arXiv:2302.03542v3 [cs.LG] 7 Jun 2023

两个损失优于 一个：使用更廉价的代理加速优化

Blake Woodworth [1] **Konstantin Mishchenko** [2] **Francis Bach** [1]

摘要

我们提出一种算法，用于最小化梯度难以计算的目标函数，并使用一个相关且更易获取的函数作为代理。该算法基于对代理进行近似近端点迭代，并结合少量来自目标函数的随机梯度。当目标函数与代理之间的差异是 δ-光滑的，我们的算法可保证以与在 δ-光滑目标函数上进行随机梯度下降相匹配的速率收敛，从而显著提高样本效率。该算法在机器学习中有许多潜在应用，并提供了一种有理论依据的方法，可利用合成数据、物理模拟器、混合使用公共数据和私有数据等。

1. 引言

在许多机器学习问题中，目标函数难以获取，例如，难以计算其梯度以用于训练。造成这种困难的原因多种多样。例如，收集带标签的训练样本可能成本高昂或耗时很长，导致训练集规模较小。在某些应用中，计算损失函数的梯度可能需要等待真实的机器人在物理环境中执行策略，这一过程可能非常缓慢。此外，实际执行还可能损坏机器人或周围环境。最后，个人数据可能受到隐私或法律限制，因此无法用于计算梯度。在所有这些情况下，我们都面临着使训练复杂化的挑战，并可能导致使用随机梯度下降（SGD）等现成算法时性能不佳。

上述问题的一个自然解决方案是寻找第二个更易获取的“代理”函数，用它替代真正的目标函数。当可用样本较少时，可以使用合成数据来近似目标函数，或者在涉及敏感数据的应用中，使用隐私保护算法训练生成模型，并利用合成样本定义代理。

[1]法国国家信息与自动化研究院、巴黎高等师范学院、巴黎文理研究大学，法国巴黎 [2]三星人工智能中心，英国剑桥。本文工作完成于法国国家科学研究中心、巴黎高等师范学院、法国国家信息与自动化研究院任职期间。通讯作者：Blake Woodworth <blakewoodworth@gmail.com>、Kon- stantin Mishchenko <konsta.mish@gmail.com>、Francis Bach< francis.bach@inria.fr>。

为了降低操作实体机器人的成本和风险，可以使用一种速度更快、成本更低且更安全的模拟器。

遗憾的是，直接优化人工设计的代理目标函数，未必能提升原始问题上的性能，因为替代损失可能会引入偏差。我们的工作旨在解决这一问题，主要贡献是一种通用算法，它利用代理损失中的曲率信息，从而更好地优化我们所关心的目标函数。

1.1. 问题

我们的最终目标是求解如下形式的无约束凸优化问题

$$\min_{\mathbf{w}\in\mathbb{R}^d} L(\mathbf{w})\,. \tag{1}$$

此处， L 可以表示由权重参数化的机器学习模型的总体损失$\mathbf{w}$；也可以是训练损失，或任何其他需要最小化的目标函数^。为求解 (1)，我们使用代理函数 L 以及用于近似 $\mathbf{g}_k$ 的随机梯度 $\nabla L(\mathbf{w}_k)$ 并进行更新

$$\mathbf{w}_{k+1} \approx \arg\min_{\mathbf{w}} \Big\{ \underbrace{\hat{L}(\mathbf{w})}_{\text{Proxy loss}} + \underbrace{\langle \mathbf{g}_k - \nabla\hat{L}(\mathbf{w}_k), \mathbf{w}-\mathbf{w}_k\rangle}_{\text{Bias correction}} + \underbrace{\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_k\|^2}_{\text{Regularization}} \Big\}. \tag{2}$$

因此，每次迭代只使用一个随机梯度^，该梯度来自 L，但需要基于 L 进行更繁重的计算。如果^我们取 $L\equiv 0$，则可知，更新式 (2) 等价于执行一步随机梯度下降

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta\mathbf{g}_k\,. \tag{3}$$

因此，随机梯度下降是比较我们的算法的自然基准，而核心问题是，使用^（2）（其中 L非零）而不是（3），是否值得付出^额外的计算成本。我们将证明，当 L与 L足够“相似”时，我们算法按照（2）进行的更新，能够以显著更少的 L 次迭代收敛到最小值

迭代次数即可，而随机梯度下降则需要更多迭代，（3）。因此，我们的算法所需的来自 L 的随机梯度数量少于随机梯度下降所需的数量。

ˆ 要使这一方法奏效，显然需要 L 以某种方式类似于 L。我们考虑的具体相似性定义是：函数 $h := L - L$ 可微，且其梯度是 δ-Lipschitz连续的。对于 L 和 L 均二阶可微的情形，这等价于要求 $\forall_{\mathbf{w}} \|\nabla^2 L(\mathbf{w}) - \nabla^2 \hat{L}(\mathbf{w})\|_{\mathrm{op}} \leq \delta$。这种度量有时称为“$\delta$-Hessian 相似性”，在优化文献中十分常见（例如参见 Mairal，2013a年；2015年；Arjevani 和 Shamir，2015年；Kovalev 等，2022年；Chayti 和 Karimireddy，2022年）。对于某些问题，我们可以轻松构造出满足相似性条件且 δ较小的代理。例如，最小二乘目标函数的 Hessian 矩阵就是特征协方差矩阵，因此可以使用任何在算子范数下 δ-准确的协方差矩阵估计来定义合适的代理。在第 3.1 节中，我们还将进一步证明，对于最小二乘和逻辑回归，无需使用任何标签即可构造具有 $\delta = 0$的代理。

1.2. 我们的贡献

我们提出了算法 1，该算法以成本较低的代理损失访问取代对目标函数本身成本高昂的访问。在第 2.2 节中，我们证明了该算法在凸情形和强凸情形下的收敛保证；在第 2.4 节中，我们表明这些保证意味着统计最优性，并且算法的表现取决于相似性参数 δ，而非 L 本身的其他性质，例如可能大得多的光滑性参数。从技术层面来说，与许多类似方法相比，我们的分析更进一步，因为它允许每次迭代中的近端点子问题以非精确方式求解，而非精确程度由一个简单的判据刻画，该判据可在执行时进行评估。此外，在第2.3 节中，我们表明，由于近端点子问题目标函数具有强凸性，因此可以高效地满足这一判据。在第 3 节中，我们讨论了该算法的潜在应用场景：目标函数访问成本高昂，但存在合适的代理。在第 4 节中，我们通过实验展示了该算法在实际问题中的有效性，其中包括一个具有非凸目标函数的问题。最后，在第5节中，我们讨论了本研究的若干可能扩展，并初步分析了该算法在非凸情形下的表现。

1.3. 背景与相关工作

用更易处理的代理函数替代实际目标损失，是统计学和机器学习领域由来已久的思想，几乎所有应用都会使用某种代理函数。例如，合页损失或逻辑损失等凸损失函数通常被用作不连续的 0-1 分类损失的代理函数。这样可以避免最小化 0-1 分类损失时的计算不可处理性（例如参见 Arora 等人，1997），但请注意，这本身并不能缓解以我们所讨论的意义访问目标函数时遇到的其他困难，例如对训练样本的访问受限。另一个经典例子是经验风险最小化（Vapnik 和 Chervonenkis，1968），其中训练损失被用作实际目标——总体损失的代理函数。这与我们的动机更加一致，因为它用难以评估的总体损失——总体损失由未知的数据分布定义——替代为我们可以直接评估、计算梯度等的训练损失。

然而，这两个示例都依赖于来自目标分布的独立同分布样本，而我们的方法可以利用额外的“辅助信息”——合成数据、模拟器等——来补充较小规模的独立同分布样本集合。此外，代理函数的经典使用方式与我们的方法不同：通常，代理函数会被直接最小化，并将所得结果作为目标函数最小值的估计；而我们则使用目标函数本身的随机梯度，通过代理函数来促进优化。因此，经典方法依赖于代理函数的最小值逼近目标函数的最小值，而我们的方法只要求代理函数与目标函数具有相似的二阶导数。这并不一定 更容易实现，但在某些情况下会更容易（例如，参见第3.1节）；而且，对于涉及合成数据、模拟器等的应用，这种方法更为适用，因为由于合成数据或模拟器的特殊性，代理损失可能具有不同的最小值。

在更为相关的工作中，Hendrikx 等人（2020年）针对强凸情形提出了一种算法 SPAG。该算法基于此前的方法 DANE（Shamir 等人，2014），与算法1的加速版本相似。其保证与我们的结果类似，只是经过了加速；然而，SPAG 和 DANE 都要求进行精确的近端点更新，而我们的算法和分析允许这些更新存在误差。受在参数子集上进行最小化这一问题的启发，Parpas（2017）假设代理子问题定义在不同的空间上，投影到该空间及投影返回时只会引入乘性误差。在其他相关工作中，Chayti 和 Karimireddy （2022年）提出了若干用于非凸优化的代理函数算法，其中一种与算法 1的动量变体相似。

Mairal （2013a年；2015年）研究了“主化最小化”元算法，该算法在每次迭代中最小化目标函数的一个上界。我们的方法可以视为主化最小化的一个实例，其中代理函数用于定义损失函数的上界。Mairal在相同的

算法 1 PROXYPROX

输入： 初始化$\mathbf{w}_0 \in \mathbb{R}^d$，步长 $\eta > 0$
对于 $k = 0, 1, \ldots, K-1$ **执行**
  采样 $\mathbf{g}_k$，使得 $\mathbb{E}[\mathbf{g}_k \mid \mathbf{w}_k] = \nabla L(\mathbf{w}_k)$
  设置 $\mathbf{w}_{k+1} \approx$ 极小值点集合$_{\mathbf{w}} \varphi_k(\mathbf{w})$，其中

$$\varphi_k(\mathbf{w}) := \langle \mathbf{g}_k, \mathbf{w} \rangle + D_{\hat{L}}(\mathbf{w}; \mathbf{w}_k) + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_k\|^2.$$

δ相似性条件下分析了这种方法，并证明了类似的收敛速率，但这项工作与我们的工作的关键区别在于，Mairal的方法要求目标函数（对于有限和结构化问题，则要求目标函数各组成部分）具有一个有保证的上界，并要求在每次迭代中精确最小化该上界；而我们只要求期望意义下的上界，并允许近似最小化。因此，我们的设定更适用于机器学习中出现的随机优化问题。

最后，在最相关的工作中，Mairal (2013b年) 研究了一种随机逐次上界最小化方法，该方法适用于我们所关注的随机优化问题。然而，Mairal的方法要求损失函数在每个样本上的取值都有一个已知上界，并要求 δ-相似性上界对每个样本几乎必然成立。相比之下，我们只要求期望意义下的上界，并且要求 δ-相似性在期望意义下成立，这使得找到合适的代理函数更容易。此外，我们的收敛保证随迭代次数增加而改善的尺度更优，K。

2. 算法及其分析

现在，我们将介绍算法，并证明该算法在凸情形和强凸情形下的收敛保证。我们还将论证，当 L 与 L足够相似时，我们的方法可以提供明显优于随机梯度下降的保证。

2.1. 设定与符号

本文中，我们用 $\|\mathbf{u}\|$ 表示向量$\mathbf{u}$的欧几里得范数。我们假设 L的最小值能够取到，并将其记为 L^*，其中 $\mathbf{w}^*$ 表示任意一个极小值点。我们用 $B^2 := \mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2$简记（可能是随机的）初始化$\mathbf{w}_0$, 与该极小值点之间的距离。

如果对所有$\mathbf{u},\mathbf{v}$，函数 f 都是 μ-强凸的

$$f(\mathbf{v}) \geq f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle + \frac{\mu}{2}\|\mathbf{u} - \mathbf{v}\|^2, \quad (4)$$

其中，若 f 不可微，则 $\nabla f(\mathbf{u})$在此表示 f的任意次梯度。当这一条件对 $\mu = 0$成立时，f 仅为凸函数。函数 f 是 H-光滑的若其可微，且 ∇f 是 H-Lipschitz连续的；或者，等价地，对所有 $\mathbf{u}$，$\mathbf{v}$

$$|f(\mathbf{v}) - f(\mathbf{u}) - \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle| \leq \frac{H}{2}\|\mathbf{u} - \mathbf{v}\|^2. \quad (5)$$

我们将势函数为 ψ 时的 Bregman 散度定义为

$$D_\psi(\mathbf{u}; \mathbf{v}) := \psi(\mathbf{u}) - \psi(\mathbf{v}) - \langle \nabla\psi(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle . \quad (6)$$

虽然 Bregman 散度通常仅针对严格凸势函数定义 ψ，但本文允许 ψ 为任意可微函数，其中 D_ψ 如（6）所示直接定义。就本文而言，Bregman 散度的关键性质是三点恒等式：[1] 对所有 $\mathbf{u}$,$\mathbf{v}$,$\mathbf{w}$

$$D_\psi(\mathbf{u}; \mathbf{v}) - D_\psi(\mathbf{u}; \mathbf{w}) - D_\psi(\mathbf{w}; \mathbf{v}) = \langle \nabla\psi(\mathbf{v}) - \nabla\psi(\mathbf{w}), \mathbf{w} - \mathbf{u} \rangle . \quad (7)$$

我们的算法依赖两个主要假设：一个刻画目标函数 L 与代理 L之间的相似性，另一个刻画随机梯度$\mathbf{g}_k$中的噪声水平。首先，我们假设

ˆ假设 2.1。函数 $h := L - L$ 可微，且 ∇h 是 δ-Lipschitz连续的。

ˆ也就是说，我们要求代理 L的曲率与目标函数 L 相近。根据（5），假设 2.1 意味着

$$|D_h(\mathbf{u}; \mathbf{v})| = \left|D_L(\mathbf{u}; \mathbf{v}) - D_{\hat{L}}(\mathbf{u}; \mathbf{v})\right| \leq \frac{\delta}{2}\|\mathbf{u} - \mathbf{v}\|^2.$$

这也说明了算法 φ_k 中的定义是合理的1，因为当 δ 较小时

$$\varphi_k(\mathbf{w}) \approx \langle \mathbf{g}_k, \mathbf{w} \rangle + D_L(\mathbf{w}; \mathbf{w}_k) + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_k\|^2,$$

这对应于直接针对目标函数 L进行（非精确的随机）近端点迭代，这种方法众所周知具有良好的性能（例如参见，Barr´e 等人，2022年）。

我们注意到，假设 2.1 可以适用于不可微的目标函数，例如，如果 $L(\mathbf{w}) = \ell(\mathbf{w}) + \psi(\mathbf{w})$ ˆ ˆ ˆ 且 $L(\mathbf{w}) = \ell(\mathbf{w}) + \psi(\mathbf{w})$，其中 ℓ 和 ℓ 可微，这是因为 ψ 不会出现在 $L - L$中。这样一来，我们的算法也兼容非光滑正则项，例如L_1 惩罚项，只要将其纳入代理中。

除了假设 2.1之外，我们还需要以下条件：

假设 2.2。对于每个 k, $\mathbb{E}[\mathbf{g}_k \,|\mathbf{w}_k] = \nabla L(\mathbf{w}_k)$ 和$\mathbb{E}\left[\|\mathbf{g}_k - \nabla L(\mathbf{w}_k)\|^2 \,|\mathbf{w}_k\right] \leq \sigma^2$。如果 L 不可微但为凸函数，则该条件对 $\partial L(\mathbf{w}_k)$中的某个次梯度成立。

随机**梯度**$_k$ 是无偏的且具有有界方差，这一假设在优化文献中非常常见（Nemirovskii 与 Yudin，1983；Dekel 等人，2012；Bubeck，2015年）。

[1] 这并不要求凸性（Chen 与 Teboulle，1993，引理 3.1）。

2.2. 收敛保证与证明概要

我们从主要结果开始：

定理 2.3. 在假设 *2.1* 和 *2.2*下，令 L 为μ-强凸，并令 $\eta \le \frac{1}{4\delta}$。若对每个 k —

$$\mathbb{E}\|\nabla\varphi_k(\mathbf{w}_{k+1})\|^2 \le \frac{\mu}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 + G^2,$$

对于某个 G，则算法 *1*的迭代点 $\bar{\mathbf{w}}_K$ *(参见(25))* 的几何加权平均满足

$$\mathbb{E}L(\bar{\mathbf{w}}_K) - L^* \le \frac{5B^2}{8\eta}\left(1+\frac{2\eta\mu}{5}\right)^{1-K} + 2\eta\sigma^2 + \frac{G^2}{\mu},$$

其中，期望是对随机梯度的随机性以及迭代点选择的随机性取的。

详细证明见附录 A 和 B。我们先对 $\nabla\varphi_k$进行直接的代数运算。回想 $h = L - \hat{L}$，可得[2]

$$\nabla\varphi_k(\mathbf{w}_{k+1}) = -\nabla h(\mathbf{w}_{k+1}) + \nabla h(\mathbf{w}_k) + \nabla L(\mathbf{w}_{k+1}) + \mathbf{g}_k - \nabla L(\mathbf{w}_k) + \frac{\mathbf{w}_{k+1}-\mathbf{w}_k}{\eta}. \quad (8)$$

接下来，(6) 意味着

$$L(\mathbf{w}_{k+1}) - L^* = \langle\nabla L(\mathbf{w}_{k+1}), \mathbf{w}_{k+1}-\mathbf{w}^*\rangle - D_L(\mathbf{w}^*;\mathbf{w}_{k+1}). \quad (9)$$

重新整理 (8) 后，我们可以将 $\nabla L(\mathbf{w}^{k+1})$ 表示为$\nabla\varphi_k(\mathbf{w}^{k+1})$ 及其他量的形式，并将其代入 $\nabla L(\mathbf{w}^{k+1})$ 在 (9) 中的位置。经过若干简单的代数运算，包括应用三点恒等式 (7)，我们得到关键恒等式

$$\begin{aligned} & L(\mathbf{w}_{k+1}) - L^* \quad (10)\\ & \left.\begin{aligned} &= \frac{1}{2\eta}\|\mathbf{w}_k-\mathbf{w}^*\|^2 - D_h(\mathbf{w}^*;\mathbf{w}_k)\\ &- \frac{1}{2\eta}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2 + D_h(\mathbf{w}^*;\mathbf{w}_{k+1}) \end{aligned}\right\}\ (\text{Telescope})\\ & \left.\begin{aligned} &+ \langle\nabla L(\mathbf{w}_k) - \mathbf{g}_k, \mathbf{w}_{k+1}-\mathbf{w}^*\rangle \end{aligned}\right\}\ (\text{Noise})\\ & \left.\begin{aligned} &+ \langle\nabla\varphi_k(\mathbf{w}_{k+1}), \mathbf{w}_{k+1}-\mathbf{w}^*\rangle \end{aligned}\right\}\ (\text{Inexact})\\ & \left.\begin{aligned} &+ D_h(\mathbf{w}_{k+1};\mathbf{w}_k) - \frac{1}{2\eta}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2\\ &- D_L(\mathbf{w}^*;\mathbf{w}_{k+1}). \end{aligned}\right\}\ (\text{Cancel}) \end{aligned}$$

尽管这个表达式乍看起来很繁琐，但它直接引导我们采用一种策略来分析算法 1。

望远镜项与凸优化算法分析中常见的项相似。在证明定理 2.3时，粗略地说，我们对 (10) 进行平均 k，使得除两项之外的所有项都相互抵消，并证明在假设 2.1 且满足 $\eta \le \frac{1}{4\delta}$的条件下，这些项对误差的总贡献很小。

噪声项 刻画了使用含噪梯度 $\mathbf{g}_k$ 而非 $\nabla L(\mathbf{w}_k)$所产生的影响。在假设 2.2下，其期望至多为 $2\eta\sigma^2 + \frac{1}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2$。重要的是，由于 $\mathbf{g}_k$ 是无偏的，我们可以用$\|\mathbf{w}_{k+1}-\mathbf{w}_k\|$ 来界定它，而不必使用 $\|\mathbf{w}_{k+1}-\mathbf{w}^*\|$。

不精确项体现了近端点更新只能近似计算所产生的影响。如果我们能直接令 $\mathbf{w}_{k+1} = \operatorname{argmin}_{\mathbf{w}}\varphi_k(\mathbf{w})$，不精确项就会为零，但这在实践中是不可能的。相反，简单应用杨氏不等式即可将该项界定为$\frac{1}{\mu}\|\nabla\varphi_k(\mathbf{w}_{k+1})\|^2 + \frac{\mu}{4}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2$。

上文除了假设 2.1 和 2.2 所蕴含的性质外，没有用到 L 的任何性质。不过，为了处理抵消项（我们利用它们消除上文引入的不需要的量），我们最终会用到在 μ-强凸情形下，$D_L(\mathbf{w}^*;\mathbf{w}_{k+1}) \ge \frac{\mu}{2}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2$。因此，当 $\mathbf{w}_{k+1}$的选取使得 $\|\nabla\varphi_k(\mathbf{w}_{k+1})\|^2$ 足够小时，我们就能抵消噪声项和不精确项中的所有内容，唯独留下 $2\eta\sigma^2$，从而产生额外的$-\frac{\mu}{4}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2$，并将其纳入望远镜项以实现几何收缩。这里，$L$ 的强凸性对于消除不精确项对 $\|\mathbf{w}_{k+1}-\mathbf{w}^*\|$的依赖至关重要——如果没有强凸性，就不清楚如何在不令 $\nabla\varphi_k(\mathbf{w}_{k+1}) = 0$的情况下控制不精确项。补全细节并使用标准论证即可完成证明。

虽然定理 2.3 要求 L 为 μ-强凸，但当 L 为凸函数时，也可以通过正则化获得类似的保证：

定理 *2.4*。 在假设 *2.1* 和 *2.2* 下，设 L 为凸函数，且令 $\eta \le \frac{1}{4\delta}$。然后，对正则化目标函数 $L^{(\mu)}(w) := L(w) + \frac{\mu}{2}\|w-w_0\|^2$和 $\hat{L}^{(\mu)}(w) := \hat{L}(w) + \frac{\mu}{2}\|w-w_0\|^2$ 应用算法 *1*，其中 $\mu = \frac{1}{\eta K}$ 且任意 $\eta \le \frac{1}{4\delta}$，可保证对于每个 k —

$$\mathbb{E}\|\nabla\varphi_k(\mathbf{w}_{k+1})\|^2 \le \frac{1}{4\eta^2 K}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 + G^2$$

对于某个 G，则

$$\mathbb{E}L\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k\right) - L^* \le \frac{9B^2}{8\eta K} + 2\eta\sigma^2 + \eta K G^2.$$

证明详见附录 B，其核心观察是 $L^{(\mu)}$ 和 $\hat{L}^{(\mu)}$ 满足定理 2.3所需的条件，较大的 μ 会带来更快的收敛。同时，当 μ 足够小时，优化 $L^{(\mu)}$ 实质上等同于优化 L 本身。最后，我们选择 μ，以在这些相互竞争的目标之间取得平衡。

[2]这里，为简便起见，假设 L 和 $\hat{L}$ 均可微，但一般情形请参见附录 A 。

最后，在强凸和凸情形下对 η 进行最优选择，可得到以下复杂度保证；证明见附录 C：

推论 2.5. 在定理 2.3的设定下，存在一个通用常数 c 使得，对于任意 $\varepsilon > 0$，当 $G^2 \le \frac{1}{2}\mu\varepsilon$ 且

$$\eta = \frac{5}{\mu(K-1)}\Big(1 + \ln\Big(\frac{5B^2\mu(K-1)}{\varepsilon}\Big)\Big),$$

那么，算法 $\hat{\mathbf{w}}_K$ 的输出 次优性至多为 $\mathbb{E}L(\hat{\mathbf{w}}) - L^* \le \varepsilon$ 时

$$K \ge c \cdot \Big(1 + \frac{\delta}{\mu} + \frac{\sigma^2}{\mu\varepsilon}\Big)\ln\Big(e + \frac{(\mu+\delta)B^2}{\varepsilon} + \frac{\sigma^2 B^2}{\varepsilon^2}\Big).$$

同样，在定理2.4的设定下，存在一个通用常数 c 使得，对于任意 $\varepsilon > 0$，当 $G^2 \le \frac{\sigma^2}{K} + \frac{\delta^2 B^2}{K^2}$ 且

$$\eta = \min\Big\{\frac{1}{4\delta}, \frac{B}{\sigma\sqrt{K}}\Big\},$$

那么，算法 $\hat{\mathbf{w}}_K$ 的输出 次优性至多为 $\mathbb{E}L(\hat{\mathbf{w}}) - L^* \le \varepsilon$ 时

$$K \ge c \cdot \Big(\frac{\delta B^2}{\varepsilon} + \frac{\sigma^2 B^2}{\varepsilon^2}\Big).$$

2.3. 使 $\nabla\varphi_k$ 变小

定理2.3 和2.4要求每次更新都满足

$$\mathbb{E}\|\nabla\varphi_k(\mathbf{w}_{k+1})\|^2 \le \frac{\mu}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 + G^2,$$

在 $\mu = \frac{1}{\eta K}$ 凸情形下。这是一项方便的不精确性标准，因为它完全由执行时可以评估的量构成；在经典近端点算法的背景下，也有研究考察了对误差的类似界限（例如，参见Rockafellar，1976；Barr´e 等人，2022年）。然而，这引出了一个问题：找到这样的$\mathbf{w}_{k+1}$有多困难。事实证明，由于子问题目标函数 $\frac{1}{\eta}$ 是 φ_k-强凸的，这件事可能相当容易。例如，我们可以使用随机梯度下降：

命题 2.6。 设 L 为凸且 H光滑，并且可获得 L 的随机梯度，其方差至多为 ρ^2。那么对于常数 c、c'，以下方法的输出$\hat{\mathbf{w}}$为，

$$T \ge c(1+H\eta)\max\Big\{\ln\Big(\frac{c'(1+H\eta)^2}{\mu\eta}\Big), \frac{\rho^2}{G^2}\Big\}$$

在初始化为$\mathbf{w}_k$ 的 φ_k 上执行随机梯度下降 的步骤满足

$$\mathbb{E}\|\nabla\varphi_k(\hat{\mathbf{w}})\|^2 \le \frac{\mu}{4\eta}\|\hat{\mathbf{w}} - \mathbf{w}_k\|^2 + G^2.$$

证明将在附录 E中给出，其基于以下观察： L 具有 H-光滑性意味着 φ_k 是 $(H+\frac{1}{\eta})$-光滑，因此 $\|\nabla\varphi_k(\hat{\mathbf{w}})\|^2 \le 2(H+\frac{1}{\eta})(\varphi_k(\hat{\mathbf{w}}) - \varphi_k^*)$，将光滑且强凸目标函数上运行 T步随机梯度下降的标准次优性保证代入后，最终便可得到所需的界。

使用随机梯度下降来计算更新并没有什么特别之处$\mathbf{w}_{k+1}$。我们方法的一个优势是，当 L光滑时，任何能够近似最小化 φ_k 的算法都足够使用，因此我们可以利用 φ_k中的结构，采用专门的优化算法。还有其他更为特殊的随机梯度下降变体，能够直接针对梯度范数较小的解（Allen-Zhu，2018；Foster 等人，2019年），并且对 ρ^2的依赖关系可能更优。

2.4. 与直接优化 L的比较

为了更好地理解定理2.3和2.4 的含义，我们将其与直接使用随机梯度来优化目标函数 L 所能达到的结果进行比较。假设2.2 成立，目标函数 L是 H-光滑的， $\|\mathbf{w}_0 - \mathbf{w}^*\|^2 = B^2$，并且 L 要么是凸的，要么是 μ-强凸的。在这种情况下，最小化 L 的一个自然且常用的方法是使用总共 K 个随机梯度执行 K 步随机梯度下降。忽略常数因子，众所周知，当ε时，该方法将得到一个Nemirovskii 与 Yudin，1983）

$$K \ge \frac{HB^2}{\varepsilon} + \frac{\sigma^2 B^2}{\varepsilon^2} \tag{11}$$

$$K \ge \frac{H}{\mu}\ln\frac{HB^2}{\varepsilon} + \frac{\sigma^2}{\mu\varepsilon}. \tag{12}$$

相反，给定一个代理 L 满足假设 2.1，我们可以使用来自 L的相同数量的随机梯度来执行 K 步算法 1。在强凸情形下忽略常数和对数因子后，推论 2.5 表明，当 ε时，这将得到一个误差为

$$K \ge \frac{\delta B^2}{\varepsilon} + \frac{\sigma^2 B^2}{\varepsilon^2} \tag{13}$$

$$K \ge \frac{\delta}{\mu}\ln\frac{\delta B^2}{\varepsilon} + \frac{\sigma^2}{\mu\varepsilon}. \tag{14}$$

比较（11）和（12）与（13）和（14），我们可以看到，算法 1 本质上将随机梯度下降对 H（即 L的光滑常数）的依赖替换为 δ（即光滑常 数 $h = L - L$）。因此，如果能够找到一个代理 L 满足 $\delta \ll H$，算法1 可能会比随机梯度下降具有更好的保证。

另一方面，我们保证中的第二部分“统计项”与随机梯度下降相匹配。这些统计项刻画了仅使用 L K 方差为 σ^2 的随机梯度进行优化时的信息论极限，而已知在最坏情况下，这些极限无法进一步改进（Nemirovskii 与 Yudin，1983），因此我们的算法在统计意义上是最优的

（在强凸情形下，相差一个对数因子）。事实上，长期以来已知的下界表明，若要获得一个精度为ε的解，使用任何仅使用 K来自 L 的随机梯度的算法，至少需要（Nemirovskii 与 Yudin，1983）

$$K \geq \frac{HB^2}{\sqrt{\varepsilon}} + \frac{\sigma^2 B^2}{\varepsilon^2}$$
$$K \geq \sqrt{\frac{H}{\mu}} \ln \frac{HB^2}{\varepsilon} + \frac{\sigma^2}{\mu\varepsilon}.$$

因此，当 $\delta \leq \frac{H}{\sqrt{\varepsilon}}$ 处于凸情形，或 $\delta \lesssim \sqrt{H\mu}$处于强凸情形时，算法1 可以通过利用代理ˆ，超越这一下界，从而获得比仅使用 L所能实现的任何算法更好的保证；而仅使用任何来自 K 的随机梯度 L。

尽管算法 1 每次迭代只需从 L 获取一个随机梯度，但每次更新都需要求出函数 φ_k的一个近似驻点，这比执行一次随机梯度下降更新的计算成本更高。话虽如此，在 μ-强凸情形下（仅凸情形下的结论也类似），若忽略所有常数因子和对数因子，推论 2.5 表明，对于$K := \frac{\delta}{\mu} + \frac{\sigma^2}{\mu\epsilon}$、 $G^2 = \mu\epsilon$和 $\eta = \frac{1}{\mu K}$，算法 1 将找到 $O(\epsilon)$-近似极小值点 L。此外，命题 2.6（取 $\rho = \sigma$）表明，每次迭代中的近端子问题可以通过在 φ_k上执行 $T := \frac{K}{\mu K} + \frac{H\sigma^2}{\mu^2 K\epsilon}$步随机梯度下降，以足够的精度求解。因此，实现算法 1 所需的梯度更新计算总数为 $TK = \frac{H}{\mu} + \frac{H}{\mu} \cdot \frac{\sigma^2}{\mu\epsilon}$。相比之下，式 (12) 意味着，直接在 L 上执行随机梯度下降，需要$S := \frac{H}{\mu} + \frac{\sigma^2}{\mu\epsilon}$次梯度更新计算，才能达到 L的$O(\epsilon)$-近似极小值点。

因此，直接对 L 使用随机梯度下降，在所需浮点运算总数方面，计算成本可能低于算法 1 ，因为 $S < TK$。然而，关键在于，算法1 只需要从 K 获取随机梯度，而这些梯度来自 L ˆ本身（其他梯度则来自 L）；而随机梯度下降则需要 $S > K$。总之，执行算法 1 所需的计算量可能高于对 L使用随机梯度下降，但它可以通过减少从 L获取的随机梯度数量来弥补这一点；当 L 较难获取时，这可能带来显著的节省。

2.5. 与预条件化的联系

理解算法的一种实用方法是1 考虑这样一个特殊情形：代理ˆ$L(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{P}\mathbf{w} + \mathbf{b}^\top \mathbf{w}$是二次函数。在这种情况下，精确求解算法中的近端子问题1相当于求解

$$\mathbf{g}_k + \nabla\hat{L}(\mathbf{w}_{k+1}) - \nabla\hat{L}(\mathbf{w}_k) + \frac{1}{\eta}(\mathbf{w}_{k+1} - \mathbf{w}_k) = 0$$
$$\iff \mathbf{w}_{k+1} = \mathbf{w}_k - \eta(\mathbf{I} + \eta\mathbf{P})^{-1}\mathbf{g}_k. \qquad (15)$$

也就是说，算法1 中的每次更新都相当于执行一次预条件随机梯度下降更新，其预条件矩阵为$(\mathbf{I} + \eta\mathbf{P})^{-1}$。当找到满足假设2.1的二次代理时，使用固定预条件矩阵进行这些更新，将能非常有效地最小化目标函数。

然而，对于许多目标函数，Hessian 矩阵 $\nabla^2 L$ 并非常数（也不接近常数），因此无法在所有位置都用矩阵 P 对其进行良好近似，固定的预条件器因而可能不起作用。在这一更一般的情况下，我们可以将算法 1 的每一步视为对局部预条件随机梯度下降更新的近似，其预条件矩阵为 $(\mathrm{I} + \eta\nabla^2\hat{L}(\mathrm{w}_k))^{-1}$。这是因为，对于足够光滑的 ˆ$L$， $\nabla\hat{L}(\mathrm{w}_{k+1}) \approx \nabla^2 L(\mathrm{w}_k)(\mathrm{w}_{k+1} - \mathrm{w}_k) + \nabla\hat{L}(\mathrm{w}_k)$，所以

$$\mathbf{g}_k + \nabla\hat{L}(\mathbf{w}_{k+1}) - \nabla\hat{L}(\mathbf{w}_k) + \frac{1}{\eta}(\mathbf{w}_{k+1} - \mathbf{w}_k) = 0$$
$$\implies \mathbf{w}_{k+1} \approx \mathbf{w}_k - \eta(\mathbf{I} + \eta\nabla^2\hat{L}(\mathbf{w}_k))^{-1}\mathbf{g}_k. \qquad (16)$$

当然，这并非完美对应，但它表明了ˆ的作用 L：它实际上会调整每次ˆ更新，使其沿着 $\mathbf{g}_k$ 方向适应 L的局部曲率。ˆ此外，当 L的曲率足够接近 L的曲率时，这也说明了为什么算法1在 L上进行的更新应比直接使用 SGD 更有效。

3. 应用

算法1 通过近似求解一系列基于 L 的强凸子问题，可减少计算目标函数ˆ所需的随机梯度数量 L。因此，当ˆ能忠实地近似 L 时，这种方法最有优势 L ，也就是说，δ ˆ较小；此外，当模型 L 比真实目标函数 L更易获取时，这种方法也很有优势。这类情况有很多，本节将讨论其中几个例子。

3.1. 标签成本高昂时的回归与分类

在监督学习中，我们希望利用输入-标签对（x， y） $\sim \mathcal{D}$进行训练。通常，我们可以轻松收集大量独立同分布的输入 x_1，x_2，.....，但为这些输入获取标签的成本可能要高得多。例如，要判断一次磁共振成像扫描中是否存在肿瘤，可能需要咨询（并很可能需要付费聘请）受过专业训练的医疗技术人员。因此，我们希望尽可能少地使用标签进行学习，而我们的算法可以为此提供一种方法。

对于输入 $x \in \mathbb{R}^d$ 和标签 $y \in \mathbb{R}$，最小二乘回归（其中 $L(\mathbf{w}) = \frac{1}{2}\mathbb{E}_{(x,y)\sim\mathcal{D}}[(\mathbf{w}^\top x - y)^2]$，）是统计分析中最常用的方法之一。值得注意的是，这个损失函数的 Hessian 矩阵 $\nabla^2 L(\mathbf{w}) = \mathbb{E}_{x\sim\mathcal{D}_x}[xx^\top]$不依赖于标签 y，因此我们可以构造一个代理ˆ $L(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbb{E}_{x\sim\mathcal{D}_x}[xx^\top]\mathbf{w}$，使其满足假设 2.1

而 $\delta = 0$ 无需使用任何标签！因此，我们每次迭代只需使用一个带标签样本 (x_k, y_k)来估计 $\mathbf{g}_k = (\mathbf{w}^\top x_k - y_k)x_k$，然后进行近似最小化只需要 x 个样本。

同样地，对于输入 $x \in \mathbb{R}^d$ 和二元标签 $y \in \{0, 1\}$，逻辑回归损失函数为

$$L(\mathbf{w}) = -\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[y \ln s(\mathbf{w}^\top x) + (1-y)\ln(1 - s(\mathbf{w}^\top x))\right],$$

其中，$s(z) = 1/(1+\exp(-z))$ 表示逻辑函数。与前一个示例一样，该损失函数的 Hessian 矩阵为

$$\nabla^2 L(\mathbf{w}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[s(\mathbf{w}^\top x)(1 - s(\mathbf{w}^\top x))xx^\top\right],$$

同样地，它不依赖于标签 y。因此，我们只需为所有点指定标签 $y = 1$（或为它们指定任意标签），即可构造代理损失：

$$\hat{L}(\mathbf{w}) := -\mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}[\ln s(\mathbf{w}^\top x_n)].$$

这确保了 $\hat{\nabla^2} L = \nabla^2 L$，因此假设 2.1 在 $\delta = 0$下成立，尽管不使用任何标签。因此，算法 1每次迭代只需使用一个标签，外加足够的 x 样本来近似最小化 φ_k。

3.2. 合成数据、模拟器与元学习

在许多值得关注的机器学习问题中，很难从目标分布中收集独立同分布样本。例如，如上一节所述，在医学成像应用中获取标签，可能需要咨询放射科医生；遇到疑难病例时，甚至可能还需要进行额外的活检。因此，获取大型训练集可能既耗费成本又耗费时间。为缓解这一问题，人们对使用合成数据作为补充或替代目标分布中高成本样本的做法产生了浓厚兴趣。生成对抗网络 (GAN)、变分自编码器、扩散模型等生成模型的最新进展，使人们有可能利用少量真实数据，生成数量几乎不受限制的合成训练数据。例如，Frid-Adar 等人（2018）使用生成对抗网络扩充了一个小型数据集，该数据集包含 182 张肝脏病变 CT 扫描图像，从而能够训练出性能更好的分类器。同样，元学习（Hospedales 等人，2022年）、多任务学习（Zhang 和 Yang，2021年）和迁移学习（Weiss 等人，2016）都是通过利用为其他相关任务收集的数据，来提升特定任务性能的努力。例如，在学习对长颈鹿图像进行分类时，通常会使用 ImageNet 这类图像数据集进行预训练或联合训练——尽管 ImageNet 并不包含长颈鹿的标注图像。

除了获取训练数据可能遇到困难之外，评估损失函数或其梯度也可能代价高昂。在机器人学中，评估给定策略所获得的奖励，可能需要等待实体机器人执行该策略一段时间；而每次函数或梯度评估都要等待数秒甚至数分钟，会大幅拖慢训练。此外，在现实世界中训练自动驾驶汽车这类系统，存在损坏车辆或伤及行人的风险。因此，人们通常会使用 MuJoCo 等模拟器（Todorov 等人，2012）和 OpenAI Gym（Brockman 等人，2016），以便在虚拟环境中训练模型，从而实现更快且无风险的处理。

合成数据和模拟器常用于定义代理损失函数，并通过直接最小化该函数来训练模型。在元学习、多任务学习或迁移学习中，通常将其他任务作为先验，引导训练过程，使模型同时在所有任务上都有良好表现。因此，所有这些方法的效果都取决于代理损失函数的极小值是否对应于目标损失函数的极小值 L 本身。事实上，近期研究对例如 ImageNet 预训练的效用提出了一些质疑（He 等人，2019年；Kornblith 等人，2019年），而在模拟器中训练的机器人往往难以泛化到现实世界（Zhao 等人，2020年），这可能是因为这些代理函数会使模型偏向更差的解。相比之下，我们的方法不依赖代理函数的极小值与目标函数的极小值之间存在任何特定对应关系；相反，我们利用代理函数的曲率来加速目标函数本身的优化。

3.3. 使用公共数据和私人数据进行学习

我们的算法可能派上用场的另一个常见场景是目标函数 L 基于私有数据。要在严格保证隐私的同时利用这些数据进行学习，一种方法是使用差分隐私算法生成合成数据集，之后便可像上一节所述那样自由使用该数据集（例如，参见 Amin 等人，2019年；Bowen 和 Snoke，2021年）

当敏感数据之外还存在一批非私有数据时，我们的方法也可能很有用。例如，一些患者可以选择不受限制地向研究人员提供其医疗数据，而另一些患者只有在隐私得到严格保护的条件下才会提供。同样，不同个人可能受不同监管制度约束——欧盟居民的数据必须按照更严格的《通用数据保护条例》规则使用，而这些规则并不适用于例如美国居民。在这些情况下，可以只使用非私有数据进行学习，从而避免使用受保护数据带来的麻烦；但如果总数据中有很大一部分是私有数据，这样做可能并非最优。作为替代，我们可以ˆ使用非私有数据来定义 L，同时采用隐私保护机制，从L 受保护样本中计算随机梯度（Dwork 等人，2014）。当

私有样本与非私有样本的分布相似时，我们预期假设2.1 成立，且δ较小，这意味着算法1 只需迭代相对较少的次数便能达到较低的误差，因此需要访问私有数据的次数也会更少。由于隐私保证会随着访问次数的增加而减弱，这种做法有望比单独使用私有数据获得更好的性能和更强的隐私保证。

3.4. 处理分布偏移

通常，我们希望在面对分布偏移时进行学习——也就是训练分布与测试分布不同的情况。例如，输入和输出的分布可能会随时间变化，而训练集是在今天收集的，模型则在明天部署。近期研究中，Lazaridou 等人(2021年) 指出，正是由于这一原因，神经语言模型的性能可能会随着时间推移而下降。为了让模型保持最新，可以定期收集新的数据集，并重新训练一个完整的新模型。然而，收集足够数量的高质量数据可能既昂贵又耗时，因此最好能利用已有的旧数据。要应用我们的算法，我们可以令 L 表示当前数据分布上的损失函数，ˆ而过时的数据则可用于定义 L。只要分布偏移不太剧烈，我们就可以预期假设2.1 在相对较小的 δ下仍然成立，因此，实现我们的方法只需要少量新数据。

即使训练分布是平稳但有偏的，也可能出现分布偏移。例如，调查数据常常受到抽样偏差、默许偏差或社会期望偏差的影响，导致调查结果与现实不符。同样，当从某个群体或类别较为罕见的数据源中学习时，我们可能会提高罕见群体对应损失函数的权重，以防模型学会忽略该群体；这相当于将目标测试分布设定为该群体占比更高的分布。无论是哪种情况，都可以如上所述应用我们的算法，只需使用相对较少的无偏样本。

4. 实验

为展示我们的方法在实践中的有效性，我们在实际问题上开展了若干实验。

逻辑回归： 首先，我们考虑一个简单的二元逻辑回归问题，使用“蘑菇”数据集，该数据集包含 8124 个 112 维样本。在此实验中，目标函数 L 是在训练数据上计算的逻辑回归损失，加上权重为 $\mu = 10^{-6}H$ 的 ℓ_2 正则化项，其中 H 是目标函数的光滑参数。 L 的随机梯度通过计算梯度得到在均匀且有放回地抽取的、大小为 256 或 1024 的小批量上进行计算。为了模拟如第3.1节所述标签难以获取的情况，我们通过为输入特征分配随机标签来定义代理 L。图1所示结果表明，与直接对目标函数执行随机梯度下降相比，我们的方法 PROXYPROX 在单次迭代性能方面具有竞争力。与Chayti 和 Karimireddy（2022年）提出的 AUXMOM 和 AUXMVR 算法相比，我们同样具有竞争力；这两种算法分别是 PROXYPROX 的动量和方差缩减类比方法。在图1中， x轴统计每种方法使用的来自 L 的随机梯度数量，因此该比较保持了对 L的访问量不变，尽管由于每次迭代都要进行近似的近端点更新，PROXYPROX、AUXMOM 和 AUXMVR 的计算成本更高。我们使用梯度下降近似求解每个 φ_k 子问题。对于实验中的每种算法，我们通过网格搜索选择 η 等超参数，以最小化 250 步后的损失，并在完整的 1000 次迭代中运行表现最佳的选项。我们将 AUXMOM 和 AUXMVR 的 $a = 0.1$设为 0.9，这对应于标准的动量设置。

ResNet-18： 在这项实验中（总结见图 2），我们在 CIFAR-10 上训练 ResNet-18 网络，使用 2560 张图像的子集定义 L，并使用包含 50000 张图像的完整训练数据集来定义 L。虽然 L 对应训练损失函数，我们也绘制了测试损失函数，以表明我们的方法不会过拟合。我们比较了使用随机梯度下降进行 20 次和 40 次迭代（相当于在 L 数据上训练 1 个训练轮次和 2 个训练轮次）来近似最小化 φ_k的效果；步长通过网格搜索调优，最终设为 0.01。图 2表明，只需在子采样数据上进行一次遍历，似乎就足以最小化 φ_k。对于 $\mathbf{g}_k$ 和随机梯度下降更新，我们均使用 128 的小批量大小来最小化 φ_k。在最初的 ~ 25个训练轮次中，我们的方法可能优于 AdamW，并且相较随机梯度下降的优势更为明显。我们采用标准步长：随机梯度下降为 0.1，AdamW 为 0.001；AdamW 的权重衰减设为 0.1，这是通过网格搜索得到的测试准确率最佳值。所有方法也都采用余弦退火。值得注意的是，仅使用 L 进行随机梯度下降训练，测试准确率在经过几次遍历后便停滞在 65，因此我们方法的提升确实来自将 L 与 L结合使用。

5. 开放问题与未来方向

在凸情形下，我们在第 2.2 节证明，在假设2.1 和2.2成立时，算法1 可以利用代理，在 δ-光滑函数上达到与随机梯度下降相当的收敛率——尽管 L 可能并不光滑；即便光滑，其光滑参数也可能远大于 δ。然而，推论2.5 中的复杂度保证要求了解问题

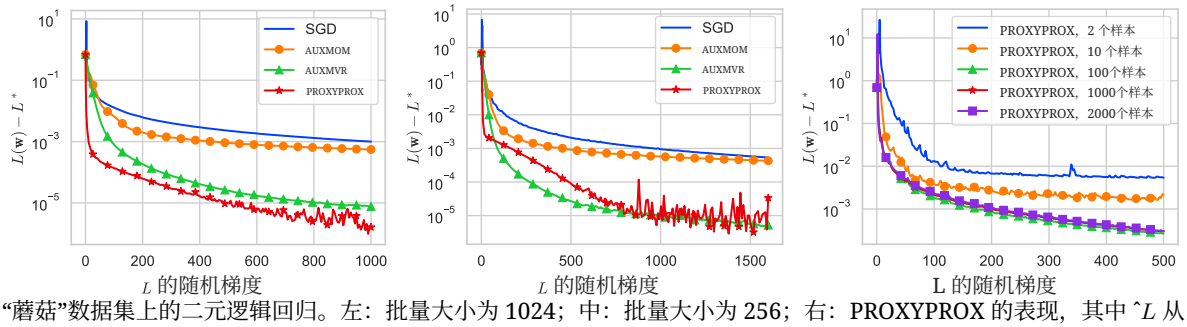


图 *1.* “蘑菇”数据集上的二元逻辑回归。左：批量大小为 1024；中：批量大小为 256；右：PROXYPROX 的表现，其中 $\hat{}L$ 从 L中进行子采样， L 的随机梯度以批量大小 256 计算。

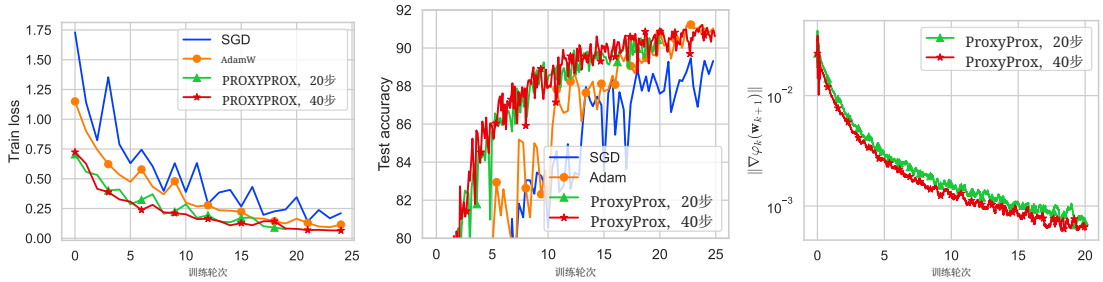


图 *2.* CIFAR-10 实验。左：训练损失， L；中：测试准确率；右：优化 φ_k 20 或 40 个随机梯度下降步骤后， $\|\nabla\varphi_k(\mathbf{w}_{k+1})\|$ 。x 轴表示遍历定义 L 的完整数据集的次数。

参数，包括 δ，而这些参数在实践中可能难以估计。因此，未来的工作或可探索能否通过一种可适应未知问题参数的方法实现类似的性能。此外，假设 2.2中对随机梯度方差的一致上界可能不切实际；在其他情境下，也有可能在更弱的假设下证明类似结果，即梯度方差仅在 $\mathbf{w}^*$处有界（Bach 和 Moulines，2011；Dragomir 等人，2021年）。

超越凸情形同样对机器学习很重要，因为许多实际使用的模型（如神经网络）都会产生非凸训练目标。 $\hat{}$从概念上讲，使用代理 L 来指导目标函数 L的优化，即使这些函数是非凸的，似乎也是合理的；我们的 CIFAR-10 实验表明，这种方法效果良好。不过，在证明理论保证方面仍有工作要做。我们可以证明，算法 1 会收敛到某个驻点的邻域：

定理 5.1。 在假设 *2.1*和*2.2*成立且 L 可微但可能非凸时，令 $\eta \leq \frac{1}{4\delta}$。若

$$\mathbb{E}\|\varphi_k(\mathbf{w}_{k+1})\|^2 \leq \frac{7}{16\eta^2}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2+\frac{1}{8}\|\nabla L(\mathbf{w}_{k+1})\|^2,$$

并且对每个 k， $\mathbb{E}\varphi_k(\mathbf{w}^{k+1}) \leq \mathbb{E}\varphi_k(\mathbf{w}_k)$ ，则

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla L(\mathbf{w}_k)\|^2 \leq \frac{48(L(\mathbf{w}_0)-L^*)}{\eta K}+8\sigma^2.$$

我们在附录D中证明这一点。取 $\eta = \frac{1}{4\delta}$，这一首项变为 $O(\delta(L(\mathbf{w}_0)-L^*)/K)$，这与在 δ-光滑目标函数上的梯度下降速率相符，与我们在凸情形下的结果相呼应。然而，无论$\Omega(\sigma^2)$ 如何，噪声项始终为 η 或 K。目前尚不清楚，这只是我们的分析存在不足，还是确实反映了算法1的最坏情况性能。据我们所知，对于非凸情形下的随机近端点方法，目前普遍缺乏令人满意的分析，因此这种情况可能并非我们的方法所独有。

致谢

我们感谢法国政府提供的支持，该支持由法国国家研究署管理，属于“未来投资计划”的一部分，项目编号 ANR-19-P3IA-0001（PRAIRIE 3IA研究所）；同时也感谢欧洲研究委员会的支持（资助编号 SEQUOIA 724063）。

参考文献

Allen-Zhu，Z. 如何以随机方式使梯度变小：更快的凸与非凸随机梯度下降。 神经信息处理系统进展，31，2018。 (引自第 5)

Amin, K., Dick, T., Kulesza, A., Munoz, A. 和 Vassil- vitskii, S. 差分隐私协方差估计。神经信息处理系统进展，32，2019年。 （引用于第7页）

Arjevani, Y. 和 Shamir, O. 的通信复杂度

分布式凸学习与优化。 神经信息处理系统进展，28，2015年。 (引文见第 2 页)

Arora, S., Babai, L., Stern, J. 和 Sweedyk, Z. 格、码和线性方程组中近似最优解的困难性。计算机与系统科学杂志，54(2):317–331，1997。（引用于第2页）

Bach, F. 和 Moulines, E. 机器学习随机逼近算法的非渐近分析。神经信息处理系统进展，24，2011。（引用于第9页）

Barr´e, M., Taylor, A. B. 和 Bach, F. 关于采用非精确近端算子的一阶方法的原理性分析与设计。《数学规划》，第 1–46 页，2022年。（引用见第 3 和第 5 页）Bowen, C. M. 和 Snoke, J. 对 NIST PSCR 差分隐私合成数据挑战赛中的差分隐私合成数据算法进行了比较研究。《隐私与保密杂志》，11:1，2021年。（引用见第 7 页）Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J. 和 Zaremba, W. OpenAI Gym。arXiv 预印本 arXiv:1606.01540，2016年。（引用见第 7 页）Bubeck, S. 凸优化：算法与复杂性。《机器学习基础与趋势》，8 (3-4):231–357，2015年。（引用见第 3 页）Chayti, E. M. 和 Karimireddy, S. P. 利用辅助信息进行优化。arXiv 预印本 arXiv:2206.00395，2022年。（引用见第 2 和第 8 页）Chen, G. 和 Teboulle, M. 使用 Bregman 函数的近端类最小化算法的收敛性分析。《SIAM 优化杂志》，3(3):538–543，1993年。（引用见第 3 页）Dekel, O., Gilad-Bachrach, R., Shamir, O. 和 Xiao, L. 使用小批量进行最优分布式在线预测。《机器学习研究杂志》，13(Jan):165–202，2012年。（引用见第 3 页）Dragomir, R. A., Even, M. 和 Hendrikx, H. 快速随机 Bregman 梯度方法：精确分析与方差缩减。载于《国际机器学习会议》，第 2815–2825 页，2021年。（引用见第 9 页）Dwork, C., Roth, A. 等。差分隐私的算法基础。《理论计算机科学基础与趋势》，9(3–4):211–407，2014年。（引用见第 7 页）

Foster, D. J., Sekhari, A., Shamir, O., Srebro, N., Sridharan, K. 和 Woodworth, B. 随机凸优化中使梯度变小的复杂性。载于学习理论会议，第 1319–1345 页，2019年。（引用见第5）

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J. 和 Greenspan, H. 使用生成对抗网络进行合成数据增强，以改进肝脏病变分类。载于国际生物医学成像研讨会（*ISBI*），第289–293页，2018。（引用于第7页）

He, K., Girshick, R. 和 Doll´ar, P. 重新思考 ImageNet 预训练。载于*IEEE/CVF*国际计算机视觉会议论文集，第4918–4927页，2019年。（引用于第7页）

Hendrikx, H.、Xiao, L.、Bubeck, S.、Bach, F. 和 Massoulie, L. 分布式优化的统计预条件加速梯度法。载于《国际机器学习会议》，第 4203–4227 页，2020年。（引用见第 2 页）Hospedales, T.、Antoniou, A.、Micaelli, P. 和 Storkey, A. 神经网络中的元学习：综述。《IEEE 模式分析与机器智能汇刊》，44(9):5149–5169，2022年。doi: 10.1109/TPAMI. 2021.3079209。（引用见第 7 页）Kornblith, S.、Shlens, J. 和 Le, Q. V. 更好的 ImageNet 模型是否具有更强的迁移能力？载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 2661–2671 页，2019年。（引用见第 7 页）Kovalev, D.、Beznosikov, A.、Borodich, E. D.、Gasnikov, A. 和 Scutari, G. 最优梯度滑动及其在相似性条件下分布式优化中的应用。载于《神经信息处理系统进展》，2022年。（引用见第 2 页）Lazaridou, A.、Kuncoro, A.、Gribovskaya, E.、Agrawal, D.、Liska, A.、Terzi, T.、Gimenez, M.、de Masson d'Autume, C.、Kocisky, T.、Ruder, S. 等。注意差距：评估神经语言模型的时间泛化。《神经信息处理系统进展》，34: 29348–29363，2021年。（引用见第 8 页）Mairal, J. 基于一阶代理函数的优化。载于《国际机器学习会议》，第 783–791 页。PMLR，2013a年。（引用见第 2 和 3 页）Mairal, J. 面向大规模优化的随机主化-最小化算法。《神经信息处理系统进展》，26，2013b年。（引用见第 3 页）Mairal, J. 增量式主化-最小化优化及其在大规模机器学习中的应用。《SIAM 优化杂志》，25(2):829–855，2015年。（引用见第 2 页）

Nemirovskii, A. S. 和 Yudin, D. B. 优化中的问题复杂度与方法效率。Wiley- Interscience，1983。(引用见第3、5、6及第19页)

Nesterov, Y. 和 Vial, J.-P. 随机规划的置信水平解。*Automatica*，44(6):1559–1568，2008。（引自第13页）

Parpas, P. 一种适用于一类复合优化问题的多层近端梯度算法。*SIAM*科学计算杂志，39(5):S681–S701，2017。（引自第2页）

Rockafellar, R. T. 增广拉格朗日函数及近端点算法在凸规划中的应用。运筹学数学，1(2):97–116，1976。（引自第5页）

Shamir, O.、Srebro, N. 和 Zhang, T. 使用近似牛顿型方法实现通信高效的分布式优化。载于国际机器学习会议，第1000–1008页，2014年。（引自第2页）

Todorov, E., Erez, T. 和 Tassa, Y. Mujoco：一种用于基于模型的控制的物理引擎。载于 国际智能机器人与系统会议，第 5026–5033 页，2012 年。（引用见第7页）

Vapnik, V. N. 和 Chervonenkis, A. Y. 事件出现频率向其概率的一致收敛。载于《科学院院刊》，第 181 卷，第 781–783 页。俄罗斯科学院，1968 年。（引用见第 2页）

Weiss, K., Khoshgoftaar, T. M. 和 Wang, D. 迁移学习综述。《大数据杂志》，3(1):1–40，2016 年。（引用见第 7 页）

Zhang, Y. 和 Yang, Q. 多任务学习综述。*IEEE*知识与数据工程汇刊，34(12):5586–5609，2021年。（引用见第 7页）

Zhao, W.、Queralta, J. P. 和 Westerlund, T. 机器人学中深度强化学习的仿真到现实迁移：综述。载于 *2020*年 *IEEE*计算智能专题研讨会系列（*SSCI*），第 737–744 页。IEEE，2020年。（引用于第 7 页）

定理 2.3 的证明

ˆ 我们的分析要求，在假设 2.1 下， $h = L - L$ 是可微的；但当 L 为凸时，并不要求L 本身可微。具体而言，假设 2.2 要求每个随机梯度 g_k 的期望必须是 L 在 w_k处的一个次梯度。因此，在以下证明中，我们将使用 $\nabla L(\mathrm{w}_k) := \mathbb{E}[\mathrm{g}_k \,|\mathrm{w}_k]$ 作为记号，表示与 g_k的期望相对应的特定次梯度 ˆ。由于 $h = L - L$ 是可微的，我们继而可以用它来定义 $\nabla L(\mathrm{w}_k) := \nabla L(\mathrm{w}_k) - \nabla h(\mathrm{w}_k)$。最后，由于 $0 \in \partial L(\mathrm{w}^*)$，即使 L 在其极小值点处不可微，我们也记 $\nabla L(\mathrm{w}^*) := 0$ 。

引理 A.1。 令 $\mathbf{w}^* \in$ 极小值点集合$_{\mathbf{w}}\, L(\mathbf{w})$，且 $\eta \leq \frac{1}{4\delta}$。那么在假设 *2.1* 和 *2.2*下，对于任意 $\mathbf{w}^{k+1}$， $\mathbf{w}_{\ k}$

$$\begin{aligned}\mathbb{E}[L(\mathbf{w}_{k+1}) - L^*] \leq \frac{1}{2\eta}\mathbb{E}\|\mathbf{w}_k - \mathbf{w}^*\|^2 - \mathbb{E}D_h(\mathbf{w}^*; \mathbf{w}_k) \\ - \frac{1}{2\eta}\mathbb{E}\|\mathbf{w}_{k+1} - \mathbf{w}^*\|^2 + \mathbb{E}D_h(\mathbf{w}^*; \mathbf{w}_{k+1}) \\ + \mathbb{E}\left\langle \nabla\varphi_k(\mathbf{w}_{k+1}),\, \mathbf{w}_{k+1} - \mathbf{w}^*\right\rangle + 2\eta\sigma^2 \\ - \frac{1}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 - \mathbb{E}D_L(\mathbf{w}^*; \mathbf{w}_{k+1}),\end{aligned}$$

其中，期望取自$\mathbf{g}_k$中的随机性。

证明。 计算 $\nabla\varphi_k$ 并重新整理可得

$$\begin{aligned}&\nabla\varphi_k(\mathbf{w}_{k+1}) = \nabla\hat{L}(\mathbf{w}_{k+1}) + \mathbf{g}_k - \nabla\hat{L}(\mathbf{w}_k) + \frac{1}{\eta}(\mathbf{w}_{k+1} - \mathbf{w}_k) \\ \implies\ &\nabla L(\mathbf{w}_{k+1}) = \nabla h(\mathbf{w}_{k+1}) - \nabla h(\mathbf{w}_k) + \nabla\varphi_k(\mathbf{w}_{k+1}) + \nabla L(\mathbf{w}_k) - \mathbf{g}_k - \frac{1}{\eta}(\mathbf{w}_{k+1} - \mathbf{w}_k). \qquad (17)\end{aligned}$$

此外，(6) 立即推出

$$L(\mathbf{w}_{k+1}) - L^* = \left\langle \nabla L(\mathbf{w}_{k+1}),\, \mathbf{w}_{k+1} - \mathbf{w}^*\right\rangle - D_L(\mathbf{w}^*; \mathbf{w}_{k+1}). \qquad (18)$$

将 (17) 代入 (18) 后，我们推导得

$$\begin{aligned}&L(\mathbf{w}_{k+1}) - L^* \\ &= \left\langle \nabla L(\mathbf{w}_{k+1}),\, \mathbf{w}_{k+1} - \mathbf{w}^*\right\rangle - D_L(\mathbf{w}^*; \mathbf{w}_{k+1}) \\ &= \left\langle \nabla h(\mathbf{w}_{k+1}) - \nabla h(\mathbf{w}_k) + \nabla\varphi_k(\mathbf{w}_{k+1}) + \nabla L(\mathbf{w}_k) - \mathbf{g}_k + \frac{1}{\eta}(\mathbf{w}_k - \mathbf{w}_{k+1}),\, \mathbf{w}_{k+1} - \mathbf{w}^*\right\rangle - D_L(\mathbf{w}^*; \mathbf{w}_{k+1}) \\ &= D_h(\mathbf{w}^*; \mathbf{w}_{k+1}) - D_h(\mathbf{w}^*; \mathbf{w}_k) + D_h(\mathbf{w}_{k+1}; \mathbf{w}_k) + \left\langle \nabla\varphi_k(\mathbf{w}_{k+1}) + \nabla L(\mathbf{w}_k) - \mathbf{g}_k,\, \mathbf{w}_{k+1} - \mathbf{w}^*\right\rangle \\ &\qquad + \frac{1}{2\eta}\|\mathbf{w}_k - \mathbf{w}^*\|^2 - \frac{1}{2\eta}\|\mathbf{w}_{k+1} - \mathbf{w}^*\|^2 - \frac{1}{2\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 - D_L(\mathbf{w}^*; \mathbf{w}_{k+1}), \qquad (19)\end{aligned}$$

其中，第三个等号利用了三点性质，其中 D_h, (7)。接下来，假设 2.2、杨氏不等式以及$\eta \leq \frac{1}{4\delta}$共同推出

$$\begin{aligned}\mathbb{E}\left\langle \nabla L(\mathbf{w}_k) - \mathbf{g}_k,\, \mathbf{w}_{k+1} - \mathbf{w}^*\right\rangle &= \mathbb{E}\left\langle \nabla L(\mathbf{w}_k) - \mathbf{g}_k,\, \mathbf{w}_{k+1} - \mathbf{w}_k\right\rangle \\ &\leq \frac{\eta\mathbb{E}\|\nabla L(\mathbf{w}_k) - \mathbf{g}_k\|^2}{1 - 2\eta\delta} + \frac{1 - 2\eta\delta}{4\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 \\ &\leq 2\eta\sigma^2 + \frac{1 - 2\eta\delta}{4\eta}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2,\end{aligned}$$

其中期望是取 关于 $\mathbf{g}_k$中的随机性。由于 h是 δ-光滑的，根据 (5)， $|D_h(\mathbf{w};\ \mathbf{w}_k)| \leq \frac{\delta}{2}\|\mathbf{w} - \mathbf{w}_k\|^2$, so

$$\mathbb{E}[D_h(\mathbf{w}_{k+1}; \mathbf{w}_k) + \left\langle \nabla L(\mathbf{w}_k) - \mathbf{g}_k,\, \mathbf{w}_{k+1} - \mathbf{w}^*\right\rangle] \leq 2\eta\sigma^2 + \frac{1}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2.$$

对（19）取期望并代入，即可完成证明。 □

定理 **定理 2.3**。 在假设 *2.1* 和 *2.2*下，设 L 是 μ-强凸函数，并设 $\eta \leq \frac{1}{4\delta}$。如果对每个 k

$$\mathbb{E}\|\nabla\varphi_k(\mathbf{w}_{k+1})\|^2 \leq \frac{\mu}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 + G^2,$$

对于某个 G，那么算法*1*的迭代点的几何加权平均值 $\bar{\mathbf{w}}_K$ （见(25)），满足

$$\mathbb{E}L(\bar{\mathbf{w}}_K) - L^* \leq \frac{5B^2}{8\eta}\left(1+\frac{2\eta\mu}{5}\right)^{1-K} + 2\eta\sigma^2 + \frac{G^2}{\mu},$$

期望是对随机梯度中的随机性以及迭代点的选择取的。

证明。为简洁起见，记 $\Delta_k := \frac{1}{2\eta}\|\mathbf{w}_k-\mathbf{w}^*\|^2 - D_h(\mathbf{w}^*;\mathbf{w}_k)$。由于 L 是强凸的，由 (4) 和 (6) 可立即得出 $D_L(\mathbf{w}^*;\mathbf{w}_{k+1}) \geq \frac{\mu}{2}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2$。此外，根据 (5)，在假设 2.1 下可得

$$|D_h(\mathbf{w}^*;\mathbf{w}_{k+1})| \leq \frac{\delta}{2}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2 \leq \frac{1}{8\eta}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2.$$

这意味着

$$0 \leq \Delta_{k+1} \leq \frac{5}{8\eta}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2 \leq \frac{5}{4\eta\mu}D_L(\mathbf{w}^*;\mathbf{w}_{k+1}).$$

因此，根据引理 A.1和杨氏不等式

$$\begin{aligned}
&\mathbb{E}[L(\mathbf{w}_{k+1}) - L^*] \\
&\leq \mathbb{E}\Delta_k - \mathbb{E}\Delta_{k+1} + 2\eta\sigma^2 + \mathbb{E}\left\langle\nabla\varphi_k(\mathbf{w}_{k+1}),\, \mathbf{w}_{k+1}-\mathbf{w}^*\right\rangle - \frac{1}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 - \mathbb{E}D_L(\mathbf{w}^*;\mathbf{w}_{k+1}) \\
&\leq \mathbb{E}\Delta_k - \left(1+\frac{2\eta\mu}{5}\right)\mathbb{E}\Delta_{k+1} + 2\eta\sigma^2 + \mathbb{E}\left\langle\nabla\varphi_k(\mathbf{w}_{k+1}),\, \mathbf{w}_{k+1}-\mathbf{w}^*\right\rangle - \frac{1}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 - \frac{\mu}{4}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2.
\end{aligned} \tag{20}$$

因此，根据杨氏不等式和假设的上界

$$\|\nabla\varphi_k(\mathbf{w}_{k+1})\|^2 \leq \frac{\mu}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 + G^2 \tag{21}$$

我们得出

$$\mathbb{E}\left\langle\nabla\varphi_k(\mathbf{w}_{k+1}),\, \mathbf{w}_{k+1}-\mathbf{w}^*\right\rangle \leq \frac{1}{\mu}\mathbb{E}\|\nabla\varphi_k(\mathbf{w}_{k+1})\| + \frac{\mu}{4}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2 \tag{22}$$

$$\leq \frac{1}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 + \frac{\mu}{4}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2 + \frac{G^2}{\mu}. \tag{23}$$

将其代入（20）可得

$$\mathbb{E}[L(\mathbf{w}_{k+1}) - L^*] \leq \mathbb{E}\Delta_k - \left(1+\frac{2\eta\mu}{5}\right)\mathbb{E}\Delta_{k+1} + 2\eta\sigma^2 + \frac{G^2}{\mu}. \tag{24}$$

与强凸随机优化中的常见做法一样，我们现在引入权重 $\alpha_k = (1+\frac{2\eta\mu}{5})^{k-1}$ (类似的加权方法见 Nesterov & Vial，2008，权重之和为 $A_K := \sum_{k=1}^K \alpha_k$ 并尝试给加权迭代点的次优性推导一个上界

$$\bar{\mathbf{w}}_K := \frac{1}{A_K}\sum_{k=1}^K \alpha_k\mathbf{w}_k. \tag{25}$$

根据 L 的凸性以及（24），

$$\begin{aligned}\mathbb{E}L(\bar{\mathbf{w}}_K) - L^* &\le \frac{1}{A_K}\sum_{k=1}^{K}\alpha_k\mathbb{E}[L(\mathbf{w}_k) - L^*] \\ &\le \frac{\mathbb{E}\Delta_0}{A_K} + 2\eta\sigma^2 + \frac{G^2}{\mu} \\ &\le \frac{5\mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2}{8\eta A_K} + 2\eta\sigma^2 + \frac{G^2}{\mu}.\end{aligned}$$

最后，我们给出下界

$$A_K = \frac{5\big(\big(1+\frac{2\eta\mu}{5}\big)^K - 1\big)}{2\eta\mu} \ge \left(1+\frac{2\eta\mu}{5}\right)^{K-1},$$

由此完成证明。 □

B. 定理 2.4 的证明

定理 *2.4*。在假设 *2.1* 和 *2.2* 下，设 L 为凸函数，且设 $\eta \le \frac{1}{4\delta}$。然后，对正则化目标函数 $L^{(\mu)}(w) := L(w) + \frac{\mu}{2}\|w - w_0\|^2$ 和 $\hat{L}^{(\mu)}(w) := \hat{L}(w) + \frac{\mu}{2}\|w - w_0\|^2$ 应用算法 *1*，其中 $\mu = \frac{1}{\eta K}$ 且任意 $\eta \le \frac{1}{4\delta}$，可确保对于每个 k

$$\mathbb{E}\|\nabla\varphi_k(\mathbf{w}_{k+1})\|^2 \le \frac{1}{4\eta^2 K}\mathbb{E}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 + G^2$$

对于某个 G，则

$$\mathbb{E}L\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k\right) - L^* \le \frac{9B^2}{8\eta K} + 2\eta\sigma^2 + \eta K G^2.$$

证明。 根据构造， $L^{(\mu)}$ 是 μ-强凸的，但 $L^{(\mu)} - \hat{L}^{(\mu)} = h$，因此这一差值仍然是 δ-光滑的。因此，我们可以将定理 2.3 中的分析应用于 $L^{(\mu)}$。记 $\mathbf{w}_\mu^* := \operatorname{argmin}(\mathbf{w})$。在证明定理2.3的过程中，我们在式（24）中证明了

$$\begin{aligned}\mathbb{E}[L^{(\mu)}(\mathbf{w}_{k+1}) - L^{(\mu)*}] &\le \mathbb{E}\Delta_k - \left(1+\frac{2\eta\mu}{5}\right)\mathbb{E}\Delta_{k+1} + 2\eta\sigma^2 + \frac{G^2}{\mu} \\ &\le \mathbb{E}\Delta_k - \mathbb{E}\Delta_{k+1} + 2\eta\sigma^2 + \frac{G^2}{\mu},\end{aligned}$$

其中 $\Delta_k := \frac{1}{2\eta}\|\mathbf{w}_k - \mathbf{w}_\mu^*\|^2 - D_h(\mathbf{w}_\mu^*; \mathbf{w}_k)$，且对于所有 k，$0 \le \Delta_k \le \frac{5}{8\eta}\|\mathbf{w}_k - \mathbf{w}_\mu^*\|^2$。此外，$L^{(\mu)}(\mathbf{w}_{k+1}) \ge L(\mathbf{w}_{k+1})$，并且

$$L^{(\mu)*} \le L^{(\mu)}(\mathbf{w}^*) = L^* + \frac{\mu}{2}\|\mathbf{w}_0 - \mathbf{w}^*\|^2.$$

因此，根据 L的凸性

$$\begin{aligned}\mathbb{E}L\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k\right) - L^* &\le \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[L(\mathbf{w}_k) - L^*] \\ &\le \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[L^{(\mu)}(\mathbf{w}_k) - L^{(\mu)*} + \frac{\mu}{2}\|\mathbf{w}_0 - \mathbf{w}^*\|^2\Big] \\ &\le \frac{\mathbb{E}\Delta_0}{K} + 2\eta\sigma^2 + \frac{G^2}{\mu} + \frac{\mu}{2}\mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2. \qquad (26)\end{aligned}$$

最后，

$$0 = \nabla L^{(\mu)}(\mathbf{w}_\mu^*) = \nabla L(\mathbf{w}_\mu^*) + \mu(\mathbf{w}_\mu^* - \mathbf{w}_0),$$

so,

$$\begin{aligned}\left\|\mathbf{w}_0-\mathbf{w}_\mu^*\right\|^2-\left\|\mathbf{w}_0-\mathbf{w}^*\right\|^2 &= -\left\|\mathbf{w}^*-\mathbf{w}_\mu^*\right\|^2+2\left\langle\mathbf{w}_0-\mathbf{w}_\mu^*,\,\mathbf{w}^*-\mathbf{w}_\mu^*\right\rangle\\ &= -\left\|\mathbf{w}^*-\mathbf{w}_\mu^*\right\|^2-\frac{2}{\mu}\left\langle\nabla L(\mathbf{w}_\mu^*),\,\mathbf{w}_\mu^*-\mathbf{w}^*\right\rangle\\ &\le 0.\end{aligned}$$

其中，最后一行用到了 L的凸性。因此，

$$\Delta_0 \le \frac{5}{8\eta}\|\mathbf{w}_0-\mathbf{w}_\mu^*\|^2 \le \frac{5}{8\eta}\|\mathbf{w}_0-\mathbf{w}^*\|^2.$$

与（26）结合后，这意味着

$$\mathbb{E}L\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k\right)-L^* \le \frac{5\mathbb{E}\|\mathbf{w}_0-\mathbf{w}^*\|^2}{8\eta K}+2\eta\sigma^2+\frac{G^2}{\mu}+\frac{\mu}{2}\mathbb{E}\|\mathbf{w}_0-\mathbf{w}^*\|^2,$$

代入我们对 $\mu = {}_{\eta K}^{1}$ 的选择，即完成证明。 □

C. 推论 2.5

推论 2.5。在此　定理 2.3的设定下，存在一个通用常数 c，使得对于任意 $\varepsilon>0$，当 $G^2 \le {}_{2}^{1}\mu\varepsilon$ an~~d~~

$$\eta=\frac{5}{\mu(K-1)}\Big(1+\ln\Big(\frac{5B^2\mu(K-1)}{\varepsilon}\Big)\Big),$$

*th*当输出 $\hat{\mathbf{w}}$（算法 *1* 的输出）满足以下条件时，其次优性至多为 $\mathbb{E}L(\hat{\mathbf{w}})-L^*\le\varepsilon$。

$$K \ge c\cdot\Big(1+\frac{\delta}{\mu}+\frac{\sigma^2}{\mu\varepsilon}\Big)\ln\Big(e+\frac{(\mu+\delta)B^2}{\varepsilon}+\frac{\sigma^2B^2}{\varepsilon^2}\Big).$$

同样，在定理 2.4的情形下，存在一个通用常数 c，使得对于任意 $\varepsilon>0$，当 $G^2 \le {}_{K}^{\sigma^2}+{}_{K^2}^{\delta^2B^2}$ ~~且~~ ——

$$\eta=\min\Big\{\frac{1}{4\delta},\,\frac{B}{\sigma\sqrt{K}}\Big\},$$

则输出 $\hat{\mathbf{w}}$（算法 *1* 的输出）满足以下条件时，其次优性至多为 $\mathbb{E}L(\hat{\mathbf{w}})-L^*\le\varepsilon$。

$$K \ge c\cdot\Big(\frac{\delta B^2}{\varepsilon}+\frac{\sigma^2B^2}{\varepsilon^2}\Big).$$

证明。凸情形下的复杂度保证可通过将所选的 η 代入定理 2.4 的保证中直接得出，然后选择足够大的 K，使期望次优性小于 ε。

接下来讨论强　　凸情形。要应用定理 2.3，我们需要 $\eta \le {}_{4\delta}^{1}$。暂且假设 η 也满足　　s

$$\eta \ge \frac{1}{2\mu(K-1)}.$$

那么，定理 2.3 给出的保证中的第一项至多为

$$\frac{5B^2}{8\eta}\left(1+\frac{2\eta\mu}{5}\right)^{1-K} \le \frac{5B^2\mu(K-1)}{4}\left(1+\frac{2\eta\mu}{5}\right)^{1-K}.$$

接下来，我们注意到，对于 $g(z)=\ln(1+z)$ 以及任意 $z\ge 0$

$$g'(z)=\frac{1}{1+z}=1-\frac{z}{1+z}\ge 1-z.$$

因此，对于 $z \geq 0$

$$\ln(1+z) = \int_0^z g'(t)dt \geq \int_0^z (1-t)dt = z - \frac{1}{2}z^2.$$

因此，

$$\left(1+\frac{2\eta\mu}{5}\right)^{1-K} = \exp\left(-(K-1)\ln\left(1+\frac{2\eta\mu}{5}\right)\right) \leq \exp\left(-\frac{2(K-1)\eta\mu}{5}\left(1-\frac{\eta\mu}{5}\right)\right)$$

因此，只要我们选择 $\eta \leq \frac{5}{2\mu}$，就可以推出

$$\frac{5B^2}{8\eta}\left(1+\frac{2\eta\mu}{5}\right)^{1-K} \leq \frac{5B^2\mu(K-1)}{4}\exp\left(-\frac{(K-1)\eta\mu}{5}\right).$$

因此，若满足关于 η 的以下所有约束：

$$\eta \leq \frac{1}{4\delta} \tag{27}$$

$$\eta \leq \frac{5}{2\mu} \tag{28}$$

$$\eta \geq \frac{1}{2\mu(K-1)} \tag{29}$$

$$\eta \geq \frac{5}{\mu(K-1)}\ln\left(\frac{5B^2\mu(K-1)}{4\varepsilon}\right) \tag{30}$$

那么，定理 2.3 中的保证的第一项有上界 ε。因此我们看到 t

$$\eta = \frac{5}{\mu(K-1)}\left(1+\ln\left(\frac{5B^2\mu(K-1)}{4\varepsilon}\right)\right). \tag{31}$$

这个值始终满足（29）和（30），接下来我们将选择 K 足够大，使其也满足（27）和（28）。具体而言，我们要求 K 满足

$$K \geq 1 + 2\max\left\{1, \frac{10\delta}{\mu}\right\}\left(1+\ln\left(\frac{5B^2\mu(K-1)}{4\varepsilon}\right)\right). \tag{32}$$

将步长（31）代入定理 2.3 中保证的第二项可得

$$2\eta\sigma^2 = \frac{10\sigma^2}{\mu(K-1)}\left(1+\ln\left(\frac{5B^2\mu(K-1)}{4\varepsilon}\right)\right).$$

因此，为了使其小于 ε，我们要求

$$K \geq 1 + \frac{10\sigma^2}{\mu\varepsilon}\left(1+\ln\left(\frac{5B^2\mu(K-1)}{4\varepsilon}\right)\right).$$

结合（32），我们得出，若 η 按照（31）设定，则

$$K \geq 10\max\left\{1, \frac{\delta}{\mu}, \frac{\sigma^2}{\mu\varepsilon}\right\}\left(1+\ln\left(\frac{5B^2\mu K}{4\varepsilon}\right)\right) \tag{33}$$

并且，此外 $G \leq \sqrt{\mu\varepsilon}$，则期望次优性至多为 3ε，因此在上文中改用 $\varepsilon' = \varepsilon/3$ 可确保准确度ε。

最后，为了完成证明，我们注意到对于 $A,\ B > 0$，不等式

$$X \geq A(1+\ln(BX))$$

通过设置即可满足

$$X \geq A(1+4\ln(e+AB)).$$

为展示鉴于此，我们首先注意到，由于 $z \mapsto \sqrt{z}$ 是 $z > 0$上的凹函数，因此有不等式 y

$$\sqrt{z} \leq \sqrt{1} + (z-1)\left(\frac{d}{dz}\sqrt{z}\Big|_{z=1} \right) = 1 + \frac{z-1}{2} = \frac{1+z}{2}.$$

这意味着，对于 $z > 0$

$$\frac{d}{dz}\sqrt{z} = \frac{1}{2\sqrt{z}} \geq \frac{1}{1+z} = \frac{d}{dz}\ln(1+z), \tag{34}$$

最后，由于 $\sqrt{0} = \ln(1+0)$，这意味着对于所有 $z \geq 0$，$\ln(1+z) \leq \sqrt{z}$。

因此，假设

$$X = \alpha A(1 + 4\ln(e + AB)). \tag{35}$$

对于某个 $\alpha \geq 1$。于是，

$$\begin{aligned} A(1+\ln(BX)) &= A(1 + \ln(\alpha) + \ln(AB) + \ln(1 + 4\ln(e+AB))) \\ &\leq A(1 + \ln(\alpha) + \ln(e+AB) + \sqrt{4\ln(e+AB)}) \\ &\leq A(1 + \ln(\alpha) + 3\ln(e+AB)) \\ &\leq \frac{1}{\alpha}X + \ln(c)A \\ &\leq \frac{1+\ln(\alpha)}{\alpha}X \\ &\leq X, \end{aligned}$$

其中我们使用了对于所有 $\alpha \geq 1$，$1 + \ln(\alpha) \leq \alpha$。

因此，对于充分大的常数 c，令

$$K = c \cdot \left(1 + \frac{\delta}{\mu} + \frac{\sigma^2}{\mu\varepsilon}\right) \ln\left(e + \left(1 + \frac{\delta}{\mu} + \frac{\sigma^2}{\mu\varepsilon}\right)\frac{\mu\mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2}{\varepsilon}\right)$$

满足（33），证明完毕。 □

定理 5.1 的证明

引理 D.1。 在假设 *2.1* 和 *2.2*下，令 $\eta \leq \frac{1}{4\delta}$。那么对于任意满足 $\mathbb{E}\varphi_k(\mathbf{w}) \leq \mathbb{E}\varphi_k(\mathbf{w}_k)$ 的 $\mathbf{w}$，

$$\mathbb{E}[L(\mathbf{w}) - L(\mathbf{w}_k)] \leq -\frac{1}{4\eta}\mathbb{E}\|\mathbf{w} - \mathbf{w}_k\|^2 + 2\eta\sigma^2.$$

^证明。 在假设 2.1下，$h = L - L$ 是 δ-光滑的。因此，

$$L(\mathbf{w}) - \hat{L}(\mathbf{w}) \leq L(\mathbf{w}_k) - \hat{L}(\mathbf{w}_k) + \frac{\delta}{2}\|\mathbf{w} - \mathbf{w}_k\|^2 + \left\langle \nabla L(\mathbf{w}_k) - \nabla\hat{L}(\mathbf{w}_k),\, \mathbf{w} - \mathbf{w}_k \right\rangle.$$

重新排列并代入 φ_k后，可得

$$L(\mathbf{w}) - L(\mathbf{w}_k) \leq \varphi_k(\mathbf{w}) - \varphi_k(\mathbf{w}_k) - \frac{1-\eta\delta}{2\eta}\|\mathbf{w} - \mathbf{w}_k\|^2 + \langle \nabla L(\mathbf{w}_k) - \mathbf{g}_k,\, \mathbf{w} - \mathbf{w}_k \rangle.$$

接下来，我们利用杨氏不等式和假设2.2 给出上界

$$\mathbb{E}\langle \nabla L(\mathbf{w}_k) - \mathbf{g}_k,\, \mathbf{w} - \mathbf{w}_k \rangle \leq \frac{\eta\mathbb{E}\|\nabla L(\mathbf{w}_k) - \mathbf{g}_k\|^2}{1-2\eta\delta} + \frac{1-2\eta\delta}{4\eta}\mathbb{E}\|\mathbf{w} - \mathbf{w}_k\|^2 \leq \frac{\eta\sigma^2}{1-2\eta\delta} + \frac{1-2\eta\delta}{4\eta}\mathbb{E}\|\mathbf{w} - \mathbf{w}_k\|^2.$$

事实是，$\eta \leq \frac{1}{4\delta}$ 和 $\mathbb{E}\varphi_k(\mathbf{w}) \leq \mathbb{E}\varphi_k(\mathbf{w}_k)$，从而完成证明。 □

引理 D.2. 在假设 2.1 和 2.2下，令 L 可微，并令 $\eta \le \frac{1}{4\delta}$。如果 $\mathbf{w}^{k+1}$ 满足

$$\mathbb{E}\|\varphi_k(\mathbf{w}_{k+1})\|^2 \le \frac{7}{16\eta^2}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 + \frac{1}{8}\|\nabla L(\mathbf{w}_{k+1})\|^2$$

then

$$-\frac{1}{4\eta}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 \le -\frac{\eta}{48}\mathbb{E}\|\nabla L(\mathbf{w}_{k+1})\|^2 + \frac{\eta\sigma^2}{6}.$$

证明。 根据松弛三角不等式，对于任意向量 a,b,c,d,

$$\|a\|^2 = \|a+b+c+d-b-c-d\|^2 \le 4(\|a+b+c+d\|^2 + \|b\|^2 + \|c\|^2 + \|d\|^2). \tag{36}$$

此外,

$$\nabla\varphi_k(\mathbf{w}_{k+1}) = \nabla\hat{L}(\mathbf{w}_{k+1}) - \nabla L(\mathbf{w}_{k+1}) - \nabla\hat{L}(\mathbf{w}_k) + \nabla L(\mathbf{w}_k) + \mathbf{g}_k - \nabla L(\mathbf{w}_k) + \nabla L(\mathbf{w}_{k+1}) + \frac{1}{\eta}(\mathbf{w}_{k+1}-\mathbf{w}_k). \tag{37}$$

因此（36）满足

$$\begin{aligned} a &= \nabla L(\mathbf{w}_{k+1}) \\ b &= -\nabla\varphi_k(\mathbf{w}_{k+1}) \\ c &= \nabla\hat{L}(\mathbf{w}_{k+1}) - \nabla L(\mathbf{w}_{k+1}) - \nabla\hat{L}(\mathbf{w}_k) + \nabla L(\mathbf{w}_k) \\ d &= \mathbf{g}_k - \nabla L(\mathbf{w}_k) \end{aligned}$$

因此 $a+b+c+d = \frac{1}{\eta}(\mathbf{w}_k - \mathbf{w}^{k+1})$，这意味着在假设 2.1 和 2.2 下

$$\begin{aligned} &\frac{1}{\eta^2}\|\mathbf{w}-\mathbf{w}_k\|^2 \\ &\ge \frac{1}{4}\mathbb{E}\|\nabla L(\mathbf{w})\|^2 - \mathbb{E}\|\nabla\varphi_k(\mathbf{w})\|^2 - \mathbb{E}\|\mathbf{g}_k - \nabla L(\mathbf{w}_k)\|^2 - \mathbb{E}\|\nabla\hat{L}(\mathbf{w}) - \nabla L(\mathbf{w}) - \nabla\hat{L}(\mathbf{w}_k) + \nabla L(\mathbf{w}_k)\|^2 \\ &\ge \frac{1}{4}\mathbb{E}\|\nabla L(\mathbf{w})\|^2 - \mathbb{E}\|\nabla\varphi_k(\mathbf{w})\|^2 - \delta^2\mathbb{E}\|\mathbf{w}-\mathbf{w}_k\|^2 - \sigma^2 \\ &\ge \frac{1}{8}\mathbb{E}\|\nabla L(\mathbf{w})\|^2 - \frac{1}{2\eta^2}\mathbb{E}\|\mathbf{w}-\mathbf{w}_k\|^2 - \sigma^2, \end{aligned}$$

其中我们使用了 $\mathbb{E}\|\nabla\varphi_k(\mathbf{w})\|^2$ 的上界以及 $\eta \le \frac{1}{4\delta}$ 这一事实。重新整理后即可完成证明。 □

定理 5.1。 在假设 2.1 和 2.2下，令 L 可微但可能是非凸的，令 $\eta \le \frac{1}{4\delta}$。 — *If*

$$\mathbb{E}\|\varphi_k(\mathbf{w}_{k+1})\|^2 \le \frac{7}{16\eta^2}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 + \frac{1}{8}\|\nabla L(\mathbf{w}_{k+1})\|^2,$$

并且对每个 k，$\mathbb{E}\varphi_k(\mathbf{w}^{k+1}) \le \mathbb{E}\varphi_k(\mathbf{w}_k)$，则

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla L(\mathbf{w}_k)\|^2 \le \frac{48(L(\mathbf{w}_0)-L^*)}{\eta K} + 8\sigma^2.$$

证明。 根据引理D.1和D.2

$$\mathbb{E}[L(\mathbf{w}_{k+1}) - L(\mathbf{w}_k)] \le -\frac{1}{4\eta}\mathbb{E}\|\mathbf{w}_{k+1}-\mathbf{w}_k\|^2 + 2\eta\sigma^2 \le -\frac{\eta}{48}\mathbb{E}\|\nabla L(\mathbf{w}_{k+1})\|^2 + \frac{\eta\sigma^2}{6}. \tag{38}$$

因此，重新整理并对 k 取平均，得到

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla L(\mathbf{w}_k)\|^2 \le \frac{48(L(\mathbf{w}_0)-L(\mathbf{w}_K))}{\eta K} + 8\sigma^2 \le \frac{48(L(\mathbf{w}_0)-L^*)}{\eta K} + 8\sigma^2 \tag{39}$$

至此，证明完成。 □

E. 命题 2.6 的证明

命题 2.6. 设$\hat{}L$为凸且 H-光滑，并且$\hat{}L$ 的随机梯度可用，其方差至多为 ρ^2。那么对于常数 c、c'，以下算法的输出$\hat{}\mathbf{w}$为

$$T \geq c(1+H\eta)\max\left\{\ln\left(\frac{c'(1+H\eta)^2}{\mu\eta}\right), \frac{\rho^2}{G^2}\right\}$$

从 φ_k 开始的随机梯度下降 步，在 φ_k 初始化于$\mathbf{w}_k$ 时满足

$$\mathbb{E}\|\nabla\varphi_k(\hat{\mathbf{w}})\|^2 \leq \frac{\mu}{4\eta}\|\hat{\mathbf{w}}-\mathbf{w}_k\|^2 + G^2.$$

证明。 由于 $\varphi_k(\mathbf{w}) = \hat{}L(\mathbf{w}) + \frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_k\|^2$（相差一个额外的仿射项），容易看出 φ_k 是 $\frac{1}{\eta}$-强凸且 $(H+\frac{1}{\eta})$-光滑的。因此，对于任意 $\mathbf{w}$

$$\|\nabla\varphi_k(\mathbf{w})\|^2 \leq 2\left(H+\frac{1}{\eta}\right)(\varphi_k(\mathbf{w}) - \varphi_k(\mathbf{w}_\varphi^*)), \tag{40}$$

其中 $\mathbf{w}_\varphi^* =$ 极小值点集合$_{\mathbf{w}}\,\varphi_k(\mathbf{w})$。此外，使用最优步长并从 $\hat{}\mathbf{w}$ 开始对 φ_k 进行随机梯度下降所得到的输出 $\mathbf{w}_k$ 对于通用常数 $c, c' \geq 1$ (Nemirovskii 与 Yudin, 1983)

$$\mathbb{E}[\varphi_k(\hat{\mathbf{w}}) - \varphi_k(\mathbf{w}_\varphi^*)] \leq c\left(H+\frac{1}{\eta}\right)\|\mathbf{w}_k-\mathbf{w}_\varphi^*\|^2\exp\left(-\frac{T}{c'(1+H\eta)}\right) + \frac{c\eta\rho^2}{T}. \tag{41}$$

因此，根据松弛三角不等式以及 φ_k 的强凸性

$$\begin{aligned}
\|\mathbf{w}_k-\mathbf{w}_\varphi^*\|^2 &\leq 2\|\hat{}\mathbf{w}-\mathbf{w}_k\|^2 + 2\|\hat{}\mathbf{w}-\mathbf{w}_\varphi^*\|^2 && (42)\\
&\leq 2\|\hat{\mathbf{w}}-\mathbf{w}_k\|^2 + 4\eta\big(\varphi_k(\hat{\mathbf{w}}) - \varphi_k(\mathbf{w}_\varphi^*)\big) && (43)\\
&\leq 2\|\hat{\mathbf{w}}-\mathbf{w}_k\|^2 + \frac{4c\eta^2\rho^2}{T} + 4c(1+H\eta)\exp\left(-\frac{T}{c'(1+H\eta)}\right)\|\mathbf{w}_k-\mathbf{w}_\varphi^*\|^2. && (44)
\end{aligned}$$

因此，若

$$T \geq c'(1+H\eta)\ln(8c(1+H\eta)) \tag{45}$$

then

$$\|\mathbf{w}_k-\mathbf{w}_\varphi^*\|^2 \leq 4\|\hat{\mathbf{w}}-\mathbf{w}_k\|^2 + \frac{8c\eta^2\rho^2}{T}, \tag{46}$$

于是

$$\begin{aligned}
\|\nabla\varphi_k(\hat{}\mathbf{w})\|^2 &\leq 2\left(H+\frac{1}{\eta}\right)(\varphi_k(\hat{}\mathbf{w}) - \varphi_k(\mathbf{w}_\varphi^*)) && (47)\\
&\leq 2c\left(H+\frac{1}{\eta}\right)^2\exp\left(-\frac{T}{c'(1+H\eta)}\right)\|\mathbf{w}_k-\mathbf{w}_\varphi^*\|^2 + \frac{2c(1+H\eta)\rho^2}{T} && (48)\\
&\leq 2c\left(H+\frac{1}{\eta}\right)^2\exp\left(-\frac{T}{c'(1+H\eta)}\right)\left(4\|\hat{\mathbf{w}}-\mathbf{w}_k\|^2 + \frac{8c\eta^2\rho^2}{T}\right) + \frac{2c(1+H\eta)\rho^2}{T} && (49)\\
&\leq 8c\left(H+\frac{1}{\eta}\right)^2\exp\left(-\frac{T}{c'(1+H\eta)}\right)\|\hat{\mathbf{w}}-\mathbf{w}_k\|^2 + \frac{4c(1+H\eta)\rho^2}{T}. && (50)
\end{aligned}$$

因此，选择

$$T \geq \max\left\{c'(1+H\eta)\ln\left(\frac{32c(1+H\eta)^2}{\mu\eta}\right), \frac{4c(1+H\eta)\rho^2}{G^2}\right\} \tag{51}$$

确保

$$\|\nabla\varphi_k(\hat{\mathbf{w}})\|^2 \leq \frac{\mu}{4\eta}\|\hat{\mathbf{w}}-\mathbf{w}_k\|^2 + G^2. \tag{52}$$

ˆ最后，我们注意到，我们对 T (51) 的选择满足条件 (45)，因为 L 为 H 光滑意味着 L 是$(H+\delta)$-光滑，因此

$$\mu \leq H + \delta \leq H + \frac{1}{\eta} \implies \frac{1 + H\eta}{\eta\mu} \geq 1.$$

□