超大规模优化问题中的效率。《SIAM优化期刊》，22(2):341–362，2012年。[40] Yurii Nesterov 和 Vladimir Spokoiny。凸函数的随机无梯度最小化。《计算数学基础》，17(2):527–566，2017年。[41] Donald J Newman。单峰曲面上最大值的位置。《ACM学报》（JACM），12(3): 395–398，1965年。[42] J. Nocedal 和 S. Wright。《数值优化》。《施普林格运筹学与金融工程丛书》。施普林格纽约，2006年。ISBN 9780387227429。URL https://books.google.ru/books?id=7wDpBwAAQBAJ.[43] Boris Polyak。《优化导论》。Optimization Software - Inc.，出版部，1987年。

[44] 邱雨扬、Uday Shanbhag 和 Farzad Yousefian。用于不可微、非凸和分层联邦优化的零阶方法。载于 A. Oh、T. Naumann、A. Globerson、K. Saenko、M. Hardt 和 S. Levine（编），《神经信息处理系统进展》，第36卷，第3425–3438页。Curran Associates, Inc.，2023年。URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0a70c9cd8179fe6f8f6135fafa2a8798-Paper-Conference.pdf.[45] Herbert Robbins 和 Sutton Monro。一种随机逼近方法。《数学统计年刊》，第400–407页，1951年。[46] Tim Salimans、Jonathan Ho、Xi Chen、Szymon Sidor 和 Ilya Sutskever。进化策略：强化学习的一种可扩展替代方案。arXiv预印本 arXiv:1703.03864，2017年。[47] Bobak Shahriari、Kevin Swersky、Ziyu Wang、Ryan P. Adams 和 Nando de Freitas。让人类退出环路：贝叶斯优化综述。《IEEE会刊》，104(1):148–175，2016年。doi: 10.1109/JPROC.2015.2494218.[48] Shai Shalev-Shwartz 和 Shai Ben-David。《理解机器学习：从理论到算法》。剑桥大学出版社，2014年。[49] Ohad Shamir。一种利用双点反馈进行多臂老虎机和零阶凸优化的最优算法。《机器学习研究期刊》，18(52):1–11，2017年。[50] Ohad Shamir 和 张彤。用于非光滑优化的随机梯度下降：收敛结果与最优平均方案。载于 Sanjoy Dasgupta 和 David McAllester（编），《第30届国际机器学习会议论文集》，《机器学习研究会论文集》第28卷，第71–79页，美国佐治亚州亚特兰大，2013年6月17日至19日。机器学习研究会论文集。URL https://proceedings.mlr.press/v28/shamir13.html.[51] Alexander Shapiro、Darinka Dentcheva 和 Andrzej Ruszczyński。《随机规划讲义》。工业与应用数学学会，2009年。doi: 10.1137/1.9780898718751。URL https://epubs.siam.org/doi/abs/10.1137/1.9780898718751.[52] Vladimir Solodkin、Andrew Veprikov 和 Aleksandr Beznosikov。马尔可夫随机性与非欧几何下优化问题的方法。arXiv预印本 arXiv:2408.01848，2024年。[53] Sebastian U. Stich。（随机）梯度法的统一最优分析，2019年。URL https://arxiv.org/abs/1907.04232.[54] Ben Taskar、Vassil Chatalbashev、Daphne Koller 和 Carlos Guestrin。学习结构化预测模型：一种大间隔方法。载于《第22届国际机器学习会议论文集》，第896–903页，2005年。[55] Alexandre B. Tsybakov。极小极大风险的下界，第77–135页。施普林格纽约，纽约州纽约市，2009年。ISBN 978-0-387-79052-7。doi: 10.1007/978-0-387-79052-7_2。URL https://doi.org/10.1007/978-0-387-79052-7_2.[56] Sharan Vaswani、Francis Bach 和 Mark Schmidt。过参数化模型的快速与更快收敛，以及一种加速感知机。载于《第22届国际人工智能与统计学会议》，第1195–1204页。机器学习研究会论文集，2019年。[57] 吴健、Saul Toscano-Palmerin、Peter I Frazier 和 Andrew Gordon Wilson。用于超参数调优的实用多保真贝叶斯优化。载于《人工智能中的不确定性》，第788–798页。机器学习研究会论文集，2020年。[58] David B Yudin 和 Arkadi S Nemirovskii。凸极值问题求解的信息复杂度与高效方法。Matekon，13(2):22–45，1976年。[59] 赵亚伟。数据联邦上的马尔可夫链镜像下降。arXiv预印本 arXiv:2309.14775，2023年。

附录

补充材料：梯度自由方法是高效应对马尔可夫随机性的关键

A 附录概述

本节介绍技术附录的整体结构。
在第 B 节中，我们介绍算法 1 所具有的额外对抗鲁棒性，并给出非光滑情形下结果的正式表述。

在第 节 C中，我们定义上界证明所用的简写符号
在第 D 节和 E节中，我们分别逐步介绍单点和双点设定下所有定理的引理与证明，涵盖光滑和非光滑问题。

在第 F节中，我们给出下界，并更详细地概述相关结果。
最后，在第 G节中，我们正式陈述所用的常识性事实。

B 附加结果

B.1 对抗噪声

除了证明主要结果表明算法在随机噪声下具有最优的缩放关系外，我们还证明了算法的鲁棒性。具体而言，本文考虑的预言机 F 返回的值可能带有加性、非随机且可能具有对抗性的误差 $\Delta(x) \leq \Delta$。

$$\hat{F}(x, Z) = F(x, Z) + \Delta(x). \tag{9}$$

我们将证明，对于 Δ 而言，只要误差处于可容忍的阈值范围内，就不会影响算法的收敛保证。该阈值在光滑情形和非光滑情形之间有所不同，但在单点设定和双点设定之间没有差异。关于 Δ的具体界见D和E节中的定理。

B.2 非光滑

在非光滑情形下，我们考虑一组类似的假设，但 f 不再必然是光滑的，甚至不一定可微。

假设 5

函数 f 在 $\mathbb{R}^d$上是 μ-强凸的，即存在常数 $\mu > 0$，使得对所有 x，$y \in \mathbb{R}^d$和 $\lambda \in [0;\ 1]$：以下不等式均成立

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \lambda(1-\lambda)\frac{\mu}{2}\|x-y\|^2$$

假设 6

函数 f 在 $\mathbb{R}^d$上是 G-利普希茨连续的，即存在常数 $G > 0$，使得对所有 x，$y \in \mathbb{R}^d$，下列不等式都成立：

$$|f(x) - f(y)| \leq G\|x-y\|.$$

同样，对于双点情形，我们需要如下推广：

假设 6′

对于所有 $Z \in \mathsf{Z}$，函数 $F(\cdot, Z)$ 在 $\mathbb{R}^d$上是 G-利普希茨连续的。

关于噪声水平，我们在单点情形下保留假设 4′。然而，在 f 不可微的情况下，双点情形下无法保留该假设。因此，我们还将使用函数无偏性。在这种情况下，我们不对噪声方差作任何额外假设，因为平滑函数的梯度已经有界，这是因为它是利普希茨连续且可微的。 G

假设 7

对于所有 $x \in \mathbb{R}^d$，均有 $\mathbb{E}_\pi F(x, Z) = f(x)$。

定理 3

设假设3 至6 成立，并考虑由算法 4 求解的问题（1）。那么，对于任意目标精度ε和批量大小乘数 B（符号说明见表 2和 3），并对 γ、t、p作适当选择后，确保 $\mathbb{E}\|x^N - x^*\|^2 \leq \varepsilon$ 所需的预言机调用次数有如下界：

$$B \cdot \tilde{\mathcal{O}}\left(\sqrt{\frac{\sqrt{d}G^2}{\mu^2\varepsilon}}\log\frac{1}{\varepsilon} + \frac{d\,(d+\tau)\,\sigma_1^2 G^2}{B\mu^4\varepsilon^3} + \frac{dG^2}{B\mu^2\varepsilon}\right) \quad \text{one-point oracle calls}.$$

我们给出以下定理。

定理 3′

假设假设 5， 6′、3 和 7 成立，并考虑由算法 4）求解的问题（1。那么，对于任意目标精度 ε 和批量大小乘数 B（符号说明见表 2 和 3 ），并在适当选择γ、t、p的情况下，确保 $\mathbb{E}\|x^N - x^*\|^2 \leq \varepsilon$ 所需的预言机调用次数有如下上界：

$$B \cdot \tilde{\mathcal{O}}\left(\sqrt{\frac{\sqrt{d}G^2}{\mu^2\varepsilon}}\log\frac{1}{\varepsilon} + \frac{(d+\tau)G^2}{B\mu^2\varepsilon}\right) \quad \text{two-point oracle calls}.$$

可以看出，只要混合时间小于问题的维度，结果就不依赖于混合时间。在采用 $\tau = 1$ 时，我们的结果与独立同分布噪声下的先前工作一致，正如表 1 中所述。

B.3 预言机复杂度界

在正文部分，我们重点关注达到精度 ε所需的预期预言机调用次数。这是衡量许多算法预言机复杂度的常用指标，但我们也可以要求更强的高概率界。通常，随着迭代次数增加，高概率界可以由中心极限定理轻松推出。然而，我们需要更加谨慎，因为我们的批量大小分布取决于迭代次数 N。我们回顾一下，批量大小 b_i 来自截断对数几何分布

$$b_i = \begin{cases} 2^{J_i} l, & 2^{J_i} < M \\ l, & \text{else} \end{cases}, \quad J_i \sim \text{Geom}(1/2)$$

并且 M随迭代次数 N按 $M \lesssim {}^{N}$对数N变化。基于此，我们对总和 $S_N = \sum b_i$应用伯恩斯坦不等式：

$$P(S_N > \alpha \mathbb{E}[S_N]) \leq \exp\left(-\frac{\alpha^2 N^2 [\mathbb{E}b_1]^2}{2N\mathbb{E}[b_1^2] + \frac{2\alpha}{3} MN[\mathbb{E}b_1]}\right) \leq$$
$$\exp\left(-c\frac{\alpha^2 N^2 l^2 (\log M)^2}{NMl^2 + \alpha MNl(\log M)}\right) \leq e^{-c(\log M)^2 \alpha}.$$

这表明，均值的归一化偏差具有次指数行为，从而证实在高概率意义下，期望值具有典型性。

C 符号与定义。

本节定义上界证明中使用的简写符号。一般符号与定义见表 2和 3。

马尔可夫误差：

$$h(x, Z) := F(x, Z) - f(x) \tag{10}$$

单样本梯度估计量：

$$\hat{g}_i := d\frac{\hat{F}(x + te_i, Z_i^{(+)}) - \hat{F}(x - te_i, Z_i^{(-)})}{2t} e_i \tag{11}$$

$$\tilde{g}_i := d\frac{F(x + te_i, Z_i^{(+)}) - F(x - te_i, Z_i^{(-)})}{2t} e_i \tag{12}$$

$$\overset{(10)}{=} d\frac{f(x + te_i) + h(x + te_i, Z_i^{(+)}) - f(x - te_i) - h(x - te_i, Z_i^{(-)})}{2t} e_i$$

$$g_i := d\frac{f(x + te_i) - f(x - te_i)}{2t} e_i \tag{13}$$

批量梯度估计量：

$$\hat{g}^j := \hat{g}_{rd}[2^j l] = \frac{1}{2^j l} \sum_{i=1}^{2^j l} \hat{g}_i \tag{14}$$

（不要与 $\hat{}g^k$ 混淆，后者是在第 k 次迭代时计算得到的 $\hat{}g_{ml}$ ）

$$\tilde{g}^j := \frac{1}{2^j l} \sum_{i=1}^{2^j l} \tilde{g}_i \tag{15}$$

$$g^j := \frac{1}{2^j l} \sum_{i=1}^{2^j l} g_i \tag{16}$$

方向梯度：

$$\nabla_{e_i} f(x_0) := d\langle \nabla f(x_0), e_i \rangle e_i \tag{17}$$

$$\nabla_{e_i} F_i := d\langle \nabla F(x, Z_i), e_i \rangle e_i \tag{18}$$

其他：

$$\mathbb{E}_e := \mathbb{E}_{e_1, e_2, \ldots, e_{2^j l}} \tag{19}$$

$\mathbb{E}_Z := \mathbb{E}_{Z_1, Z_2, \ldots, Z_{2^j l}}$, where $Z_1 \sim \xi$ - arbitrary initial distribution on $(\mathsf{Z}, \mathcal{Z})$

$\mathbb{E} := \mathbb{E}_Z \mathbb{E}_e$

$\mathcal{F}_k := \sigma(x^1, x^2, \ldots\ldots, x^k)$ - 前 k 次迭代的 σ-代数

$\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot | \mathcal{F}_k]$

$$r^N := \frac{1}{\mu}(f(x_f^N) - f(x^*)) + \left\|x^N - x^*\right\|^2 \tag{20}$$

D 单点结果的证明

D.1 马尔可夫方差缩减

引理 3 （引理 1的扩展版本）。 设假设 3 和 4 （4′）成立。则对于任意 $n \geq 1$ 和 $x \in \mathbb{R}^d$以及任意初始分布 ξ 在 （Z, $\mathcal{Z}$）上，我们有

$$\mathbb{E}_Z \left[\left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{e_i} \left[h(x + te_i, Z_i) e_i \right] \right)^2 \right] \lesssim \frac{\tau}{dn} \sigma_1^2, \tag{21}$$

$$\mathbb{E}_Z \left[\left\| \frac{1}{n} \sum_{i=1}^n \nabla F(x, Z_i) - \nabla f(x) \right\|^2 \right] \lesssim \frac{\tau}{n} \sigma_2^2, \tag{22}$$

Proof. Th证明 f of (22 can be found 在引理 1 中 o f Beznosi k$^{\text{ov et a}}$ l. [5].
假设 4 下的证明依赖于这样一个事实：上述引理 1 对随机实现只要求满足以下两个条件 $\nabla F(x, Z_i)$:

$$\begin{cases} \mathbb{E}_\pi \nabla F(x, Z_i) = \nabla f(x) \\ \|\nabla F(x, Z_i) - \nabla f(x)\|^2 \leq \sigma_2^2 \end{cases}$$

记 $h_t(x, Z_i) := \mathbb{E}_e \left[h(x + te, Z_i) e \right]$, $e \sim RS_2^d(1)$。

Thus (21) $\Leftrightarrow \mathbb{E}_Z \left[\left(\frac{1}{n} \sum_{i=1}^n h_t(x, Z_i) \right)^2 \right] \lesssim \frac{\tau}{n} \frac{\sigma_1^2}{d} \Leftrightarrow \begin{cases} \mathbb{E}_\pi h_t(x, Z_i) = 0 \\ \|h_t(x, Z_i)\|^2 \lesssim \frac{\sigma_1^2}{d} \end{cases}$

让我们证明这两个等式，先从无偏性开始:

$$\mathbb{E}_\pi h_t(x, Z_i) = \mathbb{E}_\pi \mathbb{E}_e \left[h(x + te, Z_i) e \right] = \mathbb{E}_e \mathbb{E}_\pi \left[h(x + te, Z_i) e \right] \overset{(4)}{=} \mathbb{E}_e 0 = 0$$

$$\begin{aligned} \|h_t(x, Z_i)\|^2 &= \|\mathbb{E}_e \left[h(x + te, Z_i) e \right]\|^2 \\ &= \langle \mathbb{E}_e \left[h(x + te, Z_i) e \right], h_t(x, Z_i) \rangle \\ &\overset{①}{=} \mathbb{E}_e \left[h(x + te, Z_i) \cdot \langle e, h_t(x, Z_i) \rangle \right] \\ &\overset{②}{\leq} \sqrt{\mathbb{E}_e h(x + te, Z_i)^2} \cdot \sqrt{\mathbb{E}_e \langle e, h_t(x, Z_i) \rangle^2} \\ &\overset{(81)}{=} \sqrt{\mathbb{E}_e h(x + te, Z_i)^2} \cdot \sqrt{\frac{1}{d} \|h_t(x, Z_i)\|^2} \\ &\overset{(4)}{\leq} \sqrt{\sigma_1^2} \cdot \sqrt{\frac{1}{d} \|h_t(x, Z_i)\|^2}, \end{aligned}$$

其中，1 成立是因为 $h_t(x, Z_i)$ 不依赖于 e；而 2 是对以下点积应用柯西-施瓦茨不等式所得：$\langle x(e) \; y(e) \rangle := \mathbb{E}_e [x \cdot y]$。

为完成证明，我们将得到的不等式两边平方:

$$\|h_t(x, Z_i)\|^2 \leq \frac{\sqrt{\sigma_1^2}}{\sqrt{d}} \cdot \sqrt{\|h_t(x, Z_i)\|^2} \Rightarrow \|h_t(x, Z_i)\|^2 \leq \frac{\sigma_1^2}{d}.$$

□

D.2 平滑函数的性质

以下引理阐明了 l_2-球平滑函数的关键性质

引理 4. 假设 f 是凸的。则对所有 $x \in \mathbb{R}^d$，以下结论成立

$$\begin{aligned} &\textit{If } f \textit{ is } L\textit{-smooth / } G\textit{-Lipschitz / } \mu\textit{-strongly convex [Assumptions 1, 2 and 6],} \\ &\textit{then } f_t \textit{ from (2) is also } L\textit{-smooth / } G\textit{-Lipschitz / } \mu\textit{-strongly convex.} \end{aligned} \tag{23}$$

$$\nabla f_t(x) = \mathbb{E}_e \left[g(x) \right], \tag{24}$$

$$f_t(x) \geq f(x), \tag{25}$$

如果 f 还满足 G-利普希茨条件：

$$f_t(x) \leq f(x) + Gt, \tag{26}$$

$$f_t \textit{ is } L\textit{-smooth with } L = \frac{\sqrt{d}G}{t}, \tag{27}$$

如果 f 还满足 L-光滑条件：

$$f_t(x) \leq f(x) + Lt^2, \tag{28}$$

$$\|\nabla f(x) - \nabla f_t(x)\|^2 \leq L^2 t^2, \tag{29}$$

$$\|\nabla f_t(x)\|^2 \geq \tfrac{1}{2}\|\nabla f(x)\|^2 - L^2 t^2. \tag{30}$$

证明。 证明 (23)，我们从 G-利普希茨性开始：

$$\begin{aligned} |f_t(x) - f_t(y)| &= |\mathbb{E}_r \left[f(x + tr) - f(y + tr) \right]| \\ &\overset{(80)}{\leq} \mathbb{E}_r |f(x + tr) - f(y + tr)| \\ &\overset{(6)}{\leq} \mathbb{E}_r G \|x - y\| = G\|x - y\|. \end{aligned}$$

接下来， L-光滑性的证明类似。最后， μ-强凸性以及 f_t、(24)、(25) 和 (28) 均由 [1]中的引理 A2-A3 证明。

（26）和（27）可见于 Gasnikov 等人的第 4.1 节 [21]。
我们依次证明其余不等式。
（29）的证明：

$$\begin{aligned} \|\nabla f(x) - \nabla f_t(x)\|^2 &= \|\nabla f(x) - \mathbb{E}_r \nabla f(x + tr)\|^2 \\ &= \|\mathbb{E}_r \left[\nabla f(x) - \nabla f(x + tr) \right]\|^2 \\ &\overset{(80)}{\leq} \mathbb{E}_r \|\nabla f(x) - \nabla f(x + tr)\|^2 \\ &\overset{(1)}{\leq} \mathbb{E}_r L^2 t^2 = L^2 t^2. \end{aligned}$$

（30）的证明：

$$\begin{aligned} \|\nabla f_t(x)\|^2 &= \|\nabla f(x) + [\nabla f_t(x) - \nabla f(x)]\|^2 \\ &\overset{①}{\geq} \frac{1}{2}\|\nabla f(x)\|^2 - \|\nabla f_t(x) - \nabla f(x)\|^2 \\ &\overset{(29)}{\geq} \frac{1}{2}\|\nabla f(x)\|^2 - L^2 t^2, \end{aligned}$$

其中 **1** 用到了 $\|a + b\|^2 \geq 1/2\|a\|^2 - \|b\|^2$。 □

D.3 梯度近似的不等式

引理 5. 假设假设条件*1*、假设条件 *3* 和假设条件 *4*。则对于任意初始分布 ξ 在 (Z,$\mathcal{Z}$) 以及所有 $x \in \mathbb{R}^d$:

$$\|\hat{g}^j - \tilde{g}^j\|^2 \leq \frac{d^2\Delta^2}{t^2}, \tag{31}$$

$$\mathbb{E}\|\tilde{g}_i - g_i\|^2 \leq \frac{d^2\sigma_1^2}{t^2}, \tag{32}$$

$$\mathbb{E}\|\mathbb{E}_e\left[\tilde{g}^j - g^j\right]\|^2 \leq \frac{dC_1\tau\sigma_1^2}{t^2 2^j l}, \tag{33}$$

$$\mathbb{E}\|g_i - \nabla_{e_i} f\|^2 \leq \frac{d^2L^2t^2}{4}, \tag{34}$$

$$\mathbb{E}\|\tilde{g}^j - \mathbb{E}_e\tilde{g}^j\|^2 \leq \frac{3}{2^j l}\left[\frac{d^2\sigma_1^2}{t^2} + \frac{d^2L^2t^2}{4} + d\|\nabla f\|^2\right], \tag{35}$$

$$\mathbb{E}\|\tilde{g}^j - \mathbb{E}_e g^j\|^2 \lesssim \frac{d(d+\tau)\sigma_1^2}{t^2 2^j l} + \frac{d^2L^2t^2}{2^j l} + \frac{d\|\nabla f\|^2}{2^j l}, \tag{36}$$

$$\mathbb{E}\|\hat{g}^j - \nabla f_t\|^2 \lesssim \frac{d^2\Delta^2}{t^2} + \frac{d(d+\tau)\sigma_1^2}{t^2 2^j l} + \frac{d^2L^2t^2}{2^j l} + \frac{d\|\nabla f\|^2}{2^j l}, \tag{37}$$

$$\|\mathbb{E}\hat{g}^j - \nabla f_t\|^2 \leq \frac{2d^2\Delta^2}{t^2} + \frac{2dC_1\tau\sigma_1^2}{t^2 2^j l}. \tag{38}$$

$$\tag{39}$$

证明。 我们逐一证明所有估计，从（31）开始:

$$\begin{aligned}
\left\|\hat{g}^j - \tilde{g}^j\right\|^2 &\overset{(14),(15)}{=} \left\|\frac{1}{2^j l}\sum_{i=1}^{2^j l}[\hat{g}_i - \tilde{g}_i]\right\|^2 \\
&\overset{(11),(12)}{=} \frac{d^2}{4t^2}\left\|\frac{1}{2^j l}\sum_{i=1}^{2^j l}[\hat{F}(x+te_i, Z_i^+) - \hat{F}(x-te_i, Z_i^-) \right. \\
&\qquad \left. -F(x+te_i, Z_i^+) + F(x-te_i, Z_i^-)]e_i\right\|^2 \\
&\overset{(9)}{=} \frac{d^2}{4t^2}\left\|\frac{1}{2^j l}\sum_{i=1}^{2^j l}\left[\Delta(x+te_i) - \Delta(x-te_i)\right]e_i\right\|^2 \\
&\overset{(77)}{\leq} \frac{d^2}{4t^2 2^j l}\sum_{i=1}^{2^j l}\left\|\left[\Delta(x+te_i) - \Delta(x-te_i)\right]e_i\right\|^2 \\
&\overset{\|e_i\|=1}{=} \frac{d^2}{4t^2 2^j l}\sum_{i=1}^{2^j l}\left|\Delta(x+te_i) - \Delta(x-te_i)\right|^2 \\
&\overset{(9)}{\leq} \frac{d^2}{4t^2}4\Delta^2 \\
&= \frac{d^2\Delta^2}{t^2}.
\end{aligned}$$

证明（32）:

$$\begin{aligned}
\mathbb{E}\|\tilde{g}_i - g_i\|^2 &\overset{(12),(13)}{=} \mathbb{E}\left\|d\frac{h(x+te_i, Z_i^+) - h(x-te_i, Z_i^-)}{2t}e_i\right\|^2 \\
&\overset{\|e_i\|=1}{=} \frac{d^2}{4t^2}\mathbb{E}\left[h(x+te_i, Z_i^+) - h(x-te_i, Z_i^-)\right]^2 \\
&\overset{(77),(4)}{\leq} \frac{d^2\sigma_1^2}{t^2}.
\end{aligned}$$

证明（33）：

$$
\begin{aligned}
& \mathbb{E}\left\|\mathbb{E}_e\left[\tilde{g}^j - g^j\right]\right\|^2 \\
\overset{(12),(16)}{=} & \mathbb{E}\left\|\frac{1}{2^j l}\sum_{i=1}^{2^j l}\mathbb{E}_e\left[d\frac{h(x+te_i,Z_i^+)-h(x-te_i,Z_i^-)}{2t}e_i\right]\right\|^2 \\
= & \frac{d^2}{t^2}\mathbb{E}\left\|\frac{1}{2^j l}\sum_{i=1}^{2^j l}\mathbb{E}_e\left[\frac{h(x+te_i,Z_i^+)e_i - h(x-te_i,Z_i^-)e_i}{2}\right]\right\|^2 \\
\overset{(77)}{\leq} & \frac{d^2}{t^2}\frac{1}{2}\left[\mathbb{E}\left\|\frac{1}{2^j l}\sum_{i=1}^{2^j l}\mathbb{E}_e\left[h(x+te_i,Z_i^+)e_i\right]\right\|^2 + \mathbb{E}\left\|\frac{1}{2^j l}\sum_{i=1}^{2^j l}\mathbb{E}_e\left[h(x-te_i,Z_i^-)e_i\right]\right\|^2\right] \\
\overset{(21)}{\leq} & \frac{dC_1\tau\sigma_1^2}{t^2 2^j l}.
\end{aligned}
$$

（34）的证明：

$$
\begin{aligned}
& \mathbb{E}\|g_i - \nabla_{e_i} f\|^2 \\
\overset{(13),(17)}{=} & \mathbb{E}\left\|d\frac{f(x+te_i)-f(x-te_i)}{2t}e_i - d\langle\nabla f(x),e_i\rangle e_i\right\|^2 \\
= & d^2\mathbb{E}\left|\frac{f(x+te_i)-f(x)+f(x)-f(x-te_i)-2t\langle\nabla f(x),e_i\rangle}{2t}\right|^2 \\
= & d^2\mathbb{E}\left|\frac{f(x+te_i)-f(x)-\langle\nabla f(x),te_i\rangle}{2t} + \frac{f(x)-f(x-te_i)+\langle\nabla f(x),-te_i\rangle}{2t}\right|^2 \\
\overset{①}{\leq} & \frac{2d^2}{4t^2}\left(\frac{L^2t^4}{4}+\frac{L^2t^4}{4}\right) \\
= & \frac{d^2L^2t^2}{4},
\end{aligned}
$$

其中 ① 用到了假设 1、（74）和（77）。

（35）的证明：

$$
\begin{aligned}
\mathbb{E}\left\|\tilde{g}^j - \mathbb{E}_e\tilde{g}^j\right\|^2 \overset{(15)}{=} & \mathbb{E}_Z\mathbb{E}_e\left\|\frac{1}{2^j l}\sum_{i=1}^{2^j l}\left[\tilde{g}_i - \mathbb{E}_{e_i}\tilde{g}_i\right]\right\|^2 \\
\overset{①}{=} & \mathbb{E}_Z\mathbb{E}_e\frac{1}{2^{2j}l^2}\sum_{i=1}^{2^j l}\|\tilde{g}_i - \mathbb{E}_{e_i}\tilde{g}_i\|^2 \\
\overset{(78)}{\leq} & \frac{1}{2^{2j}l^2}\sum_{i=1}^{2^j l}\mathbb{E}_Z\mathbb{E}_e\|\tilde{g}_i\|^2 \\
\overset{(77)}{\leq} & \frac{3}{2^{2j}l^2}\sum_{i=1}^{2^j l}\mathbb{E}\left[\|\tilde{g}_i - g_i\|^2 + \|g_i - \nabla_{e_i}f\|^2 + \|\nabla_{e_i}f\|^2\right] \\
\overset{(32),(34),(81)}{\leq} & \frac{3}{2^j l}\left[\frac{d^2\sigma_1^2}{t^2} + \frac{d^2L^2t^2}{4} + d\|\nabla f\|^2\right],
\end{aligned}
$$

其中 ① 成立，因为 $\tilde{g}_i$ 关于 e_i 和 $\mathbb{E}_e\left[\tilde{g}_i - \mathbb{E}_{e_i}\left[\tilde{g}_i\right]\right] = 0$相互独立。式（36）的证明：

$$
\mathbb{E}\left\|\tilde{g}^j - \mathbb{E}_e g^j\right\|^2 \overset{(77)}{\leq} 2\mathbb{E}\left[\left\|\tilde{g}^j - \mathbb{E}_e\tilde{g}^j\right\|^2 + \|\mathbb{E}_e\tilde{g}^j - \mathbb{E}_e g^j\|^2\right]
$$

$$\overset{(35),(33)}{\leq} \quad 2\left[\frac{3}{2^j l}\left[\frac{d^2\sigma_1^2}{t^2}+\frac{d^2L^2t^2}{4}+d\|\nabla f\|^2\right]+\frac{dC_1\tau\sigma_1^2}{t^2 2^j l}\right]$$
$$\lesssim \quad \frac{d\left(d+\tau\right)\sigma_1^2}{t^2 2^j l}+\frac{d^2L^2t^2}{2^j l}+\frac{d\|\nabla f\|^2}{2^j l}.$$

式（37）的证明：

$$\mathbb{E}\left\|\hat{g}^j-\nabla f_t\right\|^2 \overset{(77)}{\leq} 2\mathbb{E}\left[\left\|\hat{g}^j-\tilde{g}^j\right\|^2\|+\|\tilde{g}^j-\mathbb{E}_e g^j\|^2\right]$$
$$\overset{(31),(36)}{\lesssim} \frac{d^2\Delta^2}{t^2}+\frac{d\left(d+\tau\right)\sigma_1^2}{t^2 2^j l}+\frac{d^2L^2t^2}{2^j l}+\frac{d\|\nabla f\|^2}{2^j l}.$$

（38）的证明：

$$\left\|\mathbb{E}\hat{g}^j-\nabla f_t\right\|^2 \overset{(77)}{\leq} 2\left\|\mathbb{E}\hat{g}^j-\mathbb{E}\tilde{g}^j\right\|^2+2\|\mathbb{E}\tilde{g}^j-\nabla f_t\|^2$$
$$\overset{(24)}{=} 2\left\|\mathbb{E}\hat{g}^j-\mathbb{E}\tilde{g}^j\right\|^2+2\|\mathbb{E}_Z\mathbb{E}_e\tilde{g}^j-\mathbb{E}_e g^j\|^2$$
$$\overset{(80)}{\leq} 2\left\|\mathbb{E}\hat{g}^j-\mathbb{E}\tilde{g}^j\right\|^2+2\mathbb{E}_Z\|\mathbb{E}_e\tilde{g}^j-\mathbb{E}_e g^j\|^2$$
$$\overset{(31),(33)}{\leq} \frac{2d^2\Delta^2}{t^2}+\frac{2dC_1\tau\sigma_1^2}{t^2 2^j l}.$$

□

引理 6. 假设满足假设 6、假设 2以及假设 4。那么，对于任意定义在 ξ 上的初始分布 （Z,$\mathcal{Z}$）以及所有 $x\in\mathbb{R}^d$，以下不等式成立：

$$\mathbb{E}\|g_i\|^2 \lesssim dG^2, \tag{40}$$
$$\mathbb{E}\left\|\tilde{g}^j-\mathbb{E}_e\tilde{g}^j\right\|^2 \lesssim \tfrac{2}{2^j l}\left[\tfrac{d^2\sigma_1^2}{t^2}+dG^2\right], \tag{41}$$
$$\mathbb{E}\left\|\tilde{g}^j-\mathbb{E}_e g^j\right\|^2 \lesssim \tfrac{dC_1(d+\tau)\sigma_1^2}{t^2 2^j l}+\tfrac{dG^2}{2^j l}. \tag{42}$$
$$\mathbb{E}\left\|\hat{g}^j-\nabla f_t\right\|^2 \lesssim \tfrac{d^2\Delta^2}{t^2}+\tfrac{dC_1(d+\tau)\sigma_1^2}{t^2 2^j l}+\tfrac{dG^2}{2^j l}. \tag{43}$$

证明。
（40）的证明：

$$\mathbb{E}\|g_i\|^2 \overset{(11)}{=} \frac{d^2}{4t^2}\mathbb{E}\left|f(x+te_i)-f(x-te_i)\right|^2$$
$$\overset{(77)}{\leq} \frac{d^2}{2t^2}\mathbb{E}\left[\left|f(x+te_i)-\mathbb{E}_{e_i}f(x+te_i)\right|^2+\left|\mathbb{E}_{e_i}f(x+te_i)-f(x-te_i)\right|^2\right]$$
$$\overset{①}{\leq} \frac{d^2}{t^2}\mathbb{E}\left|f(x+te_i)-\mathbb{E}_{e_i}f(x+te_i)\right|^2$$
$$\overset{②}{\lesssim} dG^2,$$

其中，1 利用了 e_i 的分布是对称的这一事实；2 利用了如下事实：对于 f，若其为 G-Lipshitz 且满足 $e\in RS_2^d$ **(1)**，则有 $\mathbb{E}[f(e)-\mathbb{E}_e f(e)]^2\lesssim\frac{G^2}{d}$ [。推理方式与 [49], 引理 9]相同。 —

（41）的证明：

$$\mathbb{E}\left\|\tilde{g}^j-\mathbb{E}_e\tilde{g}^j\right\|^2 \overset{①}{\leq} \frac{1}{2^{2j}l^2}\sum_{i=1}^{2^j l}\mathbb{E}_Z\mathbb{E}_e\|\tilde{g}_i\|^2$$

$$\overset{(77)}{\leq} \quad \frac{2}{2^{2j}l^2}\sum_{i=1}^{2^{j}l}\mathbb{E}\left[\|\tilde{g}_i - g_i\|^2 + \|g_i\|^2\right]$$
$$\overset{(32),(40)}{\leq} \quad \frac{2}{2^{j}l}\left[\frac{d^2\sigma_1^2}{t^2} + dG^2\right],$$

其中1 与（35）类似。
（42）的证明：

$$\mathbb{E}\left\|\tilde{g}^j - \mathbb{E}_e g^j\right\|^2 \overset{(77)}{\leq} 2\mathbb{E}\left[\left\|\tilde{g}^j - \mathbb{E}_e\tilde{g}^j\right\|^2 + \|\mathbb{E}_e\tilde{g}^j - \mathbb{E}_e g^j\|^2\right]$$
$$\overset{(41),(33)}{\leq} 2\left[\frac{2}{2^{j}l}\left[\frac{d^2\sigma_1^2}{t^2} + dG^2\right] + \frac{dC_1\tau\sigma_1^2}{t^2 2^{j}l}\right]$$
$$\lesssim \frac{d\left(d+\tau\right)\sigma_1^2}{t^2 2^{j}l} + \frac{dG^2}{2^{j}l}.$$

（43）的证明：

$$\mathbb{E}\left\|\hat{g}^j - \nabla f_t\right\|^2 \overset{(77)}{\leq} 2\mathbb{E}\left[\left\|\hat{g}^j - \tilde{g}^j\right\|^2 + \left\|\mathbb{E}_e\tilde{g}^j - \nabla f_t\right\|^2\right]$$
$$\overset{(31),(42)}{\lesssim} \frac{d^2\Delta^2}{t^2} + \frac{d\left(d+\tau\right)\sigma_1^2}{t^2 2^{j}l} + \frac{dG^2}{2^{j}l}.$$

□

引理 7（引理 2）。设假设 *3* 和 *4* 成立。对于任意初始分布 ξ 在 $(\mathsf{Z},\mathcal{Z})$，梯度估计 $\hat{}g_{ml}$ 满足 $\mathbb{E}[\hat{}g_{ml}] = \mathbb{E}\left[\hat{}g_{rd}\left[2^{\lfloor\log_2 M\rfloor}l\right]\right]$。此外，

$$\|\nabla f_t(x) - \mathbb{E}[\hat{g}_{ml}]\|^2 \lesssim \frac{d^2\Delta^2}{t^2} + \frac{d\tau\sigma_1^2}{t^2 MB}. \tag{44}$$

此外，在假设 *1*

$$\mathbb{E}[\|\nabla f_t(x) - \hat{g}_{ml}\|^2] \lesssim \frac{d^2\Delta^2}{t^2} + \frac{d\left(d+\tau\right)\sigma_1^2}{t^2 B} + \frac{d^2L^2t^2}{B} + \frac{d}{B}\|\nabla f\|^2. \tag{45}$$

而在假设 *6*

$$\mathbb{E}[\|\nabla f_t(x) - \hat{g}_{ml}\|^2] \lesssim \frac{d^2\Delta^2}{t^2} + \frac{d\left(d+\tau\right)\sigma_1^2}{t^2 B} + \frac{dG^2}{B}. \tag{46}$$

证明。回忆一下，$\hat{}g_{ml}$ 是基线估计 $\hat{}g_{rd}[l]$ $\overset{(14)}{=}$ $\hat{}g^0$ 与细化项 $2^J[\hat{}g^J - \hat{}g^{J-1}]$之和。于是，为了证明$\mathbb{E}[\hat{}g_{ml}] = \mathbb{E}\hat{}g^{\lfloor\log_2 M\rfloor}$，我们使用全期望定律：

$$\mathbb{E}[\hat{g}_{ml}] = \mathbb{E}\left[\mathbb{E}_J[\hat{g}_{ml}]\right] = \mathbb{E}[\hat{g}^0] + \sum_{j=1}^{\lfloor\log_2 M\rfloor}\mathbb{P}\{J=j\}\cdot 2^j\mathbb{E}[\hat{g}^j - \hat{g}^{j-1}] \tag{47}$$
$$= \mathbb{E}[\hat{g}^0] + \sum_{j=1}^{\lfloor\log_2 M\rfloor}\mathbb{E}[\hat{g}^j - \hat{g}^{j-1}] = \mathbb{E}\hat{g}^{\lfloor\log_2 M\rfloor}.$$

这立即有助于我们证明命题（44）：

$$\|\nabla f_t(x) - \mathbb{E}\hat{g}_{ml}\|^2 = \left\|\nabla f_t(x) - \mathbb{E}\left[\hat{g}^{\lfloor\log_2 M\rfloor}\right]\right\|^2 \overset{(38)}{\leq} \frac{2d^2\Delta^2}{t^2} + \frac{2dC_1\tau\sigma_1^2}{t^2 2^{\lfloor\log_2 M\rfloor}l} \overset{l\geq B}{\lesssim} \frac{d^2\Delta^2}{t^2} + \frac{d\tau\sigma_1^2}{t^2 MB}.$$

证明命题（45）时，我们同样从全期望开始：

$$
\begin{aligned}
& \mathbb{E}[\|\nabla f(x)-\hat{g}_{ml}\|^2] \\
\overset{(77)}{\leq} \quad & 2\mathbb{E}[\left\|\nabla f(x)-\hat{g}^0\right\|^2]+2\mathbb{E}[\|\hat{g}_{ml}-\hat{g}^0\|^2] \\
= \quad & 2\mathbb{E}[\left\|\nabla f(x)-\hat{g}^0\right\|^2]+2\sum_{j=1}^{\lfloor \log_2 M\rfloor}\mathbb{P}\{J=j\}\cdot 4^j\mathbb{E}[\|\hat{g}^j-\hat{g}^{j-1}\|^2] \\
= \quad & 2\mathbb{E}[\left\|\nabla f(x)-\hat{g}^0\right\|^2]+2\sum_{j=1}^{\lfloor \log_2 M\rfloor}2^j\mathbb{E}[\|\hat{g}^j-\hat{g}^{j-1}\|^2] \\
\overset{①}{=} \quad & 2\mathbb{E}[\left\|\nabla f(x)-\hat{g}^0\right\|^2]+2\sum_{j=1}^{\lfloor \log_2 M\rfloor}2^j\mathbb{E}[\|\tilde{g}^j-\tilde{g}^{j-1}\|^2] \\
\overset{(77)}{\leq} \quad & 2\mathbb{E}[\left\|\nabla f(x)-\hat{g}^0\right\|^2]+4\sum_{j=1}^{\lfloor \log_2 M\rfloor}2^j\left(\mathbb{E}\|\tilde{g}^j-\mathbb{E}_e g^j\|^2+\mathbb{E}\|\mathbb{E}_e g^{j-1}-\tilde{g}^{j-1}\|^2\right) \\
\leq \quad & 2\mathbb{E}[\left\|\nabla f(x)-\hat{g}^0\right\|^2]+16\sum_{j=0}^{\lfloor \log_2 M\rfloor}2^j\mathbb{E}[\|\mathbb{E}_e g^j-\tilde{g}^j\|^2] \\
\overset{(37),(36)}{\lesssim} \quad & 2\left[\frac{d^2\Delta^2}{t^2}+\frac{d\left(d+\tau\right)\sigma_1^2}{t^2 l}+\frac{d^2L^2t^2}{l}+\frac{d}{l}\cdot\|\nabla f\|^2\right]+ \\
& 16\sum_{j=0}^{\lfloor \log_2 M\rfloor}2^j\left[\frac{d\left(d+\tau\right)\sigma_1^2}{t^2 2^j l}+\frac{d^2L^2t^2}{2^j l}+\frac{d\|\nabla f\|^2}{2^j l}\right] \\
\overset{l\geq \log_2 M\cdot B}{\lesssim} \quad & 2\left[\frac{d^2\Delta^2}{t^2}+\frac{d\left(d+\tau\right)\sigma_1^2}{t^2 B}+\frac{d^2L^2t^2}{B}+\frac{d}{B}\cdot\|\nabla f\|^2\right]+ \\
& 16\left[\frac{d\left(d+\tau\right)\sigma_1^2}{t^2 B}+\frac{d^2L^2t^2}{B}+\frac{d\|\nabla f\|^2}{B}\right] \\
\lesssim \quad & \frac{d^2\Delta^2}{t^2}+\frac{d\left(d+\tau\right)\sigma_1^2}{t^2 B}+\frac{d^2L^2t^2}{B}+\frac{d}{B}\|\nabla f\|^2,
\end{aligned}
$$

其中，1 利用了 $\hat{}g^j-\hat{}g^{j-1}=\tilde{g}^j-\tilde{g}^{j-1}$，因为 $\tilde{g}^j-\hat{}g^j \overset{(31)}{=} \tilde{g}^{j-1}-\hat{}g^{j-1}$。

(46) 的证明完全相同，只需将 (37) 和 (36) 替换为 (43) 和 (42)

$$
\mathbb{E}[\|\nabla f(x)-\hat{g}_{ml}\|^2]\lesssim\frac{d^2\Delta^2}{t^2}+\frac{d\left(d+\tau\right)\sigma_1^2}{t^2 B}+\frac{dG^2}{B}.
$$

□

D.4 定理 1 的证明

定理1 的证明需要两个技术性引理。

引理 8. 假设满足假设 *1* 和 *2*。那么，对于算法 *1* 在 $\theta=(p\eta^{-1}-1)/(\beta p\eta^{-1}-1)$、$\theta>0$、 $\eta\geq 1$、 $p>0$ 以及任意 $\alpha>0$ 的迭代点，以下结论成立：

$$
\begin{aligned}
\mathbb{E}_k[\left\|x^{k+1}-x^*\right\|^2] \leq & (1+\alpha p\gamma\eta)(1-\beta)\left\|x^k-x^*\right\|^2+(1+\alpha p\gamma\eta)\beta\|x_g^k-x^*\|^2 \\
& +(1+\alpha p\gamma\eta)(\beta^2-\beta)\left\|x^k-x_g^k\right\|^2+p^2\eta^2\gamma^2\mathbb{E}_k[\left\|\hat{g}^k\right\|^2] \\
& -2\eta^2\gamma\langle\nabla f(x_g^k),x_g^k+(p\eta^{-1}-1)x_f^k-\eta^{-1}px^*\rangle \\
& +\frac{p\eta\gamma}{\alpha}\left\|\mathbb{E}_k\left[\hat{g}^k\right]-\nabla f(x_g^k)\right\|^2.
\end{aligned}
\tag{48}
$$

引理 9。 假设满足假设 *1* 和 *2*。问题 (4) 由算法 *1*求解。于是，对于任意 $u\in\mathbb{R}^d$，我们有

$$
\mathbb{E}_k\left[f(x_f^{k+1})\right]\leq f(u)-\langle\nabla f(x_g^k),u-x_g^k\rangle-\frac{\mu}{2}\left\|u-x_g^k\right\|^2-\frac{\gamma}{2}\|\nabla f(x_g^k)\|^2
$$

$$+\frac{\gamma}{2}\left\|\mathbb{E}_k\left[\hat{g}^k\right]-\nabla f(x_g^k)\right\|^2+\frac{L\gamma^2}{2}\mathbb{E}_k\left[\|\hat{g}^k\|^2\right].$$

这些结果由 Beznosikov 等人在 [5] 中以引理 5 和引理 6 的形式证明，所用记号略有不同：$\hat{}\ f$ 对应于 f而 $\hat{}g$ 对应于 g。

引理 10 (步长调优)。 给定经过 N 次迭代后的优化误差上界为

$$r^N \le \exp(-N\Gamma a)r^0+\Gamma b$$

以及步长 $\Gamma \le \frac{1}{u}$ 的上界，存在一个常数步长 $\Gamma_0 \le \frac{1}{u}$，使得

$$r^N=\tilde{\mathcal{O}}\left(\exp\left(-\frac{Na}{u}\right)r^0+\frac{b}{aN}\right)$$

等价地，达到 $r^N \lesssim \varepsilon$所需的迭代次数：

$$N=\tilde{\mathcal{O}}\left(\frac{u}{a}\ln\varepsilon^{-1}+\frac{b}{a\varepsilon}\right) \tag{49}$$

证明。 该设定是第 4 节 Stich [53] 所考虑设定的一个更简单版本，因此我们将以类似于其引理 2 的方式调优 Γ：

$$\Gamma:=\min\left(\frac{\ln\max(2,ar^0N/b)}{aN},\frac{1}{u}\right)$$

如果 $\frac{1}{u}\le\frac{\ln\max(2,\ ar^0N/b)}{aN}$，则 $\Gamma:=\frac{1}{u}$。

$$r^N\le\exp\left(\frac{-Na}{u}\right)r^0+\frac{b}{u}\le\exp\left(\frac{-Na}{u}\right)r^0+\frac{b\ln(\ldots)}{aN}=\tilde{\mathcal{O}}\left(\exp\left(-\frac{Na}{u}\right)r^0+\frac{b}{aN}\right)$$

否则 $\frac{\ln\max(2,\ ar^0N/b}{aN}\le\frac{1}{u}$，且 $\Gamma:=\frac{\ln\max(2,ar^0N/b)}{aN}$，立即得到 $\Gamma b=\tilde{\mathcal{O}}(\frac{b}{aN})$。

$$\exp(-N\Gamma a)r^0=\exp\left(-\ln\max(2,ar^0N/b)\right)=\frac{1}{\max(2,ar^0N/b)}$$

若 $ar^0N/b>2$，我们还可得到 $\tilde{\mathcal{O}}(\frac{b}{aN})$；否则 $\frac{1}{2}\le\frac{b}{aNr^0}$，我们同样可得到 $\tilde{\mathcal{O}}(\frac{b}{aN})$。

为完成证明，我们应当消除步长目前依赖于迭代次数这一问题。这可以通过重启过程轻松实现：该过程将以步长 $\Gamma(N)$运行算法 $N=1,2,4,\ldots$ 次迭代。

□

定理 4 (定理 *1*)

设假设 1 至 4 成立，并考虑由算法 4求解的问题 (1。则对于适当选择的隐藏参数（其中 $p\simeq\frac{B}{B+d}$），以及任意选择的自由参数（见表 3），有：

$$\mathbb{E}r^N\lesssim\exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right)r^0+\frac{p\sqrt{\gamma}}{\mu^{3/2}}\cdot\left[\sigma_1^2\frac{d\,(d+\tau)}{t^2B}+t^2\frac{L^2d^2}{B}\right]+\frac{\Delta^2d^2}{\mu^2t^2}+\frac{Lt^2}{\mu}$$

此外，对于任意的 $\varepsilon\gtrsim\frac{d\Delta\sqrt{L}}{\mu^{3/2}}$以及适当选择的 t 和 γ，确保 $r^N\lesssim\varepsilon$ 所需的预言机调用次数有如下界：

$$B\cdot\tilde{\mathcal{O}}\left(\max\left[1,\frac{d}{B}\right]\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\frac{Ld\,(d+\tau)\,\sigma_1^2}{B\mu^3\varepsilon^2}\right)\quad\text{one-point oracle calls}.$$

证明。 应用引理 9 ， 取 $u=x^*$ （对于任意的 x^*）以及 $u=x_f^k$ ， 对 f_t， 可得：

$$\begin{aligned}\mathbb{E}_k\Big[f_t(x_f^{k+1})\Big] \leq & f_t(x^*)-\langle\nabla f_t(x_g^k),x^*-x_g^k\rangle-\frac{\mu}{2}\left\|x^*-x_g^k\right\|^2-\frac{p\gamma}{2}\|\nabla f_t(x_g^k)\|^2 \\ & +\frac{p\gamma}{2}\left\|\mathbb{E}_k\left[\hat{g}^k\right]-\nabla f_t(x_g^k)\right\|^2+\frac{Lp^2\gamma^2}{2}\mathbb{E}_k\left[\|\hat{g}^k\|^2\right],\end{aligned} \tag{50}$$

$$\begin{aligned}\mathbb{E}_k\Big[f_t(x_f^{k+1})\Big] \leq & f_t(x_f^k)-\langle\nabla f_t(x_g^k),x_f^k-x_g^k\rangle-\frac{\mu}{2}\left\|x_f^k-x_g^k\right\|^2-\frac{p\gamma}{2}\|\nabla f_t(x_g^k)\|^2 \\ & +\frac{p\gamma}{2}\left\|\mathbb{E}_k\left[\hat{g}^k\right]-\nabla f_t(x_g^k)\right\|^2+\frac{Lp^2\gamma^2}{2}\mathbb{E}_k\left[\|\hat{g}^k\|^2\right].\end{aligned} \tag{51}$$

结合 $2p\gamma\eta\cdot(50)+2\gamma\eta(\eta-p)\cdot(51)+(48)$ 可得：

$$\begin{aligned}&\mathbb{E}_k[\left\|x^{k+1}-x^*\right\|^2+2\gamma\eta^2 f_t(x_f^{k+1})] \\ \leq&(1+\alpha p\gamma\eta)(1-\beta)\left\|x^k-x^*\right\|^2+(1+\alpha p\gamma\eta)\beta\|x_g^k-x^*\|^2 \\ &+(1+\alpha p\gamma\eta)(\beta^2-\beta)\left\|x^k-x_g^k\right\|^2-2\eta^2\gamma\langle\nabla f_t(x_g^k),x_g^k+(p\eta^{-1}-1)x_f^k-\eta^{-1}px^*\rangle \\ &+p^2\eta^2\gamma^2\mathbb{E}_k[\left\|\hat{g}^k\right\|^2]+\frac{p\eta\gamma}{\alpha}\|\mathbb{E}_k\left[\hat{g}^k\right]-\nabla f_t(x_g^k)\|^2 \\ &+2p\gamma\eta\Big(f_t(x^*)-\langle\nabla f_t(x_g^k),x^*-x_g^k\rangle-\frac{\mu}{2}\left\|x^*-x_g^k\right\|^2-\frac{p\gamma}{2}\|\nabla f_t(x_g^k)\|^2 \\ &+\frac{p\gamma}{2}\left\|\mathbb{E}_k\left[\hat{g}^k\right]-\nabla f_t(x_g^k)\right\|^2+\frac{Lp^2\gamma^2}{2}\mathbb{E}_k\left[\|\hat{g}^k\|^2\right]\Big) \\ &+2\gamma\eta(\eta-p)\Big(f_t(x_f^k)-\langle\nabla f_t(x_g^k),x_f^k-x_g^k\rangle-\frac{\mu}{2}\left\|x_f^k-x_g^k\right\|^2-\frac{p\gamma}{2}\|\nabla f_t(x_g^k)\|^2 \\ &+\frac{p\gamma}{2}\left\|\mathbb{E}_k\left[\hat{g}^k\right]-\nabla f_t(x_g^k)\right\|^2+\frac{Lp^2\gamma^2}{2}\mathbb{E}_k\left[\|\hat{g}^k\|^2\right]\Big) \\ =&(1+\alpha p\gamma\eta)(1-\beta)\left\|x^k-x^*\right\|^2+2\gamma\eta\left(\eta-p\right)f_t(x_f^k)+2p\gamma\eta f_t(x^*) \\ &+\left((1+\alpha p\gamma\eta)\beta-p\gamma\eta\mu\right)\left\|x_g^k-x^*\right\|^2 \\ &+(1+\alpha p\gamma\eta)(\beta^2-\beta)\left\|x^k-x_g^k\right\|^2-p\gamma^2\eta^2\left\|\nabla f_t(x_g^k)\right\|^2 \\ &+\left(\frac{p\eta\gamma}{\alpha}+p\gamma^2\eta^2\right)\left\|\mathbb{E}_k\left[\hat{g}^k\right]-\nabla f_t(x_g^k)\right\|^2+\left(p^2\eta^2\gamma^2+p^2\gamma^3\eta^2L\right)\mathbb{E}_k\left[\|\hat{g}^k\|^2\right] \\ \overset{(76)}{\leq}&(1+\alpha p\gamma\eta)(1-\beta)\left\|x^k-x^*\right\|^2+2\gamma\eta\left(\eta-p\right)f_t(x_f^k)+2p\gamma\eta f_t(x^*) \\ &+\left((1+\alpha p\gamma\eta)\beta-p\gamma\eta\mu\right)\left\|x_g^k-x^*\right\|^2 \\ &+(1+\alpha p\gamma\eta)(\beta^2-\beta)\left\|x^k-x_g^k\right\|^2-p\gamma^2\eta^2\left\|\nabla f_t(x_g^k)\right\|^2 \\ &+p\eta\gamma\left(\frac{1}{\alpha}+\gamma\eta\right)\left\|\mathbb{E}_k\left[\hat{g}^k\right]-\nabla f_t(x_g^k)\right\|^2+2p^2\eta^2\gamma^2\left(1+\gamma L\right)\mathbb{E}_k\left[\|\hat{g}^k-\nabla f_t(x_g^k)\|^2\right] \\ &+2p^2\eta^2\gamma^2\left(1+\gamma L\right)\mathbb{E}_k\left[\underbrace{\left\|\nabla f_t(x_g^k)\right\|^2}_{x_g^k\in\mathcal{F}_k}\right].\end{aligned}$$

选择 $\alpha=\frac{\beta}{2p\eta\gamma}$ 可得：

$$\beta=\sqrt{4p^2\mu\gamma/3}\overset{\gamma\leq\frac{3}{4L}}{\leq}\sqrt{p^2\mu/L}<1,$$

$$
\begin{aligned}
&(1+\alpha p\eta\gamma)(1-\beta) = \left(1+\frac{\beta}{2}\right)(1-\beta) \leq \left(1-\frac{\beta}{2}\right), \\
&((1+\alpha p\eta\gamma)\beta - p\mu\gamma\eta) = \left(\beta + \frac{\beta^2}{2} - p\mu\gamma\eta\right) \overset{\beta<1}{<} \left(\frac{3\beta}{2} - p\mu\gamma\eta\right) \overset{p\mu\gamma\eta=3\beta/2}{\leq} 0.
\end{aligned}
$$

因此：

$$
\begin{aligned}
&\mathbb{E}_k[\left\|x^{k+1}-x^*\right\|^2 + 2\gamma\eta^2 f_t(x_f^{k+1})] \\
&\qquad \leq (1-\beta/2)\left\|x^k-x^*\right\|^2 + 2\gamma\eta\left(\eta-p\right) f_t(x_f^k) + 2p\gamma\eta f_t(x^*) \\
&\qquad\quad + p\eta^2\gamma^2\left(1+2p/\beta\right)\left\|\mathbb{E}_k\left[\hat{g}^k\right] - \nabla f_t(x_g^k)\right\|^2 \\
&\qquad\quad + 2p^2\eta^2\gamma^2\left(1+\gamma L\right)\mathbb{E}_k\left[\left\|\hat{g}^k - \nabla f_t(x_g^k)\right\|^2\right] \\
&\qquad\quad - p\gamma^2\eta^2(1-2p(1+\gamma L))\left\|\nabla f_t(x_g^k)\right\|^2.
\end{aligned}
$$

从等式两边减去 $2\gamma\eta^2 f_t(x^*)$，得到：

$$
\begin{aligned}
&\mathbb{E}_k[\left\|x^{k+1}-x^*\right\|^2 + 2\gamma\eta^2(f_t(x_f^{k+1}) - f_t(x^*))] \\
&\qquad \leq (1-\beta/2)\left\|x^k-x^*\right\|^2 + (1-p/\eta)\cdot 2\gamma\eta^2(f_t(x_f^k) - f_t(x^*)) \\
&\qquad\quad + p\eta^2\gamma^2\left(1+2p/\beta\right)\left\|\mathbb{E}_k\left[\hat{g}^k\right] - \nabla f_t(x_g^k)\right\|^2 \\
&\qquad\quad + 2p^2\eta^2\gamma^2\left(1+\gamma L\right)\mathbb{E}_k\left[\left\|\hat{g}^k - \nabla f_t(x_g^k)\right\|^2\right] \\
&\qquad\quad - p\gamma^2\eta^2(1-2p(1+\gamma L))\left\|\nabla f_t(x_g^k)\right\|^2 \\
&\qquad \overset{\beta/2=p/\eta}{=} (1-\beta/2)\left[\left\|x^k-x^*\right\|^2 + 2\gamma\eta^2(f_t(x_f^k) - f_t(x^*))\right] \\
&\qquad\quad + p\eta^2\gamma^2\left(1+2p/\beta\right)\left\|\mathbb{E}_k\left[\hat{g}^k\right] - \nabla f_t(x_g^k)\right\|^2 \\
&\qquad\quad + 2p^2\eta^2\gamma^2\left(1+\gamma L\right)\mathbb{E}_k\left[\left\|\hat{g}^k - \nabla f_t(x_g^k)\right\|^2\right] \\
&\qquad\quad - p\gamma^2\eta^2(1-2p(1+\gamma L))\left\|\nabla f_t(x_g^k)\right\|^2.
\end{aligned}
$$

应用引理 7，可得：

$$
\begin{aligned}
&\mathbb{E}_k[\left\|x^{k+1}-x^*\right\|^2 + 2\gamma\eta^2(f_t(x_f^{k+1}) - f_t(x^*))] \\
&\qquad \lesssim (1-\beta/2)\left[\left\|x^k-x^*\right\|^2 + 2\gamma\eta^2(f_t(x_f^k) - f_t(x^*))\right] \\
&\qquad\quad + p\eta^2\gamma^2\left(1+2p/\beta\right)\cdot\left[\frac{d^2\Delta^2}{t^2} + \frac{d\tau\sigma_1^2}{t^2 MB}\right] \\
&\qquad\quad + 2p^2\eta^2\gamma^2\left(1+\gamma L\right)\cdot\left[\frac{d^2\Delta^2}{t^2} + \frac{d\left(d+\tau\right)\sigma_1^2}{t^2 B} + \frac{d^2L^2t^2}{B} + \frac{d}{B}\left\|\nabla f(x_g^k)\right\|^2\right] \\
&\qquad\quad - p\gamma^2\eta^2(1-2p(1+\gamma L))\left\|\nabla f_t(x_g^k)\right\|^2 \\
&\qquad = \Big/ \frac{1}{M} = p(1+2p/\beta)^{-1} \Big/ \\
&\qquad = (1-\beta/2)\left[\left\|x^k-x^*\right\|^2 + 2\gamma\eta^2(f_t(x_f^k) - f_t(x^*))\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ p^2\eta^2\gamma^2 \cdot \left[\frac{d^2\Delta^2 M}{t^2} + \frac{d\tau\sigma_1^2}{t^2 B}\right] \\
&+ 2p^2\eta^2\gamma^2\left(1+\gamma L\right) \cdot \left[\frac{d^2\Delta^2}{t^2} + \frac{d\left(d+\tau\right)\sigma_1^2}{t^2 B} + \frac{d^2L^2t^2}{B} + \frac{d}{B}\left\|\nabla f(x_g^k)\right\|^2\right] \\
&- p\gamma^2\eta^2(1-2p(1+\gamma L))\left\|\nabla f_t(x_g^k)\right\|^2 \\
\overset{(30)}{\lesssim}\ & (1-\beta/2)\left[\left\|x^k - x^*\right\|^2 + 2\gamma\eta^2(f_t(x_f^k) - f_t(x^*))\right] \\
&+ \Delta^2 \cdot \left[\frac{p^2\eta^2\gamma^2 d^2 M + p^2\eta^2\gamma^2(1+\gamma L)d^2}{t^2}\right] \\
&+ \left\|\nabla f_t(x_g^k)\right\|^2 \cdot \left[p^2\eta^2\gamma^2(1+\gamma L)\frac{d}{B} - p\gamma^2\eta^2(1-2p(1+\gamma L))\right] \\
&+ \sigma_1^2 \cdot \left[\frac{p^2\eta^2\gamma^2 d\tau + p^2\eta^2\gamma^2(1+\gamma L)d\left(d+\tau\right)}{t^2 B}\right] \\
&+ \frac{t^2}{B} \cdot p^2\eta^2\gamma^2(1+\gamma L)L^2(d^2+d) \\
\overset{\gamma L<1}{\lesssim}\ & (1-\beta/2)\left[\left\|x^k - x^*\right\|^2 + 2\gamma\eta^2(f_t(x_f^k) - f_t(x^*))\right] \\
&+ p^2\eta^2\gamma^2 \cdot \left[\sigma_1^2\frac{d\left(d+\tau\right)}{t^2 B} + t^2\frac{L^2d^2}{B} + \Delta^2\frac{d^2M}{t^2}\right] \\
&+ \left\|\nabla f(x_g^k)\right\|^2 \cdot p\gamma^2\eta^2 \underbrace{\left[-1 + p\left(1+\gamma L\right)\left(1+\frac{d}{B}\right)\right]}_{=0 \text{ for } p\simeq\frac{B}{B+d}} \\
\overset{p\eta\gamma=3\beta/(2\mu)}{\lesssim}\ & (1-\beta/2)\left[\left\|x^k - x^*\right\|^2 + 2\gamma\eta^2(f_t(x_f^k) - f_t(x^*))\right] \\
&+ \frac{\beta^2}{\mu^2} \cdot \left[\sigma_1^2\frac{d\left(d+\tau\right)}{t^2 B} + t^2\frac{L^2d^2}{B} + \Delta^2\frac{d^2M}{t^2}\right].
\end{aligned}
$$

最后，我们执行递归，并代入 $\beta = \sqrt{4p^2\mu\gamma/3}$, $\eta = \sqrt{\frac{3}{\mu\gamma}}$, $r_t^N = \left\|x^N - x^*\right\|^2 + \frac{1}{\mu}(f_t(x_f^N) - f_t(x^*))$:

$$
\begin{aligned}
\mathbb{E}r_t^N &\lesssim \left(1 - \sqrt{\frac{p^2\mu\gamma}{3}}\right)^N r_t^0 \\
&\quad + \frac{\beta}{\mu^2} \cdot \left[\sigma_1^2\frac{d\left(d+\tau\right)}{t^2 B} + t^2\frac{L^2d^2}{B} + \Delta^2\frac{d^2M}{t^2}\right] \\
&\lesssim \exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right) r_t^0 \\
&\quad + \frac{p\sqrt{\gamma}}{\mu^{3/2}} \cdot \left[\sigma_1^2\frac{d\left(d+\tau\right)}{t^2 B} + t^2\frac{L^2d^2}{B} + \Delta^2\frac{d^2M}{t^2}\right] \\
&\overset{①}{\lesssim} \exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right) r_t^0 \\
&\quad + \frac{p\sqrt{\gamma}}{\mu^{3/2}} \cdot \left[\sigma_1^2\frac{d\left(d+\tau\right)}{t^2 B} + t^2\frac{L^2d^2}{B}\right]
\end{aligned}
$$

$$+\frac{\Delta^2 d^2}{\mu^2 t^2},$$

其中，1 利用了 $M \simeq \frac{1}{p}\left(1+\frac{1}{\sqrt{\mu\gamma}}\right) \Rightarrow Mp\sqrt{\gamma} \eqsim \sqrt{\gamma} \pm \frac{1}{\sqrt{\mu}} \leq \frac{1}{\sqrt{L}} + \frac{1}{\sqrt{\mu}} \lesssim \frac{1}{\sqrt{\mu}}$。回想一下， x^* 是任意的。因此，令 $x^* = \arg\min f(x)$，我们可以界定非光滑的 f:

$$\begin{aligned} r^N &= \left\|x^N - x^*\right\|^2 + \frac{6}{\mu}(f(x_f^N) - f(x^*)) \\ &= \|x^N - x^*\|^2 + \frac{6}{\mu}(\underbrace{f(x_f^N) - f_t(x_f^N)}_{\leq 0 \text{ (28)}} \underbrace{-f(x^*) + f_t(x^*))}_{\leq Lt^2 \text{ (28)}} + \frac{6}{\mu}(f_t(x_f^N) - f_t(x^*)) \\ &\leq r_t^N + 6\frac{Lt^2}{\mu} \end{aligned}$$

由此可得

$$\begin{aligned} \mathbb{E}r^N \lesssim & \exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right) r^0 \\ & + \frac{p\sqrt{\gamma}}{\mu^{3/2}} \cdot \left[\sigma_1^2 \frac{d\,(d+\tau)}{t^2 B} + t^2 \frac{L^2 d^2}{B}\right] \\ & + \frac{\Delta^2 d^2}{\mu^2 t^2} + \frac{Lt^2}{\mu} \end{aligned}$$

为了完成分析，我们需要定义 t 和 γ，以及噪声的可容忍水平 Δ。目前我们得到的表达式如下:

$$\mathbb{E}r^N \lesssim \exp(-N\Gamma a)r^0 + \Gamma b + c, \Gamma \leq \frac{1}{u}$$

with

$$\begin{aligned} \Gamma &= \sqrt{\gamma} \\ u &\simeq \sqrt{L} \\ a &\simeq p\sqrt{\mu} \\ b &\simeq \frac{p}{\mu^{3/2}} \cdot \left[\sigma_1^2 \frac{d\,(d+\tau)}{t^2 B} + t^2 \frac{L^2 d^2}{B}\right] \\ c &= \frac{\Delta^2 d^2}{\mu^2 t^2} + \frac{Lt^2}{\mu} \end{aligned}$$

为了得到 $c \lesssim \varepsilon$，我们必须界定 t:

$$\frac{d\Delta}{\mu\sqrt{\varepsilon}} \lesssim t \lesssim \frac{\sqrt{\mu\varepsilon}}{\sqrt{L}}$$

因此，我们对对抗噪声 $\varepsilon \gtrsim \frac{d\Delta\sqrt{L}}{\mu^{3/2}} \Leftrightarrow \Delta \lesssim \frac{\varepsilon\mu^{3/2}}{d\sqrt{L}}$ 进行界定。

应用引理 10，要得到 $r^N \lesssim \varepsilon$ 需要 N 迭代次数:

$$N = \tilde{\mathcal{O}}\left(\frac{1}{p}\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon} + \frac{d}{B\mu^2\varepsilon}\left[\frac{(d+\tau)\,\sigma_1^2}{t^2} + L^2 t^2 d\right]\right) \tag{52}$$

回顾 $p \simeq \frac{B}{B+d}$，并将 t 设置为其上界后，我们得到迭代总次数:

$$N = \tilde{\mathcal{O}}\left(\left[1+\frac{d}{B}\right]\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon} + \frac{Ld\,(d+\tau)\,\sigma_1^2}{\mu^3\varepsilon^2 B}\right)$$

˜最后，正如第 2.1 节所述，每个 ˆg_{ml} 需要 $\mathcal{O}(B)$ 次预言机调用，因此预言机复杂度为：

$$B \cdot \tilde{\mathcal{O}}\left(\max\left[1, \frac{d}{B}\right]\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon} + \frac{Ld\left(d+\tau\right)\sigma_1^2}{B\mu^3\varepsilon^2}\right) \quad \text{one-point oracle calls}.$$

□

D.5 定理的证明 3

定理5（定理 *3*）

假设 2 至 4 和 6 成立，并考虑由算法（4）求解的问题1。那么，对于合适的隐藏参数选择（其中 $p \simeq 1$），以及任意的自由参数选择（见表 3），有：

$$\mathbb{E}r^N \lesssim \exp\left(-\sqrt{\frac{\mu\gamma N^2}{3}}\right) r^0 + \frac{\sqrt{\gamma}}{\mu^{3/2}} \cdot \left[\sigma_1^2\frac{dC_1\left(d+\tau\right)}{t^2B} + \frac{G^2d}{B}\right] + \frac{\Delta^2d^2}{\mu^2t^2} + \frac{Gt}{\mu}$$

此外，对于任意的 $\varepsilon \gtrsim \left[\frac{d\Delta G}{\mu^2}\right]^{2/3}$以及适当选择的 t 和 γ，确保 $r^N \lesssim \varepsilon$ 所需的预言机调用次数有如下界：

$$B \cdot \tilde{\mathcal{O}}\left[\sqrt{\frac{\sqrt{d}G^2}{\mu^2\varepsilon}}\log\frac{1}{\varepsilon} + \frac{d\left(d+\tau\right)\sigma_1^2G^2}{\mu^4\varepsilon^3B} + \frac{G^2d}{B\mu^2\varepsilon}\right] \quad \text{one-point oracle calls}.$$

证明。 该证明与光滑情形几乎完全相同。区别在于我们使用（46）而非（45）。因此$p \simeq 1$就足够了，因为含有$\left\|\nabla f(x_g^k)\right\|$的项不再存在。此外，$\frac{d^2L^2t^2}{B} \to \frac{dG^2}{B}$。最后，我们可以使用引理9，因为平滑函数确实是光滑的（27）。

$$\begin{aligned}
\mathbb{E}r^N \lesssim& \exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right) r_t^0 \\
&+ \frac{p\sqrt{\gamma}}{\mu^{3/2}} \cdot \left[\sigma_1^2\frac{dC_1\left(d+\tau\right)}{t^2B} + \frac{G^2d}{B}\right] \\
&+ \frac{\Delta^2d^2}{\mu^2t^2} + \underbrace{\frac{Gt}{\mu}}_{(26)} \\
\overset{p\simeq 1}{\simeq}& \exp\left(-\sqrt{\frac{\mu\gamma N^2}{3}}\right) r_t^0 \\
&+ \frac{\sqrt{\gamma}}{\mu^{3/2}} \cdot \left[\sigma_1^2\frac{dC_1\left(d+\tau\right)}{t^2B} + \frac{G^2d}{B}\right] \\
&+ \frac{\Delta^2d^2}{\mu^2t^2} + \frac{Gt}{\mu}
\end{aligned}$$

应用引理 10 时：

$$\Gamma = \sqrt{\gamma}$$

$$
\begin{aligned}
u &\simeq \sqrt{L} \overset{(27)}{\simeq} \sqrt{\frac{\sqrt{d}G}{t}} \\
a &= \sqrt{\mu} \\
b &= \frac{1}{\mu^{3/2}} \cdot \left[\sigma_1^2 \frac{dC_1\left(d+\tau\right)}{t^2 B} + \frac{G^2 d}{B}\right]
\end{aligned}
$$

我们得出， $r^N \lesssim \varepsilon$ 需要进行 N 次迭代:

$$
N = \tilde{\mathcal{O}}\left(\sqrt{\frac{\sqrt{d}G}{t\mu}} \log\frac{1}{\varepsilon} + \frac{d}{B\mu^2\varepsilon}\left[\frac{\left(d+\tau\right)\sigma_1^2}{t^2} + G^2\right]\right).
$$

要得到 $c \lesssim \varepsilon$ ，我们需要对 t 进行界定:

$$
\frac{d\Delta}{\mu\sqrt{\varepsilon}} \lesssim t \lesssim \frac{\mu\varepsilon}{G}
$$

因此，我们对对抗噪声 $\varepsilon \gtrsim \left[\frac{d\Delta G}{\mu^2}\right]^{2/3} \Leftrightarrow \Delta \underset{\sim}{\lesssim} \frac{\varepsilon^{3/2}\mu^2}{dG}$ 进行界定。

代入 $L = \frac{\sqrt{d}G}{t}$，并将 t 设为其上界，我们得到迭代总次数:

$$
N = \tilde{\mathcal{O}}\left(\sqrt{\frac{\sqrt{d}G^2}{\mu^2\varepsilon}} \log\frac{1}{\varepsilon} + \frac{d\left(d+\tau\right)\sigma_1^2 G^2}{\mu^4\varepsilon^3 B} + \frac{G^2 d}{B\mu^2\varepsilon}\right).
$$

以及预言机复杂度:

$$
B \cdot \tilde{\mathcal{O}}\left(\sqrt{\frac{\sqrt{d}G^2}{\mu^2\varepsilon}} \log\frac{1}{\varepsilon} + \frac{d\left(d+\tau\right)\sigma_1^2 G^2}{\mu^4\varepsilon^3 B} + \frac{G^2 d}{B\mu^2\varepsilon}\right) \quad \text{one-point oracle calls}.
$$

□

E 两点结果的证明

单点和双点反馈的证明在实质上仅在引理 5和引理 7中有所不同，其余部分则会沿用。

E.1 梯度近似的不等式

引理 5′。 假设满足假设 1、假设 3 和假设4。那么，对于任意定义在 ξ 上的初始分布 (Z,$\mathcal{Z}$) 以及所有 $x \in \mathbb{R}^d$，以下不等式成立:

$$
\begin{aligned}
&\mathbb{E}[\|\tilde{g}_i - \nabla_{e_i} F_i\|^2] \le \tfrac{L^2 d^2 t^2}{4}, && (53)\\
&\mathbb{E}\|\nabla_{e_i} F_i - \nabla_{e_i} f\|^2 \le d\sigma_2^2, && (54)\\
&\mathbb{E}\left\|\tilde{g}^j - \mathbb{E}_e \tilde{g}^j\right\|^2 \le \tfrac{3}{2^j l}\left[d^2 t^2 L^2/4 + d\sigma_2^2 + d\|\nabla f\|^2\right], && (55)\\
&\mathbb{E}_Z\left\|\mathbb{E}_e \tilde{g}^j - \nabla f_t\right\|^2 \lesssim \tfrac{\tau}{2^j l}\sigma_2^2, && (56)\\
&\left\|\mathbb{E}\hat{g}^j - \nabla f_t\right\|^2 \lesssim \tfrac{d^2\Delta^2}{t^2} + \tfrac{\tau}{2^j l}\sigma_2^2, && (57)\\
&\mathbb{E}\left\|\tilde{g}^j - \nabla f_t\right\|^2 \lesssim \tfrac{d^2 t^2 L^2}{2^j l} + \tfrac{d+\tau}{2^j l}\sigma_2^2 + \tfrac{d}{2^j l}\|\nabla f\|^2. && (58)\\
&\mathbb{E}\left\|\hat{g}^j - \nabla f_t\right\|^2 \lesssim \tfrac{d^2\Delta^2}{t^2} + \tfrac{d^2 t^2 L^2}{2^j l} + \tfrac{d+\tau}{2^j l}\sigma_2^2 + \tfrac{d}{2^j l}\|\nabla f\|^2. && (59)
\end{aligned}
$$

证明。
我们从（53）开始，逐一证明所有估计：

$$\begin{aligned}
&\mathbb{E}[\|\tilde{g}_i - \nabla_{e_i} F_i\|^2] \\
\overset{(12),(18)}{=} \quad & d^2 \mathbb{E}\left[\left\| \frac{F(x+te_i, Z_i) - F(x-te_i, Z_i)}{2t} e_i - \langle \nabla F(x, Z_i), e_i \rangle e_i \right\|^2\right] \\
\overset{(1')+(74)}{\leq} \quad & \frac{L^2 d^2 t^2}{4}.
\end{aligned}$$

式（54）的证明：

$$\mathbb{E}\|\nabla_{e_i} F_i - \nabla_{e_i} f\|^2 = \mathbb{E}_Z \mathbb{E}_{e_i} \|\nabla_{e_i} F_i - \nabla_{e_i} f\|^2 \overset{(17),(81)}{=} d\mathbb{E}_Z \|\nabla F_i - \nabla f\|^2 \overset{(4')}{\leq} d\sigma_2^2.$$

式（55）的证明：

$$\begin{aligned}
\mathbb{E}\left\|\tilde{g}^j - \mathbb{E}_e \tilde{g}^j\right\|^2 \quad & \overset{\text{like (35)}}{\leq} \quad \frac{1}{2^{2j} l^2} \sum_{i=1}^{2^j l} \mathbb{E}\|\tilde{g}_i\|^2 \\
& \overset{(77)}{\leq} \quad \frac{3}{2^{2j} l^2} \sum_{i=1}^{2^j l} \Big[\mathbb{E}\left[\|\tilde{g}_i - \nabla_{e_i} F_i\|^2\right] + \\
& \qquad\qquad \mathbb{E}[\|\nabla_{e_i} F_i - \nabla_{e_i} f\|^2] + \mathbb{E}[\|\nabla_{e_i} f\|^2] \Big] \\
& \overset{(53),(54),(81)}{=} \quad \frac{3}{2^j l} \left[d^2 t^2 L^2 / 4 + d\sigma_2^2 + d\|\nabla f\|^2 \right].
\end{aligned}$$

式（56）的证明：

$$\begin{aligned}
\mathbb{E}_Z \left\|\mathbb{E}_e \tilde{g}^j - \nabla f_t\right\|^2 \quad & \overset{(15)}{=} \quad \mathbb{E}_Z \left\| \mathbb{E}_e \left[\frac{1}{2^j l} \sum_{i=1}^{2^j l} \tilde{g}_i \right] - \nabla f_t \right\|^2 \\
& \overset{(24)}{\leq} \quad \mathbb{E}_Z \left\| \frac{1}{2^j l} \sum_{i=1}^{2^j l} \nabla F_t(x, Z_i) - \nabla f_t(x) \right\|^2 \\
& \overset{①}{\lesssim} \quad \frac{\tau}{2^j l} \sigma_2^2,
\end{aligned}$$

其中 1 对 22）应用（ ∇F_t， ∇f_t。下面验证假设 $4'$ 成立：无偏性：

$$\mathbb{E}_\pi \nabla F_t(x, \quad Z) = \mathbb{E}_\pi \nabla [\mathbb{E}_r F(x + tr, \quad Z)] =$$
$$\mathbb{E}_\pi \mathbb{E}_r \nabla F(x + tr, Z) = \mathbb{E}_r \mathbb{E}_\pi \nabla F(x + tr, Z) = \mathbb{E}_r \nabla f(x + tr) = \nabla f_t(x).$$

方差：

$$\|\nabla F_t(x, Z) - \nabla f_t(x)\|^2 = \|\mathbb{E}_r \nabla F(x + tr, Z) - \nabla f(x + tr)\|^2 \overset{(80)}{\leq}$$
$$\mathbb{E}_r \|\nabla F(x + tr, Z) - \nabla f(x + tr)\|^2 \overset{(4')}{\leq} \mathbb{E}_r \sigma_2^2 = \sigma_2^2.$$

证明（57）：

$$\left\|\mathbb{E}\hat{g}^j - \nabla f_t\right\|^2 \quad \overset{(77)}{\leq} \quad 2\left\|\mathbb{E}\hat{g}^j - \mathbb{E}\tilde{g}^j\right\|^2 + 2\|\mathbb{E}\tilde{g}^j - \nabla f_t\|^2$$

$$\begin{aligned} &\overset{(80)}{\leq} && 2\left\|\mathbb{E}\hat{g}^j - \mathbb{E}\tilde{g}^j\right\|^2 + 2\mathbb{E}_Z\|\mathbb{E}_e\tilde{g}^j - \nabla f_t\|^2 \\ &\overset{(31),(56)}{\lesssim} && \frac{d^2\Delta^2}{t^2} + \frac{\tau}{2^j l}\sigma_2^2. \end{aligned}$$

证明（58）：

$$\begin{aligned} \mathbb{E}\left\|\tilde{g}^j - \nabla f_t\right\|^2 &\overset{(77)}{\leq} && 2\mathbb{E}\left\|\tilde{g}^j - \mathbb{E}_e\tilde{g}^j\right\|^2 + 2\mathbb{E}\|\mathbb{E}_e\tilde{g}^j - \nabla f_t\|^2 \\ &\overset{(55),(56)}{\lesssim} && \frac{1}{2^j l}\left[d^2t^2L^2 + d\sigma_2^2 + d\|\nabla f\|^2\right] + \frac{\tau}{2^j l}\sigma_2^2 \\ &\lesssim && \frac{d^2t^2L^2}{2^j l} + \frac{d+\tau}{2^j l}\sigma_2^2 + \frac{d}{2^j l}\|\nabla f\|^2. \end{aligned}$$

（59）的证明：

$$\begin{aligned} \mathbb{E}\left\|\hat{g}^j - \nabla f_t\right\|^2 &\overset{(77)}{\leq} && 2\mathbb{E}\left\|\hat{g}^j - \tilde{g}^j\right\|^2 + 2\mathbb{E}\left\|\tilde{g}^j - \nabla f_t\right\|^2 \\ &\overset{(31),(58)}{\lesssim} && \frac{d^2\Delta^2}{t^2} + \frac{d^2t^2L^2}{2^j l} + \frac{d+\tau}{2^j l}\sigma_2^2 + \frac{d}{2^j l}\|\nabla f\|^2. \end{aligned}$$

□

引理 11. 假设条件 6′、假设条件 2、假设条件 7成立。那么，对于任意初始分布 ξ 在 $(\mathsf{Z}, \mathcal{Z})$ 上，以及所有 $x \in \mathbb{R}^d$，以下不等式均成立：

$$\mathbb{E}\|\tilde{g}_i\|^2 \lesssim dG^2\,, \tag{60}$$
$$\mathbb{E}\left\|\tilde{g}^j - \mathbb{E}_e\tilde{g}^j\right\|^2 \leq \tfrac{dG^2}{2^j l}\,, \tag{61}$$
$$\mathbb{E}\left\|\mathbb{E}_e\tilde{g}^j - \nabla f_t\right\|^2 \leq \tfrac{4C_1\tau G^2}{2^j l}\,, \tag{62}$$
$$\left\|\mathbb{E}\hat{g}^j - \nabla f_t\right\|^2 \lesssim \tfrac{d^2\Delta^2}{t^2} + \tfrac{\tau G^2}{2^j l}\,, \tag{63}$$
$$\mathbb{E}\left\|\hat{g}^j - \nabla f_t\right\|^2 \lesssim \tfrac{d^2\Delta^2}{t^2} + \tfrac{(d+\tau)G^2}{2^j l}\,. \tag{64}$$

证明。

（60）的证明：

$$\mathbb{E}\|\tilde{g}_i\|^2 \overset{(11)}{=} \frac{d^2}{4t^2}\mathbb{E}\left|F(x+te_i, Z_i) - F(x-te_i, Z_i)\right|^2 \overset{\text{like }(40)}{\lesssim} dG^2.$$

（61）的证明：

$$\mathbb{E}\left\|\tilde{g}^j - \mathbb{E}_e\tilde{g}^j\right\|^2 \overset{\text{like }(35)}{\leq} \frac{1}{2^{2j}l^2}\sum_{i=1}^{2^j l}\mathbb{E}_Z\mathbb{E}_e\|\tilde{g}_i\|^2 \overset{(60)}{\leq} \frac{dG^2}{2^j l}.$$

（62）的证明：

$$\begin{aligned} \mathbb{E}\left\|\mathbb{E}_e\tilde{g}^j - \nabla f_t\right\|^2 &\overset{(15)}{=} \mathbb{E}\left\|\mathbb{E}_e\left[\frac{1}{2^j l}\sum_{i=1}^{2^j l}\tilde{g}_i\right] - \nabla f_t\right\|^2 \\ &\overset{(24)}{=} \mathbb{E}\left\|\mathbb{E}_e\left[\frac{1}{2^j l}\sum_{i=1}^{2^j l}\nabla F_t(x, Z_i)\right] - \nabla f_t\right\|^2 \end{aligned}$$

$$\overset{①}{\leq} \quad \frac{4C_1\tau G^2}{2^j l},$$

其中 1 使用带有 $\sigma_2^2 = 4G^2$ 的（ 4′ 22）。下面验证假设条件成立：无偏性：

$$\begin{aligned}\mathbb{E}_Z[\nabla F_t(x,Z)] &\overset{(24)}{=} \mathbb{E}_Z\mathbb{E}_e\left[d\frac{F(x+te,Z)-F(x-te,Z)}{2t}e\right]\\ &= \mathbb{E}_e\mathbb{E}_Z\left[d\frac{F(x+te,Z)-F(x-te,Z)}{2t}e\right]\\ &\overset{(7)}{=} \mathbb{E}_e\left[d\frac{f(x+te)-f(x-te)}{2t}e\right] \overset{(24)}{=} \nabla f_t(x).\end{aligned}$$

方差： $\|\nabla F_t(x,Z) - \nabla f_t(x)\| \overset{(77)}{\leq} 2\|\nabla F_t(x,\ Z)\|^2 + 2\|\nabla f_t(x)\|^2 \overset{②}{\leq} 4G^2$,
其中 2 利用了引理 4 平滑后的 f_t 和 F_t 均可微， G-利普希茨，因此其梯度范数有界于 G。

(63)的证明:

$$\begin{aligned}\left\|\mathbb{E}\hat{g}^j - \nabla f_t\right\|^2 &\overset{(77),(80)}{\leq} 2\left[\mathbb{E}\left\|\hat{g}^j - \tilde{g}^j\right\|^2 + \mathbb{E}_Z\|\mathbb{E}_e\tilde{g}^j - \nabla f_t\|^2\right]\\ &\overset{(31),(61)}{\lesssim} \frac{d^2\Delta^2}{t^2} + \frac{\tau G^2}{2^j l}.\end{aligned}$$

(64)的证明:

$$\begin{aligned}\mathbb{E}\left\|\hat{g}^j - \nabla f_t\right\|^2 &\overset{(77)}{\leq} 3\mathbb{E}\left[\left\|\hat{g}^j - \tilde{g}^j\right\|^2 + \|\tilde{g}^j - \mathbb{E}_e\tilde{g}^j\|^2 + \|\mathbb{E}_e\tilde{g}^j - \nabla f_t\|^2\right]\\ &\overset{(31),(61),(62)}{\lesssim} \frac{d^2\Delta^2}{t^2} + \frac{(d+\tau)G^2}{2^j l}.\end{aligned}$$

□

引理 7′。 设假设 3 和 4′ 成立。对于任意定义在 ξ 上的初始分布 (Z,$\mathcal{Z}$)， 以下不等式成立：在假设 1下：

$$\mathbb{E}[\|\nabla f_t(x) - \hat{g}_{ml}\|^2] \lesssim \frac{d^2\Delta^2}{t^2} + \frac{d^2t^2L^2}{B} + \frac{d+\tau}{B}\sigma_2^2 + \frac{d}{B}\|\nabla f\|^2. \tag{65}$$

$$\|\nabla f_t(x) - \mathbb{E}[\hat{g}_{ml}]\|^2 \lesssim \frac{d^2\Delta^2}{t^2} + \frac{\tau}{MB}\sigma_2^2. \tag{66}$$

在假设 6下：

$$\mathbb{E}[\|\nabla f_t(x) - \hat{g}_{ml}\|^2] \lesssim \frac{d^2\Delta^2}{t^2} + \frac{(d+\tau)G^2}{2^j l}. \tag{67}$$

$$\|\nabla f_t(x) - \mathbb{E}[\hat{g}_{ml}]\|^2 \lesssim \frac{d^2\Delta^2}{t^2} + \frac{\tau}{MB}G^2. \tag{68}$$

证明。 证明与引理7几乎完全相同，因此我们只列出计算过程。证明（66）：

$$\begin{aligned}\|\nabla f_t(x) - \mathbb{E}[\hat{g}_{ml}]\|^2 &\overset{(47)}{=} \left\|\nabla f_t(x) - \mathbb{E}\left[\hat{g}^{\lfloor\log_2 M\rfloor}\right]\right\|^2\\ &\overset{(57)}{\lesssim} \frac{d^2\Delta^2}{t^2} + \frac{\tau}{MB}\sigma_2^2.\end{aligned}$$

证明（68）：

$$\|\nabla f_t(x) - \mathbb{E}[\hat{g}_{ml}]\|^2 \overset{(47)}{=} \left\|\nabla f_t(x) - \mathbb{E}\left[\hat{g}^{\lfloor \log_2 M \rfloor}\right]\right\|^2 \overset{(63)}{\lesssim} \frac{d^2\Delta^2}{t^2} + \frac{\tau}{MB}G^2.$$

（65）的证明：

$$\begin{aligned}
& \mathbb{E}[\|\nabla f_t(x) - \hat{g}_{ml}\|^2] \\
\leq\ & 2\mathbb{E}[\left\|\nabla f_t(x) - \hat{g}^0\right\|^2] + 2\sum_{j=1}^{\lfloor \log_2 M \rfloor} 2^j \mathbb{E}[\|\tilde{g}^j - \tilde{g}^{j-1}\|^2] \\
\leq\ & 2\mathbb{E}[\left\|\nabla f_t(x) - \hat{g}^0\right\|^2] + 4\sum_{j=1}^{\lfloor \log_2 M \rfloor} 2^j \left(\mathbb{E}\|\tilde{g}^j - \nabla f_t(x)\|^2 + \mathbb{E}\|\nabla f_t(x) - \tilde{g}^{j-1}\|^2\right) \\
\leq\ & 2\mathbb{E}[\left\|\nabla f_t(x) - \hat{g}^0\right\|^2] + 16\sum_{j=0}^{\lfloor \log_2 M \rfloor} 2^j \mathbb{E}[\|\nabla f_t(x) - \hat{g}^j\|^2] \\
\overset{(59),(58)}{\lesssim}\ & \frac{d^2\Delta^2}{t^2} + \frac{d^2t^2L^2}{B} + \frac{d+\tau}{B}\sigma_2^2 + \frac{d}{B}\|\nabla f\|^2.
\end{aligned}$$

（67）的证明：

$$\mathbb{E}[\|\nabla f_t(x) - \hat{g}_{ml}\|^2] \overset{(64)}{\lesssim} \frac{d^2\Delta^2}{t^2} + \frac{(d+\tau)G^2}{2^j l}.$$

□

E.2 定理 1′ 的证明

定理 1′

设假设 1′ 至 4′ 成立，并考虑由算法（4）求解的问题，算法 1。那么，对于适当选择的隐藏参数（满足 $p \simeq \frac{B}{B+d}$），以及任意选择的自由参数（见表 3），有：

$$\mathbb{E}r^N \lesssim \exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right) r^0 + \frac{p\sqrt{\gamma}}{\mu^{3/2}} \cdot \left[\sigma_2^2\frac{d+\tau}{B} + t^2\frac{L^2d^2}{B}\right] + \frac{\Delta^2 d^2}{\mu^2 t^2} + \frac{Lt^2}{\mu}$$

此外，对于任意的 $\varepsilon \gtrsim \frac{d\Delta\sqrt{L}}{\mu^{3/2}}$，并适当选择 t 和 γ，确保 $r^N \lesssim \varepsilon$ 所需的预言机调用次数上界为

$$B \cdot \tilde{\mathcal{O}}\left(\max\left[1, \frac{d}{B}\right]\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon} + \frac{(d+\tau)\sigma_2^2}{B\mu^2\varepsilon}\right) \quad \text{two-point oracle calls}.$$

证明。 将引理7 代入定理 7′ 的证明中1，得到：

$$\begin{aligned}
\mathbb{E}r^N \lesssim & \exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right) r^0 \\
& + \frac{p\sqrt{\gamma}}{\mu^{3/2}} \cdot \left[\sigma_2^2\frac{d+\tau}{B} + t^2\frac{L^2d^2}{B}\right] \\
& + \frac{\Delta^2 d^2}{\mu^2 t^2} + \frac{Lt^2}{\mu}
\end{aligned}$$

应用引理10，其中：

$$\begin{aligned}\Gamma &= \sqrt{\gamma}\\ u &\simeq \sqrt{L}\\ a &\simeq p\sqrt{\mu}\\ b &\simeq \frac{p}{\mu^{3/2}}\cdot\left[\sigma_2^2\frac{d+\tau}{B}+t^2\frac{L^2d^2}{B}\right]\end{aligned}$$

我们得到 $r^N \lesssim \varepsilon$ 需要 N 次迭代：

$$N=\tilde{\mathcal{O}}\left(\frac{1}{p}\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\frac{1}{B\mu^2\varepsilon}\left[(d+\tau)\sigma_2^2+L^2t^2d^2\right]\right)$$

Δ、 t 和 p 的界沿用如下： $\varepsilon \gtrsim \frac{d\Delta\sqrt{L}}{\mu^{3/2}} \Leftrightarrow \Delta \lesssim \frac{\varepsilon\mu^{3/2}}{d\sqrt{L}}$、 $t \simeq \frac{\sqrt{\mu\varepsilon}}{\sqrt{L}}$、 $p \eqsim \frac{B}{B+d^\circ}$ 因此，迭代总次数为：

$$\tilde{\mathcal{O}}\left(\left[1+\frac{d}{B}\right]\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\frac{(d+\tau)\sigma_2^2}{B\mu^2\varepsilon}\right)$$

最后，预言机复杂度为：

$$B\cdot\tilde{\mathcal{O}}\left(\max\left[1,\frac{d}{B}\right]\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\frac{(d+\tau)\sigma_2^2}{B\mu^2\varepsilon}\right)\quad\text{two-point oracle calls}.$$

□

E.3 定理 3′的证明

定理 3′

令假设 1′ 至 4′ 成立，并考虑由算法（4）求解的问题1。那么，对于隐藏参数的适当选择（满足 $p\simeq 1$），以及自由参数的任意选择（见表3），有：

$$\mathbb{E}r^N \lesssim \exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right)r^0+\frac{p\sqrt{\gamma}}{\mu^{3/2}}\cdot G^2\frac{d+\tau}{B}+\frac{\Delta^2d^2}{\mu^2t^2}+\frac{Gt}{\mu}$$

此外，对于任意的 $\varepsilon \gtrsim \frac{d\Delta\sqrt{L}}{\mu^{3/2}}$ 以及适当选择的 t 和 γ，为确保 $r^N \lesssim \varepsilon$ 所需的预言机调用次数有以下界：

$$B\cdot\tilde{\mathcal{O}}\left(\sqrt{\frac{\sqrt{d}G^2}{\mu^2\varepsilon}}\log\frac{1}{\varepsilon}+\frac{(d+\tau)G^2}{B\mu^2\varepsilon}\right)\quad\text{two-point oracle calls}.$$

证明。 在光滑情形的证明中，将（65）和（66）替换为（67）和（68），可得：

$$\begin{aligned}\mathbb{E}r^N \lesssim &\exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right)r^0\\ &+\frac{p\sqrt{\gamma}}{\mu^{3/2}}\cdot G^2\frac{d+\tau}{B}\\ &+\frac{\Delta^2d^2}{\mu^2t^2}+\frac{Gt}{\mu}\end{aligned}$$

应用引理 10，其中：

$$\begin{aligned}\Gamma &= \sqrt{\gamma}\\ u &\simeq \sqrt{L} \overset{Lemma\ 4}{\simeq} \sqrt{\frac{\sqrt{d}G}{t}}\\ a &\simeq p\sqrt{\mu}\\ b &\simeq \frac{p(d+\tau)G^2}{\mu^{3/2}B}\end{aligned}$$

我们得到， $r^N \lesssim \varepsilon$ 需要经过 N 次迭代：

$$N = \tilde{\mathcal{O}}\left(\sqrt{\frac{\sqrt{d}G}{t\mu}}\log\frac{1}{\varepsilon} + \frac{(d+\tau)G^2}{B\mu^2\varepsilon}\right)$$

Δ、 t 和 p 的界继承自： $\varepsilon \gtrsim \left[\frac{d\Delta G}{\mu^2}\right]^{2/3} \Leftrightarrow \Delta \lesssim \frac{\varepsilon^{3/2}\mu^2}{dG}$、 $t \simeq \frac{\mu\varepsilon}{G}$、 $p \simeq 1$。因此，迭代总次数为：

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{\sqrt{d}G^2}{\mu^2\varepsilon}}\log\frac{1}{\varepsilon} + \frac{(d+\tau)G^2}{B\mu^2\varepsilon}\right)$$

最后，预言机复杂度为：

$$B\cdot\tilde{\mathcal{O}}\left(\sqrt{\frac{\sqrt{d}G^2}{\mu^2\varepsilon}}\log\frac{1}{\varepsilon} + \frac{(d+\tau)G^2}{B\mu^2\varepsilon}\right) \quad \text{two-point oracle calls}.$$

□

F 下界

F.1 主要定理

首先，我们介绍一些结果，以证实我们在二阶矩界条件下的分析是最优的。具体而言，我们检验

$$\mathbb{E}_\pi|F(x,Z) - f(x)|^2 < \sigma_1^2$$

而不是假设4 以及

$$\mathbb{E}_\pi\|\nabla F(x,Z) - \nabla f(x)\|^2 < \sigma_2^2$$

而不是假设 4′。
然后，我们展示如何在困难实例问题的构造中使用裁剪技术，以在对数因子范围内保持下界。

我们的主要结果是以下两个定理。它们表明，我们的方法和分析在单点和双点情形下都具有理论最优性。

定理 6（单点反馈）

对于任何解决该问题的（可能具有随机性的）算法（4），存在一个函数 f 满足假设 1 至4，使得

$$\mathbb{E}\|\hat{x}_N - x^*\|^2 \gtrsim \frac{\sqrt{d(d+\tau)\sigma_1^2}}{\mu\sqrt{N}} \quad \text{as } N \to \infty.$$

因此，使用单点反馈时，要达到解的 ε 邻域，算法至少需要

$$N = \Omega\left(\frac{d(d+\tau)\sigma_1^2}{\mu^2\varepsilon^2}\right) \quad \text{one-point oracle calls.}$$

定理 7（双点反馈）

对于任何能够解决问题（4）的（可能是随机化的）算法，都存在一个函数 f 满足假设 $1'$ 至 $4'$，使得

$$\mathbb{E}\|\hat{x}_N - x^*\|^2 \gtrsim \frac{(d+\tau)\sigma_2^2}{\mu^2 N} \quad \text{as } N \to \infty.$$

因此，使用双点反馈时，要达到解的 ε 邻域，至少需要

$$N = \Omega\left(\frac{(d+\tau)\sigma_2^2}{\mu^2\varepsilon}\right) \quad \text{two-point oracle calls.}$$

我们注意到，由于最优速率具有两部分结构，自然的做法是在其中一部分占主导的情形下分别证明这两部分。我们引入以下情形：

- $\tau \geq d$ — 高相关区域
- $\tau \leq d$ — 高维区域

接下来，我们总结声称在各个区域内成立的下界：显然，只有 1 个

表 4： 强凸情形，下界

	高相关		**高维**	
零阶单点	$\frac{d\tau\sigma_1^2}{\mu^2\varepsilon^2}$	（**新**，定理8）	$\frac{d^2\sigma_1^2}{\mu^2\varepsilon^2}$	Akhavan 等人 [2]（我们的定理9）
零阶双点	$\frac{\tau\sigma_2^2}{\mu^2\varepsilon}$	Beznosikov 等人 [5]（即使对于**一阶**）	$\frac{d\sigma_2^2}{\mu^2\varepsilon}$	Duchi 等人 [15]（我们的定理10）

4 个界同时取决于维度和混合时间。对于其他情形，我们可以使用已有的构造，这些构造分别处理混合和零阶信息，并将其调整以适用于我们的假设。结合全部四个界，我们在单点和双点设定下都得到了紧下界。下面讨论一些重要的相关结果。

Akhavan 等人 [2] 研究了单点反馈的一种特殊情形，其中噪声变量不依赖于查询点——这使得他们的下界适用于我们的情形。他们未考虑的唯一因素是 σ_1^2，然而，如果使用缩放高斯噪声，该因素也会从他们的证明中出现，此外还有 μ^2 因子；相关结果见我们的定理 9 。Beznosikov 等人 [5], 考虑了一阶预言机，但困难实例问题是一维二次问题，这使得一阶信息和零阶信息等价。Duchi 等人 [15]考虑了双点设定下一般凸情形，并给出了紧下界。然而，通过对难以区分的函数族中的每个线性函数添加公共二次部分，他们的证明可以转换为强凸问题。关于更形式化的归约，请参见定理 10。

最后，据我们所知，我们给出了一阶反馈和高相关区域中的一个新颖下界。

定理 8（单点、高相关）

在定理 6的条件下，以下界成立：

$$\mathbb{E}\|\hat{x}_N - x^*\|^2 \gtrsim \frac{\sqrt{d\tau\sigma_1^2}}{\mu\sqrt{N}}.$$

证明。 考虑函数族

$$f_\omega(x) = \frac{\mu}{2}\|x\|^2 + \langle S(x), \omega\rangle$$

其中 $\omega \in \{\pm 1\}^d$ 以及 $S:\ \mathbb{R}^d \to \mathbb{R}^d$ 稍后确定。对于相同的 ω 值，考虑零阶预言机

$$F_\omega(x, Z) = \frac{\mu}{2}\|x\|^2 + \langle S(x), Z + \omega\rangle = f_\omega(x) + \langle S(x), Z\rangle$$

且离散　时间马尔可夫过程，其转移概率由公式确定

$$Z_{t+1} = \begin{cases} \xi_{t+1}, & \text{w.p. } 1/\tau, \\ Z_t, & \text{w.p. } 1 - 1/\tau, \end{cases}$$

其中$\{\xi_t\}_{t=1}^\infty$ 相互独立，且

$$\xi_t \sim \pi = \mathcal{N}(0, s^2 I_d).$$

如此选取 Z_t 后，显然假设 3 得到满足，且

$$\mathbb{E}_\pi F_\omega(x, Z) = f_\omega(x).$$

现在，我们将证明，所有算法都无法在短时间内区分 f_ω 。首先，注意到

$$\|\hat{x} - x_\omega^*\|^2 \geq \frac{1}{4}\|x_{\omega'}^* - x_\omega^*\|^2 \tag{69}$$

其中 $\omega' = \arg\min_{\tilde{\omega}} \|x - x_{\tilde{\omega}}^*\|^2$。稍后我们将使用汉明距离 $\rho(\omega', \omega)$来界定 $\|x_{\omega'}^* - x_\omega^*\|$。但首先，我们先界定该距离本身。
应用阿苏阿德引理 [55] ，我们得到

$$\max_\omega \mathbb{E}_\omega \rho(\omega', \omega) \geq \frac{d}{2}\left(1 - \max_{\rho(\omega_1,\omega_2)=1} \|P_{\omega_1} - P_{\omega_2}\|_{TV}\right) \tag{70}$$

其中， P_ω 表示算法针对顺序查询所生成的 F_ω 输出的联合分布。$\hat{x}$ 是算法经过 N 步后的输出。现在我们来界定相邻分布之间的全变差。首先，我们使用 Pinsker 不等式：

$$2\|P_{\omega_1} - P_{\omega_2}\|_{TV}^2 \le D_{KL}\left(\mathrm{Law}(\{\omega_1 + Z_i\}_{i=1}^N),\ \mathrm{Law}(\{\omega_2 + Z_i\}_{i=1}^N)\right) = \Big/$$

然后，利用全概率公式，我们考虑一组固定索引所对应的条件 KL-散度，这些索引引入了新的样本。如果已知链的状态未发生变化，则一步 KL 等于 0。在其他步骤中，它等于均值为 ω_1 和 ω_2的高斯分布之间的 KL。我们按照状态切换次数 k对各项进行分组。

$$\Big/ = \sum_{k=0}^{N} D_{KL}\Big(\mathrm{Law}(\{\mathcal{N}(\omega_1,s^2I)\}_{i=1}^k), \mathrm{Law}(\{\mathcal{N}(\omega_2,s^2I)\}_{i=1}^k)\Big)\,\mathbb{P}(|\{1 \le t \le N:\ Z_t = \xi_t\}| = k)$$

利用 $\rho(\omega_1,\ \omega_2) = 1$，我们化简得到

$$= \sum_{k=0}^{N} D_{KL}\left(\mathcal{N}(1, s^2 I_k),\ \mathcal{N}(-1, s^2 I_k)\right)\mathbb{P}(|\{1 \le t \le N:\ Z_t = \xi_t\}| = k) =$$

$$\sum_{k=0}^{N} \frac{2k}{s^2}\mathbb{P}(|\{1 \le t \le N:\ Z_t = \xi_t\}| = k) = \frac{2}{s^2}\sum_{k=0}^{N} k\mathbb{P}(|\{1 \le t \le N:\ Z_t = \xi_t\}| = k) =$$

$$\frac{2}{s^2}\mathbb{E}(|\{1 \le t \le N:\ Z_t = \xi_t\}|) = \frac{2}{s^2}\sum_{t=1}^{N}\mathbb{E}(I_{Z_t=\xi_t}) = \frac{2N}{s^2\tau}.$$

取 $s^2 = \frac{8N}{\tau}$，我们得到

$$\|P_{\omega_1} - P_{\omega_2}\|_{TV} \le \sqrt{\frac{2N\tau}{8N\tau}} = \frac{1}{2}. \tag{71}$$

现在我们声称，存在这样一个 $S(x)$，满足假设1 至4以及

$$\|x^*_{\omega'} - x^*_{\omega}\|^2 \ge \frac{1}{2}\frac{\sqrt{\frac{\sigma_1^2}{9}\tau}}{\sqrt{4\mu^2 dN}}\rho(\omega',\omega) = \frac{1}{12}\frac{\sqrt{\sigma_1^2\tau}}{\sqrt{\mu^2 dN}}\rho(\omega',\omega). \tag{72}$$

结合（69）、（72）、（70）、（71），我们得出结论

$$\max_{\omega}\mathbb{E}_\omega\|\hat{x} - x^*_\omega\|^2 \ge \frac{1}{96}\frac{d\sqrt{\sigma_1^2\tau}}{\sqrt{\mu^2 dN}} = \frac{1}{96}\frac{\sqrt{d\tau\sigma_1^2}}{\mu\sqrt{N}}.$$

现在我们应引入 $S(x)$ 并检查（72）以及假设 1 至 4。
记作 $\delta = \sqrt[4]{\frac{\sigma_1^2\tau}{\mu^2 dN}}$。设 S(x) 是可分的，且

$$S(x)_i = \frac{\mu}{4}s(x_i) = \frac{\mu}{4}\cdot\begin{cases} 2\delta x_i, & 0 \le x_i \le \delta, \\ 3\delta^2 - (x_i - 2\delta)^2, & \delta \le x_i \le 2\delta, \\ 3\delta^2, & 2\delta \le x_i. \end{cases}$$

并且 $s(x_i)$ 关于零对称。可以直接验证， $s(x_i)$ 是 2-光滑的。为了检验 f_w 的强凸性和光滑性，我们注意到

$$\nabla f_\omega(x) = \mu x + \nabla\langle S(x), \omega\rangle = \mu x + \nabla S(x)\odot\omega,$$

其中 $\odot$ 是逐坐标乘积。第二项的 Lipschitz 常数有界

$$\|\nabla S(x)\odot\omega-\nabla S(y)\odot\omega\|=\|\nabla S(x)-\nabla S(y)\|\le\frac{\mu}{2}\|x-y\|.$$

这意味着函数的强凸常数 μ 和梯度 Lipschitz 常数 L f_ω 均处于范围$[\frac{\mu}{2};\ \frac{3\mu}{2}]$。因此，为了得到完全严格的界，我们在（2$\pi$72）中使用 μ，而不是。
通过驻点条件也可以直接验证 $x^*_\omega=-\frac{1}{2}\omega\delta$，并且（72）随之成立。这里我们还注意到 $\|x^*_\omega\|^2=\frac{1}{2}d\delta^2<1$当 N足够大时，因此,⁻ 当所需精度足够高时，函数的极小化点位于标准单位球内。

最后，我们需要检查有界噪声假设（4）。在当前设定下，我们可以保证相对于平稳分布的方差有界

$$\mathbb{E}_\pi h^2(x,Z)=\mathbb{E}_\pi\langle S(x),Z\rangle^2=s^2\|S(x)\|^2\le\frac{9s^2\mu^2d\delta^4}{16}=\frac{9N}{2\tau}\frac{\sigma_1^2\tau}{N}\le 9\sigma_1^2. \tag{73}$$

因此，为了得到完全严格的界，我们在（ $\sigma_1^2/9$）中使用72，而不是 σ_1^2。通过截断可以得到恰当的一致界，参见第F.2节。 □

F.2 关于截断的注

然而，我们还必须处理另一个问题——目前对于噪声，我们只有二阶矩界，正如其他使用独立同分布噪声而非马尔可夫型噪声的下界中一样。处理独立同分布噪声的一致有界性十分直接——由于噪声分布是高斯分布，我们可以利用尾界，将噪声截断在 $[-\sigma\log N;\ \sigma\log N]$对于所有查询点，其概率为 $1-o(1/N)$。对于定理9 和10，这给出了除对数因子外所需的界。
然而，在定理8的设定中，这一技巧将无法奏效，因为算法可以故意在某个点调用预言机，从而使下一步产生较大的噪声。为了解决这一问题，我们截断预言机，而不是截断噪声。对于某个 $t>1$ (t 将关于 N呈对数增长）我们引入

$$\hat F(x,Z)=\max\left(\min_\omega f_\omega(x)-t\sigma_1,\ \min(F(x,Z),\ \max_\omega f_\omega(x)+t\sigma_1)\right).$$

根据构造

$$|\hat F(x,Z)-\mathbb{E}_\pi\hat F(x,Z)|^2\le 2t^2\sigma_1^2+2|\max_\omega f_\omega(x)-\min_\omega f_\omega(x)|^2=$$
$$2t^2\sigma_1^2+2\|S(x)\|_1^2\le 2t\sigma_1^2+8d^2\mu^2\delta^4=2t^2\sigma_1^2+\frac{8d^2\sigma_1^2\tau}{N}.$$

请注意，当 N 足够大时，第二项会变得可以忽略不计。现在，截断会引入如下形式的偏差

$$|\mathbb{E}_\pi F(x,Z)-\mathbb{E}_\pi\hat F(x,Z)|\le|\mathbb{E}_\pi h(x,Z)I_{h(x,Z)>t\sigma}|\le$$
$$\le\int_{t\sigma}^{\infty}xe^{-\frac{x^2}{2\sigma_1^2}}dx=\sigma_1^2\int_t^\infty xe^{-\frac{x^2}{2}}dx=\sigma_1^2e^{-\frac{t^2}{2}}.$$

选择 $t\sim\log N$可使该偏差在 N 中小到超多项式级别，即 $\lesssim poly(\frac{1}{N})$，从而使其处于可接受的对抗性偏差水平 $\Delta\lesssim\frac{\varepsilon\mu^{3/2}}{dL}$ 以内。最后这一步会引入偏差，但可以通过仔细调整证明中使用的高斯分布来避免，使相互截断不会导致期望值发生变化。这是可行的，因为受截断影响的总概率质量呈指数级小，因此经过任何涉及该部分概率质量的变换后，全变差距离仍然很大。

F.3 单点高维情形

Akhavan 等人讨论了独立同分布的单点设定。 [1], 作者考虑了目标函数具有更一般的高阶光滑性的情况，并给出了任意加性噪声分布的下界。我们的视角有所不同——我们采用通常的光滑函数，考察 $N \to \infty$ 时的渐近行为，并且可以自由选择噪声结构。不过，我们也声称得到了更强的结果——我们的界还显示出额外的 $\mu^2\sigma_1^2$ 缩放关系，并且根据定理1，该界在渐近意义下是紧的。

定理 9（单点、高维）

在定理 6的条件下，以下界成立：

$$\mathbb{E}\|\hat{x}_N - x^*\|^2 \gtrsim \frac{\sqrt{d^2\sigma_1^2}}{\mu\sqrt{N}}.$$

证明。 仔细考察后可知，该证明重复、简化并扩展了 Akhavan 等人的构造。 [1],在我们的假设下。不过，为了便于表述，我们将采用定理 8中的自有记号。我们考虑相同的函数族 f_ω，但噪声为独立同分布、与点无关的高斯噪声，其方差为σ_1^2。这需要重新定义 δ 并修订（71）和（72）。对于这种噪声，我们使用与 Akhavan 等人类似的、关于相邻分布之间 KL散度的界。 [1, 定理 6。1]。我们还利用了高斯分布满足 $I_0 = \frac{1}{2\sigma_1^2}$ 这一性质。于是得到

$$D_{KL}(P_{\omega_1}, P_{\omega_2}) \le \frac{N}{2\sigma_1^2}\|f_{\omega_1} - f_{\omega_2}\|_\infty^2 < \frac{N\mu^2\delta^4}{2\sigma_1^2}.$$

重新定义 $\delta = \sqrt[4]{\frac{\sigma_1^2}{\mu^2 N}}$后，我们验证（71）成立。于是（72）可变换为

$$\|x^*_{\omega'} - x^*_\omega\|^2 \ge \frac{1}{2}\sqrt{\frac{\sigma_1^2}{4\mu^2 N}}\rho(\omega', \omega).$$

结合（69）、（72）、（70）、（71），我们得出

$$\max_\omega \mathbb{E}_\omega\|\hat{x} - x^*_\omega\|^2 \ge \frac{1}{16}\frac{d\sqrt{\sigma_1^2}}{\sqrt{\mu^2 N}}.$$

□

F.4 双点高维情形

定理 10 给出了从 Duchi 等人的下界推导到强凸目标函数的过程。结合上述截断技术，即可完成本节所有证明。 [15]

定理 10（双点、高维）

在定理 7的条件下，以下界成立：

$$\mathbb{E}\|\hat{x}_N - x^*\|^2 \gtrsim \frac{d\sigma_2^2}{\mu^2 N}.$$

证明。 考虑用于 $v \in \{\pm 1\}^d$ 的函数族

$$f_v(x) = \frac{\mu}{2}\|x\|^2 + \delta\langle x, v\rangle$$

以及相应的预言机

$$F_v(x, Z) = \frac{\mu}{2}\|x\|^2 + \langle x, \delta v + Z\rangle.$$

噪声序列 Z_i 没有任何马尔可夫性，反而我们选择令其服从独立同分布 $\sim \mathcal{N}(0,\ s^2 I_d)$。这个函数族很容易满足假设 $1'$ 至 $4'$，参数为 $\sigma_2^2 \geq \mathbb{E}\|Z\|^2 = ds^2$。同样，此处我们只考虑二阶矩界，如上所述。

这种构造与 Duchi 等人在证明 [15, 命题 1], 时使用的构造类似，但由于我们研究的是强凸问题，因此加入了一个确定性二次部分。因此，函数总是存在一个全局极小值点

$$x_v^* = \arg\min f_v(x) = -\frac{\delta}{\mu} v.$$

通常，我们可以用估计值与最优解的符号之间的汉明距离来界定其与最优解之间的距离

$$\max_v \mathbb{E}\|\hat{x}_N - x_v^*\|^2 \geq \frac{\delta^2}{\mu^2} \sum_{i=1}^{d} \mathbb{P}(\operatorname{sign}(\hat{x}_N^i) \neq -\operatorname{sign}(v^i)).$$

Duchi 等人 [15] 证明了这类概率之和的一个下界

$$\sum_{i=1}^{d} \mathbb{P}(\operatorname{sign}(\hat{x}_N^i) \neq -\operatorname{sign}(v^i)) \geq d\left(1 - \sqrt{\frac{2N\delta^2}{ds^2}}\right).$$

由于这些函数仅相差一个共同的确定性函数，这个不等式也适用于我们的函数集。因此，我们得到

$$\max_v \mathbb{E}\|\hat{x}_N - x_v^*\|^2 \geq \frac{d\delta^2}{\mu^2}\left(1 - \sqrt{\frac{2N\delta^2}{ds^2}}\right).$$

取 $s^2 = \frac{\sigma_2^2}{d}$ 和 $\delta^2 = \frac{\sigma_2^2}{4N}$ 即可得到所需结果

$$\mathbb{E}\|\hat{x}_N - x^*\|^2 \gtrsim \frac{d\sigma_2^2}{\mu^2 N}.$$

□

G 基本事实

引理 12. 若 f 在 $\mathbb{R}^d$中是 L-光滑的，则对于任意 $x, y \in \mathbb{R}^d$

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{L}{2}\|x - y\|^2. \tag{74}$$

引理 13（柯西-施瓦茨不等式 等式）。 对于任意 a、 b、 x_1、...、 $x_n \in \mathbb{R}^d$ 和 $c > 0$，以下不等式成立 :

$$2\langle a, b \rangle \leq \frac{\|a\|^2}{c} + c\|b\|^2, \tag{75}$$

$$\|a + b\|^2 \leq \left(1 + \frac{1}{c}\right)\|a\|^2 + (1 + c)\|b\|^2, \tag{76}$$

$$\left\|\sum_{i=1}^n x_i\right\|^2 \leq n \cdot \sum_{i=1}^n \|x_i\|^2. \tag{77}$$

引理 14。 对于具有有限二阶矩的随机变量 ξ :

$$\mathbb{E}\|\xi - \mathbb{E}\xi\|^2 \leq \mathbb{E}\|\xi\|^2 . \tag{78}$$

引理 15（詹森不等式）。 如果 f 是凸函数，那么对于任意 $n \in \mathbb{N}^*$ 和 x_1、...、 $x_n \in \mathbb{R}^d$，以下不等式成立:

$$f\left(\frac{1}{n}\sum_{i=1}^n x_i\right) \leq \frac{1}{n}\sum_{i=1}^n f(x_i) . \tag{79}$$

概率形式:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] .$$

应用于 $f(X) = \|X\|^2$:

$$\|\mathbb{E}[X]\|^2 \leq \mathbb{E}\left[\|X\|^2\right] . \tag{80}$$

Lemma 16 (随机投影的范数)。 对于 $e \sim RS_2^d(1)$ ，有如下等式成立 *s:*

$$\mathbb{E}_e\langle v, e \rangle^2 = \|v\|^2 \cdot 1/d . \tag{81}$$

证明。

$$\mathbb{E}\langle v, e \rangle^2 = \|v\|^2\mathbb{E}\langle v/\|v\|, e \rangle^2 = \|v\|^2\mathbb{E}\langle (1, 0, \ldots, 0), \tilde{e} \rangle^2 = \|v\|^2\mathbb{E}[\tilde{e}_1]^2 \overset{①}{=} \|v\|^2 \cdot 1/d,$$

其中 1 使用了 $\sum_i \mathbb{E}[\tilde{e}_i]^2 = 1$ 和 $E[\tilde{e}_1]^2 = \mathbb{E}[\tilde{e}_2]^2 = \ldots$

□