进行随机优化，感受无噪声：通过双动量机制实现稳定的随机优化

Tehila Dahan以色列理工学院电子与计算机工程系 海法，以色列
t.dahan@campus.technion.ac.il

Kfir Y. Levy电子与计算机工程系 以色列理工学院 海法，以色列 kfirylevy@technion.ac.il

arXiv:2304.04172v2 [cs.LG] 5 Mar 2025

摘要

优化方法对机器学习的成功至关重要，随机梯度下降（SGD）是训练模型的基础算法。然而，SGD 往往对学习率的选择十分敏感，因此需要进行大量超参数调优。在这项工作中，我们提出了一种新的 SGD 变体，在两个关键方面增强了稳定性。**首先**，我们的方法允许使用相同的固定学习率，在**无论噪声大小如何**的情况下都达到最优收敛速率，从而无需在无噪声和含噪环境之间调整学习率。**其次**，我们的方法在**较宽的学习率范围内**都能达到这些最优速率，相较于需要精确选择学习率的标准 SGD，大幅降低了敏感性。我们的核心创新是一种基于双动量机制的新型梯度估计器，它结合了近期提出的两种基于动量的技术。利用这一估计器，我们设计了标准算法和加速算法，使其对学习率的选择具有鲁棒性。具体而言，我们的方法在无噪声和含噪的随机凸优化场景中均能达到最优收敛速率，而无需进行学习率衰减或精细调优。我们还证明，该方法在广泛的学习率范围内都能保持最优性能，凸显了其稳定性和实用性。实证研究进一步验证了该方法的鲁棒性和增强的稳定性。

1 引言

随机凸优化（SCO）是一个基础框架，涵盖了若干经典的机器学习（ML）问题，例如线性回归、逻辑回归和支持向量机等。在过去二十年中，随机凸优化得到了广泛研究，并对机器学习领域产生了深远影响：它使随机梯度下降（SGD）成为训练机器学习模型的标准工具；例如，参见 Shalev-Shwartz 等人（2007）；Welling 和 Teh（2011）；Mairal 等人（2009）；Recht 等人（2011）；此外，这也推动了复杂随机梯度下降变体的设计，它们在训练现代大规模模型中发挥着核心作用（Duchi 等人，2011；Kingma 和 Ba，2015）。应用 SGD 类方法时的一个实际难点是需要调节学习率及其他超参数。众所周知，这类算法的性能在很大程度上取决于学习率的选择是否恰当。自适应随机梯度下降的变体（如 AdaGrad 和 Adam（Duchi 等人，2011年；Kingma & Ba，2015；Levy 等人，2018；Kavis 等人，2019年；Jacobsen & Cutkosky，2022年）旨在通过在训练过程中调整学习率来缓解这一问题。然而，尽管自适应方法减少了对超参数调优的需求，却可能增加额外的复杂性，也未必总能带来更好的泛化性能。在许多应用中，实践者仍倾向于采用标准 SGD，因为与自适应方法相比，它往往能取得更低的测试误差和更好的泛化效果（Wu 等人，2016年；Ruder，2016年）。因此，设计出对学习率选择具有鲁棒性的 SGD 变体，同时保留标准 SGD 的简洁性和泛化优势，在实践中将大有裨益。

为应对这些挑战，我们提出了一种新方法，它保留了标准 SGD 的简洁性和泛化优势，同时显著增强了对学习率选择的鲁棒性。此外，由于我们的方法侧重于稳定梯度估计，而非调整学习率，因此它与自适应技术相互独立，并有望与之结合以取得更佳性能。

聚焦于SCO设置： 本文聚焦于普遍的SCO设置，其中目标函数（期望损失）是光滑损失上的期望（SCO-EOS）；例如，线性回归和逻辑回归问题就属于这种情况（但支持向量机不属于）。在这种情况下，众所周知，为了获得最优性能，SGD需要仔细调整学习率。例如，在无噪声情况下，SGD（或此时的GD）应采用 $\eta^{\text{Offline}} = 1/L$ 的学习率，其中 L 是目标函数的光滑性参数。然而，如果我们在含噪设置中应用这一 η^{Offline}，SGD的保证就会变得毫无意义。为了获得最优的SGD保证，我们大致应将学习率降低 $\sigma\sqrt{T}$ 倍，其中 T 是SGD迭代点（及样本）的总数，并且 σ 是梯度估计中的噪声方差。这说明了SGD对 η 选择的敏感性，这一挑战同样会影响随机加速方法，例如 Lan（2012）、Hu等人（2009年）以及Xiao（2010）。

贡献。 我们为 SCO-EOS 问题引入了一种新颖的梯度估计器，每次迭代仅使用一个样本，并证明其平方误差 $\|\epsilon_t\|^2$，会随着更新次数的增加而缩小，即$\|\epsilon_t\|^2 \propto 1/t$，其中 t 表示迭代点。相比之下，对于标准 SGD 估计器，通常有 $\|\epsilon_t\|^2 = \text{方差}_t = O(1)$。我们的新估计器融合了两种与动量概念相关的近期机制：由 Cutkosky (2019年) 提出的 Anytime Averaging；以及一种校正动量技术（Cutkosky & Orabona，2019年）。因此，我们将该估计器记为 μ^2，其代表**动量**^{2}。

如下文所述，我们的新估计器使我们能够“进行随机（优化），却感受不到噪声”，即允许我们使用与无噪声情况下 GD 所采用的类似工具。具体而言，**(i)** 无论噪声大小如何，我们都可以使用与 GD 中完全相同的固定学习率；**(ii)** 它使我们能够使用梯度估计的范数作为停止准则，这在 GD 中是一种常见做法（Beck，2014年）。最后，**(iii)** 它使我们能够设计出对学习率选择极其稳健的新型 SGD 变体，相较于标准 SGD 显著降低对学习率的敏感性。
具体而言，我们设计了一种称为 μ^2-SGD 的 SGD 变体，以及一种称为μ^2 –ExtraSGD 的加速版本。它们采用了我们新提出的估计量，并证明了其稳定性不受学习率选择的影响 η。我们展示了以下结果：

- **对于 μ^2-SGD：** 采用与 $\eta_{\text{离线}} = 1/8LT$ 完全相同的**学习率**（其中 T 是迭代点/数据样本的总数），在无噪声情形下，μ^2-SGD 的收敛速率为 $O(L/T)$；在含噪情形下，收敛速率为 $O(L/T + \tilde{\sigma}/\sqrt{T})$。此外，在含噪情形下，对于**广泛范围的学习率选择**，即 $\eta \in [\eta_{\text{最小值}}\, \eta_{\text{最大值}}]$, 且比值 $\eta_{\text{最大值}}/\eta_{\text{最小值}} \approx (\tilde{\sigma}/L)\sqrt{T}$，$\mu^2$-SGD 的收敛速率与最优 SGD 相同 $O(L/T + \tilde{\sigma}/\sqrt{T})$。

- 对于 μ^2 – ExtraSGD：在采用与 $\eta_{\text{离线}} = 1/2L$ 完全相同的学习率时，μ^2 – ExtraSGD 在无噪声情形下可达到最优收敛速率 $O(L/T^2)$，在含噪情形下可达到最优速率 $O(L/T^2 + \tilde{\sigma}/\sqrt{T})$。此外，在含噪情形下，对于极其广泛的学习率选择范围，即 $\eta \in [\eta_{\min}\, \eta_{\max}]$,，且比值 $\eta_{\max}/\eta_{\min} \approx (\tilde{\sigma}/L)T^{3/2}$，$\mu^2$ – ExtraSGD 仍可达到相同的最优收敛速率 $O(L/T^2 + \tilde{\sigma}/\sqrt{T})$。上述最优速率对于 SCO-EOS 问题也是紧的，参见 Cutkosky（2022a）中的定理 16.7。

这些比率显著大于标准SGD 对应的比率，而标准SGD 只有在 $\eta^{\max}/\eta^{\min} \approx O(1)$时才能实现最优收敛。这表明，与标准 SGD 相比，我们的方法在稳定性方面有了显著提升（详见附录A 中的详细讨论）。

我们通过实验证明，与各种基线方法相比，我们的方法具有更高的稳定性和更优的性能，从而证实了该方法在理论和实践上的优势。
从技术层面来看，值得注意的是，我们组合使用的每一种动量技术单独都无法确保其恰当组合后能够确保的稳定性，也就是说，无法确保 μ^2-随机梯度下降和 μ^2 –ExtraSGD 的稳定性。此外，我们的加速版本 μ^2 –ExtraSGD 需要以恰当交织的方式，谨慎而精细地融合多种技术，因此分析虽然简洁，却也相当精细。

相关工作： 梯度下降（GD）算法及其随机对应算法随机梯度下降（SGD（Robbins& Monro, 1951）是机器学习和优化的基石。它们在各个领域的应用推动了众多优雅且实用的变体的发展（Duchi etal., 2011年; Kingma &Ba, 2015; Ge et al., 2015）。有趣的是，许多用于实际训练非凸学习模型的随机梯度下降变体，最初都是在随机凸优化的框架下设计的。

正如我们在引言中所提到的，随机梯度下降的性能在很大程度上取决于学习率的选择。大量研究致力于设计能够在整个学习过程中隐式且自适应地优化学习率的方法（Duchi 等人，2011年; Kingma 与 Ba，2015; Kavis 等人，2019年; Antonakopoulos 等人，2022年; Jacobsen 与 Cutkosky，2022年; Ivgi 等人，2023年; Defazio 与 Mishchenko，2023年）；此类方法也被从业者广泛采用。然而，在一些实际场景中，标准（非自适应）随机梯度下降已被证明比自适应变体具有更好的泛化能力，但仍需要找到合适的学习率（例如，参见Giladi 等人（2019年）。

动量（Polyak，1964年）是另一种广泛应用于实践的技术（Sutskever 等人，2013），并且与 Nesterov 加速法（Nesterov，1983）有着有趣的联系。后者是一种开创性方法，可使光滑凸目标函数的收敛速率快于梯度下降。尽管 Nesterov 加速法容易受到噪声影响，Lan （2012）；Hu 等人（2009年）；Xiao （2010）设计了随机加速变体，使其收敛速率能够在无噪声情况下的快速速率与含噪情况下的标准随机梯度下降速率之间进行插值（具体取决于噪声大小）。

我们的工作基于两种与动量概念相关的近期机制：**(i)** Anytime 平均机制Cutkosky (2019年），该机制通过对梯度预言机的查询点取平均来实现。以及**(ii)** 校正动量技术Cutkosky & Orabona(2019年），该技术在整个学习过程中对梯度本身取平均（同时引入校正）。值得注意的是，Anytime 机制已被证明在设计自适应和加速方法方面极为有用（Cutkosky，2019年; Kavis 等人，2019年; Antonakopoulos 等人，2022年）。校正动量机制主要用于为随机非凸问题设计最优且自适应的算法（Cutkosky &Orabona，2019年; Levy 等人，2021年）。

2 设定

考虑目标函数为凸函数 f 的随机优化问题： $\mathcal{K} \mapsto \mathbb{R}$ 满足，

$$f(x) := \mathbf{E}_{z\sim\mathcal{D}} f(x; z) \,, \tag{1}$$

其中， $\mathcal{K} \subseteq \mathbb{R}^d$ 是一个紧凸集， $\mathcal{D}$ 是一个未知分布，我们可以从中抽取独立同分布样本 $\{z_t \sim \mathcal{D}\}_t$。我们考虑一阶优化方法，这些方法迭代地使用此类样本生成一系列查询点，并最终输出一个解 $x_{\text{output}} \in \mathcal{K}$。我们的目标是近似最小化 $f(\cdot)$，因此我们的性能度量是期望超额损失，

$$\text{ExcessLoss} := \mathbf{E}[f(x_{\text{output}})] - \min_{x\in\mathcal{K}} f(x) \,,$$

其中期望是针对样本随机化而言的。

更具体地说，在每次迭代 t 中，此类方法都会维护一个查询点 $x_t \in \mathbb{R}^d$，该点根据过去的查询点和过去的样本 $\{z_1, \ldots, z_{t-1}\}$计算得出。然后，基于 x_t，并利用通过独立于过去样本抽取一个新样本$z_t \sim \mathcal{D}$而得到的梯度估计 g_t，计算下一个查询点 x_{t+1}，即 $g_t := \nabla f(x_t; z_t)$. 注意，该导数是相对于 x 求得的。样本之间的独立性意味着 g_t 是 $\nabla f(x_t)$ 的无偏估计，其含义如下：$\mathbf{E}[g_t|x_t] = \nabla f(x_t)$. 通常，可以将$g_t = \nabla f(x_t; z_t)$ 的计算视为一个（含噪）梯度预言机：它接收一个查询点 $x_t \in \mathcal{K}$，并输出一个向量 $g_t \in \mathbb{R}^d$，该向量是 $\nabla f(x_t)$ 的无偏估计。

假设。我们作出以下假设，**有界直径：** 存在 $D > 0$ 使得：$\max_{x,y\in\mathcal{K}} \|x - y\| \le D$。**有界方差：** 存在 $\sigma > 0$ 使得，

$$\mathbf{E}\|\nabla f(x; z) - \nabla f(x)\|^2 \le \sigma^2 \,, \forall x \in \mathcal{K} \tag{2}$$

光滑函数上的期望： 存在 $L > 0$ 这样的 $\forall x,\ y \in \mathcal{K}\ ,\ z \in$ **支持**$\{\mathcal{D}\}$,

$$\|\nabla f(x;z) - \nabla f(y;z)\| \leq L\|x-y\|\ , \tag{3}$$

这意味着期望损失 $f(\cdot)$ 是 L 光滑的。

有界光滑性方差。 上述假设意味着存在 $\sigma_L^2 \in [0, L^2]$ 使得,

$$\mathbf{E}\left\|(\nabla f(x;z) - \nabla f(x)) - (\nabla f(y;z) - \nabla f(y))\right\|^2 \leq \sigma_L^2\|x-y\|^2\ ,\quad \forall x, y \in \mathcal{K}\ . \tag{4}$$

显然，在确定性设定下，$f(x;\ z) = f(x)\quad \forall z \in$ **支持**$\{\mathcal{D}\}$，我们有 $\sigma_L = 0$。在附录B中，我们证明了公式（3）如何推出公式（4）。

记号： 记号 $\nabla f(x;\ z)$ 表示关于 x 的梯度，即 $\nabla := \nabla_x$。我们用 $\|\cdot\|$ 表示欧几里得范数。给定序列 $\{y_t\}_t$，记 $y_{t_1:\ t_2} := \sum_{\tau=t_1}^{t_2} y_\tau$。对于正整数 N，记 $[N] := \{1, \ldots, N\}$。我们还令 $\Pi_{\mathcal{K}}$: $\mathbb{R}^d \mapsto \mathcal{K}$ 表示到 $\mathcal{K}$ 的正交投影，即 $\Pi_{\mathcal{K}}(x) = \arg\min_{y\in\mathcal{K}} \|y-x\|^2,\quad \forall x \in \mathbb{R}^d$。我们还记 $\tilde{\sigma}^2 := 32D^2\sigma_L^2 + 2\sigma^2$。

3 动量机制

在这里，我们介绍与动量概念相关的两种机制背景。有趣的是，这些方法都与对学习算法不同组成部分进行平均有关。我们在第 4 节提出的方法结合了上述两种机制。

3.1 机制 I：Anytime-GD

第一种机制与对含噪梯度预言机的**查询点**进行平均有关。在标准随机梯度下降中，我们在所计算的迭代点处查询梯度；而在 Anytime-SGD（Cutkosky，2019年）中，我们在所计算迭代点的加权平均处查询梯度。

更具体地说，我们在公式（5）和（6）中描述的 Anytime-SGD 算法（Cutkosky，2019年；Kavis 等人，2019年）采用学习率 $\eta > 0$ 和一系列非负权重$\{\alpha_t\}_t$。该算法维护两个序列 $\{w_t\}_t$，$\{x_t\}_t$。初始化时 $x_1 = w_1$，并且，

$$w_{t+1} = w_t - \eta\alpha_t g_t\ , \forall t \in [T]\ , \text{where } g_t = \nabla f(x_t; z_t)\ , \tag{5}$$

其中 $z_t \sim \mathcal{D}$。随后，Anytime-SGD 更新，

$$x_{t+1} = \frac{\alpha_{1:t}}{\alpha_{1:t+1}} x_t + \frac{\alpha_{t+1}}{\alpha_{1:t+1}} w_{t+1}\ . \tag{6}$$

上述内容意味着，x_t是 w_t的加权平均，即 $x_{t+1} = \frac{1}{\alpha_{1:t+1}} \sum_{\tau=1}^{t+1} \alpha_\tau w_\tau$。因此，在每个迭代点处，都会在 x_t 处查询梯度 g_t，而 x_t 是过去迭代点的加权平均；随后，w_{t+1} 的更新方式与梯度下降类似，梯度 g_t的权重为 α_t。

有趣的是，Wang 等人 (2021年)（见第 4.5.1节），一种与 Anytime-GD 非常相似的算法可以与经典的重球法（Polyak，1964年）联系起来，而后者在其迭代点中融入了动量。 Cutkosky (2019年) 已证明，对于凸损失函数（包括光滑和非光滑情形），Anytime-SGD 获得了与随机梯度下降相同的收敛速率。而且，人们发现这一技术对于设计通用加速方法极其有用（Cutkosky，2019年； Kavis 等人，2019年）。

下一个定理对于分析 Anytime-SGD 至关重要，而且实际上适用范围更广，

定理 3.1（改述自 Cutkosky （2019年）中的定理 1）。 设 f： $\mathcal{K} \mapsto \mathbb{R}$ 是一个凸函数，其最小值 $w^* \in \arg\min_{w\in\mathcal{K}} f\ (w)$。另设 $\{\alpha_t \geq 0\}_t$，以及 $\{w_t \in \mathcal{K}\}_t$，$\{x_t \in \mathcal{K}\}_t$，满足 $\{x_t\}_t$ 是 $\{\alpha_t\}_t$ 的加权平均 $\{w_t\}_t$，即满足 $x_1 = w_1$，并且对于任意 $t \geq 1$，$x_{t+1} = \frac{1}{\alpha_{1:t+1}} \sum_{\tau=1}^{t+1} \alpha_\tau w_\tau$. 则对于任意 $t \geq 1$，以下结论成立：

$$\alpha_{1:t}\left(f(x_t) - f(w^*)\right) \leq \sum_{\tau=1}^{t} \alpha_\tau \nabla f(x_\tau)\cdot(w_\tau - w^*)\ .$$

上述定理对任意序列 $\{w_t \in \mathcal{K}\}_t$ 均成立，作为一个特例，它也适用于 Anytime-梯度下降算法。因此，上述定理将计算序列 $\{w_t \in \mathcal{K}\}_t$、$\{x_t \in \mathcal{K}\}_t$的给定算法的超额损失与其加权遗憾 $\mathcal{R}_t := \sum_{\tau=1}^{t} \alpha_\tau \nabla f(x_\tau)\cdot(w_\tau - w^*)$。联系起来。

3.2 机制 II：递归校正平均

第二种机制与对训练过程中计算得到的**梯度估计**进行平均有关，这是实际应用中一种常见且至关重要的技术。虽然直接取平均可能会给梯度估计带来偏差，Cutkosky &Orabona（2019年提出加入一种名为STORM （随机递归动量）的偏差校正机制。研究表明，这种机制能够显著降低方差。

具体而言，STORM 维护一个估计量 d_t，它是过去随机梯度的校正后的加权平均，然后执行一个随机梯度下降式更新步骤，

$$w_{t+1} = w_t - \eta d_t \,. \tag{7}$$

校正动量估计的更新方式如下，

$$d_t = \nabla f(w_t; z_t) + (1-\beta_t)(d_{t-1} - \nabla f(w_{t-1}, z_t)) \,, \tag{8}$$

对于某些 $\beta_t \in [0,1]$。上述结论意味着 $\mathbf{E}[d_t] = \mathbf{E}\nabla f(w_t)$。然而，一般而言 $\mathbf{E}[d_t|w_t] \neq \nabla f(w_t)$，因此 d_t 是有条件偏差的（与标准 SGD 估计不同）。此外，在上述表达式的两项中选择相同的样本 z_t，对于方差缩减效果至关重要。

4 我们的方法：双动量机制

我们的方法是将上述两种动量机制结合起来。因此，我们的算法命名为 μ^2-随机梯度下降（动量^2-随机梯度下降），并在算法 1 中对其进行描述。直观来说，这些动量（平均）技术都能稳定算法，而将它们结合起来，得到的方法几乎与离线梯度下降一样稳定。请注意，只有正确组合这些技术才能取得我们的结果；单独采用其中一种而不结合另一种，无法达到这一效果。

我们首先描述算法 1，然后在定理 4.1中给出我们的主要结果，表明我们方法的梯度估计误差 会逐渐减小。这表明我们可以将梯度估计的范数用作停止准则，这在梯度下降中是一种常见做法（Beck，2014）。另一个优点在定理 4.2中得到体现，即 μ^2-随机梯度下降在非常宽泛的学习率范围内都能获得与标准 SGD 相同的收敛速率（与 SGD 相比）。

接下来，我们将详细介绍算法1中的组成部分：

更新规则：请注意，为了保证一般性，我们允许算法1的方程（10）采用广泛的更新规则。对更新规则唯一的限制是，其迭代点 $\{w_t\}_t$ 始终属于 $\mathcal{K}$。稍后，我们将具体分析自然的 SGD 式更新规则，

$$w_{t+1} = \Pi_{\mathcal{K}}(w_t - \eta\alpha_t d_t) \,. \tag{9}$$

动量计算：从算法1中的方程（12）可以看出，动量 d_t 的更新方式与方程（8）中的STORM更新类似，但有两个主要区别：**（i）**首先，我们将重要性权重 $\{\alpha_t\}_t$ 纳入STORM ，并递归地更新加权动量 $\alpha_t d_t$。更重要的是，**（ii）**我们在平均值 x_t处查询含噪梯度，而不是在迭代点本身处查询；这些平均值（方程（11））按照 Anytime-SGD 的思路计算。这一点可以从 g_{t+1} 和 $\tilde{g}_t$ 的计算中看出：它们在平均值而非迭代点处查询梯度。因此，正如我们所承诺的，我们的算法结合了两种不同的动量机制。

接下来，我们给出主要结果，表明算法 1得到的估计误差极小。式（10）中的更新规则唯一的限制是，其迭代点 $\{w_t\}_t$ 始终属于 $\mathcal{K}$。

定理 4.1。设 $f\colon \mathcal{K} \mapsto \mathbb{R}$，并假设 $\mathcal{K}$ 是凸的，直径为 D，且公式 (2)、(3)、(4) 中的假设成立。然后调用算法 *1*，取 $\{\alpha_t = t+1\}_t$ 和 $\{\beta_t = 1/\alpha_t\}$，可确保：

$$\mathbf{E}\|\epsilon_t\|^2 := \mathbf{E}\|d_t - \nabla f(x_t)\|^2 \leq \tilde{\sigma}^2/t \,,$$

here $\epsilon_t := d_t - \nabla f(x_t)$，以及 $\tilde{\sigma}^2 := 32D^2\sigma_L^2 + 2\sigma^2$。 $\star$ 在附录 *D.1* 我们提供高概率界。

因此，根据上述定理，随着我们的推进， d_t 相对于精确梯度 $\nabla f(x_t)$会缩小 。相反，在标准 SGD（以及 Anytime-SGD）中，期望平方误差是固定的，即$\mathbf{E}\|g_t - \nabla f(w_t)\|^2 \leq O(\sigma^2)$。

基于定理 4.1，我们可以分析算法1 ，其中采用了公式（9）中给出的特定 SGD 型更新规则。

算法 1 μ^2-随机梯度下降

#迭代次数 T，初始化 x_1、$\eta > 0$，权重 $\{\alpha_t\}_t$，校正动量权重 $\{\beta_t\}_t$
初始化： 设置 $w_1 = x_1$，抽取 $z_1 \sim \mathcal{D}$ 并设置 $d_1 = \nabla f(x_1, z_1)$
针对 $t = 1, \ldots, T$ **do**
迭代更新：

$$\text{Use an update rule to compute } w_{t+1} \in \mathcal{K} \tag{10}$$

查询更新（Anytime-SGD style）：

$$x_{t+1} = \frac{\alpha_{1:t}}{\alpha_{1:t+1}} x_t + \frac{\alpha_{t+1}}{\alpha_{1:t+1}} w_{t+1}\,, \tag{11}$$

Up校正动量（STORM style）日期：
抽取 $z_{t+1} \sim \mathcal{D}$，计算 $g_{t+1} := \nabla f(x_{t+1}; z_{t+1})$ 和 $\tilde{g}_t := \nabla f(x_t; z_{t+1})$，并进行更新，

$$d_{t+1} = g_{t+1} + (1 - \beta_{t+1})(d_t - \tilde{g}_t) \tag{12}$$

结束循环
输出： x_T

定理 4.2 (μ^2-随机梯度下降保证)。设 $f\colon \mathbb{R}^d \mapsto \mathbb{R}$ 为凸函数，并假设 $w^* \in \arg\min_{w\in\mathcal{K}} f(w)$ 也是其在 $\mathbb{R}^d$ 中的全局最小值。此外，作出与定理 *4.1*相同的假设。然后，调用算法 *1*，其中 $\{\alpha_t = t+1\}_t$ 且 $\{\beta_t = 1/\alpha_t\}_t$，并在算法 (9) 的式 $\eta \le 1/8LT$ 中采用学习率(10) 所示的 *SGD* 型更新规则，可保证*1*：

$$\mathbf{E}(f(x_T) - f(w^*)) = \mathbf{E}\Delta_T \le O\left(\frac{D^2}{\eta T^2} + 2\eta\tilde{\sigma}^2 + \frac{4D\tilde{\sigma}}{\sqrt{T}}\right)\,,$$

其中 $\Delta_t := f(x_t) - f(w^*)$，且 $\tilde{\sigma}^2 := 32D^2\sigma_L^2 + 2\sigma^2$。

的稳定性-随机梯度下降。上述引理表明，μ^2-随机梯度下降在离线（无噪声）和含噪情况下，η^{Offline} 使用相同的固定学习率 $\frac{1}{8LT}$，$\tilde{\sigma}$ 即可获得随机梯度下降的最优收敛速率，而该学习率不依赖于噪声 。这与随机梯度下降形成对比，后者要么需要将离线学习率缩小 $\sigma\sqrt{T}$；要么使用复杂的自适应学习率（Duchi 等人，2011年；Levy 等人，2018）。

此外，令 $\eta^{\text{Noisy}} = 1/(8LT + \tilde{\sigma}T^{3/2}/D)$，则可以看出，在含噪情况下，我们的方法能够在极其宽泛的 $[\eta^{\text{Noisy}}, \eta^{\text{Offline}}]$ 范围内采用学习率；同时仍可获得相同的最优随机梯度下降收敛速率。事实上，注意到 $\eta^{\text{Offline}}/\eta^{\text{Noisy}} \approx (\tilde{\sigma}/L)T^{1/2}$。

注： 请注意，在定理 4.2 中，我们假设 $f(\cdot)$ 的全局最小值属于 $\mathcal{K}$。在下一节中，我们将介绍 μ^2 − ExtraSGD——一种加速版的 μ^2-SGD，它不需要这一假设，并能实现更快的收敛速率和更好的稳定性。

4.1 扩展

误差一致较小。我们已经证明，每次迭代只使用一个样本时，我们的方法能够得到梯度估计 d_t，其平方误差为 $O(1/t)$且逐渐缩小。不过，如果我们希望采用提前停止策略，那么理想情况下，所有迭代点上的误差都应一致保持较小，即为 $O(1/T)$。我们在附录中证明了这是可行的，代价只是总体样本复杂度按对数增长。其思路是在我们的方法中引入逐渐减小的批大小：在第 t 轮，我们建议使用批大小 $b_t \propto T/t$。因此，经过 T 轮后，我们总共使用$\sum_{t=1}^{T} b_t = O(T\log\ T)$个样本。这一改动使所有迭代点上的平方误差都能一致保持较小，即为 $O(1/T)$。

一般算子的精确估计。我们已经证明，我们的方法能够对查询点的梯度生成$O(1/t)$估计 $\{x_t\}_{t\in[T]}$。类似地，这一方法可以推广到为其他算子生成 $O(1/t)$ 估计。例如，如果我们希望估计海森矩阵$\nabla^2 f(x_t)$，并假设海森矩阵是利普希茨连续的，且具有与式（2）、（3）、（4）类似的有界方差，那么我们便可以按照式（12）的思路保持良好的海森矩阵估计

无权重版本的 μ^2-随机梯度下降。自然会产生这样的问题：我们是否可以采用更为标准的

均匀权重 $\{\alpha_t = 1\}_{t\in[T]}$，用于 μ^2-随机梯度下降。沿用与我们对定理4.2的证明非常相似的思路，我们证明了确实可以做到这一点，同时使用 $\eta \propto 1/L\log T$，并得到类似的界，但会在 T中产生对数因子。

4.2 两种机制的必要性

这里我们讨论将 Anytime机制与 STORM机制相结合对于使 μ^2-SGD 获得误差收缩性质（见定理4.1）的重要性，这是我们其他结果的关键。具体而言，方程 (5)(6) 中出现的 Anytime机制（不含STORM）采用了如下形式的梯度估计 $g_t := \nabla f(x_t; z_t)$，因此显然，这些梯度估计的误差 $\|\epsilon_t^{\text{Anytime}}\|^2 := \|g_t - \nabla f(x_t)\|^2$ 是 $O(1)$，这与标准SGD类似。相反，方程 (7)(8) 中出现的 STORM 机制（不含Anytime）维持估计d_t。STORM 更新规则实现了方差缩减，该方差缩减取决于连续查询点之间的距离，即 $\|w_t - w_{t-1}\|^2$，而后者又将取决于学习率（正如原始 STORM 论文中所述）。这使方差缩减机制与学习率相互耦合，因此无法实现对学习率的稳健性（例如：在标准 STORM 方法中使用固定学习率将无法收敛）。因此，将这些技术结合起来对于获得定理 4.1 中所证实的误差收缩性质至关重要；该性质独立于学习率，因此能够获得稳定性，正如我们在定理 4.2 和 5.2 中所证实的那样。

4.3 证明 概要 定理。 4.1

证明。 首先注意， x_t 始终属于 $\mathcal{K}$，因为它们是 $\{w_t \in \mathcal{K}\}_t$ 的加权平均。我们的第一步是界定连续查询之间的差异。 x_t 的定义意味着：$\alpha_{1:t-1}(x_t - x_{t-1}) = \alpha_t(w_t - x_t)$， 从而得到，

$$\|x_t - x_{t-1}\|^2 = (\alpha_t/\alpha_{1:t-1})^2 \|w_t - x_t\|^2 \le (16/\alpha_{t-1}^2)D^2 \,. \tag{13}$$

其中我们使用了 $\alpha_t = t+1$，这意味着对于任意 $t \ge 2$，都有 $\alpha_t/\alpha_{1:t-1} \le 4/\alpha_{t-1}$；我们还使用了 $\|w_t - x_t\| \le D$。记号：在继续证明之前，我们需要一些记号。我们将记 $\bar{g}_t := \nabla f(x_t)$，并回顾算法 1中的以下记号： $g_t := \nabla f(x_t, z_t)$； $\tilde{g}_{t-1} := \nabla f(x_{t-1}, z_t)$。我们还将记作， $\epsilon_t := d_t - \bar{g}_t$.

Rec由方程（12），并结合上述 ϵ_t 的定义，可以推导出以下结果 ，

$$\alpha_t\epsilon_t = \beta_t\alpha_t(g_t - \bar{g}_t) + (1-\beta_t)(\alpha_t Z_t + \frac{\alpha_t}{\alpha_{t-1}}\alpha_{t-1}\epsilon_{t-1}) \,,$$

其中，我们记 $Z_t := (g_t - \bar{g}_t) - (\tilde{g}_{t-1} - \bar{g}_{t-1})$。现在，利用 $\alpha_t = t+1$ 和 $\beta_t = 1/(t+1)$，可以证明 $\alpha_t\beta_t = 1$，且 $\alpha_t(1-\beta_t) = \alpha_t - 1 := \alpha_{t-1}$。此外，$(1-\beta_t)\frac{\alpha_t}{\alpha_{t-1}} = \frac{\alpha_t - 1}{\alpha_{t-1}} = 1$。代入上述结果可得，

$$\alpha_t\epsilon_t = \alpha_{t-1}Z_t + \alpha_{t-1}\epsilon_{t-1} + (g_t - \bar{g}_t) = M_t + \alpha_{t-1}\epsilon_{t-1} \,. \tag{14}$$

其中，对于任意 $t > 1$，我们记 $M_t := \alpha_{t-1}Z_t + (g_t - \bar{g}_t)$，以及 $M_1 = g_1 - \bar{g}_1$。展开上述方程，可得任意 $t \in [T]$ 的显式表达式： $\alpha_t\epsilon_t = \sum_{\tau=1}^t M_\tau$.
注意到，序列 $\{M_t\}_t$ 是关于由截至时间 t 的样本历史所诱导的自然滤过 $\{\mathcal{F}_t\}_t$ 的鞅差序列；这意味着，

$$\mathbf{E}\|\alpha_t\epsilon_t\|^2 = \left\|\sum_{\tau=1}^t M_\tau\right\|^2 = \sum_{\tau=1}^t \mathbf{E}\|M_\tau\|^2 \le 2\sum_{\tau=1}^t \alpha_{t-1}^2\mathbf{E}\|Z_t\|^2 + 2\sum_{\tau=1}^t \mathbf{E}\|g_t - \bar{g}_t\|^2 \,. \tag{15}$$

利用方程（ 4 $\sigma_L^2\|x_t - x_{t-1}\|^2 \le 16\sigma_L^2 D^2/\alpha_{t-1}^2$ 13 $\|Z_t\|^2 = \|\ g_t - \bar{g}_t\ -\ \tilde{g}_{t-1} - \bar{g}_{t-1}\ \|^2 \le$ ），结合方程（ ）可以得到界，**E**E()()，我们还可以对其进行界定**E**。将其代回方程（15），并求和即可证明该定理。 $\|g_t - \bar{g}_t\|^2 \le \sigma^2$ □

5 加速版本： μ^2 − EXTRA随机梯度下降

在此，我们提出一种加速版本，它采用双重动量机制，与我们用于 μ^2-随机梯度下降的方法相同。我们的方法依赖于一种名为超梯度（Korpelevich,

1976；Nemirovski，2004；Juditsky 等人，2011年）的方法。该技术已与 Anytime-SGD 机制相结合，见 Cutkosky （2019年）；Kavis 等人 （2019年），并表明它能够带来加速。在此，我们进一步融入一种适当的 STORM 机制，由此得到一种新方法，命名为 μ^2 -ExtraSGD。我们的主要结果见定理 5.2。

在技术层面，我们的加速版本需要以恰当的方式谨慎而精细地融合多种技术，从而得出简洁而细致的分析。

乐观 OGD、超梯度法和 UnixGrad：超梯度技术与一种名为乐观在线梯度下降（乐观 OGD）的算法模板相关（Rakhlin 与 Sridharan，2013）。在该算法中，我们以在线方式接收一系列（可能任意的）损失向量 $\{d_t \in \mathbb{R}^d\}_{t\in[T]}$。我们的目标是计算一系列迭代点（或决策点） $\{w_t \in \mathcal{K}\}_t$，其中 $\mathcal{K}$ 是给定的凸集。请注意，我们只能依据过去的信息来选择 w_t $\{d_1, \ldots, d_{t-1}\}$。我们的目标是确保对任意 $w \in \mathcal{K}$都具有较低的加权遗憾，其中后者定义如下：

$$\mathcal{R}_T(w) := \sum_{t=1}^{T} \alpha_t d_t \cdot (w_t - w^*) \,,$$

以及 $\{\alpha_t > 0\}$ 是一个预定义权重序列。在乐观设置下，我们假设可以访问一系列“提示向量” $\{\hat{} d_t \in \mathbb{R}^d\}_t$，并且在选取 w_t 之前，我们还可以$\hat{}$ $\hat{}$访问 $\{d_1, \ldots, d_t\}$。Rakhlin 和 Sridharan （2013）表明，如果这些提示是有用的，即在$\hat{}$满足 $d_t \approx d_t$的意义下，那么通过适当地纳入这些提示，可以降低遗憾值。因此，在**Optimistic-OGD** 中，我们维护两个序列：决策点序列 $\{w_t\}_t$ 和辅助序列 $\{y_t\}$，更新方式如下：

乐观梯度下降：

$$w_t = \arg\min_{w\in\mathcal{K}} \alpha_t \hat{d}_t \cdot w + \frac{1}{2\eta}\|w - y_{t-1}\|^2 \;\&\; y_t = \text{求最小值}_{y\in\mathcal{K}} \alpha_t d_t \cdot y + \frac{1}{2\eta}\|y - y_{t-1}\|^2 \qquad (16)$$

Rakhlin与 Sridharan 证明了 (2013年)；Kavis 等人(2019年)表明，上述算法对任意 $w \in \mathcal{K}$，遗憾界如下

定理 5.1 （例如参见 Kavis 等人 （2019年）中的定理 1 的证明）。 设 $\eta > 0$， $\{\alpha_t \geq 0\}_{t\in[T]}$ 和 $\mathcal{K}$ 是一个具有有界直径 D 的凸集合。那么，对于任意 $w \in \mathcal{K}$，乐观 *OGD* 都能保证以下结果

$$\sum_{t=1}^{T} \alpha_t d_t \cdot (w_t - w) \leq \frac{4D^2}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \alpha_t^2 \|d_t - \hat{d}_t\|^2 - \frac{1}{2\eta}\sum_{t=1}^{T} \|w_t - y_{t-1}\|^2 \,.$$

该**外梯度算法** (Nemirovski，2004）旨在最小化凸函数 f： $\mathcal{K} \mapsto \mathbb{R}$。为此，它采用乐观 OGD 模板，并作出以下损失和提示向量选择：$\hat{} d_t = \nabla f(y_{t-1})$，以及 d_t： $= \nabla f(w_t)$。

该**UnixGrad 算法**（Kavis 等人，2019年）可以看作 Extragradient 的 Anytime 版本，其中我们同样旨在最小化凸函数 f： $\mathcal{K} \mapsto \mathbb{R}$按照 Anytime-GD 的思路，UnixGrad 维护两个基于加权平均 $\{x_t, \hat{} x_t\}_t$ 的序列 $\{w_t, y_t\}_t$，

$$\hat{x}_t = \frac{\alpha_{1:t-1}}{\alpha_{1:t}} x_{t-1} + \frac{\alpha_t}{\alpha_{1:t}} y_{t-1} \,,\; x_t = \frac{\alpha_{1:t-1}}{\alpha_{1:t}} x_{t-1} + \frac{\alpha_t}{\alpha_{1:t}} w_t \qquad (17)$$

$\hat{}$然后，基于上述平均值，Unixgrad 将损失和提示向量设置如下： $d_t = \nabla f(\hat{} x_t)$，以及 d_t： $= \nabla f(x_t)$。注意，上述平均规则意味着 x_t's是 w_t's 的加权平均，即 $x_t = \frac{1}{\alpha_{1:t}} \sum_{\tau=1}^{t} \alpha_\tau w_\tau$。后者使得能够利用定理 3.1 的任意时刻保证。

The上述方法也存在随机版本，在这些版本中，我们只能查询含噪梯度 s.

我们的方法。 我们建议将乐观 OGD 模板与STORM 机制结合起来，构建在 Unixgrad 所采用的 Anytime 机制之上。具体而言，我们保持与公式 (17) 中相同的加权平均，并将动量估计定义如下：在第 t轮次抽取一个新样本 $z_t \sim \mathcal{D}$，并计算

$$\tilde{g}_{t-1} = \nabla f(x_{t-1}; z_t) \;,\; \hat{g}_t = \nabla f(\hat{x}_t; z_t) \;,\; g_t = \nabla f(x_t; z_t) \,. \qquad (18)$$

基于上述内容，按如下方式计算（校正动量）损失和提示向量，

$$\alpha_t \hat{d}_t = \alpha_t \hat{g}_t + (1-\beta_t)\alpha_t(d_{t-1} - \tilde{g}_{t-1}) \;\&\; \alpha_t d_t \;\; = \alpha_t g_t + (1-\beta_t)\alpha_t(d_{t-1} - \tilde{g}_{t-1}) \qquad (19)$$

算法 2 μ^2 –ExtraSGD

#迭代次数 T，初始化 $y_0, \eta > 0$，权重 $\{\alpha_t\}_t$，校正动量权重 $\{\beta_t\}_t$ˆ 初始化：设置 $x_0 = 0$，以及 ˆ$x_1 = y_0$，抽取 $z_1 \sim \mathcal{D}$并设置 $d_0 = \tilde{g}_0 = d_1 = \nabla f(\hat{x}_1, z_1)$ 对于 $t = 1, \ldots, T$ 执行 计算：$w_t = \arg\min_{w \in \mathcal{K}} \alpha_t \hat{d}_t \cdot w + \frac{1}{2\eta}\|w - y_{t-1}\|^2$，& 更新：$x_t = \frac{\alpha_{1:t-1}}{\alpha_{1:t}} x_{t-1} + \frac{\alpha_t}{\alpha_{1:t}} w_t$计算：$g_t = \nabla f(x_t; z_t)$ & 更新：$d_t = g_t + (1-\beta_t)(d_{t-1} - \tilde{g}_{t-1})$ 计算：$y_t = \arg\min_{y \in \mathcal{K}} \alpha_t d_t \cdot y + \frac{1}{2\eta}\|y - y_{t-1}\|^2$ & 更新：ˆ$x_{t+1} = \frac{\alpha_{1:t}}{\alpha_{1:t+1}} x_t + \frac{\alpha_{t+1}}{\alpha_{1:t+1}} y_t$ 抽取一个新样本 $z_{t+1} \sim \mathcal{D}$并计算，$\tilde{g}_t = \nabla f(x_t; z_{t+1})$，ˆ$g_{t+1} = \nabla f(\hat{x}_{t+1}, z_{t+1})$ ˆ更新：$d_{t+1} = \hat{g}_{t+1} + (1-\beta_{t+1})(d_t - \tilde{g}_t)$ 循环结束 输出：x_T

然后，根据式（16）中的乐观 OGD 进行更新。注意，$\alpha_t d_t$ 的更新规则与算法1中使用的更新完全相同；此外，正如我们已经指出的，式（x_t 17）中的序列是 $\{\alpha_t\}_t$ 序列的加权平均 $\{w_t\}_t$。因此，若选择 $\alpha_t = t + 1$和 $\beta_t = 1/\alpha_t$，则可应用定理4.1，从而得到：

$$\mathbf{E}\|\epsilon_t\|^2 := \mathbf{E}\|d_t - \nabla f(x_t)\|^2 \le \tilde{\sigma}^2/t\,, \quad \text{where } \tilde{\sigma}^2 := 32D^2\sigma_L^2 + 2\sigma^2\,.$$

算法2 中的伪代码展示了我们的 μ^2–ExtraSGD 算法及其适当的计算顺序。可以看出，该算法结合了乐观 OGD 更新（式（16））、适当的ˆ任意时刻平均（式（17）），以及针对STORM 的更新 d_t，d_t（式（19））。

现在我们可以陈述 μ^2 –ExtraSGD 的保证了，

定理 5.2（μ^2 –ExtraSGD）。设 f：$\mathcal{K} \mapsto \mathbb{R}$ 是一个凸函数，$\mathcal{K}$ 是一个直径为 D 的凸集，并记 $w^* \in \arg\min_{w \in \mathcal{K}} f(w)$。那么，在满足方程 (2)、(3)、(4) 中假设的情况下，采用 $\{\alpha_t = t+1\}_t$ 和 $\{\beta_t = 1/\alpha_t\}_t$ 调用算法2，并且 $\eta \le 1/2L$，可保证：

$$\mathbf{E}(f(x_T) - f(w^*)) := \mathbf{E}\Delta_T \le O\left(\frac{D^2}{\eta T^2} + \frac{\tilde{\sigma} D}{\sqrt{T}}\right)\,.$$

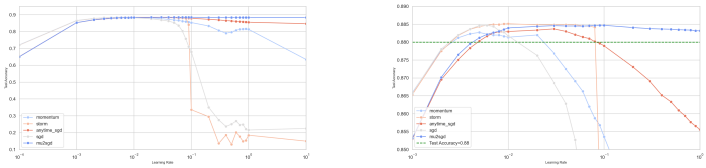
正如算法2中所示，恰当的 μ^2 加速版本需要以正确交织的方式，谨慎而精细地融合上述技术。

μ^2 – ExtraSGD 的稳定性。上述引理表明，对于离线（无噪声）和含噪情形，μ^2 – ExtraSGD 采用相同的固定学习率 $\eta^{\text{Offline}} = 1/2L$，即可达到最优速率。这与现有的加速方法不同：它们要么需要将离线学习率降低–$\sigma\sqrt{T}$ 倍（Xiao，2010年）；要么需要采用复杂的自适应学习率（Cutkosky，2019年；Kavis 等，2019年）。此外，令 $\eta^{\text{Noisy}} := 1/(2L + \tilde{\sigma}T^{3/2}/D)$，则可以看出，在含噪情形下，我们的方法允许采用范围极广的学习率 $[\eta^{\text{Noisy}}, \eta^{\text{Offline}}]$，仍能达到相同的最优收敛速率 $O\left(LD^2/T^2 + \tilde{\sigma}D/\sqrt{T}\right)$。事实上，注意到比率 $\eta^{\text{Offline}}/\eta^{\text{Noisy}} \approx (\tilde{\sigma}/L)T^{3/2}$。此外，与要求 $w^* \in \arg\min_{w \in \mathcal{K}} f(w)$ 同时也是 $f(\cdot)$ 的全局最小值的定理 4.2 相反，定理 5.2 不要求这一假设。

6 实验

我们首先在 μ^2-随机梯度下降算法的凸设置下评估所提出的方法，其中模型权重在每次梯度更新后都被投影到单位球上。评估在 MNIST 数据集（LeCun 等人，2010）上进行，使用逻辑回归模型。我们将采用理论框架所建议参数的方法（$\alpha_t = t$，$\beta_t = \frac{1}{t}$）与多个基线优化器进行比较。其中包括 μ^2-随机梯度下降算法的各个组成部分——STORM 和 Anytime-SGD——所有方法均使用相同的参数设置进行测试。如图1 所示，μ^2-随机梯度下降算法在广泛的学习率范围内始终展现出更优的稳定性。值得注意的是，

尽管 AnyTime-SGD 算法在广泛的学习率范围内保持了很强的稳定性，STORM 在较高学习率下却遇到了困难。通过结合这两种方法，μ^2-随机梯度下降算法实现了更高的稳定性和性能，优于单独使用任一方法的表现。此外，当考虑该设置下更为典型的学习率范围时（见图1b），μ^2-随机梯度下降算法未必总能取得绝对最佳的结果，相较于其他算法亦是如此。尽管如此，它在更广泛的学习率范围内始终保持较高的性能和稳健性，大幅减少了为寻找高性能学习率而进行广泛搜索的需求。此外，通过利用基于理论框架的动量参数，μ^2-随机梯度下降算法进一步消除了超参数调优的需求，而这一过程可能会带来极高的计算成本。



a 在广泛的学习率范围内。　　b 在典型学习率范围内。

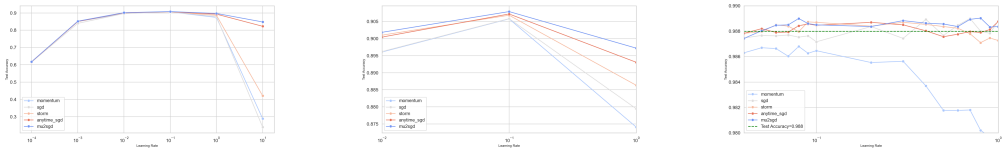
图1：MNIST：凸设置下不同学习率的测试准确率（↑越好）。

深度学习变体。 我们在 非凸 设置下，使用 MNIST 数据集上的 2层卷积网络和 CIFAR-10 数据集上的 ResNet-18（Krizhevsky 等人，2014）。首先，我们将最初定义为$x_t := \frac{\alpha_t w_t + \alpha_{1:t-1} x_{t-1}}{\alpha_{1:t}}$ 的 AnyTime 更新重新表述为数学上等价的基于动量的方法：

$$x_t = \gamma_t w_t + (1 - \gamma_t) x_{t-1}$$

其中 $\gamma_t := \frac{\alpha_t}{\alpha_{1:t}}$。对于非凸模型，在迭代次数中衰减动量参数可能过于激进；因此，我们提出了一种使用固定动量参数的启发式方法，以提高适应性。请注意，通过设置 $\alpha_t = C\alpha_{1:t-1}$，其中 $C > 0$是一个常数，我们推导出$\gamma_t = \frac{C}{C+1}$，使其在所有时间步 $t \geq 1$ 上都保持不变。

我们表明，在非凸设置中使用固定的动量参数（$\gamma_t = 0.1$，$\beta_t = 0.9$）能够确保较高的稳定性和出色的性能**(i)** 在广泛的学习率范围内，以及**(ii)**在不同随机种子下，正如图2 和附录H.3所示。MNIST 和 CIFAR-10 上均观察到一致的结果，详见附录H.3。这些发现凸显了我们方法的稳健性与适应性，使其成为优化非凸模型的可靠选择。



a CIFAR-10：广泛范围。　　b CIFAR-10：典型搜索范围。　　c MNIST：典型搜索范围。

图 2：非凸设置下不同学习率范围内的测试准确率（↑ 越高越好）。

所有实验均使用 PyTorch 框架进行。凸实验在 Apple M2 芯片上运行，而非凸实验则在 NVIDIA A30 GPU 上执行。结果取三个不同随机种子的平均值。更多详情请参阅附录H以及我们的 GitHub 代码仓库。[1]

7 结论

通过精心融合两种近期的动量技术，我们针对 SCO-EOS 场景设计了一种新的误差递减型梯度估计。基于此，我们提出了两种分别采用随机梯度下降和超梯度模板的算法，并证明它们对学习率的选择具有显著的稳定性，从而实现更稳健的训练。未来，进一步探索我们提出的非凸启发式方法在超大规模模型上的适用性将十分有趣，因为这类模型需要更多计算资源。此外，探究我们能否为非凸问题设计一种具有与我们的方法类似理论性质的算法，也将十分有趣。

[1] https://github.com/dahan198/mu2sgd

致谢

本研究部分得到了以色列高等教育委员会规划与预算委员会、Technion 人工智能中心（Tech.AI）以及以色列科学基金会（资助编号：3109/24）的支持。

参考文献

Kimon Antonakopoulos、Dong Quan Vu、Volkan Cevher、Kfir Levy 和 PanayotisMertikopoulos。Undergrad：一种具有近乎无维数依赖的收敛速率保证的通用黑箱优化方法。载国际机器学习会议，第 772–795 页。机器学习研究论文集，2022年。

Amir Beck. 非线性优化导论：理论、算法与 *MATLAB* 应用。工业与应用数学学会，2014 年。

Ashok Cutkosky。随时在线到批量、乐观算法与加速。载于国际机器学习会议，第 1446–1454 页。机器学习研究论文集，2019年。

Ashok Cutkosky。EC525：机器学习优化课程讲义。*EC525*：机器学习优化，2022年a。

Ashok Cutkosky。EC525：机器学习优化课程讲义。*EC525*：机器学习优化，2022年b。

Ashok Cutkosky 和 Francesco Orabona。基于动量的非凸随机梯度下降方差缩减。神经信息处理系统进展，32，2019年。

Aaron Defazio 和 Konstantin Mishchenko。通过 D 自适应实现无学习率学习。收录于 Andreas Krause、Emma Brunskill、Kyunghyun Cho、Barbara Engelhardt、Sivan Sabato 和 Jonathan Scarlett（编），第 *40* 届国际机器学习会议论文集，第 202 卷，机器学习研究论文集，第 7449–7479 页。机器学习研究论文集（PMLR），2023年7月23–29日。网址：https://proceedings.mlr.press/v202/defazio23a.html。

John Duchi、Elad Hazan 和 Yoram Singer。《用于在线学习和随机优化的自适应次梯度方法》。《机器学习研究期刊》，12(7)，2011年。

Rong Ge、Furong Huang、Chi Jin 和 Yang Yuan。从鞍点逃逸——用于张量分解的在线随机梯度。载于第 *28* 届学习理论大会论文集，第 797–842 页，2015 年。

Niv Giladi、Mor Shpigel Nacson、Elad Hoffer 和 Daniel Soudry。处于稳定性的边缘：如何调整超参数，以在神经网络的异步训练中保留极小值选择？载于国际学习表征会议，2019年。

Elad Hazan等人。《在线凸优化导论》。优化领域的基础与趋势®，2(3-4):157–325，2016年。

Chonghai Hu、WeikePan和James T Kwok。随机优化和在线学习的加速梯度方法。载于神经信息处理系统进展，第 781–789 页，2009年。

Maor Ivgi、Oliver Hinder 和 Yair Carmon。DOG 是随机梯度下降的最佳搭档：无参数动态步长计划。发表于第 *40* 届国际机器学习会议（*ICML*）论文集，2023年。

Andrew Jacobsen and Ashok Cutkosky. 无参数镜像下降。*arXiv*预印本 *arXiv:2203.00444*，2022年。Anatoli Juditsky、ArkadiNemirovski和Claire Tauvel。使用随机镜像近端算法求解变分不等式。随机系统，1(1)：17–58，2011年。

Ali Kavis、Kfir Y Levy、Francis Bach 和 Volkan Cevher。Unixgrad：一种用于约束优化、具有最优保证的通用自适应算法。载于第*33*届神经信息处理系统国际会议论文集，第6260–6269页，2019年。

Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于 Yoshua Bengio 和 Yann LeCun（编），第3届国际学习表征会议（ICLR 2015），美国加利福尼亚州圣迭戈，2015年5月7日至9日，会议赛道论文集，2015年。网址：http: //arxiv.org/abs/1412.6980。G. M. Korpelevich。用于求解鞍点及其他问题的外梯度法。《Matecon》，12：747–756，1976年。Alex Krizhevsky、Vinod Nair 和 Geoffrey Hinton。CIFAR-10 数据集。在线资源：http://www. cs. toronto. edu/kriz/cifar. html，55(5)，2014年。Guanghui Lan。一种随机复合优化的最优方法。《数学规划》，133(1-2)：365–397，2012年。Yann LeCun、Corinna Cortes、Chris Burges 等。MNIST 手写数字数据库，2010年。Kfir Yehuda Levy、Alp Yurtsever 和 Volkan Cevher。在线自适应方法、普适性与加速。载于 NeurIPS，2018年。Kfir Yehuda Levy、Ali Kavis 和 Volkan Cevher。Storm+：采用递归动量的全自适应随机梯度下降，用于非凸优化。载于 NeurIPS，2021年。Julien Mairal、Francis R. Bach、Jean Ponce 和 Guillermo Sapiro。用于稀疏编码的在线字典学习。载于 Andrea Pohoreckyj Danyluk、Léon Bottou 和 Michael L. Littman（编），第26届年度国际机器学习会议论文集（ICML 2009），加拿大魁北克省蒙特利尔，2009年6月14日至18日，ACM 国际会议论文集系列第382卷，第689–696页。ACM，2009年。doi：10.1145/1553374.1553463。网址：https: //doi.org/10.1145/1553374.1553463。Stanislav Minsker。论伯恩斯坦不等式在自伴算子上的若干扩展。《统计与概率快报》，127：111–119，2017年。Arkadi Nemirovski。用于求解利普希茨连续单调算子变分不等式及光滑凸凹鞍点问题的近端法，收敛速率为 $O(1/t)$。《SIAM 优化杂志》，15(1)：229–251，2004年。Yurii Nesterov。一种求解凸规划问题的方法，收敛速率为 $O(1/k^2)$。载于《苏联数学文献》，第27卷，第372–376页，1983年。Boris T Polyak。若干加速迭代方法收敛的途径。《苏联计算数学与数学物理》，4(5)：1–17，1964年。Sasha Rakhlin 和 Karthik Sridharan。基于可预测序列的优化、学习与博弈。载于《神经信息处理系统进展》，第3066–3074页，2013年。Benjamin Recht、Christopher Ré、Stephen J. Wright 和 Feng Niu。Hogwild：一种用于并行化随机梯度下降的无锁方法。载于 John Shawe-Taylor、Richard S. Zemel、Peter L. Bartlett、Fernando C. N. Pereira 和 Kilian Q. Weinberger（编），《神经信息处理系统进展24：第25届神经信息处理系统年度会议》，2011年。会议于2011年12月12日至14日在西班牙格拉纳达举行，论文集，第693–701页，2011年。网址：
h
t
t
p
s://
p
r
o
c
e
edings.neurips.cc/paper/2011/hash/218a0aefd1d1a4be65601cc6ddc1520e-Abstract.html。Herbert Robbins 和 Sutton Monro。一种随机逼近法。《数学统计年刊》，第400–407页，1951年。Sebastian Ruder。梯度下降优化算法综述。arXiv 预印本 arXiv:1609.04747，2016年。Shai Shalev-Shwartz、Yoram Singer 和 Nathan Srebro。Pegasos：用于支持向量机的原始估计次梯度求解器。载于第24届国际机器学习会议论文集，第807–814页，2007年。

IlyaSutskever、James Martens、George Dahl 和 Geoffrey Hinton。论深度学习中初始化和动量的重要性。载于 国际机器学习会议，第 1139–1147 页。机器学习研究论文集，2013年。

Jun-Kun Wang、Jacob Abernethy 和 Kfir Y Levy。Fenchel 博弈中的无悔动态：算法凸优化的统一框架。*arXiv* 预印本 *arXiv:2111.11309*，2021年。

Max Welling 和 Yee W Teh。通过随机梯度 Langevin 动态进行贝叶斯学习。载于第 *28* 届国际机器学习会议（*ICML-11*）论文集，第 681–688 页。Citeseer，2011年。

Yonghui Wu、Mike Schuster、Zhifeng Chen、Quoc V Le、Mohammad Norouzi、Wolfgang Macherey、Maxim Krikun、Yuan Cao、Qin Gao、Klaus Macherey 等。Google 的神经机器翻译系统：弥合人工翻译与机器翻译之间的差距。*arXiv* 预印本 *arXiv:1609.08144*，2016年。

Lin Xiao. 用于正则化随机学习和在线优化的双重平均方法。机器学习研究期刊,11(Oct):2543–2596, 2010。

关于学习率选择的稳定性

这里我们正式解释，当我们讨论算法 $\mathcal{A}$ 相对于学习率选择 η的稳定性时，具体指的是什么。

给定一个随机凸优化算法 $\mathcal{A}$，我们通常可以将其泛化界表示如下：对于给定的 $0 \leq \eta \leq \bar{\eta}$ 选择，算法 $\mathcal{A}$ 可保证：

$$\textbf{Excess-Loss} \leq \mathcal{R}_{\mathcal{A}}(\eta) \,,$$

其中， $\mathcal{R}$： $\mathbb{R}_+ \mapsto \mathbb{R}_+$ 是学习率 η的函数，即收敛速率。该描述适用于我们论文中提到的所有方法。

在这种情况下，我们可以如下定义最优学习率：

$$\eta^* := \min_{\eta \in [0,\bar{\eta}]} \mathcal{R}_{\mathcal{A}}(\eta) \,.$$

我们将 $\mathcal{A}$ 的阶数最优学习率范围定义为：

$$\text{Range}_{\mathcal{A}} := \{\eta \in [0,\bar{\eta}] : \text{R}_{\mathcal{A}}(\eta) \leq 2\text{R}_{\mathcal{A}}(\eta^*)\} \,.$$

换句话说，该集合由所有能够使收敛速率达到最优值（相差一个乘法因子2 [2]）的学习率组成。

通常，集合Range$_{\mathcal{A}}$ 是 $\mathbb{R}_+$中的一条线段，因此我们可以写成 $\text{Range}_{\mathcal{A}} = [\eta^{\min}, \eta^{\max}]$。此外，我们还可以记作，

$$\textbf{ratio}_{\mathcal{A}} = \eta^{\max}/\eta^{\min} \,.$$

因此，比率越高， $\mathcal{A}$相对于学习率 η选择的稳定性就越好。接下来，我们针对论文中提到的方法，对 SCO-EOS 设定下的稳定性进行比较，该设定见第 2 节。并证实 μ^2-随机梯度下降以及 μ^2 −Extra**SGD** 的稳定性优于标准 SGD 和加速随机梯度下降。

标准 SGD 的稳定性：对于标准 SGD，众所周知，当选择 $\eta \in (0, \frac{1}{2L}]$ 时，其收敛速率为，

$$\mathcal{R}_{\text{SGD}}(\eta) := \frac{D^2}{\eta T} + \eta\sigma^2$$

因此，在典型情况下，即 $\frac{D}{\sigma\sqrt{T}} \leqq \frac{1}{2L}$（也就是说噪声不可忽略时），我们有 $\eta^* = \frac{D}{\sigma\sqrt{T}}$，并且 $\mathcal{R}_{\text{SGD}}(\eta^*) = \frac{2D\sigma}{\sqrt{T}}$。因此可以验证，

$$\textbf{ratio}_{\text{SGD}} \leq 15 \,.$$

反之，**(ii)** 在非典型情况下，我们有 $\frac{D}{\sigma\sqrt{T}} > \frac{1}{2L}$ $\eta^* = \frac{1}{2L}$。在这种情况下，我们可以再次验证，

$$\textbf{ratio}_{\text{SGD}} \leq 15 \,.$$

这证实，对于标准 SGD，我们有稳定性比率为**比率**$_{\text{SGD}} \approx O(1)$。

加速随机梯度下降的稳定性：加速随机梯度下降有若干变体。对于这类算法，当学习率取为 $\eta \in (0, \frac{1}{2L}]$时，这类方法具有收敛速率（例如参见定理2，见 Lan （2012）），

$$\mathcal{R}_{\text{Accel-SGD}}(\eta) := \frac{D^2}{\eta T^2} + \eta\sigma^2 T$$

因此，类似于我们对标准 SGD的分析，可以验证，对于此类方法，我们有，**比率**$_{\text{Accel-SGD}} \approx O(1)$。

[2]对 2 的选择相当任意，我们同样可以选择任何充分大于 1 的因子 .

稳定性 μ^2-随机梯度下降： 正如我们在定理4.2 中所证明的， μ^2-随机梯度下降方法确保对于 $\eta \in (0, \frac{1}{8LT}]$，其享有如下收敛速率：

$$\mathcal{R}_{\mu^2-\mathrm{SGD}}(\eta) := \frac{D^2}{\eta T^2} + 2\eta\tilde{\sigma}^2 + \frac{4D\tilde{\sigma}}{\sqrt{T}} .$$

因此，在这种情况下，我们有 $\eta^* = \frac{D}{\sqrt{2}\tilde{\sigma}T}$，并且 $\mathcal{R}_{\mu^2-\text{随机梯度下降}}(\eta^*) = \frac{2\sqrt{2}D\tilde{\sigma}}{T} + \frac{4D\tilde{\sigma}}{\sqrt{T}}$。因此可以验证，在这种情况下，我们有 $\eta^{\min} \approx \frac{D}{\tilde{\sigma}T^{3/2}}$ 并且 $\eta^{\max} = \frac{1}{8LT}$，因此，

$$\textbf{ratio}_{\mu^2-\mathrm{SGD}} \approx \frac{\tilde{\sigma}}{LD}\sqrt{T} .$$

稳定性 μ^2 – 额外随机梯度下降： 正如我们在定理中所证明的5.2 我们提出的 μ^2 – 额外随机梯度下降方法保证，对于 $\eta \in (0, \frac{1}{2L}]$，其收敛速率为：

$$\mathcal{R}_{\mu^2\text{ – ExtraSGD}}(\eta) := \frac{D^2}{\eta T^2} + \frac{\tilde{\sigma}D}{\sqrt{T}} .$$

因此，在这种情况下，我们有 $\eta^* = \frac{1}{2L}$，并且有 $\mathcal{R}_{\mu^2\text{ – 额外内容随机梯度下降}}(\eta^*) = \frac{2LD^2}{T^2} + \frac{\tilde{\sigma}D}{\sqrt{T}}$。因此可以验证，在这种情况下，我们有 $\eta^{\min} \approx \frac{D}{\tilde{\sigma}T^{3/2}}$和 $\eta^{\max} = \frac{1}{2L}$，因此，

$$\textbf{ratio}_{\mu^2\text{ – ExtraSGD}} \approx \frac{\tilde{\sigma}}{LD}T^{3/2} .$$

B 解释有界光滑度方差假设

这里我们证明，式（3）意味着对于某个 $\sigma_L^2 \in [0, L]$，式（4）成立。

固定 $x,\ y \in \mathcal{K}$，则式（3）意味着，对于任意 $z \in \textbf{Support}\{\mathcal{D}\}$，都存在 $L_{x,\ y;z} \in [0, L]$，使得：

$$\|\nabla f(x;z) - \nabla f(y;z)\|^2 = L_{x,y;z}^2\|x-y\|^2 .$$

类似地，存在 $L_{x,\ y} \in [0, L]$，使得：

$$\|\nabla f(x) - \nabla f(y)\|^2 = L_{x,y}^2\|x-y\|^2 .$$

显然，在确定性情形下，我们有 $L_{x,\ y;z} = L_{x,\ y}$，$\forall z \in \textbf{Support}\{\mathcal{D}\}$。因此，

$$\begin{aligned}\mathbf{E}\|(\nabla f(x;z) - \nabla f(x)) - (\nabla f(y;z) - \nabla f(y))\|^2 &= \mathbf{E}\|\nabla f(x;z) - \nabla f(y;z)\|^2 - \|\nabla f(x) - \nabla f(y))\|^2 \\ &= \mathbf{E}(L_{x,y;z}^2 - L_{x,y}^2)\cdot\|x-y\|^2 = \sigma_L^2\{x,y\}\cdot\|x-y\|^2 ,\end{aligned}$$

其中我们用到了 $\mathbf{E}(\nabla f(x;\ z) - \nabla f(y;\ z)) = (\nabla f(x) - \nabla f(y))$，并记 $\sigma_L^2\{x,y\} := \mathbf{E}(L_{x,\ y;z}^2 - L_{x,\ y}^2)$。该记号意味着 $\sigma_L^2\{x,y\} \in [0, L^2]$；显然，在确定性情形下，对所有 $x,\ y \in \mathcal{K}$，$\sigma_L^2\{x,y\} = 0$。因此，若记为：

$$\sigma_L^2 := \sup_{x,y\in\mathcal{K}} \sigma_L^2\{x,y\} ,$$

那么 $\sigma_L^2 \in [0, L^2]$ 满足式（4），并且在确定性（无噪声）情形下等于 0。

C 证明定理. 4.1

定理 *4.1* 的证明。 首先注意到， x_t始终属于 $\mathcal{K}$，因为它们是 $\{w_t \in \mathcal{K}\}_t$ 的加权平均。接下来，我们对连续两次查询之间的差值进行界定。根据定义，

$$\alpha_{1:t-1}(x_t - x_{t-1}) = \alpha_t(w_t - x_t) ,$$

这意味着，

$$\|x_t - x_{t-1}\|^2 = (\alpha_t/\alpha_{1:t-1})^2\|w_t - x_t\|^2 \le (16/t^2)D^2 = (16/\alpha_{t-1}^2)D^2 . \tag{20}$$

其中我们用到了 $\alpha_t = t+1$，这意味着对任意 $t \geq 2$，$\alpha_t/\alpha_{1:t-1} \leq 4/t$；我们还用到了 $\|w_t - x_t\| \leq D$，这一结论成立是因为 $w_t, x_t \in \mathcal{K}$；最后我们使用 $\alpha_{t-1} = t$。

记号： 在继续进行之前，我们需要一些记号。我们记 $\bar{g}_t := \nabla f(x_t)$，并回顾算法 1 中的以下记号：$g_t := \nabla f(x_t, z_t)$；$\tilde{g}_{t-1} := \nabla f(x_{t-1}, z_t)$，我们还记 $\bar{g}_t := \nabla f(x_t)$，以及

$$\epsilon_t := d_t - \bar{g}_t \,.$$

现在，回顾公式（12），

$$\alpha_t d_t = \alpha_t g_t + (1-\beta_t)\alpha_t(d_{t-1} - \tilde{g}_{t-1}) \,.$$

结合上述内容与 ϵ_t 的定义，可得如下递归关系，

$$\begin{aligned}
\alpha_t\epsilon_t &:= \alpha_t d_t - \alpha_t \bar{g}_t \\
&= \alpha_t(g_t - \bar{g}_t) + (1-\beta_t)\alpha_t(d_{t-1} - \tilde{g}_{t-1}) \\
&= \beta_t\alpha_t(g_t - \bar{g}_t) + (1-\beta_t)\alpha_t(d_{t-1} - \tilde{g}_{t-1} + g_t - \bar{g}_t) \\
&= \beta_t\alpha_t(g_t - \bar{g}_t) + (1-\beta_t)\alpha_t(\bar{g}_{t-1} - \tilde{g}_{t-1} + g_t - \bar{g}_t) + (1-\beta_t)\alpha_t(d_{t-1} - \bar{g}_{t-1}) \\
&= \beta_t\alpha_t(g_t - \bar{g}_t) + (1-\beta_t)\alpha_t((g_t - \bar{g}_t) - (\tilde{g}_{t-1} - \bar{g}_{t-1})) + (1-\beta_t)\alpha_t(d_{t-1} - \bar{g}_{t-1}) \\
&= \beta_t\alpha_t(g_t - \bar{g}_t) + (1-\beta_t)\alpha_t Z_t + (1-\beta_t)\frac{\alpha_t}{\alpha_{t-1}}\alpha_{t-1}\epsilon_{t-1}
\end{aligned}$$

其中，我们将 $Z_t := (g_t - \bar{g}_t) - (\tilde{g}_{t-1} - \bar{g}_{t-1})$ 记为。现在，利用 $\alpha_t = t+1$ 和 $\beta_t = 1/(t+1)$，可以证明 $\alpha_t\beta_t = 1$，且 $\alpha_t(1-\beta_t) = \alpha_t - 1 := \alpha_{t-1}$。此外，$(1-\beta_t)\frac{\alpha_t}{\alpha_{t-1}} = \frac{\alpha_t - 1}{\alpha_{t-1}} = 1$。将这些关系代入上面的等式可得，

$$\alpha_t\epsilon_t = \alpha_{t-1}Z_t + \alpha_{t-1}\epsilon_{t-1} + (g_t - \bar{g}_t) = M_t + \alpha_{t-1}\epsilon_{t-1} \,. \tag{21}$$

其中，对任意 $t > 1$，我们记 $M_t := \alpha_{t-1}Z_t + (g_t - \bar{g}_t)$，以及 $M_1 = g_1 - \bar{g}_1$。展开上述方程，可得到任意 $t \in [T]$ 的显式表达式，

$$\alpha_t\epsilon_t = \sum_{\tau=1}^{t} M_\tau \,. \tag{22}$$

现在请注意，序列 $\{M_t\}_t$ 相对于由截至时间 t 的样本历史所诱导的自然滤过 $\{\mathcal{F}_t\}_t$，是一个鞅差序列。事实上，

$$\mathbf{E}[M_t|\mathcal{F}_{t-1}] = \mathbf{E}[(g_t - \bar{g}_t)|\mathcal{F}_{t-1}] + \alpha_{t-1}\mathbf{E}[Z_t|\mathcal{F}_{t-1}] = \mathbf{E}[(g_t - \bar{g}_t)|x_t] + \alpha_{t-1}\mathbf{E}[Z_t|x_{t-1}, x_t] = 0 \,.$$

因此，利用下方的引理 C.1可得，

$$\begin{aligned}
\mathbf{E}\|\alpha_t\epsilon_t\|^2 = \left\|\sum_{\tau=1}^{t} M_\tau\right\|^2 &= \sum_{\tau=1}^{t}\mathbf{E}\|M_\tau\|^2 = \sum_{\tau=1}^{t}\mathbf{E}\|\alpha_{t-1}Z_t + (g_t - \bar{g}_t)\|^2 \\
&\leq 2\sum_{\tau=1}^{t}\alpha_{t-1}^2\mathbf{E}\|Z_t\|^2 + 2\sum_{\tau=1}^{t}\mathbf{E}\|g_t - \bar{g}_t\|^2 \\
&\leq 2\sum_{\tau=1}^{t}\alpha_{t-1}^2\mathbf{E}\|(g_t - \bar{g}_t) - (\tilde{g}_{t-1} - \bar{g}_{t-1})\|^2 + 2\sum_{\tau=1}^{t}\sigma^2 \\
&= 2\sum_{\tau=1}^{t}\alpha_{t-1}^2\mathbf{E}\|(\nabla f(x_t; z_t) - \nabla f(x_t)) - (\nabla f(x_{t-1}; z_t) - \nabla f(x_{t-1}))\|^2 + 2t\sigma^2 \\
&\leq 2\sum_{\tau=1}^{t}\alpha_{t-1}^2\sigma_L^2\|x_t - x_{t-1}\|^2 + 2t\sigma^2 \\
&\leq 32D^2\sigma_L^2\sum_{\tau=1}^{t}(\alpha_{t-1}^2/\alpha_{t-1}^2) + 2t\sigma^2 \\
&= (32D^2\sigma_L^2 + 2\sigma^2)\cdot t \\
&= \tilde{\sigma}^2 \cdot t \,.
\end{aligned} \tag{23}$$

这里第一个不等式用到了 $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$，该结论对任意 a、$b \in \mathbb{R}^d$ 均成立；第二个不等式用到了有界方差假设；第三个不等式用到了式（4），最后一个不等式用到了式（20）。

将上述不等式除以 $\alpha_t^2 = (t+1)^2$，引理得证，

$$\mathbf{E}\|d_t - \nabla f(x_t)\|^2 = \mathbf{E}\|\epsilon_t\|^2 = \mathbf{E}\|\alpha_t \epsilon_t\|^2/\alpha_t^2 \le \tilde{\sigma}^2 t/(t+1^2) \le \tilde{\sigma}^2/(t+1)\,.$$

Lemma C.1. 设 $\{M_t\}_t$ 为关于滤子 $\{\mathcal{F}_t\}_t$ 的鞅差序列，则 *n*
the 以下结论对任意 t均成立，

$$\mathbf{E}\left\|\sum_{\tau=1}^{t} M_\tau\right\|^2 = \sum_{\tau=1}^{t} \mathbf{E}\left\|M_\tau\right\|^2 \,.$$

□

C.1 证明引理 C.1

*Pro*引理的*C.1*。我们将通过对 t进行归纳法来证明该引理。基础情形为 $t =$ 1
clea确实成立。

现在，对于归纳步骤，假设等式对 $t \ge 1$ 成立，并证明它对 $t+1$ 成立。事实上，

$$\begin{aligned}
\mathbf{E}\left\|\sum_{\tau=1}^{t+1} M_\tau\right\|^2 &= \mathbf{E}\left\|M_{t+1} + \sum_{\tau=1}^{t} M_\tau\right\|^2 \\
&= \mathbf{E}\left\|\sum_{\tau=1}^{t} M_\tau\right\|^2 + \mathbf{E}\|M_{t+1}\|^2 + 2\mathbf{E}\left(\sum_{\tau=1}^{t} M_\tau\right)\cdot M_{t+1} \\
&= \sum_{\tau=1}^{t+1} \mathbf{E}\left\|M_\tau\right\|^2 + 2\mathbf{E}\left[\mathbf{E}\left[\left(\sum_{\tau=1}^{t} M_\tau\right)\cdot M_{t+1}|\mathcal{F}_t\right]\right] \\
&= \sum_{\tau=1}^{t+1} \mathbf{E}\left\|M_\tau\right\|^2 + 2\mathbf{E}\left[\left(\sum_{\tau=1}^{t} M_\tau\right)\cdot \mathbf{E}\left[M_{t+1}|\mathcal{F}_t\right]\right] \\
&= \sum_{\tau=1}^{t+1} \mathbf{E}\left\|M_\tau\right\|^2 + 0 \\
&= \sum_{\tau=1}^{t+1} \mathbf{E}\left\|M_\tau\right\|^2 \,,
\end{aligned}$$

其中，第三行由归纳假设以及全期望法则得出；第四行是因为 $\{M_\tau\}_{\tau=1}^t$ 关于 $\mathcal{F}_t$ 可测；第五行是因为$\mathbf{E}[M_{t+1}|\mathcal{F}_t] = 0$。因此，我们完成了归纳步骤，从而引理成立。

□

D 扩展

这里我们为定理提供了若干扩展和补充4.1。

- 在第 D.1 节中，我们为 $\|\epsilon_t\|^2$提供高概率界。
- 在第 D.2 节中，我们展示了如何以额外付出 $O(\log T)$ 倍总样本复杂度为代价，获得 $\|\epsilon_t\|^2 \le O(1/T)$。（也就是说，我们证明这需要总计 $O(T\log T)$ 个样本，而非 $O(T)$ 个样本。）

- 我们对 SGD 的一个更标准变体的扩展采用了均匀权重和 $\eta \propto 1/L$对数T学习率，详见第G 节。

D.1 高概率界

为了获得高概率界，我们还需要另一项假设：$\mathcal{K}$ 中的随机梯度有界，即存在 $G > 0$ 使得 $\|\nabla f(x, z)\| \leq G$，$\forall x \in \mathcal{K}$，$z \in$ **支持**$\{\mathcal{D}\}$。现在我们可以证明，以概率 $\geq 1-\delta$，对所有 $t \in [T]$，均有，

$$\|\epsilon_t\|^2 \leq O\left(\frac{\tilde{\sigma}^2}{t} \cdot \log(T/\delta) + \frac{U_{\max}^2}{t^2} \cdot \log(T/\delta)\right),$$

其中，我们记 $U_{\max} := 8LD + 2G$。

回顾定理4.1 的证明中，我们在式（22）中证明了以下内容:

$$\alpha_t \epsilon_t = \sum_{\tau=1}^{t} M_\tau := M_{1:t}. \tag{24}$$

其中 $M_t := \alpha_{t-1} Z_t + (g_t - \bar{g}_t)$，以及 $M_1 = g_1 - \bar{g}_1$，其中我们记 $Z_t := (g_t - \bar{g}_t) - (\tilde{g}_{t-1} - \bar{g}_{t-1})$。因此 M_t 是关于即将到来的样本所诱导的自然滤过的鞅差序列。并且 M_t 以概率 1 有界，因为

$$\begin{aligned}
\|M_t\| &= \|\alpha_{t-1} Z_t + (g_t - \bar{g}_t)\| \\
&\leq \|\alpha_{t-1} Z_t\| + \|(g_t - \bar{g}_t)\| \\
&\leq \|\alpha_{t-1}(g_t - \tilde{g}_{t-1})\| + \|\alpha_{t-1}(\bar{g}_t - \bar{g}_{t-1})\| + \|(g_t - \bar{g}_t)\| \\
&\leq L\alpha_{t-1}\|x_t - x_{t-1}\| + L\alpha_{t-1}\|x_t - x_{t-1}\| + 2G \\
&\leq 2L\alpha_{t-1} \cdot 4D/\alpha_{t-1} + 2G \\
&= 8LD + 2G \\
&:= U_{\max}.
\end{aligned}$$

whe在此我们使用了.....的光滑性 $\nabla f(\cdot,\ z)$ 以及公式（20）。我们还记 $U_{\max} := 8LD + 2G$.

最后，类似于定理4.1 的证明，我们可以证明，

$$\mathbf{E}_{t-1}\|M_t\|^2 = \tilde{\sigma}^2$$

其中，$\mathbf{E}_{t-1}$ 表示以截至并包括第 $t-1$ 轮的样本（随机化）历史为条件的条件期望。

现在，由于 $\{M_t\}_t$ 是相对于由优化过程诱导的自然滤过的鞅序列，并且它有界，条件二阶矩也有界，因此我们可以使用 Minsker (2017) [3] 来证明，以概率 $\geq 1-\delta$，对于所有 $t \in [T]$，我们有:

$$\|M_{1:t}\|^2 \leq O\left(\tilde{\sigma}^2 t \cdot \log(T/\delta) + U_{\max}^2 \log(T/\delta)\right)$$

其中，对数中的 T 来自并集界的使用。

因此，基于上述内容以及公式（24），我们立即得出结论：以概率 $\geq 1-\delta$，对于所有$t \in [T]$，我们有:

$$\|\epsilon_t\|^2 = \frac{1}{\alpha_t^2}\|M_{1:t}\|^2 \leq O\left(\frac{\tilde{\sigma}^2}{t} \cdot \log(T/\delta) + \frac{U_{\max}^2}{t^2} \cdot \log(T/\delta)\right)$$

其中我们使用了 $\alpha_t = t$。证明至此完成。

D.2 一致小误差界

这里我们展示，将样本复杂度增加log T 倍后，即可获得 $\mathbf{E}\|\epsilon_t\|^2 \leq \tilde{\sigma}^2/T$，$\forall t \in [T]$ 的一致小界。

为此，我们将在第 t轮使用大小为 $b_t = \lceil T/t \rceil$ 的批大小。

[3]实际上 推论 4.1 出自Minsker (2017)，它是其中定理 3.1 的推论，适用于独立矩阵之和。尽管如此，我们仍可从Minsker (2017) 中的定理 3.2 得到关于向量鞅差情形的推论，该定理适用于这种鞅差情形。

总样本数。因此，我们在所有轮次中使用的样本总数为 $\sum_{t=1}^{T} b_t =$ $O\,T \quad T$ T (log)。因此，样本复杂度仅增加了 log 因子。

误差分析。 使用批大小 b_t 后，我们的估计量在第 t 轮次中的方差降低了 b_t倍。因此，完全按照公式（23）中的相同步骤，我们可以证明如下结论。

$$\mathbf{E}\|\alpha_t \epsilon_t\|^2 \le \sum_{\tau=1}^{t} \frac{\tilde{\sigma}^2}{b_\tau} \le \frac{\tilde{\sigma}^2}{T} \sum_{\tau=1}^{t} \tau \le \tilde{\sigma}^2 \frac{t^2}{T} .$$

回忆 $\alpha_t = t$ 并除以 α_t^2，得到，

$$\mathbf{E}\|\epsilon_t\|^2 \le \frac{\tilde{\sigma}^2}{T} ,$$

这就确立了误差的一致小量界。

证明定理 4. 2

定理的证明 *4.2*。 该证明直接结合了定理 4.1 以及在线梯度下降（OGD）的标准遗憾界，进而能够利用定理 3.1。

第 1 部分：遗憾界。对式（9）中更新规则的标准遗憾分析表明，对于任意 t （例如参见（Hazan 等人，2016年），以及（15.1 ）中的定理Cutkosky，2022b）），

$$\sum_{\tau=1}^{t} \alpha_\tau d_\tau \cdot (w_\tau - w^*) \le \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{\tau=1}^{t} \alpha_\tau^2 \|d_\tau\|^2 . \tag{25}$$

第 2 部分：任意时刻保证。由于 x_t 是 w_t 的加权平均，我们可以应用定理 3.1，它保证对于任意 $t \in [T]$,

$$\alpha_{1:t}\Delta_t = \alpha_{1:t}(f(x_t) - f(w^*)) \le \sum_{\tau=1}^{t} \alpha_\tau \nabla f(x_\tau) \cdot (w_\tau - w^*) ,$$

其中我们记 $\Delta_t := f(x_t) - f(w^*)$。

第3部分：结合保证。将上述任意时刻保证与公式（25）中的界限结合起来，可得:

$$\begin{aligned}
\alpha_{1:t}\Delta_t &\le \sum_{\tau=1}^{t} \alpha_\tau \nabla f(x_\tau) \cdot (w_\tau - w^*) \\
&= \sum_{\tau=1}^{t} \alpha_\tau d_\tau \cdot (w_\tau - w^*) + \sum_{\tau=1}^{t} \alpha_\tau (\nabla f(x_\tau) - d_\tau) \cdot (w_\tau - w^*) \\
&= \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{\tau=1}^{t} \alpha_\tau^2 \|d_\tau\|^2 - \sum_{\tau=1}^{t} \alpha_\tau \epsilon_\tau \cdot (w_\tau - w^*) \\
&\le \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{\tau=1}^{t} \alpha_\tau^2 \|\nabla f(x_\tau) + \epsilon_\tau\|^2 + \sum_{\tau=1}^{t} \|\alpha_\tau \epsilon_\tau\| \cdot \|w_\tau - w^*\| \\
&\le \frac{D^2}{2\eta} + \eta \sum_{\tau=1}^{t} \alpha_\tau^2 \|\nabla f(x_\tau)\|^2 + \eta \sum_{\tau=1}^{t} \alpha_\tau^2 \|\epsilon_\tau\|^2 + D \sum_{\tau=1}^{t} \|\alpha_\tau \epsilon_\tau\| \\
&\le \frac{D^2}{2\eta} + 2\eta L \sum_{\tau=1}^{t} \alpha_\tau^2 \Delta_\tau + \eta \sum_{\tau=1}^{t} \alpha_\tau^2 \|\epsilon_\tau\|^2 + D \sum_{\tau=1}^{t} \|\alpha_\tau \epsilon_\tau\| \\
&\le \frac{D^2}{2\eta} + 4\eta L \sum_{\tau=1}^{t} \alpha_{1:\tau} \Delta_\tau + \eta \sum_{\tau=1}^{t} \alpha_\tau^2 \|\epsilon_\tau\|^2 + D \sum_{\tau=1}^{t} \|\alpha_\tau \epsilon_\tau\| ,
\end{aligned} \tag{26}$$

其中，第一个不等式由柯西-施瓦茨不等式得出；第二个不等式成立是因为 $\|w_t - w^*\| \leq D$ $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ $a\ b \in \mathbb{R}^d$，且使用了对任意 ,, 第三个不等式由光滑函数的自界性质得出（见下文引理 E.1），这意味着 $\|\nabla f(x_\tau)\|^2 \leq 2L(f(x_\tau) - f(w^*)) := 2L\Delta_\tau$；第四个不等式则由 $\alpha_\tau^2 \leq 2\alpha_{1:\tau}$得出，因为 $\alpha_\tau = \tau + 1$。

引理 E.1. *(例如参见 (Levy 等人，2018；Cutkosky，2019))* 设 $F\colon\ \mathbb{R}^d \mapsto \mathbb{R}$ 是一个 L-光滑函数，且具有全局最小值 x^*，则对于任意 $x \in \mathbb{R}^d$，有：

$$\|\nabla F(x)\|^2 \leq 2L(F(x) - F(w^*))\ .$$

接下来，我们将对公式 (26) 取期望，得到：

$$\begin{aligned}
\alpha_{1:t}\mathbf{E}\Delta_t &\leq \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\alpha_{1:\tau}\mathbf{E}\Delta_\tau + \eta\sum_{\tau=1}^{t}\alpha_\tau^2\mathbf{E}\|\epsilon_\tau\|^2 + D\sum_{\tau=1}^{t}\mathbf{E}\|\alpha_\tau\epsilon_\tau\| \\
&\leq \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\alpha_{1:\tau}\mathbf{E}\Delta_\tau + \eta\sum_{\tau=1}^{t}\alpha_\tau^2\mathbf{E}\|\epsilon_\tau\|^2 + D\sum_{\tau=1}^{t}\sqrt{\alpha_\tau^2\mathbf{E}\|\epsilon_\tau\|^2} \\
&\leq \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\alpha_{1:\tau}\mathbf{E}\Delta_\tau + \eta\sum_{\tau=1}^{t}\alpha_\tau^2\cdot\tilde{\sigma}^2/\alpha_\tau + D\sum_{\tau=1}^{t}\sqrt{\alpha_\tau^2\cdot\tilde{\sigma}^2/\alpha_\tau} \\
&\leq \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\alpha_{1:\tau}\mathbf{E}\Delta_\tau + \eta\tilde{\sigma}^2\sum_{\tau=1}^{t}\alpha_\tau + D\tilde{\sigma}\sum_{\tau=1}^{t}\sqrt{\alpha_\tau} \\
&\leq \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{T}\alpha_{1:\tau}\mathbf{E}\Delta_\tau + \eta\tilde{\sigma}^2\sum_{\tau=1}^{T}\alpha_\tau + D\tilde{\sigma}\sum_{\tau=1}^{T}\sqrt{\alpha_\tau} \\
&\leq \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{T}\alpha_{1:\tau}\mathbf{E}\Delta_\tau + \eta\tilde{\sigma}^2\alpha_{1:T} + 2D\tilde{\sigma}T^{3/2} \\
&\leq \frac{D^2}{2\eta} + \frac{1}{2T}\sum_{\tau=1}^{T}\alpha_{1:\tau}\mathbf{E}\Delta_\tau + \eta\tilde{\sigma}^2\alpha_{1:T} + 2D\tilde{\sigma}T^{3/2}\ , \qquad (27)
\end{aligned}$$

其中，第二行根据 Jensen不等式得出，即 $\mathbf{E}X \leq \sqrt{\mathbf{E}X^2}$ 对任意随机变量 X均成立；第三行由 $\mathbf{E}\|\epsilon_t\|^2 \leq \tilde{\sigma}^2/\alpha_t$ 得出，该结论由定理 4.1保证；第五行成立是因为 $t \leq T$；第六行由 $\sum_{t=1}^{T}\sqrt{\alpha_t} \leq 2T^{3/2}$得出，而最后一行是因为我们取$\eta \leq 1/8LT$。

为了得到最终界，我们将把下面的引理应用于公式 (27)，

引理 E.2. 设 $\{A_t\}_{t\in[T]}$ 为一列非负元素和 $\mathcal{B} \in \mathbb{R}$，并假设对于任意 $t \leq T$，

$$A_t \leq \mathcal{B} + \frac{1}{2T}\sum_{t=1}^{T}A_t\ ,$$

则有以下界限，

$$A_T \leq 2\mathcal{B}\ .$$

取 $A_t \leftarrow \alpha_{1:t}\mathbf{E}\Delta_t$ 和 $\mathcal{B} \leftarrow \frac{D^2}{2\eta} + \eta\tilde{\sigma}^2\alpha_{1:T} + 2D\tilde{\sigma}T^{3/2}$ 可得以下显式界，

$$\alpha_{1:T}\mathbf{E}\Delta_T \leq \frac{D^2}{\eta} + 2\eta\tilde{\sigma}^2\alpha_{1:T} + 4D\tilde{\sigma}T^{3/2}$$

除以 $\alpha_{1:T}$并回顾 $\alpha_{1:T} = \Theta(T^2)$，可得，

$$\mathbf{E}(f(x_T) - f(w^*)) = \mathbf{E}\Delta_T \leq O\left(\frac{D^2}{\eta T^2} + 2\eta\tilde{\sigma}^2 + \frac{4D\tilde{\sigma}}{\sqrt{T}}\right)\ ,$$

从而证明完毕。 □

E.1 证明引理 E.2

引理的证明 *E.2*。 对不等式求和 $A_t \le \mathcal{B} + \frac{1}{2T}\sum_{t=1}^{T} A_t$ 得到 t，

$$A_{1:T} \le T\mathcal{B} + T\frac{1}{2T}A_{1:T} = T\mathcal{B} + \frac{1}{2}A_{1:T} ,$$

重新排列后可得，

$$A_{1:T} \le 2T\mathcal{B} .$$

将其代回原不等式并取 $t = T$，可得，

$$A_T \le \mathcal{B} + \frac{1}{2T}A_{1:T} \le 2\mathcal{B} .$$

由此完成证明。 □

F 证明定理. 5.2

定理 *5.2* 的证明。 证明根据以下技术进行分解： μ^2 –额外的随机梯度下降所采用的方法。

第一部分：任意时刻保证。 由于 x_t 是 $\{\alpha_t\}_t$ 的加权平均 $\{w_t\}_t$，我们可以应用定理 3.1，从而得到：

$$\alpha_{1:T}(f(x_T) - f(w^*)) \le \sum_{t=1}^{T} \alpha_t \nabla f(x_t)\cdot(w_t - w^*) = \sum_{t=1}^{T} \alpha_t d_t \cdot (w_t - w^*) - \sum_{t=1}^{T} \alpha_t \epsilon_t \cdot (w_t - w^*) . \tag{28}$$

其中我们记 $\epsilon_t := d_t - \nabla f(x_t)$。

第二部分：乐观 OGD 保证。 由于 $\{w_t, y_t\}_t$ 的更新规则满足 ˆ乐观 OGD 模板（相对于损失序列和提示向量序列） $\{d_t, d_t\}_t$，我们可以应用引理 5.1 来界定式（28）中的加权遗憾，如下所示：

$$\begin{aligned}
&\alpha_{1:T}(f(x_T) - f(w^*)) \\
&\quad\le \frac{4D^2}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\alpha_t^2\|d_t - \hat{d}_t\|^2 - \frac{1}{2\eta}\sum_{t=1}^{T}\|w_t - y_{t-1}\|^2 - \sum_{t=1}^{T}\alpha_t\epsilon_t\cdot(w_t - w^*) \\
&\quad\le \frac{4D^2}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\alpha_t^2\|g_t - \hat{g}_t\|^2 - \frac{1}{2\eta}\sum_{t=1}^{T}\|w_t - y_{t-1}\|^2 + \sum_{t=1}^{T}\|\alpha_t\epsilon_t\|\cdot\|w_t - w^*\| \\
&\quad\le \frac{4D^2}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\alpha_t^2\|\nabla f(x_t; z_t) - \nabla f(\hat{x}_t; z_t)\|^2 - \frac{1}{2\eta}\sum_{t=1}^{T}\|w_t - y_{t-1}\|^2 + D\sum_{t=1}^{T}\|\alpha_t\epsilon_t\| \\
&\quad\le \frac{4D^2}{\eta} + \frac{\eta L^2}{2}\sum_{t=1}^{T}\alpha_t^2\|x_t - \hat{x}_t\|^2 - \frac{1}{2\eta}\sum_{t=1}^{T}\|w_t - y_{t-1}\|^2 + D\sum_{t=1}^{T}\|\alpha_t\epsilon_t\| \\
&\quad\le \frac{4D^2}{\eta} + \frac{\eta L^2}{2}\sum_{t=1}^{T}\alpha_t^2\left(\frac{\alpha_t}{\alpha_{1:t}}\right)^2\|w_t - y_{t-1}\|^2 - \frac{1}{2\eta}\sum_{t=1}^{T}\|w_t - y_{t-1}\|^2 + D\sum_{t=1}^{T}\|\alpha_t\epsilon_t\| \\
&\quad\le \frac{4D^2}{\eta} + \frac{4\eta L^2}{2}\sum_{t=1}^{T}\|w_t - y_{t-1}\|^2 - \frac{1}{2\eta}\sum_{t=1}^{T}\|w_t - y_{t-1}\|^2 + D\sum_{t=1}^{T}\|\alpha_t\epsilon_t\| \\
&\quad\le \frac{4D^2}{\eta} + D\sum_{t=1}^{T}\|\alpha_t\epsilon_t\| ,
\end{aligned} \tag{29}$$

ˆ其中，第一行使用公式（28）以及定理 5.1；第二行使用 $d_t - d_t = g_t - \hat{g}_t$，由公式（19）可得；第三行根据 g_t、$\hat{g}_t$ 的定义，以及$\|w_t - w^*\| \le D$，这成立是因为 w_t，$w^* \in \mathcal{K}$；第四行根据我们在公式（3）中的假设可得；

第五行成立是因为 $x_t - \hat{x}_t = (\alpha_t/\alpha_{1:t})\ (w_t - y_{t-1})$，这由公式（17）可得；第六行成立是因为 $\alpha_t^4/(\alpha_{1:t})^2 \le 4$，$\forall t \ge 1$；最后一行成立是因为 $2\eta L^2 - 1/(2\eta) \le 0$，这成立是因为我们假设 $\eta \le 1/2L$。

第三部分：μ^2 保证。注意，我们对 w_t，x_t，α_t，β_t 和 d_t 的定义满足定理 4.1 的完全相同条件。这立即意味着 $\mathbf{E}\|\epsilon_t\|^2 \le \tilde{\sigma}^2/t$，$\forall t$。利用这一点，并对式（29）取期望，可得：

$$\begin{aligned}\alpha_{1:T}\mathbf{E}(f(x_T) - f(w^*)) &\le \frac{4D^2}{\eta} + D\sum_{t=1}^{T}\mathbf{E}\|\alpha_t\epsilon_t\| \le \frac{4D^2}{\eta} + D\sum_{t=1}^{T}\sqrt{\mathbf{E}\|\alpha_t\epsilon_t\|^2} \\ &\le \frac{4D^2}{\eta} + D\tilde{\sigma}\sum_{t=1}^{T}\sqrt{\alpha_t^2/t} \le \frac{4D^2}{\eta} + 2T^{3/2}D\tilde{\sigma}\,. \end{aligned} \tag{30}$$

其中，第二个不等式利用了 Jensen不等式：$\mathbf{E}X \le \sqrt{\mathbf{E}X^2}$，该式对任意随机变量 X 均成立；最后一个不等式由 $\alpha_t^2/t \le 2t$ 得出，这意味着 $\sum_{t=1}^{T}\sqrt{\alpha_t^2/t} \le 2T^{3/2}$。
将上式两边除以 $\alpha_{1:T}$，并回想起 $\alpha_{1:T} = \Theta(T^2)$，即可完成证明。 □

G 将 μ_2-SGD 扩展至均匀权重及 $\eta \propto 1/L \log T$

这里我们展示，采用以下更标准的 $\alpha_t = 1$ 取值以及 $\eta \propto 1/L\log T$，在算法 1 中，可以获得与定理 4.2 相同的保证；不过，收敛速率会受到 $\log T$ 因子的影响。

下一个定理是定理 4.1 的一个变体，它表明即使选择均匀权重，我们也能得到 $\mathbf{E}\|\epsilon_t\|^2 \le O(\tilde{\sigma}^2/t)$。

定理G.1。设 $f: \mathcal{K} \mapsto \mathbb{R}$，并假设 $\mathcal{K}$ 是凸的，直径为 D，且公式(2)、(3)、(4) 中的假设均成立。那么，调用算法 *1* 并取 $\{\alpha_t = 1\}_t$ 和 $\{\beta_t = 1/t\}$，可得：

$$\mathbf{E}\|\epsilon_t\|^2 := \mathbf{E}\|d_t - \nabla f(x_t)\|^2 \le \tilde{\sigma}^2/t\,,$$

其中 $\epsilon_t := d_t - \nabla f(x_t)$，且 $\tilde{\sigma}^2 = 32D^2\hat{\sigma}_L^2 + 2\sigma^2$。

接下来我们给出证明概要。具体证明与定理 4. 的证明思路相同。 1.

定理 *G.1* 的证明概要。首先注意到 x_t 始终属于 $\mathcal{K}$，因为它们是 $\{w_t \in \mathcal{K}\}_t$ 的加权平均。我们的第一步是界定相邻查询之间的差异。根据 x_t 的定义，可得：

$$\alpha_{1:t-1}(x_t - x_{t-1}) = \alpha_t(w_t - x_t)\,,$$

从而产生，

$$\|x_t - x_{t-1}\|^2 = (\alpha_t/\alpha_{1:t-1})^2\|w_t - x_t\|^2 \le \frac{1}{(t-1)^2}D^2\,. \tag{31}$$

其中我们使用了 $\alpha_t = 1$ 和 $\alpha_{1:t-1} = t-1$；我们还使用了 $\|w_t - x_t\| \le D$，这成立是因为$w_t, x_t \in \mathcal{K}$。

记号：在继续进行证明之前，我们需要一些记号。我们将记 $\bar{g}_t := \nabla f(x_t)$，并回顾算法 1中的以下记号：$g_t := \nabla f(x_t, z_t)$；$\tilde{g}_{t-1} := \nabla f(x_{t-1}, z_t)$。我们还将记作：$\epsilon_t := d_t - \bar{g}_t$。

现在，回顾式（12），利用 $\alpha_t = 1$，并将其与上述 ϵ_t 的定义结合起来，可以推导出以下递归关系：

$$\begin{aligned}\epsilon_t &= \beta_t(g_t - \bar{g}_t) + (1-\beta_t)Z_t + (1-\beta_t)\epsilon_{t-1} \\ &= \frac{1}{t}(g_t - \bar{g}_t) + \frac{t-1}{t}Z_t + \frac{t-1}{t}\epsilon_{t-1}\,,\end{aligned}$$

其中，我们记 $Z_t := (g_t - \bar{g}_t) - (\tilde{g}_{t-1} - \bar{g}_{t-1})$，并使用了 $\beta_t = 1/t$。现在，将上式乘以 t，得到：

$$\begin{aligned}t\epsilon_t &= (g_t - \bar{g}_t) + (t-1)Z_t + (t-1)\epsilon_{t-1} \\ &= M_t + (t-1)\epsilon_{t-1}\,.\end{aligned}$$

其中，对于任意 $t > 1$，我们记 $M_t := (t-1)Z_t + (g_t - \bar{g}_t)$，以及 $M_1 = g_1 - \bar{g}_1$。展开上式，可得到任意 $t \in [T]$ 的显式表达式：

$$t\epsilon_t = \sum_{\tau=1}^{t} M_\tau := M_{1:t} \, .$$

注意到序列 $\{M_t\}_t$ 是关于由截至时间 t 的样本历史所诱导的自然滤过 $\{\mathcal{F}_t\}_t$ 的鞅差序列；因此可以如下界定，

$$\begin{aligned}
\mathbf{E}\|t\epsilon_t\|^2 = \left\|\sum_{\tau=1}^{t} M_\tau\right\|^2 &= \sum_{\tau=1}^{t} \mathbf{E}\|M_\tau\|^2 = \sum_{\tau=1}^{t} \mathbf{E}\|(t-1)Z_t + (g_t - \bar{g}_t)\|^2 \\
&\le 2\sum_{\tau=1}^{t}(t-1)^2\mathbf{E}\|Z_t\|^2 + 2\sum_{\tau=1}^{t}\mathbf{E}\|g_t - \bar{g}_t\|^2 \\
&\le 2\sum_{\tau=1}^{t}(t-1)^2\mathbf{E}\|(g_t - \bar{g}_t) - (\tilde{g}_{t-1} - \bar{g}_{t-1})\|^2 + 2\sum_{\tau=1}^{t}\sigma^2 \, . \qquad (32)
\end{aligned}$$

现在，结合式（4）与式（31），可得以下界：

$$\mathbf{E}\|(g_t - \bar{g}_t) - (\tilde{g}_{t-1} - \bar{g}_{t-1})\|^2 \le \sigma_L^2\|x_t - x_{t-1}\|^2 \le \sigma_L^2 D^2/(t-1)^2 \, .$$

将上述结果代回式（32）并求和即可证明该定理。 □

μ^2-随机梯度下降，使用 $\eta \propto 1/L\log T$基于上述定理，我们现在可以陈述以下保证：一种采用 $\alpha_t = 1$ 和 $\eta \propto 1/L\log\ T$ 标准选择的 μ^2- 随机梯度下降版本。

定理 G.2（μ^2-随机梯度下降保证）。设 $f\colon\ \mathbb{R}^d \mapsto \mathbb{R}$ 为一个凸函数，并假设 $w^* \in arg\ min_{w\in\mathcal{K}}\ f(w)$也是其在 $\mathbb{R}^d$中的全局最小值。此外，令我们作出与定理 4.1中相同的假设。然后，在 $\{\alpha_t = 1\}_t$和 $\{\beta_t = 1/t\}_t$的情况下调用算法 1，并在算法 1 的公式 (10) 中使用公式 (9) 的SGD型更新规则及学习率 $\eta \le \frac{1}{16L\ 1+(\ log\ T)}$，可保证，

$$\mathbf{E}(f(x_T) - f(w^*)) = \mathbf{E}\Delta_T \le \tilde{O}\left(\frac{D^2}{\eta T} + 2\eta\frac{\tilde{\sigma}^2}{T} + \frac{4D\tilde{\sigma}}{\sqrt{T}}\right) \, ,$$

其中 $\Delta_t := f(x_t) - f(w^*)$，且 $\tilde{\sigma}^2 := 32D^2\sigma_L^2 + 2\sigma^2$。

请注意，这表明该 μ^2-随机梯度下降变体具有相同的稳定性，这与我们在正文和定理中讨论的变体的稳定性类似。4.2。

接下来我们给出证明。

定理证明 *G.2*。 该证明是定理 G.1 与在线梯度下降（OGD）的标准遗憾界的直接结合，进而能够利用定理 3.1所述的任意时刻保证。

第 1 部分：遗憾界。对式（9）中更新规则的标准遗憾分析表明，对于任意 t（Hazan 等人，2016年），

$$\sum_{\tau=1}^{t}\alpha_\tau d_\tau \cdot (w_\tau - w^*) \le \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{\tau=1}^{t}\alpha_\tau^2\|d_\tau\|^2 \, . \qquad (33)$$

第 2 部分：任意时刻保证。 由于 x_t是 w_t 的加权平均值，我们可以调用定理 3.1，该定理保证对于任意 $t \in [T]$,

$$\alpha_{1:t}\Delta_t = \alpha_{1:t}(f(x_t) - f(w^*)) \le \sum_{\tau=1}^{t}\alpha_\tau\nabla f(x_\tau)\cdot(w_\tau - w^*) \, ,$$

其中，我们记 $\Delta_t := f(x_t) - f(w^*)$。

第 3 部分：结合保证。将上述任意时刻保证与公式（33）中的界结合起来，可得：

$$\begin{aligned}
\alpha_{1:t}\Delta_t &\le \sum_{\tau=1}^{t}\alpha_\tau \nabla f(x_\tau)\cdot(w_\tau - w^*)\\
&= \sum_{\tau=1}^{t}\alpha_\tau d_\tau\cdot(w_\tau - w^*) + \sum_{\tau=1}^{t}\alpha_\tau(\nabla f(x_\tau)-d_\tau)\cdot(w_\tau - w^*)\\
&= \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{\tau=1}^{t}\alpha_\tau^2\|d_\tau\|^2 - \sum_{\tau=1}^{t}\alpha_\tau\epsilon_\tau\cdot(w_\tau - w^*)\\
&\le \frac{D^2}{2\eta} + \frac{\eta}{2}\sum_{\tau=1}^{t}\alpha_\tau^2\|\nabla f(x_\tau)+\epsilon_\tau\|^2 + \sum_{\tau=1}^{t}\|\alpha_\tau\epsilon_\tau\|\cdot\|w_\tau - w^*\|\\
&\le \frac{D^2}{2\eta} + \eta\sum_{\tau=1}^{t}\alpha_\tau^2\|\nabla f(x_\tau)\|^2 + \eta\sum_{\tau=1}^{t}\alpha_\tau^2\|\epsilon_\tau\|^2 + D\sum_{\tau=1}^{t}\|\alpha_\tau\epsilon_\tau\|\\
&\le \frac{D^2}{2\eta} + 2\eta L\sum_{\tau=1}^{t}\alpha_\tau^2\Delta_\tau + \eta\sum_{\tau=1}^{t}\alpha_\tau^2\|\epsilon_\tau\|^2 + D\sum_{\tau=1}^{t}\|\alpha_\tau\epsilon_\tau\|\\
&\le \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\Delta_\tau + \eta\sum_{\tau=1}^{t}\|\epsilon_\tau\|^2 + D\sum_{\tau=1}^{t}\|\epsilon_\tau\|\,, \qquad (34)
\end{aligned}$$

其中，第一个不等式由柯西-施瓦茨不等式得出；第二个不等式成立是因为 $\|w_t - w^*\| \le D$ $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ $a\ b \in \mathbb{R}^d$，并且利用了对任意 ,, 第三个不等式根据光滑函数的自界性质（见引理 E.1）可得，即$\|\nabla f(x_\tau)\|^2 \le 2L(f(x_\tau) - f(w^*))$: $= 2L\Delta_\tau$；第四个不等式由 $\alpha_\tau = 1$ 得出。

接下来，我们将对式 (34) 取期望，得到：

$$\begin{aligned}
t\mathbf{E}\Delta_t = \alpha_{1:t}\mathbf{E}\Delta_t &\le \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\mathbf{E}\Delta_\tau + \eta\sum_{\tau=1}^{t}\mathbf{E}\|\epsilon_\tau\|^2 + D\sum_{\tau=1}^{t}\mathbf{E}\|\epsilon_\tau\|\\
&\le \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\mathbf{E}\Delta_\tau + \eta\sum_{\tau=1}^{t}\mathbf{E}\|\epsilon_\tau\|^2 + D\sum_{\tau=1}^{t}\sqrt{\mathbf{E}\|\epsilon_\tau\|^2}\\
&\le \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\mathbf{E}\Delta_\tau + \eta\sum_{\tau=1}^{t}\tilde{\sigma}^2/\tau + D\sum_{\tau=1}^{t}\sqrt{\tilde{\sigma}^2/\tau}\\
&\le \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\mathbf{E}\Delta_\tau + \eta\tilde{\sigma}^2\sum_{\tau=1}^{t}1/\tau + D\tilde{\sigma}\sum_{\tau=1}^{t}\sqrt{1/\tau}\\
&\le \frac{D^2}{2\eta} + 4\eta L\sum_{\tau=1}^{t}\mathbf{E}\Delta_\tau + \eta\tilde{\sigma}^2(1+\log t) + 2D\tilde{\sigma}\sqrt{t}\\
&\le \frac{D^2}{2\eta} + \frac{1}{4(1+\log T)}\sum_{\tau=1}^{t}\mathbf{E}\Delta_\tau + \eta\tilde{\sigma}^2(1+\log t) + 2D\tilde{\sigma}\sqrt{t}\,, \qquad (35)
\end{aligned}$$

其中第二行是由于 Jensen不等式意味着 $\mathbf{E}X \le \sqrt{\mathbf{E}X^2}$ 对于任意随机变量 X；第三行由 $\mathbf{E}\|\epsilon_t\|^2 \le \tilde{\sigma}^2/t$ 得出，该式由定理 G.1成立。我们还使用了$\sum_{\tau=1}^{t}1/\sqrt{\tau} \le 2\sqrt{t}$ 以及 $\sum_{\tau=1}^{t}1/\tau \le 1+\log t$。最后，我们使用了对 η的选择。

为了得到最终界，我们将应用下面的引理G.3 于公式（35）。

引理G.3.设 $T > 2$，且 $\{A_t\}_{t\in[T]}$ 是一个非负元素序列，$\{\mathcal{B}_t \in \mathbb{R}\}_{t\in[T]}$ 是一个单调递增的非负元素序列，并假设对于任意 $t \le T$,

$$tA_t \le \mathcal{B}_t + \frac{1}{4(1+\log T)}\sum_{t=1}^{T}A_t\,,$$

那么，以下界成立 $\forall t \in [T]$,

$$A_t \leq 2\mathcal{B}_t/t \, .$$

取 $A_t \leftarrow \mathbf{E}\Delta_t$ 和 $\mathcal{B}_t \leftarrow \frac{D^2}{2\eta} + \eta\tilde{\sigma}^2(1+\log t) + 2D\tilde{\sigma}\sqrt{t}$，可得以下显式界，

$$\mathbf{E}(f(x_T) - f(w^*)) = \mathbf{E}\Delta_T \leq \tilde{O}\left(\frac{D^2}{\eta T} + 2\eta\frac{\tilde{\sigma}^2}{T} + \frac{4D\tilde{\sigma}}{\sqrt{T}}\right) ,$$

从而完成证明。 □

G.1 证明引理 G.3

证明。我们将通过归纳法证明该引理。对于基础情形，我们有，

$$A_1 \leq \mathcal{B}_1 + \frac{1}{4(1+\log T)}A_1 \leq \mathcal{B}_1 + \frac{1}{4}A_1$$

这直接意味着 $A_1 \leq 4\mathcal{B}_1/3 \leq 2\mathcal{B}_1$，从而建立了基础情形。

对于归纳步骤，假设引理对任意 $\tau \leq t$ 都成立，并证明它对 $t+1$ 也成立。事实上，利用归纳假设可得，

$$\begin{aligned}
(t+1)A_{t+1} &\leq \mathcal{B}_{t+1} + \frac{1}{4(1+\log T)}\left(A_{t+1} + \sum_{\tau=1}^{t} A_\tau\right) \\
&\leq \mathcal{B}_{t+1} + \frac{1}{4}A_{t+1} + \frac{1}{4(1+\log T)}\sum_{\tau=1}^{t}\frac{2\mathcal{B}_\tau}{\tau} \\
&\leq \mathcal{B}_{t+1} + \frac{1}{4}A_{t+1} + \frac{2\mathcal{B}_{t+1}}{4(1+\log T)}\sum_{\tau=1}^{t}\frac{1}{\tau} \\
&\leq \mathcal{B}_{t+1} + \frac{1}{4}A_{t+1} + \frac{2\mathcal{B}_{t+1}}{4(1+\log T)}\sum_{\tau=1}^{t}\frac{1}{\tau} \\
&\leq \mathcal{B}_{t+1} + \frac{1}{4}A_{t+1} + \frac{1}{2}\mathcal{B}_{t+1} \, .
\end{aligned}$$

这里用到了 $\mathcal{B}_t$ 序列的单调性，以及对于任意 $t \leq T$，我们有 $\sum_{\tau=1}^{t}\frac{1}{\tau} \leq 1 + \log t \leq 1 + \log T$ 这一事实。重新整理上式可得 $(t+1-\frac{1}{4})A_{t+1} \leq \frac{3}{2}\mathcal{B}_{t+1}$。由于 $t \geq 1$，因此 $t+1-\frac{1}{4} \geq \frac{3}{4}(t+1)$。结合这一点以及 A_{t+1} 的非负性，可直接推出，

$$A_{t+1} \leq \frac{3\mathcal{B}_{t+1}/2}{t+1-\frac{1}{4}} \leq \frac{3\mathcal{B}_{t+1}/2}{3(t+1)/4} = 2\mathcal{B}_{t+1} \, .$$

这就确立了归纳步骤，进而完成归纳证明。

□

H 实验

H.1 技术细节

我们在一系列固定学习率下比较了以下优化算法[4]:

- μ^2-随机梯度
下降。
- 基于动量的随机梯度下降，使用 $\mu = 0.9$ 和 $\tau = 0.9$ （参见 PyTorch 文档。[5]）。
- 标准 SGD。
- STORM。
- Anytime-SGD。

我们在以下数据集上评估了我们的方法:

- **CIFAR-10 数据集**。CIFAR-10 数据集（Krizhevsky 等人，2014）包含 60,000 张彩色图像，分属十个类别，分辨率为 32x32 像素。
- **MNIST-10 数据集**。MNIST 数据集（LeCun 等人， 2010）包含70,000 张分辨率为 28x28 像素的手写数字（0-9）灰度图像。

H.2 凸设置

在本实验中，我们针对 MNIST 数据集上的逻辑回归问题展开研究。训练和测试阶段均采用大小为 64 的小批次，并对数据集进行一次完整遍历（一个训练周期）。
此外，模型权重被限制在单位球内，将解空间限制为一个紧凸集:

$$\mathcal{K} = \{w \in \mathbb{R}^d : \|w\|_2 \leq 1\},$$

其中 $w \in \mathbb{R}^d$ 表示模型权重。
在每次迭代中，都会使用投影函数$\Pi_{\mathcal{K}}(w)$将权重投影回单位球，从而确保 w 的范数保持有界。

以下算法采用各自的参数设置进行了评估： μ^2-随机梯度下降，使用$\alpha_t = t$ 和 $\beta_t = 1/t$；STORM，使用 $\beta_t = 1/t$；以及 Anytime-SGD，使用 $\alpha_t = t$。

H.2.1 不同学习率下的测试准确率

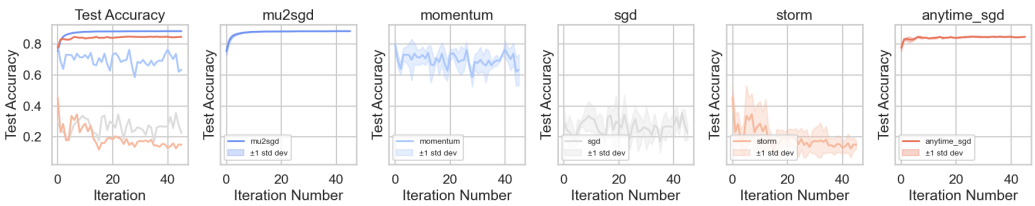


图 3：学习率为 = 10 时的测试准确率（↑ 越好）。

[4]请注意，在不同优化器之间比较固定学习率与我们的理论发现一致；我们的理论表明，对于 $\eta = \alpha_t \eta_{\mu^2\text{-SGD}} = O(t/T)$， $\simeq O(1)$，[5]https://pytorch.org/docs/stable/generated/torch.optim.SGD.html

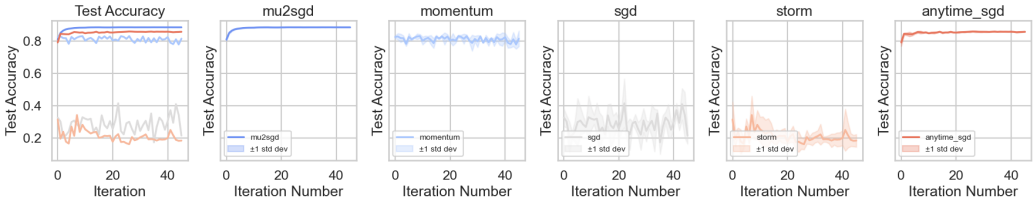


图 4：学习率为 = 1 时的测试准确率（↑ 越高越好）。

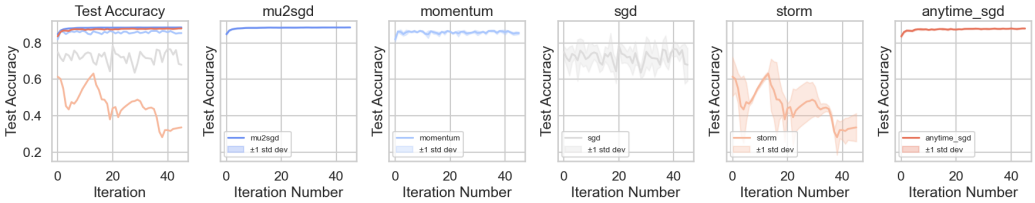


图 5：学习率为 = 0.1 时的测试准确率（↑ 越高越好）。

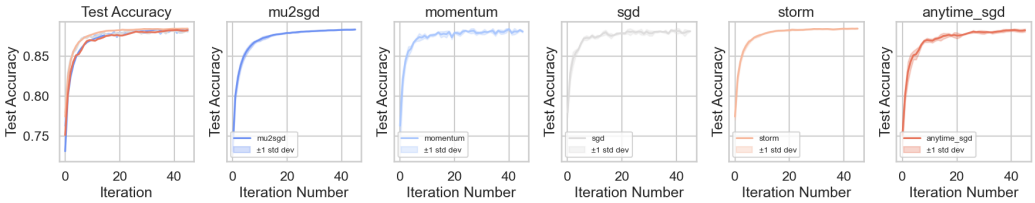


图 6：学习率为 = 0.01 时的测试准确率（↑ 越高越好）。

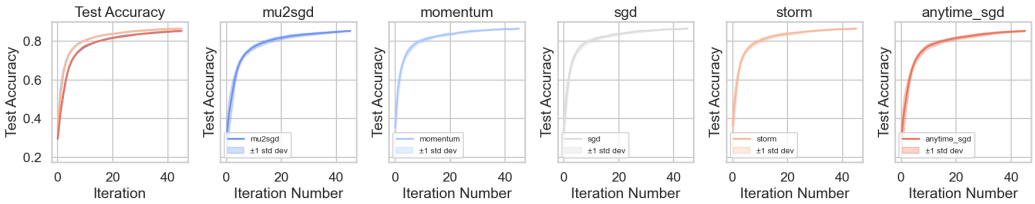


图 7：学习率为 = 0.001 时的测试准确率（↑ 越高越好）。

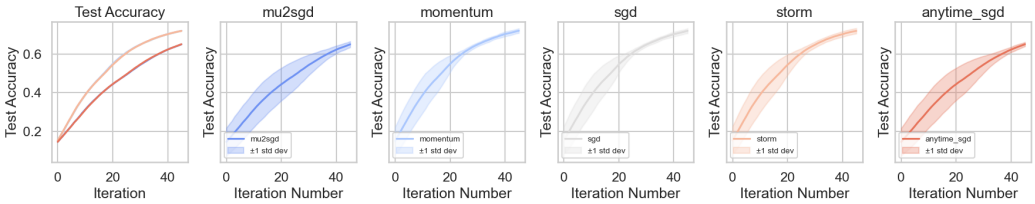


图 8：学习率为 = 0.0001 时的测试准确率（↑ 越高越好）。

H.2.2 不同学习率下的测试损失

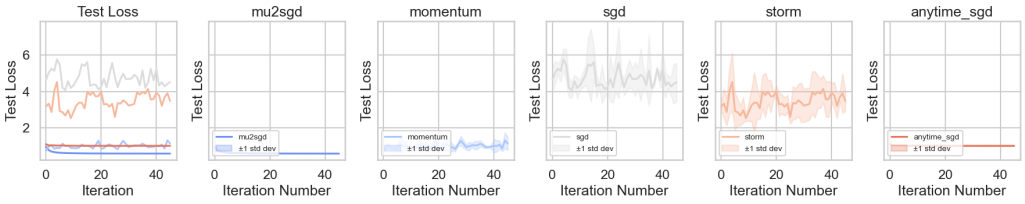


图9：学习率为 = 10 时的测试损失（↓ 越好）。

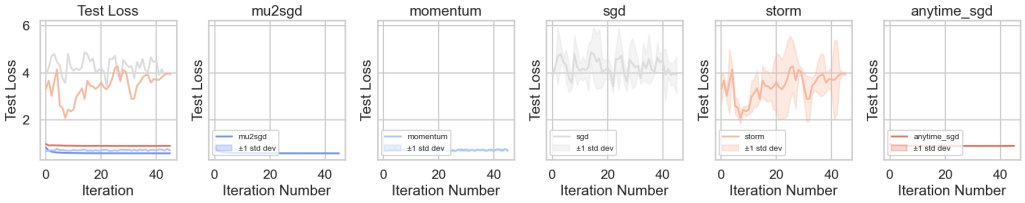


图10：学习率为 = 1 时的测试损失（↓ 越好）。

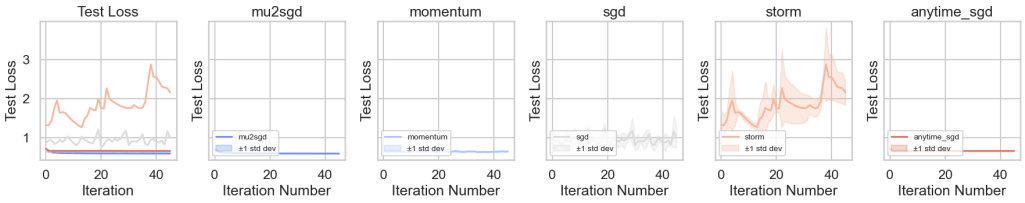


图11：学习率为 = 0.1 时的测试损失（↓ 越好）。

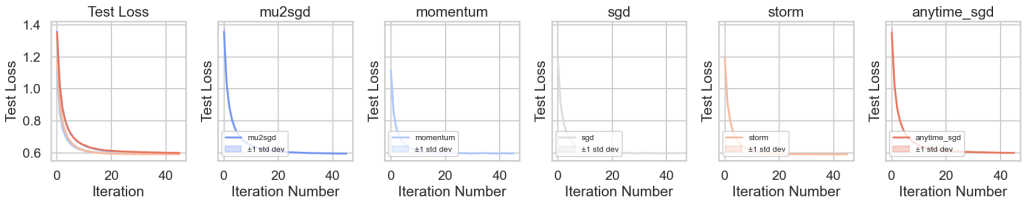


图12：学习率为 = 0.01 时的测试损失（↓ 越好）。

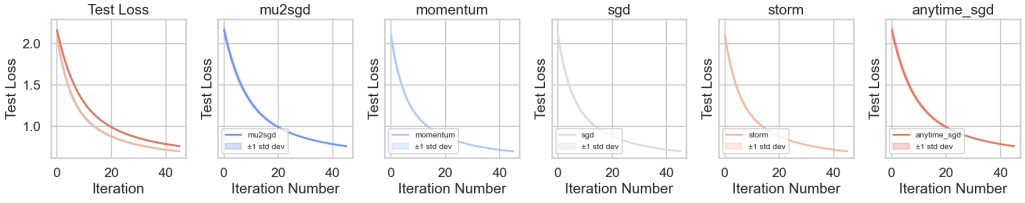


图13：学习率为 = 0.001 时的测试损失（↓ 越低越好）。

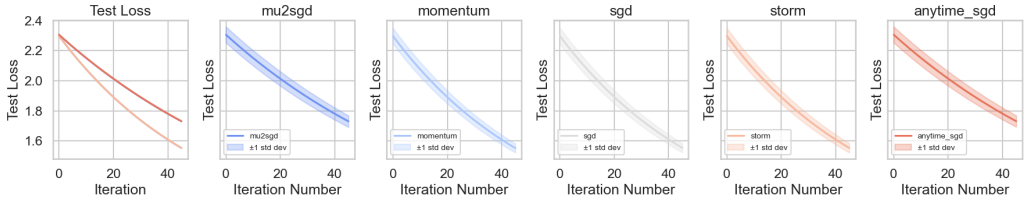


图14：学习率为 = 0.0001 时的测试损失（↓ 越好）。

H.3 非凸设置

CIFAR-10 实验。 我们使用大小为 32 的小批次训练 ResNet-18，共训练 25 轮。训练过程中使用随机裁剪（32×32，填充=2，p=0.5）和随机水平翻转（p=0.5）进行数据增强。

MNIST-10 实验。 我们训练了一个具有 ReLU 激活、最大池化和两个全连接层的双层 CNN。对第一个全连接层应用了批归一化。模型使用大小为 64 的小批次进行训练，遍历数据集一次。

使用各自的固定 参数设置评估以下算法： μ^2-随机梯度下降，参数为 $\gamma_t = 0.1$ 和 $\beta_t = 0.9$；STORM，参数为 $\beta_t = 0.9$；以及 Anytime-SGD，参数为 $\gamma_t = 0.1$。这些固定的参数选择有助于减轻非凸深度学习模型中较小动量参数的敏感性，这种敏感性源于两个关键因素：

- 历史权重的相关性：在非凸优化中，过度依赖过去的权重（例如 $\gamma_t \sim 1/t$）可能会阻碍优化进展，因为旧权重可能不再相关，或会通过鞍点和局部极小值误导优化。
- 数值稳定性：在参数量较大的模型或大型数据集上（此时 t 会变大），非常小的动量参数可能导致数值不稳定，从而影响优化的效果和可靠性。

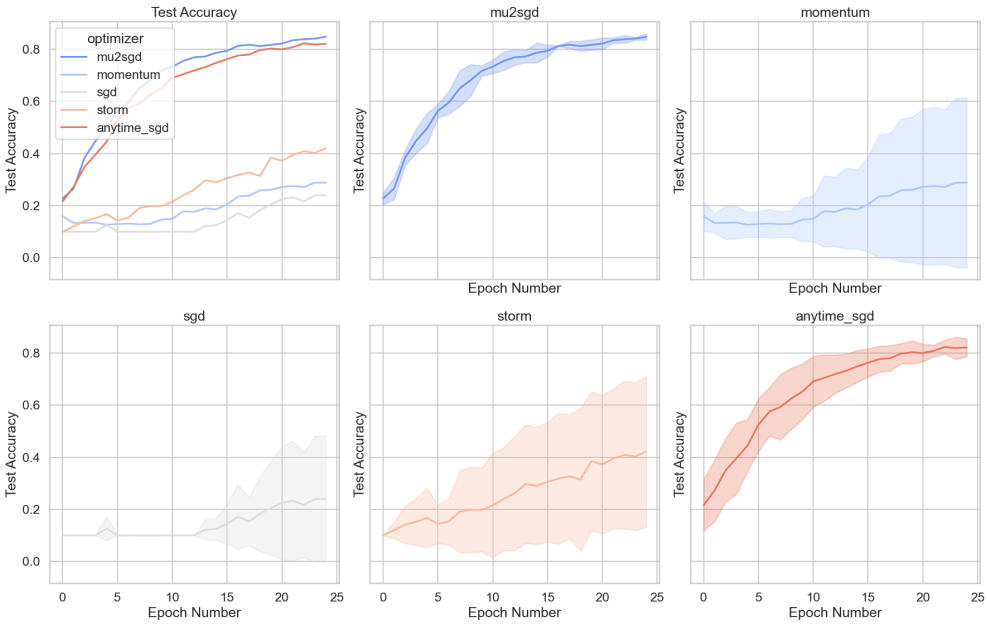


图15：CIFAR-10：学习率为 = 10 时的轮次与测试准确率关系（↑ 越大越好）。

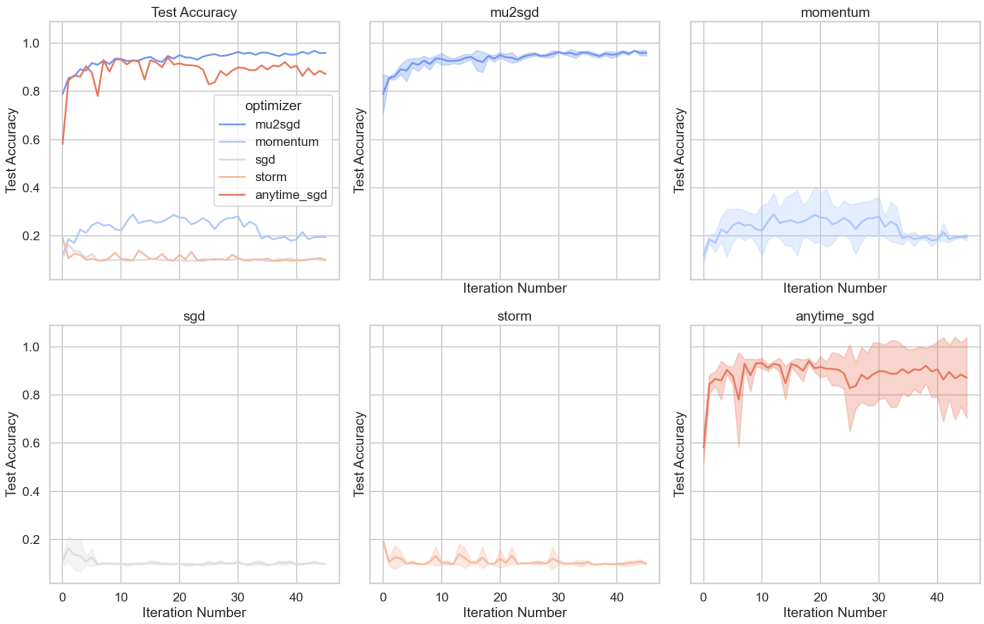


图16：MNIST：学习率=10 下的迭代次数测试准确率（↑ 越大越好）。

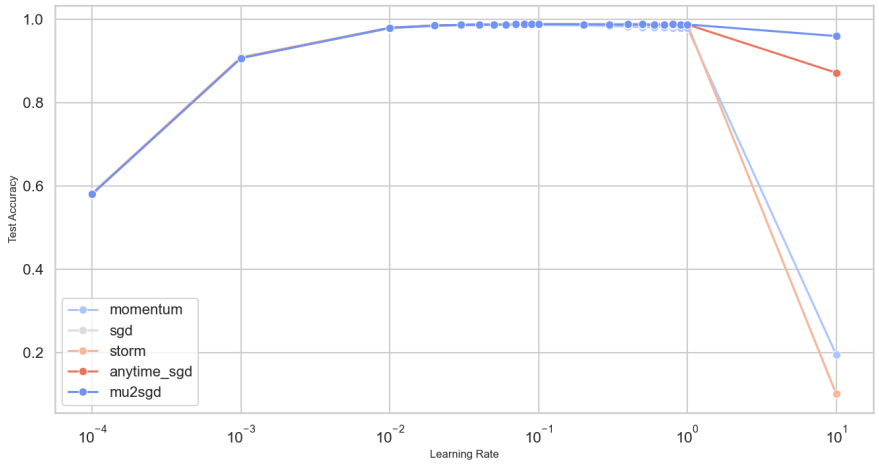


图17：MNIST：学习率从0.0001到10时的测试准确率（↑ 效果更好）。

H.3.1 CIFAR-10：不同学习率下的测试准确率

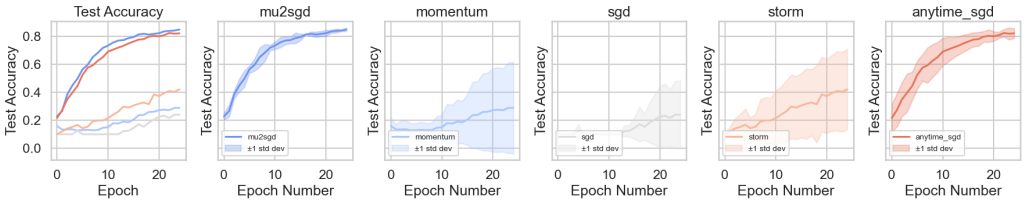


图18：学习率为 $= 10$（↑ 更好）。

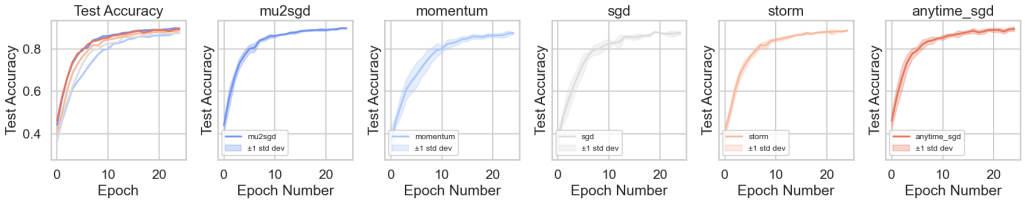


图19：学习率为 $= 1$ 时的测试准确率（↑ 越好）。

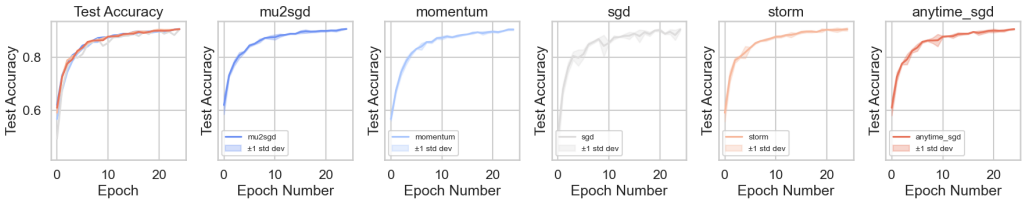


图 20：学习率为 $= 0.1$ 时的测试准确率（↑ 越高越好）。

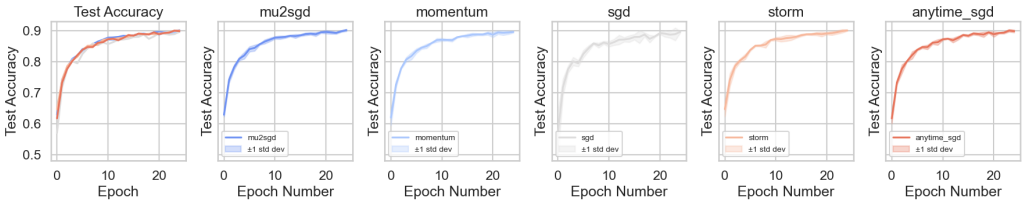


图 21：学习率 $= 0.01$ 下的测试准确率（↑ 越高越好）。

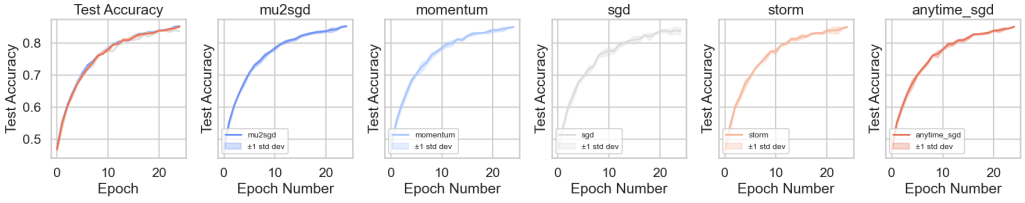


图 22：学习率为 $= 0.001$ 时的测试准确率（↑ 越大越好）。

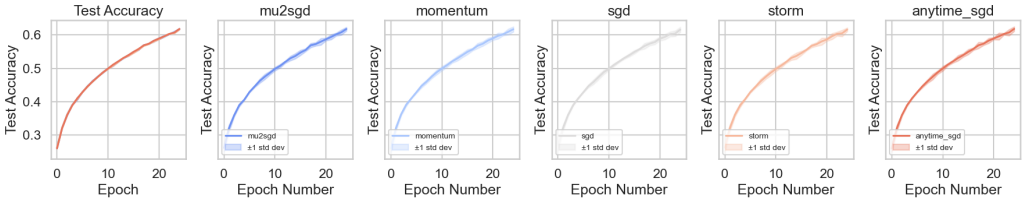


图 23：学习率为 = 0.0001 时的测试准确率（↑ 越高越好）。

H.3.2 CIFAR-10：不同学习率下的测试损失

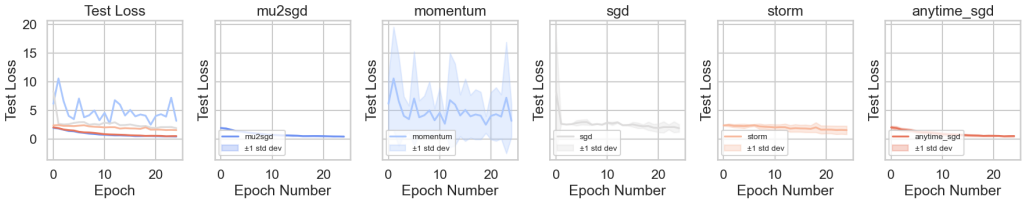


图 24：学习率为 = 10 （↓ 越低越好）[6]。

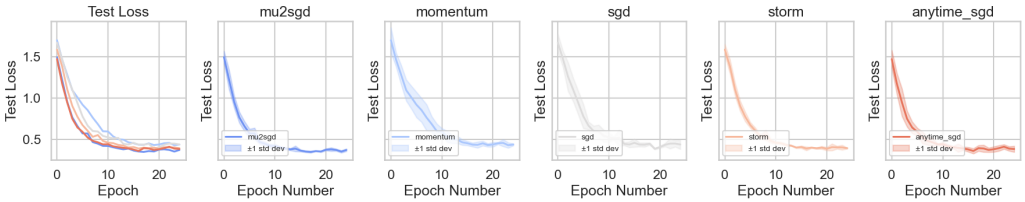


图 25：学习率为 = 1 时的测试损失（↓ 越低越好）。

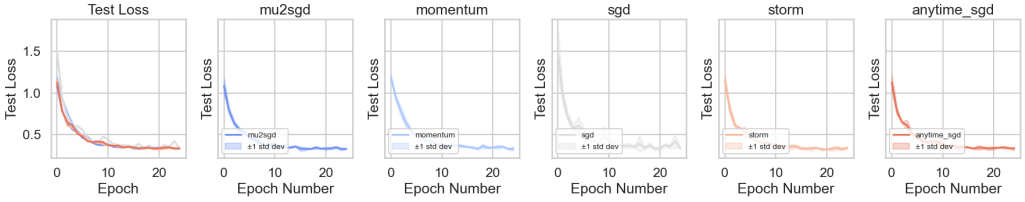


图 26：学习率为 = 0.1 时的测试损失（↓ 越低越好）。

[6]损失值为便于可视化会被裁剪到最大值 20。

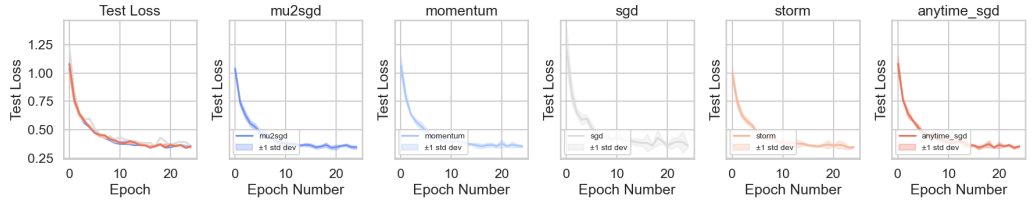


图 27：学习率为 $= 0.01$ 时的测试损失（↓ 越低越好）。

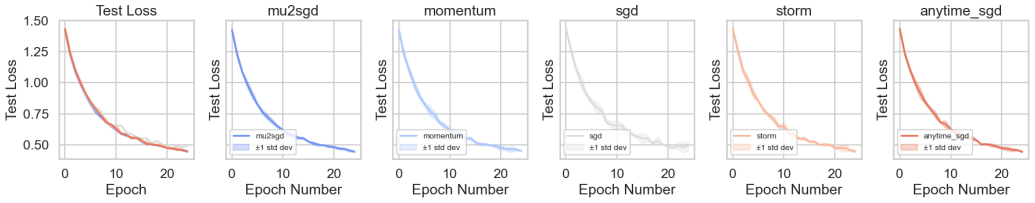


图 28：学习率为 $= 0.001$ 时的测试损失（↓ 越低越好）。

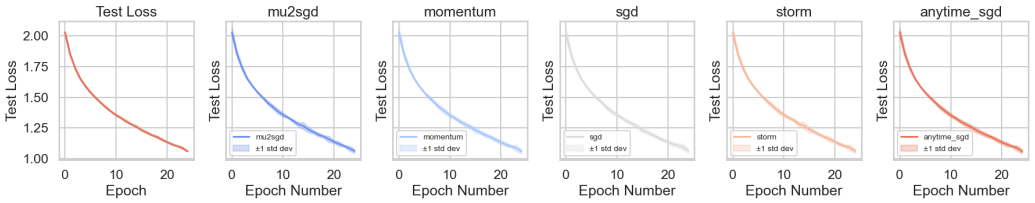


图 29：学习率为 $= 0.0001$ 时的测试损失（↓ 越低越好）。

H.3.3 MNIST：不同学习率下的测试准确率

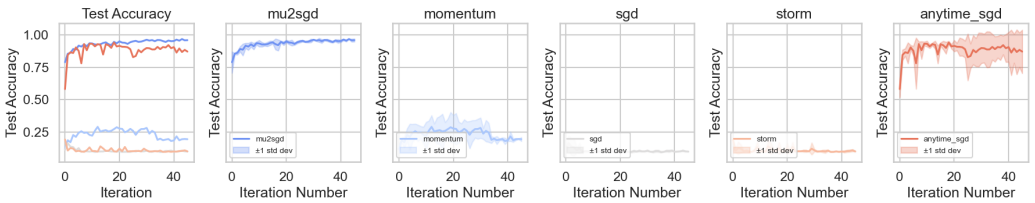


图 30：学习率为 $= 10$ 时的测试准确率（↑ 越高越好）。

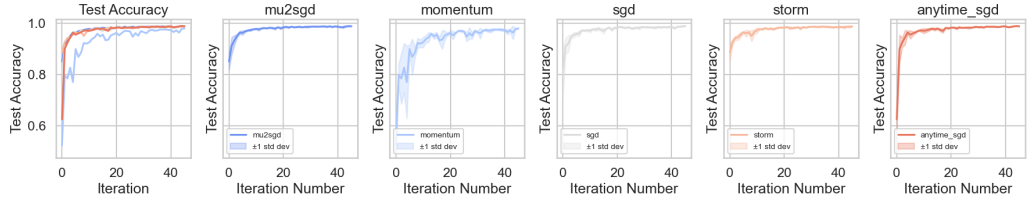


图 31：学习率为 $= 1$ 时的测试准确率（↑ 越大越好）。

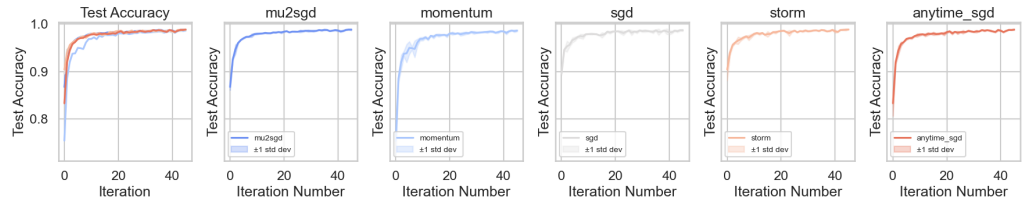


图 32：学习率 = 0.1 下的测试准确率（↑ 更好）。

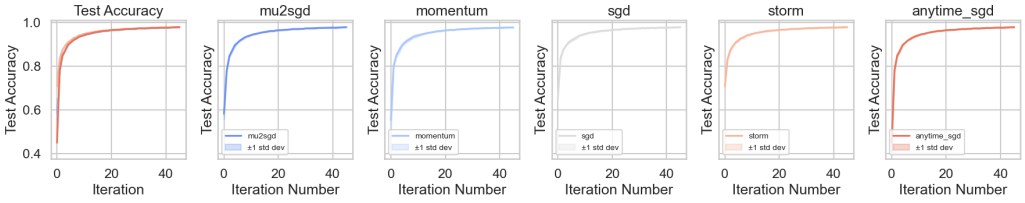


图33：学习率 = 0.01时的测试准确率（↑ 越好）。

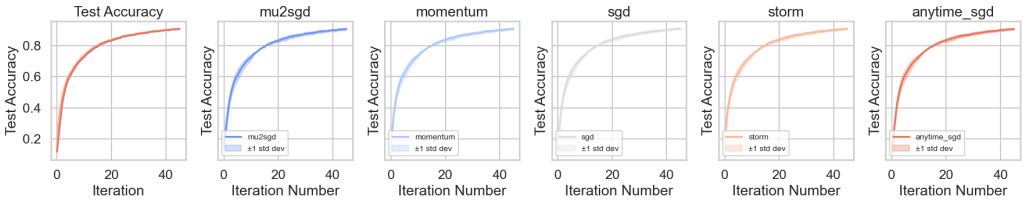


图 34：学习率 = 0.001 下的测试准确率（↑ 越高越好）。

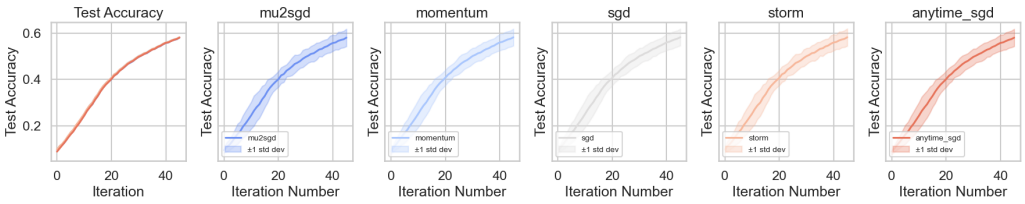


图35：学习率 = 0.0001下的测试准确率（↑ 更好）。

H.3.4 MNIST：不同学习率下的测试损失

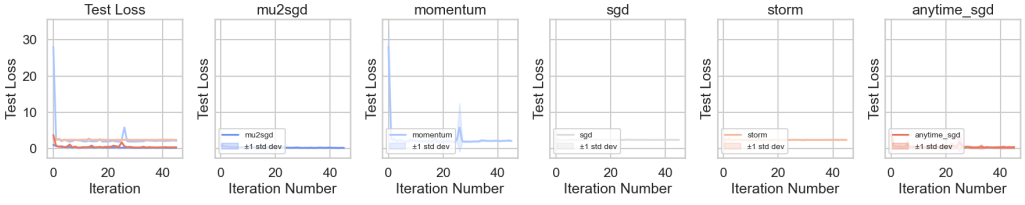


图 36：学习率为 = 10 时的测试损失（↓ 越小越好）。

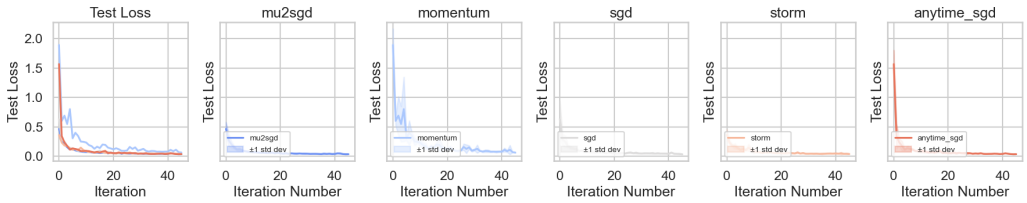


图 37：学习率为 $= 1$ 时的测试损失（↓ 越低越好）。

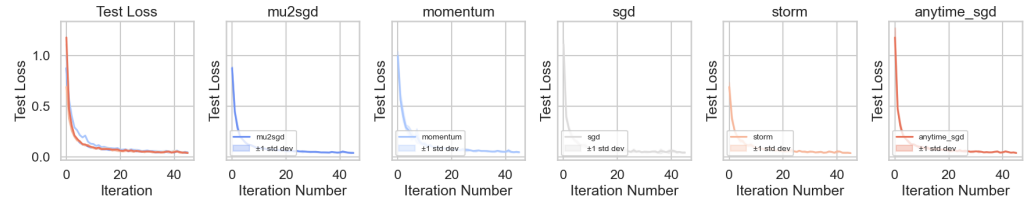


图 38：学习率为 $= 0.1$ 时的测试损失（↓ 越低越好）。

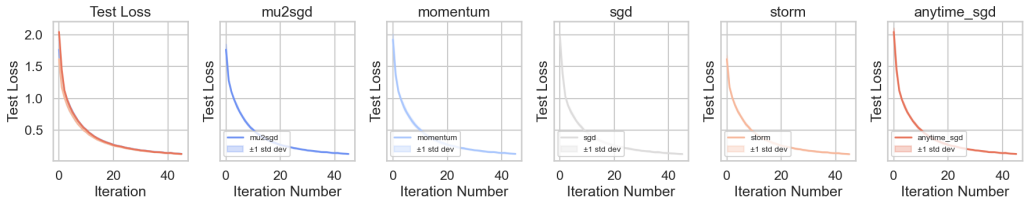


图 39：学习率为 $= 0.01$ 时的测试损失（↓ 越高越好）。

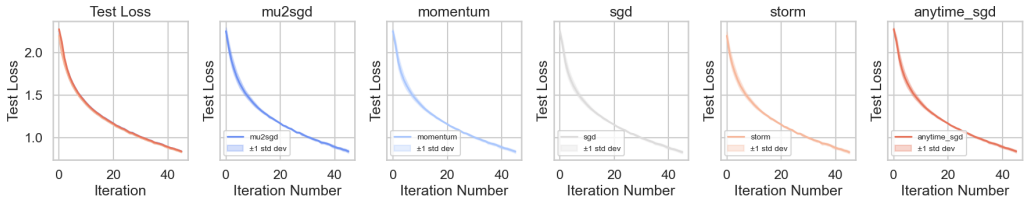


图 40：学习率为 $= 0.001$ 时的测试损失（↓ 越好）。

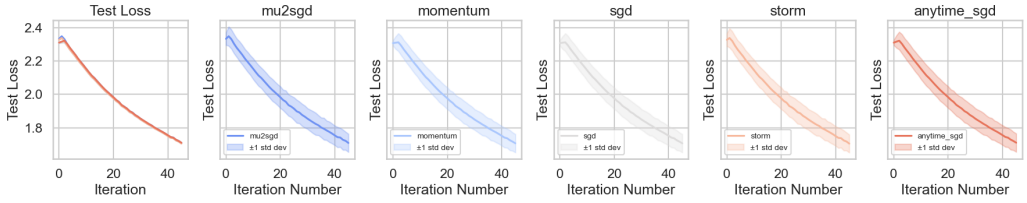


图 41：学习率为 $= 0.0001$ 时的测试损失（↓ 越好）。