arXiv:2302.03550v2 [math.OC] 6 Nov 2024

过参数化情形下动量随机梯度下降的指数收敛速率

BENJAMIN GESS 和 SEBASTIAN KASSING

ABSTRACT. 我们针对非凸优化问题，证明了动量随机梯度下降（MSGD）方案在任意固定超参数（学习率、摩擦参数）下，以及其连续时间对应过程的指数收敛速率的显式界。在小步长情形，以及极小值平坦或噪声强度较大的情况下，这些界表明，MSGD 比普通随机梯度下降（SGD）收敛更快。对于满足局部 Polyak-Łojasiewicz 不等式的目标函数，并在 MSGD 方差满足过参数化情形下成立的假设时，我们证明了上述结果。此外，我们分析了摩擦参数的最优选择，并证明 MSGD 过程几乎必然收敛到一个局部极小值。

1. INTRODUCTION

许多机器学习任务都涉及最小化函数 $f\colon \mathbb{R}^d \to \mathbb{R}$，该函数表示为关于随机变量 Γ 和非负损失 g 的期望 $f(x) = \mathbb{E}[g(x, \Gamma)]$。例如，在监督学习中，目标是最小化固定训练数据集上的平均损失。实际上，所用数据集规模庞大，因此需要采用随机优化方法，例如随机梯度下降（SGD）。这类方法在每次迭代中使用梯度 $\nabla f(x)$ 的随机近似，例如通过对 $\nabla g(x, \Gamma)$ 进行独立同分布抽样来实现。

机器学习中随机优化算法理论分析面临的第二个主要挑战是损失景观的非凸性。具体而言，使用神经网络进行监督学习时，目标函数通常具有丰富且非离散的全局极小值集合，参见例如 [Coo21,FGJ20, DK22b]。

经验观察 [SMDH13, GPS18, SGD21] 促使人们长期以来猜想，引入动量可以提升随机优化算法的性能。近年来，人们提出了大量优化算法，将动量的各种变体与自适应步长、预条件化和批归一化等其他技术结合起来 [Nes83, Qia99, DHS11, KB15]。然而，证明这些方法具有优势的理论结果寥寥无几。事实上，已知结果仅适用于确定性的连续时间系统 [Pol64, ADR22b, ADR22a, AGV22]，或具有强凸目标函数的确定性系统 [Pol64, GFJ15]。对于随机动量算法，现有文献仅给出定性结论 [GPS18, LY23]，并在凸情形下恢复了随机梯度下降的收敛速率 [GPS18, SGD21]。因此，在机器学习中常见的非凸损失景观下，为离散时间动量随机梯度下降（MSGD）的收敛速率推导显式界限，仍是一个开放问题。本文解决了这一问题。

2020 *Mathematics Subject Classification.* 主要分类 90C15；次要分类 68T07、90C26、62L20。
Key words and phrases. 动量随机梯度下降； Łojasiewicz 不等式；几乎必然收敛；过参数化；阻尼。

更确切地说，我们考虑 MSGD算法

$$\begin{aligned} X_{n+1} &= X_n + \gamma_{n+1} V_{n+1}, \\ V_{n+1} &= V_n - \gamma_{n+1}\mu V_n - \gamma_{n+1}\nabla g(X_n, \Gamma_{n+1}), \end{aligned} \tag{1}$$

对于初始值 X_0， $V_0 \in \mathbb{R}^d$、严格正实数序列 $(\gamma_n)_{n\in\mathbb{N}}$、摩擦参数 $\mu > 0$ 以及独立同分布序列 $(\Gamma_n)_{n\in\mathbb{N}}$，并推导出 $(f(X_n))_{n\in\mathbb{N}_0}$ 指数收敛速率的显式界。在小步长情形下，这些结果严格证明了如下猜想：在过参数化情形下，对于平坦极小值，引入动量会加速收敛，相较于随机梯度下降 [Woj23]，即若 $\min_{x\in\mathbb{R}^d} f(x) = 0$.[1]

事实上，我们处理了更一般的情形，并将 (1) 作为特例：我们始终假设 $f\colon \mathbb{R}^d \to \mathbb{R}$ 是一个可微函数，且其梯度为 C_L-Lipschitz 连续，[2] 对某个常数 $C_L \geq 0$ 成立，并满足 $\inf_{x\in\mathbb{R}^d} f(x) = 0$。设 $(\Omega, (\mathcal{F}_n)_{n\in\mathbb{N}_0}, \mathcal{F}, \mathbb{P})$ 为一个滤波概率空间，并设 $(X_n)_{n\in\mathbb{N}_0}$、$(V_n)_{n\in\mathbb{N}_0}$ 是 $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$-适应过程，对所有 $n \in \mathbb{N}_0$ 均满足

$$\begin{aligned} X_{n+1} &= X_n + \gamma_{n+1} V_{n+1}, \\ V_{n+1} &= V_n - \gamma_{n+1}\mu V_n - \gamma_{n+1}\nabla f(X_n) + \gamma_{n+1} D_{n+1}, \end{aligned} \tag{2}$$

其中 X_0， $V_0 \in L^2(\Omega\ \mathcal{F}_0)$，$(\gamma_n)_{n\in\mathbb{N}}$ 是一个严格为正的实数序列， $\mu > 0$ 且 $(D_n)_{n\in\mathbb{N}}$ 是关于滤波 $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$的 L^2-鞅差分序列。下文中，我们也将由式 (2) 给出的 $(X_n)_{n\in\mathbb{N}_0}$ 称为步长为 $(\gamma_n)_{n\in\mathbb{N}}$ 且摩擦参数为μ的 MSGD 方案。取 $(D_n)_{n\in\mathbb{N}} = (\nabla f(X_{n-1}) - \nabla g(X_{n-1}, \Gamma_n))_{n\in\mathbb{N}}$ 即可恢复算法 (1)。

我们在常步长情形下陈述主要结果的一个简化版本。

定理1.1。（见定理 *3.1* 和定理 *3.4*）设 $\gamma_n \equiv \gamma > 0$。设 $L > 0$ 且 $\sigma \geq 0$。设$\mathcal{D} \subset \mathbb{R}^d$ 为一个开集，并假设对于所有 $x \in \mathcal{D}$

$$|\nabla f(x)|^2 \geq 2Lf(x). \tag{3}$$

此外，对于 $n \in \mathbb{N}_0$，令 $\mathbb{A}_n = \{X_i \in \mathcal{D}$ 对所有 $i = 0, \ldots, n\}$ 成立，并假设

$$\mathbb{E}[|D_{n+1}|^2 | \mathcal{F}_n] \leq \sigma f(X_n), \quad on\ \mathbb{A}_n. \tag{4}$$

如果存在 参数 a、$b \geq 0$，使得 (11) 中的所有不等式均得到满足，那么 :

(i) 对于所有 $\varepsilon > 0$，有

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] = o((r_{\mathrm{MSGD}} - \varepsilon)^{-n}),$$

其中 r_{MSGD}：= 最小值$(1 + a\gamma、\ \delta^{-1})$，且 δ 由 (11)给出。

(ii) 如果 $\delta < 1$，则过程 $(X_n)_{n\in\mathbb{N}_0}$ 在 $\mathbb{A}_\infty$：$= \bigcap_{n\in\mathbb{N}_0} \mathbb{A}_n$上几乎必然收敛。

此外，对于固定的 μ 以及充分小的 γ，存在常数 a、b ，使得上述假设成立，且 $\delta < 1$。

定理1.1在两个主要假设下，对动量随机梯度下降的收敛速率进行了局部分析。首先，我们不采用凸性假设，而是使用局部梯度不等式（3），该不等式通常称为 Polyak-Łojasiewicz 不等式（PL 不等式）。其次，我们假设，随着过程趋近于临界点，随机扰动的方差趋于零。下文第 2 节将说明，在过参数化监督学习中，这些假设均成立。

请注意，本文的设定与近期其他研究 [LP16、YFL23、GTD23]有着根本性的不同。优化理论结果通常会比较在

[1]关于监督学习情形的讨论，参见第 2 节。[2]我们将在下文的注释 3.6 和注释 4.5 中讨论全局 Lipschitz 连续性的必要性。

超参数取最优值时的收敛速率。有人可能会认为，在实践中不可能选择最优超参数，因为问题参数 L，C_L 和 σ 是未知的。受此启发，我们分析固定超参数下的动量随机梯度下降。在下文的注3.5中，我们分析定理1.1中得到的严格收敛速率，并考察通常作为默认值选取的步长区间。为确保优化过程的稳健性，步长通常会选得较小。因此，我们将研究结果置于小步长情形下，并将定理1.1推导出的动量随机梯度下降收敛速率与随机梯度下降的收敛速率进行比较。

由于假设 (3) 和 (4) 仅被假定在局部成立，因此收敛速率以优化动力学始终处于 $\mathcal{D}$ 内这一事件为条件。然而，在假设 MSGD 从临界点附近且以较小的初始速度进行初始化的情况下，定理1.1中得到的估计可用于界定离开该定义域的概率，参见推论 3.3。此外，在集合 $\mathbb{A}_\infty = \bigcap_{n\in\mathbb{N}} \mathbb{A}_n$ 上，定理 3.1 证明了目标函数值几乎必然指数收敛至零，并且 $(X_n)_{n\in\mathbb{N}}$ 几乎必然指数收敛至一个临界点。

与定性收敛结果相比，推导收敛速率的显式界需要仔细选择合适的李雅普诺夫函数（见下文 (13)），并对超参数（例如摩擦参数 μ）以及定义李雅普诺夫函数的其他技术参数进行约束优化（见引理 3.11）。此外，定理 3.1 中假设的局部化依赖于对离开定义域 $\mathcal{D}$ 这一事件的详细控制，参见例如 (22) 和引理3.7。

在本文的第二部分，我们研究动量随机梯度下降方法的连续时间对应过程。进一步假设，f 二次连续可微，并且令 $\Sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d'}$ 为 Lipschitz 连续函数。设 $(\Omega, (\mathcal{F}_t)_{t\geq 0}, \mathcal{F}, \mathbb{P})$ 是满足通常条件的滤波概率空间，并考虑以下随机微分方程组

$$\begin{aligned} dX_t &= V_t\, dt, \\ dV_t &= -(\mu V_t + \nabla f(X_t))\, dt + \Sigma(X_t)\, dW_t, \end{aligned} \tag{5}$$

其中 $V_0, X_0 \in L^4(\Omega\ \mathcal{F}_0)$，$\mu > 0$ 且 $(W_t)_{t\geq 0}$ 是一个标准的 $\mathbb{R}^{d'}$ 值 $(\mathcal{F}_t)_{t\geq 0}$-布朗运动。

∇f 和 Σ 的 Lipschitz 连续性意味着存在唯一的连续 $\mathbb{R}^{2d}$ 值半鞅 $(X_t, V_t)_{t\geq 0}$ 满足 (5)。此外，对所有 $T \geq 0$，存在常数 $C > 0$，使得 $\mathbb{E}[\sup_{t\in[0,T]}(|X_t|^4 + |V_t|^4)] < C(1 + \mathbb{E}[|X_0|^4 + |V_0|^4])$, 参见例如 [LTE19], 中的定理19，因此对所有 $t \geq 0$，$\nabla f(X_t) \in L^4(\Omega)$ 且 $f(X_t) \in L^2(\Omega)$。我们证明，当目标函数 f 在开集 $\mathcal{D}$ 上满足 PL 条件时，$(f(X_t))_{t\geq 0}$ 指数收敛。对于适当选取的摩擦参数 μ，我们估计波动对最优收敛速率的影响，并将其与 [AGV22] 中重球常微分方程所得的收敛速率进行比较。关于系统 (5) 与随机梯度下降的连续时间版本 (29) 的收敛速率比较，请参见注4.3。

定理 1.2。（见定理 *4.2*）设 $L > 0$、$C_L^* = C_L \vee \frac{9}{8}L$ 和 $0 < \sigma < 4\sqrt{\frac{L}{C_L^*}}$。设 ~~$\mathcal{D}$~~ $\subset \mathbb{R}^d$ 是一个开集，使得对所有 $x \in \mathcal{D}$

$$|\nabla f(x)|^2 \geq 2Lf(x) \quad \textit{and} \quad \|\Sigma(x)\|_F^2 \leq \sigma f(x).$$

并选择

$$\mu = 2\sqrt{C_L^*} - \sqrt{C_L^* - L + \frac{1}{4}\sqrt{C_L^*}\sigma}.$$

那么，存在一个 $C \geq 0$ 使得

$$\mathbb{E}[\mathbb{1}_{\{T>t\}} f(X_t)] \leq C\exp(-mt), \quad \textit{for all } t \geq 0,$$

其中 $T =$ 下确界$\{t \geq 0 : \ X_t \notin \mathcal{D}\}$ 并且

$$m = 2\left(\sqrt{C_L^*} - \sqrt{C_L^* - L + \frac{1}{4}\sqrt{C_L^*}\sigma}\right).$$

概述 **相关文献：** 重球常微分方程解的收敛速率，即

$$\dot{x}_t = v_t, \quad \dot{v}_t = -\mu v_t - \nabla f(x_t), \tag{6}$$

对于具有 $\mu > 0$ 的情形，已有文献在损失景观的各种假设下推导出相关结果，最早可追溯至 Polyak [Pol63] 的工作。Polyak 证明，对于 L-强凸且二阶可微的函数 f，$(f(x_t))_{t\geq 0}$ 以速率 $\mu - \sqrt{\max(0, \mu^2 - 4L)}$ 收敛。选择$\mu = 2\sqrt{L}$ 可得到收敛速率 $2\sqrt{L}$。相比之下，对于梯度流常微分方程的解，即

$$\dot{y}_t = -\nabla f(y_t), \tag{7}$$

$(f(y_t))_{t\geq 0}$ 以速率 $2L$ 指数收敛。因此，对于在全局最小值附近相对平坦的目标函数 f，选择优化动力学 (6) 而非 (7) 更有利。1963 年，Polyak [Pol63] 和 Łojasiewicz [Loj63] 独立提出了梯度不等式 (3)，它是对强凸性假设的一种放宽。事实证明，对于 f，梯度满足利普希茨假设，再结合 (3)，仍足以证明梯度流 (7) 的解 $(f(y_t))_{t\geq 0}$ 以及重球常微分方程 (6) 的解 $(f(x_t))_{t\geq 0}$ 指数收敛，参见 [PS17]。后一个结果的证明依赖于一个李雅普诺夫函数，该函数包含动力系统的势能与动能之和，以及两者的一个交叉项。 [ADR22a] 在 L 准强凸函数这一情形下，针对具有唯一孤立极小值且梯度 Lipschitz 连续的函数，得到了摩擦参数 $\mu = 3\sqrt{L/2}$ 对应的收敛速率 $\sqrt{2L}$。此外，他们证明，对每个参数 $\mu < 3\sqrt{L/2}$，都存在一个 L-强凸目标函数 f（其梯度仅为 Hölder 连续），使得 $(f(x_t))_{t\geq 0}$ 至多以速率 $\frac{2}{3}\mu$ 收敛。注意，准强凸性蕴含 PL 不等式，参见 [ADR22b]。在 [AGV22], 中，研究者推导出满足 PL 不等式的函数的指数收敛速率，并证明，对于平坦的强凸函数，重球动力学相较于梯度流动力学也具有类似的优势。 [CEG07] 研究了具有时变摩擦参数的重球常微分方程。他们给出了摩擦衰减速率的充分必要条件，以保证过程收敛，并确定了凸目标函数下过程的 f 值。

近年来，PL 不等式因其简洁性、对几何学的深刻影响，以及适用于机器学习中出现的目标函数而受到广泛关注，例如参见 [KNS16,DK24,ADR22b, KSA23,Gar23,RB24,Woj23]。

对于离散时间重球算法，情况要复杂得多。需要区分两种根本不同的问题设置：第一，针对最优选择的超参数的收敛速率；第二，针对任意固定超参数的收敛速率。对于第一类问题，Polyak [Pol64] 的开创性工作证明，在优化二次函数并选择最优参数 γ，$\beta > 0$时，确定性重球方法比梯度下降具有更快的收敛速度。相反， [LP16、GTD23] 中给出的反例表明，对于更广泛的强凸目标函数类，即使选择最优步长，重球法也无法实现加速。此外， [YFL23] 证明，对于满足带参数 L 的 PL 不等式的目标函数类，不存在任何一阶方法能够通过选择最优步长实现加速。然而，工作 [DKP20]

找到了参数 γ 和 β，使得在 PL函数类上，重球法恢复了梯度下降所能达到的最佳收敛速率。

在这项工作中，我们考虑第二种截然不同的情形，即随机梯度和小步长情形。我们证明，在保守地选择步长时，也就是在小步长区间内，当算法收敛到平坦极小值时，以及在噪声强度较大时，MSGD 都能加速收敛，见注释3.5。正如上文所指出的，通常并不知道常数C_L，L 和 σ，因此实践者会选择足够小且随时间递减的步长，以至少保证收敛。

请注意，MSGD 过程是随机重球法（SHB）的一种轻微变体。SHB 通过加入随机噪声，将 Polyak 的重球法推广到随机情形。根据 [GPS18]，SHB 过程由以下迭代方案定义：

$$
\begin{aligned}
X_{n+1} &= X_n + \gamma_{n+1} V_{n+1}, \\
V_{n+1} &= V_n - \gamma_{n+1}\mu(\nabla f(X_n) + V_n - D_{n+1}).
\end{aligned} \tag{8}
$$

可以证明，式(8)是式(6)的一种离散化，并附加了一个扰动；其中，步长为 γ_n 的一次迭代对应于式(6)经过时间 $\sqrt{\gamma_n/\mu}$后的位置。因此，与MSGD方案(2)中直接进行的时间离散化相比，SHB过程在小步长下加快了相应常微分方程的时间进程。Nesterov加速中也会出现类似现象。在 [EBB+21] 中，作者提出了一种采用指数停止时间的连续化过程，从而无需额外的时间变换，便可将离散过程与相应的连续时间对应过程进行比较。在凸情形下，SHB的收敛速率可参见 [GPS18、SGD21]。特别是，[GPS18] 在欠参数化情形下，对比SGD [RM51]，针对更广泛的一类步长恢复了最优的 $\mathcal{O}(1/n)$收敛速率；类似现象也见于Ruppert-Polyak平均化 [DK23]。[LR17、LR20] 推导出SHB求解带随机范数的线性系统时的（加速）指数收敛速率。在这种情形下，随着SHB趋近最优点，随机梯度会趋于零，这与我们的假设(4)相当。与随机梯度下降类似，SHB能够避开严格鞍点 [LY23]，并且在经典噪声假设下，对解析目标函数收敛 [DK24]。

在 [LTE19] 中，已有研究表明，对于适当选取的扩散矩阵 Σ，随机微分方程 (5) 在有限时间区间内是 MSGD 过程的弱近似。对于随机梯度下降的连续时间对应过程，Wojtowytsch [Woj24] 表明，在过参数化情形下，噪声的特殊结构会使该过程倾向于选择一个平坦极小值。人们普遍认为，平坦极小值具有更好的泛化能力；例如，参见 [KMN+16] 关于泛化差距和极小值尖锐度的数值实验。在平均场尺度下，已有研究表明，随机梯度下降的动力学会收敛到保守型随机偏微分方程的解，参见 [GGK22, GKK24]。Hu等人 [HLZ19] 研究了类似于 (5) 中定义的随机微分方程在严格鞍点附近的行为。

本文结构如下：第 2 节基于过参数化监督学习，阐述目标函数及随机噪声大小相关假设的动机。第 3 节证明关于离散时间下 MSGD 过程的结果。第 4 节证明关于 (5) 中定义的连续时间对应过程的结果。

记号：我们用 $v^\dagger$ 表示向量 $v \in \mathbb{R}^d$的转置，用 $A^\dagger$ 表示矩阵$A \in \mathbb{R}^{n\times k}$ 的转置，并分别用 $\|A\|_F$和 $\|A\|$表示A的 Frobenius 范数和算子范数。此外，$|\cdot|$ 表示标准欧几里得范数，$\langle \cdot, \cdot \rangle$ 表示欧几里得空间上的标准标量积。

2. Loss landscape and noise in empirical risk minimization

在本节中，我们将针对机器学习应用，阐述损失景观和随机噪声的主要假设。具体而言，我们考虑监督学习中的一个二次损失函数回归问题。设 $(\theta_1,\zeta_1)\dots,(\theta_N,\zeta_N)\in\mathbb{R}^{d_{\text{in}}}\times\mathbb{R}^{d_{\text{out}}}$ 为给定的训练数据集。我们选择一个参数化假设空间 $\mathcal{S}:=\{\mathfrak{N}^x(\cdot):\ x\in\mathbb{R}^d\}$，其由函数 $\mathfrak{N}^x(\cdot):\mathbb{R}^{d_{\text{in}}}\to\mathbb{R}^{d_{\text{out}}}$ 组成，并且对于所有 $i=1,\dots,N,\quad x\mapsto\mathfrak{N}^x(\theta_i)$均可微。例如，可以将 $\mathcal{S}$ 选择为具有固定架构的全连接前馈神经网络的响应函数空间。风险最小化（相对于平方损失）的目标是选择一个合适的模型 $\mathfrak{N}^x(\cdot)$，使经验风险最小化

$$f(x)=\frac{1}{2N}\sum_{i=1}^{N}|\mathfrak{N}^x(\theta_i)-\zeta_i|^2,\quad x\in\mathbb{R}^d.$$

为了推导出如 (2) 中的动力系统，我们选取确定性的初始值 $X_0,\ V_0\in\mathbb{R}^d$、一个严格正实数序列 $(\gamma_n)_{n\in\mathbb{N}}$、一个独立同分布序列 $(I_n)_{n\in\mathbb{N}}$，使得 I_n 在 $\{1,\dots,N\}$ 上均匀分布，并考虑如下动力系统

$$\begin{aligned}X_{n+1}&=X_n+\gamma_{n+1}V_{n+1},\\ V_{n+1}&=V_n-\gamma_{n+1}\mu V_n-\frac{1}{2}\gamma_{n+1}\nabla\big(|\mathfrak{N}^x(\theta_{I_{n+1}})-\zeta_{I_{n+1}}|^2\big)\big|_{x=X_n}.\end{aligned}$$

通过如下选择，我们可以得到 (2)

$$D_{n+1}=\nabla f(X_n)-\frac{1}{2}\nabla\big(|\mathfrak{N}^x(\theta_{I_{n+1}})-\zeta_{I_{n+1}}|^2\big)\big|_{x=X_n}.$$

我们令 $(\mathcal{F}_n)_{n\in\mathbb{N}_0}=(\sigma(I_1,\ \dots,\ I_n))_{n\in\mathbb{N}_0}$，并注意到，对于所有 $n\in\mathbb{N},\ \mathbb{E}[D_{n+1}|\mathcal{F}_n]=0$ 以及

$$\mathbb{E}[|D_{n+1}|^2|\mathcal{F}_n]\le C\sum_{i=1}^{N}|(\mathfrak{N}^{X_n}(\theta_i)-\zeta_i)\nabla\mathfrak{N}^x(\theta_i)\big|_{x=X_n}|^2,$$

其中 $C>0$为常数。在定义域 $\mathcal{D}\subset\mathbb{R}^d$ 上，若梯度 $\nabla_x\mathfrak{N}^x(\theta_i)$ 对所有 $i=1,\ \dots,\ N$均有界，则这意味着

$$\mathbb{E}[|D_{n+1}|^2|\mathcal{F}_n]\le\sigma f(X_n),$$

对于某个常数 $\sigma\ge0$。类似地，梯度 ∇f满足 $|\nabla f(x)|^2\le Cf(x)$，即逆 PL 不等式，其中常数为 $C\ge0$，且定义域相同，为 $\mathcal{D}$。

接下来我们说明 PL 不等式的动机。若存在一个 $y\in\mathbb{R}^d$，使得 $f(y)=0$，则称该回归问题为过参数化回归问题。Cooper [Coo21] 证明了以下结果：对于满足 $d>Nd_{\text{out}}$，且 $\mathfrak{N}(p)$ 为 C^k-光滑的过参数化回归问题，其中 $k\ge d-Nd_{\text{out}}+1$ 且对所有 $p\in\mathbb{R}^{d_{\text{in}}}$均成立：对于几乎所有训练数据元组（除去一个勒贝格零测集），全局极小值集合 $\mathcal{M}:=\{x\in\mathbb{R}^d:\ f(x)=0\}$ 构成 $\mathbb{R}^d$中的一个闭合的（$d-Nd_{\text{out}}$）维 C^k子流形。如果这样的 $\mathcal{M}$ 是一个 C^2-流形，并且对于某个 $y\in\mathcal{M}$，有 $\dim(\text{Hess } f(y))=Nd_{\text{out}}$，那么 [Fee19] 中的定理 2.1 表明，存在 y的一个邻域 $U\subset\mathbb{R}^d$，使得 PL 不等式在 U上成立，即存在一个 $L>0$，使得对所有 $x\in U$均有 $2Lf(x)\le|\nabla f(x)|^2$。

本节的最后一个结果是针对具有 Lipschitz连续梯度的函数$f:\mathbb{R}^d\to\mathbb{R}$ 的逆 PL 不等式的一个一般版本。这一观察已由[Woj23], 提出，参见其中的引理B.1。我们通过仅假设局部极小点周围一个球形区域内满足局部Lipschitz连续性，放宽了这些假设。我们将在后续各节中反复使用这一引理。

引理 2.1。 设 $r > 0, y \in \mathbb{R}^d$，并假设 ∇f 在 $B_r(y)$ 上是 C_L-*Lipschitz* 连续的，且$\inf_{x\in B_r(y)} f(x) = f(y)$。那么，对所有 $x \in B_{r/2}(y)$，均有

$$|\nabla f(x)|^2 \le 2C_L(f(x) - f(y)). \tag{9}$$

证明。 由于 y 是 f 的临界点，因此对所有 $x \in B_{r/2}(y)$

$$|\nabla f(x)| = |\nabla f(x) - \nabla f(y)| \le \frac{C_L r}{2}.$$

如果 $\nabla f(x) = 0$，则该结论显然成立。如果 $\nabla f(x) \neq 0$，考虑函数

$$g(t) = f\Big(x - t\frac{\nabla f(x)}{|\nabla f(x)|}\Big).$$

注意，对于 $x \in B_{r/2}(y)$ 以及所有 $t \in [0, \frac{|\nabla f(x)|}{C_L}]$，都有 $x - t\frac{\nabla f(x)}{|\nabla f(x)|} \in B_r(y)$。因此，利用 ∇f 的 Lipschitz 连续性，并且由于 y 是局部极小值

$$\begin{aligned} f(y) - f(x) &\le g\Big(\frac{|\nabla f(x)|}{C_L}\Big) - g(0) = \int_0^{\frac{|\nabla f(x)|}{C_L}} g'(s)\, ds \\ &\le \frac{|\nabla f(x)|}{C_L} g'(0) + \frac{|\nabla f(x)|^2}{2C_L} = -\frac{|\nabla f(x)|^2}{2C_L}. \end{aligned}$$

□

注记 2.2。 对于满足 PL 不等式且具有 C_L-Lipschitz 连续梯度的函数 $f : \mathbb{R}^d \to \mathbb{R}$，由引理 2.1 可得

$$2L(f(x) - f(y)) \le |\nabla f(x)|^2 \le 2C_L(f(x) - f(y)), \tag{10}$$

其中 $x \in \mathbb{R}^d$ 和 $y \in \mathbb{R}^d$ 是 f 的全局最小值，且 $f(y) = 0$。因此，我们立即得到 $C_L \ge L$。在严格凸情形下

$$f(x) = \frac{1}{2}x^\dagger A x,$$

对于正定矩阵 $A \in \mathbb{R}^{d\times d}$，(10) 中的常数 C_L 和 L 分别对应 A 的最大特征值和最小特征值。

3. Momentum stochastic gradient descent in discrete time

本节中，我们考虑在 (2) 中引入的 MSGD 方案$(X_n)_{n\in\mathbb{N}_0}$。我们陈述本节的主要结果。首先，我们证明，对于足够小的步长，目标函数值在数值时间 $(t_n)_{n\in\mathbb{N}_0} = (\sum_{i=1}^n \gamma_i)_{n\in\mathbb{N}_0}$ 上指数收敛。这意味着 MSGD 过程本身几乎必然收敛。

定理 3.1. 设 $L > 0, \sigma \ge 0$。设 $\mathcal{D} \subset \mathbb{R}^d$ 为一个开集，并假设

$$|\nabla f(x)|^2 \ge 2Lf(x)$$

对所有 $x \in \mathcal{D}$ 均成立。此外，对于 $n \in \mathbb{N}_0$，令 $\mathbb{A}_n = \{X_i(\omega) \in \mathcal{D}$ 对所有 $i = 0, \dots, n\}$ 均成立，并假设

$$\mathbb{E}[|D_{n+1}|^2|\mathcal{F}_n] \le \sigma f(X_n), \quad on\ \mathbb{A}_n.$$

存在 $\bar\gamma > 0$，使得若 上确界$_{n\in\mathbb{N}} \gamma_n \le \bar\gamma$，则有：

(i) 存在 C、$m > 0$，使得对所有 $n \in \mathbb{N}$，我们都有

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \le C\exp(-mt_n),$$

where $t_n = \sum_{i=1}^n \gamma_i$.

(ii) 令 $m' < m$，并假设 $\sum_{i=0}^{\infty} \exp((m'-m)t_i) < \infty$。那么，在 $\mathbb{A}_\infty = \bigcap_{n\in\mathbb{N}_0} \mathbb{A}_n$ 上，我们有 $\exp(m't_n)f(X_n) \to 0$，几乎必然成立。(iii) 过程 $(X_n)_{n\in\mathbb{N}_0}$ 在 $\mathbb{A}_\infty$上几乎必然收敛。

对于步长 $(\gamma_n)_{n\in\mathbb{N}}$ 满足 $\gamma_n \to 0$的情形，存在 $N \in \mathbb{N}$，使得 $\sup_{n>N} \gamma_n$ 足够小，从而可以对从时间 N开始的系统 $(X_n)_{n\geq N}$ 应用定理 3.1。然而，只要 γ 足够小，定理 3.1 同样适用于常数步长序列 $\gamma_n \equiv \gamma$。注意，由于 X_0，$V_0 \in L^2(\Omega\ \mathcal{F}_0)$，$\nabla f$的 Lipschitz 连续性以及关于过程 $(D_n)_{n\in\mathbb{N}}$ 的假设意味着，对于所有 $n \in \mathbb{N}$，我们都有 $1\!\!1_{\mathbb{A}_{n-1}}X_n$、$1\!\!1_{\mathbb{A}_{n-1}}V_n$、$1\!\!1_{\mathbb{A}_{n-1}}\nabla f(X_n) \in L^2(\Omega)$ 和 $1\!\!1_{\mathbb{A}_{n-1}}f(X_n) \in L^1(\Omega)$。定理 3.1 中的收敛速率 m取决于 L，C_L，σ，μ 以及 $\sup_{n\in\mathbb{N}} \gamma_n$。

接下来，我们在小步长区间内对 μ 所有摩擦参数进行优化。我们以数值时间 $t_n = \sum_{i=1}^{n} \gamma_i$ 为变量，重新得到 [AGV22] 推导出的重球常微分方程 (6) 的收敛速率。由于随机梯度下降 (2) 与重球常微分方程 (6) 在非凸目标函数上的比较结果仅在有限时间区间内成立（例如，参见 [WKM23、LTE19、GG22],），因此连续时间结果无法推广到离散时间设置。事实上，在分析动量随机梯度下降时，由于其离散特性，会出现额外的误差项。因此，证明需要进行最坏情况分析，对允许的步长范围内这些误差项进行界定。定理 3.2 促使我们在小学习率区间内比较动量随机梯度下降与随机梯度下降，参见注3.5。

定理 3.2。设$\kappa = \frac{C_L}{L}$。令—

$$\mu \in \begin{cases} \left[\frac{1}{\sqrt{8}}\left(5-\sqrt{9-8\kappa}\right)\sqrt{L}, \frac{1}{\sqrt{8}}\left(5+\sqrt{9-8\kappa}\right)\sqrt{L}\right], & \text{if } \kappa < \frac{9}{8}, \\ \left\{(2\sqrt{\kappa}-\sqrt{\kappa-1})\sqrt{L}\right\}, & \text{if } \kappa \geq \frac{9}{8}. \end{cases}$$

那么，在定理 *3.1* 的假设下，对每个 $\varepsilon > 0$，存在 C、$\bar{\gamma} \geq 0$，使得若$\sup_{n\in\mathbb{N}} \gamma_n \leq \bar{\gamma}$，则有

$$\mathbb{E}[1\!\!1_{\mathbb{A}_{n-1}} f(X_n)] \leq C\exp(-(m-\varepsilon)t_n), \quad \text{for all } n \in \mathbb{N},$$

其中

$$m = \begin{cases} \sqrt{2L}, & \text{if } \kappa < \frac{9}{8}, \\ 2(\sqrt{\kappa}-\sqrt{\kappa-1})\sqrt{L}, & \text{if } \kappa \geq \frac{9}{8}. \end{cases}$$

利用定理 3.1 证明中的估计，我们可以对 $(X_n)_{n\in\mathbb{N}_0}$ 离开定义域 $\mathcal{D}$ 的概率进行界定，前提是其初始位置接近某个全局最小值，且初始速度较小。

推论 3.3. 设 $y \in \mathcal{D}$ 满足 $f(y) = 0$。那么，在定理 *3.1* 的假设下，对每个 $\varepsilon > 0$，存在一个 $r_0 > 0$，使得若 $X_0 \in B_{r_0}(y)$ 几乎必然成立，且 $\mathbb{E}[|V_0|^2] \leq r_0$，则有

$$\mathbb{P}(\mathbb{A}_\infty^c) \leq \varepsilon.$$

我们的结果基于以下定理，该定理推导了在求解约束优化问题的条件下的指数收敛速率。

定理 3.4。 设 $\gamma_n \equiv \gamma$ 对于一个 $\gamma > 0$。设 $a,\ b \geq 0$ 并假设

$$
\begin{aligned}
0 \geq & -1+\gamma\Big(\frac{b}{2}-a\Big)+\gamma^2 C_L+\gamma^3\frac{C_L a}{2},\\
0 \geq & a\mu+ab-a^2-2L+\gamma\Big(\frac{b\sigma}{2}+a(2C_L-\mu b+\mu a)-2L(a-\frac{b}{2})\Big)\\
& +\gamma^2 C_L\Big(\sigma+a^2-2a\mu+2L\Big)+\gamma^3 C_L a\Big(\frac{\sigma}{2}-a\mu+L\Big),\\
0 \geq & C_L-\frac{b}{2}(\mu+a-b)+\gamma\Big(\frac{b\mu^2}{2}+\frac{C_L a}{2}-2C_L\mu+\frac{ba\mu}{2}-\frac{b^2\mu}{2}+C_L b\Big)\\
& +\gamma^2 C_L\Big(\mu^2-a\mu+\frac{ba}{2}-b\mu\Big)+\gamma^3\frac{C_L a\mu}{2}(\mu-b),\\
0 \leq & \delta := 1+\gamma(a-\mu-b)+\gamma^2(b\mu-a\mu-2C_L)+\gamma^3(2C_L\mu-C_L a)+\gamma^4 C_L a\mu,\\
& ab \geq C_L \quad \textit{and} \quad \gamma \leq \frac{b}{C_L}.
\end{aligned}
\tag{11}
$$

于是，在定理 *3.1* 的假设下，存在常数 $C \geq 0$，使得

$$
\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \leq \begin{cases} C(1+a\gamma)^{-n}, & \textit{if } 1+a\gamma < \delta^{-1} \\ C\delta^n, & \textit{if } 1+a\gamma > \delta^{-1} \\ C(1+a\gamma)^{-n} n, & \textit{if } 1+a\gamma = \delta^{-1} \end{cases}, \quad \textit{for all } n \in \mathbb{N}.
$$

对于固定的 $\mu > 0$ 和充分小的 γ，可以选择参数 $a,\ b > 0$，使得上述所有不等式均得到满足，且 $\delta < 1$。我们将在下文的分析中对此进行精确说明，详见命题 3.9 和引理 3.10。

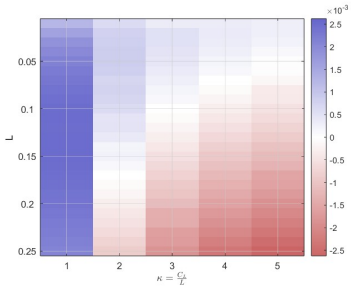
注3.5。 我们将定理 3.4 中证明的 MSGD 收敛速率与 [Woj23] 中得到的 SGD 收敛速率进行比较；后者与 [KNS16] 中无噪声情形 $\sigma = 0$ 的结果一致（另见 [VBS19, KR23]）。定理 3.4 表明，对于所有 $\varepsilon > 0$，都有

$$
\limsup_{n\to\infty}(r_{\mathrm{MSGD}}-\varepsilon)^n \mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] = 0, \tag{12}
$$

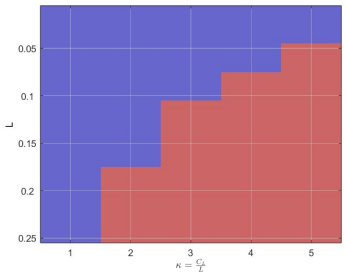
其中，r_{MSGD} 是 $r(a,\ b) := \min(1+a\gamma,\ \delta^{-1})$ 的最大值，适用于所有满足 (11) 的 $a,\ b \geq 0$。[Woj23] 在目标函数和随机噪声满足相同假设的条件下，即满足 (12) 的意义下，给出了 SGD 的收敛速率 $r_{\mathrm{SGD}} = 1-2L\gamma+\gamma^2 C_L(2L+\sigma)^2$，适用于所有满足 $\gamma < \frac{2L}{2L+\sigma}\frac{2}{C_L}$ 的步长。

首先，我们固定 $\gamma = 0.01$，这是步长 [Ben12]，的常用默认值，并比较无噪声情形下（图 1）随机梯度下降的收敛速率 r_{SGD} 与采用最优摩擦参数 μ^* 的动量随机梯度下降的收敛速率 r_{MSGD}，以及高噪声强度下的情况（图 2）。

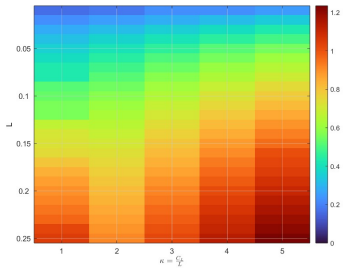
我们观察到，在无噪声情形 $\sigma = 0$下，对于平坦目标函数（即 L较小的情况），动量随机梯度下降的表现优于随机梯度下降。噪声强度 $\sigma = 100$较高时，动量随机梯度下降更加稳健。随机梯度下降仅在条件数较小（$\kappa < 4$）时收敛，而动量随机梯度下降能够适应噪声强度，并在所有给定情形下以指数速率收敛。



$r_{\mathrm{MSGD}} - r_{\mathrm{SGD}}$

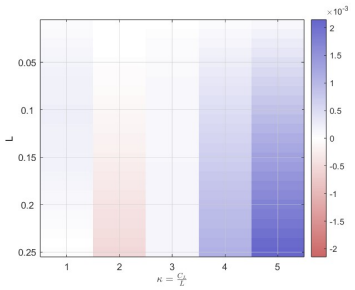


蓝色： $r_{\mathrm{MSGD}} > r_{\mathrm{SGD}}$
红色： $r_{\mathrm{MSGD}} \le r_{\mathrm{SGD}}$

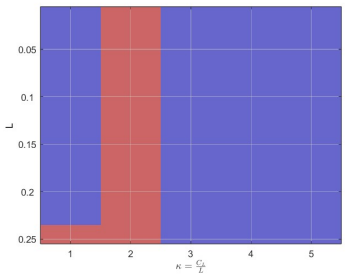


最优摩擦参数 μ^*

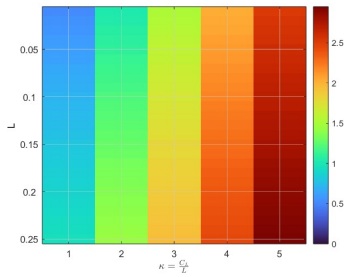
FIGURE 1. MSGD 与 SGD 收敛速率的比较 r_{MSGD}，其中 MSGD 的收敛速率 r_{SGD} 和 SGD 的收敛速率按照式 (12) 定义，固定 $\gamma = 0.01$ 和 $\sigma = 0$，取不同的 L (y轴) 和$\kappa = \frac{C_L}{L}$ (x轴)，并采用最优摩擦参数 μ^*。蓝色表示 MSGD 表现更优，红色表示 SGD 表现更优。



$r_{\mathrm{MSGD}} - r_{\mathrm{SGD}}$



蓝色： $r_{\mathrm{MSGD}} > r_{\mathrm{SGD}}$
红色： $r_{\mathrm{MSGD}} \le r_{\mathrm{SGD}}$



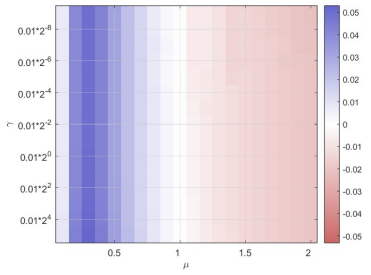
最优摩擦参数 μ^*

FIGURE 2. 在固定 $\gamma = 0.01$ 和 $\sigma = 100$ 的情况下，比较 r_{MSGD} （动量随机梯度下降）与 r_{SGD} （随机梯度下降）按照（12）定义的收敛速率；其中 L （y 轴）和 $\kappa = \frac{C_L}{L}$ （x 轴）取不同值，摩擦参数 μ^* 取最优值。对于 $\kappa \ge 4$，有 $r_{\mathrm{SGD}} < 1$，因此随机梯度下降不收敛。

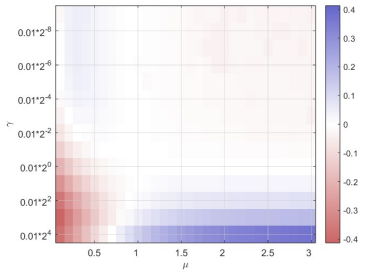
请注意， r_{MSGD} 是对 $\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)]$收敛速率严格的理论上界。为了推导 r_{MSGD} ，必须求解一个约束优化任务，参见定理 3.4。该约束优化任务通过 Matlab 中使用内点法的 *fmincon* 函数执行。因此，图 1 和图2中 r_{MSGD} 可能被低估。

我们还在问题参数 $L = \frac{1}{50}$、 $C_L = \frac{3}{50}$ 和 $\sigma = 0$ 分别固定为 $\sigma = 100$ 的情况下，比较 r_{MSGD} 与 r_{SGD}，见图 3。我们观察到，在小步长情形且没有随机噪声（$\sigma = 0$）时，存在一个较大的摩擦参数区间，在此区间内动量随机梯度下降优于随机梯度下降。当噪声强度较大（$\sigma = 100$）时，动量随机梯度下降在中等步长区间的优势最为显著。

特别地，较大的步长会导致相应连续时间模型中的噪声强度较大。研究表明，在这种情况下，该模型优于连续时间随机梯度下降，参见备注 4.3。然而，只有步长足够小时，这种启发式比较才可行。



$\frac{r_{\mathrm{MSGD}}-r_{\mathrm{SGD}}}{\gamma}$ 对于 $\sigma=0$



$\frac{r_{\mathrm{MSGD}}-r_{\mathrm{SGD}}}{\gamma}$ 对于 $\sigma=100$

FIGURE 3. 在式 (12) 的意义下，对固定的 $L=\frac{1}{50}$、$C_L=\frac{3}{50}$，比较 MSGD 的收敛速率 r_{MSGD} 与 SGD 的收敛速率 r_{SGD}，其中 γ（y 轴）和 μ（x 轴）取不同值。图中显示的是 $(r_{\mathrm{MSGD}}-r_{\mathrm{SGD}})/\gamma$ 的值。

备注 3.6. 我们讨论当随机噪声几乎必然有界时，如何弱化关于 ∇f 的全局利普希茨假设。设 $(\mathbb{A}_n)_{n\in\mathbb{N}_0}$ 由 $\mathbb{A}_n=\{X_i\in B_r(y) \text{ 对所有 } i=0,\dots,n\}$ 给出，其中 $y\in\mathbb{R}^d$ 为全局最小值，并假设存在一个 $\bar\gamma>0$，使得 $\sup_{n\in\mathbb{N}}\gamma_n\le\bar\gamma$ 且 $\bar\gamma\mu\le 1$。此外，假设存在确定性常数 $C_f, C_D\ge 0$，使得 $|\nabla f(x)|\le C_f$ 对所有 $x\in B_r(y)$ 均成立，并且对所有 $n\in\mathbb{N}_0$，我们几乎必然有 $|D_{n+1}|\le C_D$，且该结论在 $\{X_n\in B_r(y)\}$ 上成立。令 $C_V=\frac{C_f+C_D}{\mu}$，并假设 $|V_0|\le C_V$ 几乎必然成立。那么，在上述所有结论中，只需假设 ∇f 在 $\overline{B_{(r+\bar\gamma C_V)\vee 2r}(y)}$ 上是 C_L-Lipschitz 连续的，并且对所有 $x\in B_{r+\bar\gamma C_V}(y)$，$0\le f(x)$。

事实上，简单的归纳论证表明，对于所有 $n\in\mathbb{N}_0$，$|V_{n+1}|\le C_V$，因此，在事件 $\mathbb{A}_{n-1}$ 上，$X_{n+1}\in B_{r+\bar\gamma C_V}(y)$ 几乎必然成立。现在，所有泰勒估计（例如见式 (17)）都成立，前提是 ∇f 在 $B_{r+\bar\gamma C_V}(y)$ 上是 C_L-Lipschitz 连续的。此外，∇f 在 $B_{2r}(y)$ 上的 Lipschitz 连续性意味着 $B_r(y)$ 上的逆 PL 不等式，见引理 2.1。

3.1. 李雅普诺夫估计。 令 a，$b>0$ 并令 $(E_n)_{n\in\mathbb{N}_0}$ 为 $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$-适应随机过程，该过程定义为

$$E_n=af(X_n)+\langle\nabla f(X_n),V_n\rangle+\frac{b}{2}|V_n|^2. \tag{13}$$

在我们的设定中，$(E_n)_{n\in\mathbb{N}_0}$ 扮演随机李雅普诺夫函数的角色。尽管一般而言，$(E_n)_{n\in\mathbb{N}_0}$ 可能取负值，但假设逆PL条件，则存在 a 和 b 的取值，使得 $(E_n)_{n\in\mathbb{N}_0}$ 是一个非负过程。

引理 3.7。 令 a，b，$C_L>0$ 以及

$$E(x,y)=af(x)+\langle\nabla f(x),y\rangle+\frac{b}{2}|y|^2.$$

若 $ab\ge C_L$，则对于满足 $2C_Lf(x)\ge|\nabla f(x)|^2$ 的所有 $x\in\mathbb{R}^d$，我们都有 $E(x,y)\ge 0$，对于所有 $y\in\mathbb{R}^d$。

证明。若 $\nabla f(x)=0$，则结论显然成立。若 $\nabla f(x)\ne 0$，记 $\rho=\frac{|y|}{|\nabla f(x)|}$，可得

$$E(x,y)\ge\Big(\frac{a}{2C_L}-\rho+\frac{b}{2}\rho^2\Big)|\nabla f(x)|^2.$$

二次函数 $\varphi(\rho) = \frac{a}{2C_L} - \rho + \frac{b}{2}\rho^2$ 在 $\rho = \frac{1}{b}$ 处取得全局最小值，利用$ab \geq C_L$ 可推得

$$\varphi\Big(\frac{1}{b}\Big) = \frac{a}{2C_L} - \frac{1}{2b} \geq 0.$$

□

在下一个命题中，我们利用李雅普诺夫过程$(E_n)_{n\in\mathbb{N}_0}$，推导 MSGD 方案的收敛结论。

命题 3.8. 设 $L, a, b > 0, \sigma \geq 0$。设 $(\mathbb{A}_n)_{n\in\mathbb{N}_0}$ 为递减事件序列，使得对所有 $n \in \mathbb{N}_0$, $\mathbb{A}_n \in \mathcal{F}_n$，在 $\mathbb{A}_n$ 上均有

$$|\nabla f(X_n)|^2 \geq 2Lf(X_n) \quad and \quad \mathbb{E}[|D_{n+1}|^2|\mathcal{F}_n] \leq \sigma f(X_n). \tag{14}$$

设 $(\alpha_n)_{n\in\mathbb{N}}$,$(\beta_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}}$ 和 $(\epsilon_n)_{n\in\mathbb{N}}$ 由 *(20)* 给出。假设 $(\beta_n)_{n\in\mathbb{N}}$, $(\alpha_{n+1} - a\delta_{n+1} + 2\beta_{n+1}L)_{n\in\mathbb{N}_0}$, $(\epsilon_{n+1} - \frac{b}{2}\delta_{n+1})_{n\in\mathbb{N}_0}$ 和 $(\frac{C_L}{2}\gamma_n^2 - \frac{b}{2}\gamma_n)_{n\in\mathbb{N}_0}$ 均为非正实数序列，而 $(\delta_n)_{n\in\mathbb{N}}$ 为非负实数序列。此外，若 $\mathbb{P}(\bigcap_{n\in\mathbb{N}_0}\mathbb{A}_n) < \mathbb{P}(\mathbb{A}_0)$，还需假设 $ab \geq C_L$。则对所有 $n \in \mathbb{N}$，均有

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \leq{}& \Big(\prod_{i=1}^{n}(1+a\gamma_i)^{-1}\Big) \\ &\Big(\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} f(X_0)] + \sum_{i=1}^{n}\frac{\gamma_i}{1+a\gamma_i}\Big(\prod_{j=1}^{i}(1+a\gamma_j)\Big)\Big(\prod_{j=1}^{i}\delta_j\Big)\mathbb{E}[\mathbb{1}_{\mathbb{A}_0}E_0]\Big).\end{aligned} \tag{15}$$

证明。 首先，我们推导李雅普诺夫过程 $(E_n)_{n\in\mathbb{N}_0}$ 的期望的收敛速率。为此，我们分别考察式 (13) 中三个加项的时间演化。

首先，我们考察 $(f(X_n))_{n\in\mathbb{N}_0}$ 的演化。令 x, $y \in \mathbb{R}^d$ 并注意到，利用 ∇f 的 Lipschitz 连续性，我们得到

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{C_L}{2}|y - x|^2. \tag{16}$$

现在，对于 $n \in \mathbb{N}_0$，我们结合 $x = X_n$、 $y = X_{n+1}$ 和式 (2) 使用式 (16)，得到

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\mathbb{A}_n} f(X_{n+1})] \leq{}& \mathbb{E}\Big[\mathbb{1}_{\mathbb{A}_n}\big(f(X_n) - \gamma_{n+1}^2|\nabla f(X_n)|^2 \\ &+ (\gamma_{n+1} - \gamma_{n+1}^2\mu)\langle \nabla f(X_n), V_n\rangle + \frac{C_L}{2}\gamma_{n+1}^2\big|V_{n+1}\big|^2\big)\Big].\end{aligned} \tag{17}$$

接下来，我们控制 $(|V_n|^2)_{n\in\mathbb{N}_0}$ 的演化。利用式 (14)，得到

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}|V_{n+1}|^2] \leq{}& \mathbb{E}[\mathbb{1}_{\mathbb{A}_n}((1 - 2\gamma_{n+1}\mu + \gamma_{n+1}^2\mu^2)|V_n|^2 + \gamma_{n+1}^2|\nabla f(X_n)|^2 \\ &- 2(\gamma_{n+1} - \gamma_{n+1}^2\mu)\langle V_n, \nabla f(X_n)\rangle + \gamma_{n+1}^2\sigma f(X_n))].\end{aligned} \tag{18}$$

最后，

$$\begin{aligned}\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}\langle \nabla f(X_{n+1}), V_{n+1}\rangle] &= \mathbb{E}[\mathbb{1}_{\mathbb{A}_n}(\langle \nabla f(X_n), V_{n+1}\rangle + \langle \nabla f(X_{n+1}) - \nabla f(X_n), V_{n+1}\rangle)] \\ &\leq \mathbb{E}[\mathbb{1}_{\mathbb{A}_n}((1 - \gamma_{n+1}\mu)\langle \nabla f(X_n), V_n\rangle - \gamma_{n+1}|\nabla f(X_n)|^2 + C_L\gamma_{n+1}|V_{n+1}|^2)].\end{aligned} \tag{19}$$

综合估计式 (17)–(19)，我们得到

$$\begin{aligned}&\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}E_{n+1}] \\ &\leq \mathbb{E}[\mathbb{1}_{\mathbb{A}_n}(\alpha_{n+1}f(X_n) + \beta_{n+1}|\nabla f(X_n)|^2 + \delta_{n+1}\langle \nabla f(X_n), V_n\rangle + \epsilon_{n+1}|V_n|^2)],\end{aligned}$$

其中

$$
\begin{aligned}
\alpha_{n+1} =& a + \Big(\frac{b}{2} + C_L\gamma_{n+1}\Big(1 + \frac{\gamma_{n+1}a}{2}\Big)\Big)\gamma_{n+1}^2\sigma,\\
\beta_{n+1} =& -\gamma_{n+1} - a\gamma_{n+1}^2 + \Big(\frac{b}{2} + C_L\gamma_{n+1}\Big(1 + \frac{\gamma_{n+1}a}{2}\Big)\Big)\gamma_{n+1}^2,\\
\delta_{n+1} =& 1 - \mu\gamma_{n+1} + a(\gamma_{n+1} - \gamma_{n+1}^2\mu)\\
&- 2\Big(\frac{b}{2} + C_L\gamma_{n+1}\Big(1 + \frac{\gamma_{n+1}a}{2}\Big)\Big)(\gamma_{n+1} - \gamma_{n+1}^2\mu),\\
\epsilon_{n+1} =& \Big(\frac{b}{2} + C_L\gamma_{n+1}\Big(1 + \frac{\gamma_{n+1}a}{2}\Big)\Big)(1 - 2\gamma_{n+1}\mu + \gamma_{n+1}^2\mu^2).
\end{aligned} \tag{20}
$$

根据定义 $\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}\langle\nabla f(X_n)\ V_n\rangle] = \mathbb{E}[\mathbb{1}_{\mathbb{A}_n}(E_n - af(X_n) - \frac{b}{2}|V_n|^2)]$,，利用 PL 不等式 (14) 以及 $(\beta_n)_{n\in\mathbb{N}}$ 非正这一事实，可得

$$
\begin{aligned}
&\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}E_{n+1}]\\
&\le \mathbb{E}[\mathbb{1}_{\mathbb{A}_n}(\delta_{n+1}E_n + (\alpha_{n+1} - a\delta_{n+1} + 2\beta_{n+1}L)f(X_n) + (\epsilon_{n+1} - \frac{b}{2}\delta_{n+1})|V_n|^2)].
\end{aligned} \tag{21}
$$

根据关于 $(\alpha_{n+1}-a\delta_{n+1}+2\beta_{n+1}L)_{n\in\mathbb{N}_0}$ 和 $(\epsilon_{n+1}-\frac{b}{2}\delta_{n+1})_{n\in\mathbb{N}_0}$ 的假设，我们有 $\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}E_{n+1}] \le \delta_{n+1}\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}E_n]$. 。对于 $n \in \mathbb{N}$，我们利用引理 3.7 和 $(\mathbb{A}_n)_{n\in\mathbb{N}_0}$ 的单调性来证明 $\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}E_n] \le \mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}}E_n]$，于是，迭代可得

$$
\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}}E_n] \le \Big(\prod_{i=1}^{n}\delta_i\Big)\mathbb{E}[\mathbb{1}_{\mathbb{A}_0}E_0]. \tag{22}
$$

接下来，我们使用 (22) 对 $(f(X_n))_{n\in\mathbb{N}_0}$ 的期望进行界定。与 (17) 类似，对于所有$n \in \mathbb{N}_0$，我们有

$$
\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}f(X_{n+1})] \le \mathbb{E}\Big[\mathbb{1}_{\mathbb{A}_n}\big(f(X_n) + \gamma_{n+1}\langle\nabla f(X_{n+1}), V_{n+1}\rangle + \frac{C_L}{2}\gamma_{n+1}^2\big|V_{n+1}\big|^2\big)\Big]
$$

因此，根据 E_{n+1} 的定义，

$$
\begin{aligned}
&(1 + a\gamma_{n+1})\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}f(X_{n+1})]\\
&\le \mathbb{E}\Big[\mathbb{1}_{\mathbb{A}_n}\big(f(X_n) + \gamma_{n+1}E_{n+1} + (\frac{C_L}{2}\gamma_{n+1}^2 - \frac{b}{2}\gamma_{n+1})\big|V_{n+1}\big|^2\big)\Big].
\end{aligned}
$$

根据假设，$(\frac{C_L}{2}\gamma_n^2 = \frac{b}{2}\gamma_n)_{n\in\mathbb{N}}$是一个非正实数序列。因此，我们可以忽略上述上界中的最后一项，得到

$$
\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}f(X_{n+1})] \le (1 + a\gamma_{n+1})^{-1}\mathbb{E}[\mathbb{1}_{\mathbb{A}_n}f(X_n)] + \frac{\gamma_{n+1}}{1 + a\gamma_{n+1}}\Big(\prod_{i=1}^{n+1}\delta_i\Big)\mathbb{E}[\mathbb{1}_{\mathbb{A}_0}E_0]. \tag{23}
$$

利用 $f(X_n)$ 的非负性、$(\mathbb{A}_n)_{n\in\mathbb{N}_0}$ 的单调性以及 (23)，可以归纳证明，对所有 $n \in \mathbb{N}$,

$$
\begin{aligned}
&\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}}f(X_n)]\\
&\le \Big(\prod_{i=1}^{n}(1 + a\gamma_i)^{-1}\Big)\Big(\mathbb{E}[\mathbb{1}_{\mathbb{A}_0}f(X_0)] + \sum_{i=1}^{n}\frac{\gamma_i}{1 + a\gamma_i}\Big(\prod_{j=1}^{i}(1 + a\gamma_j)\Big)\Big(\prod_{j=1}^{i}\delta_j\Big)\mathbb{E}[\mathbb{1}_{\mathbb{A}_0}E_0]\Big).
\end{aligned}
$$

□

定理 *3.4* 的证明。在常数步长序列$\gamma_n \equiv \gamma > 0$的情况下应用命题 3.8，关于参数的假设恰好与定理 3.4 中的相同。现在，对于满足定理 3.4 中所有不等式且满足命题 3.8 其余假设的参数 a, b, μ, γ，对所有 $n \in \mathbb{N}$，我们有

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \leq (1+a\gamma)^{-n}\Big(\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} f(X_0)] + \frac{\gamma}{1+a\gamma}\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} E_0]\sum_{i=1}^{n}((1+a\gamma)\delta)^i\Big).$$

因此，如果（$1+a\gamma$）$\delta = 1$，则对于常数 $C \geq 0$，我们有

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \leq C(1+a\gamma)^{-n} n.$$

如果（$1+a\gamma$）$\delta \neq 1$，则有

$$\sum_{i=1}^{n}((1+a\gamma)\delta)^i = \frac{1-\big((1+a\gamma)\delta\big)^{n+1}}{1-(1+a\gamma)\delta}.$$

□

3.2. **小步长情形**。在本节中，我们考虑步长充分小的情形（γ_n）$_{n\in\mathbb{N}}$，并证明 MSGD 过程的主要结果，即定理 3.1 和定理 3.2。

命题 3.9。设 L、a、b、$\mu > 0$ 和 $\sigma \geq 0$。设 $(\mathbb{A}_n)^{n\in\mathbb{N}_0}$ 为递减事件序列，使得对于所有 $n \in \mathbb{N}_0$、$\mathbb{A}_n \in \mathcal{F}_n$，在 $\mathbb{A}_n$ 上都有

$$|\nabla f(X_n)|^2 \geq 2Lf(X_n) \quad and \quad \mathbb{E}[|D_{n+1}|^2|\mathcal{F}_n] \leq \sigma f(X_n).$$

假设

$$\mu - a + b > 0 \quad , \quad a\mu - a^2 + ab - 2L < 0 \quad and \quad C_L - \frac{b}{2}(\mu + a - b) < 0. \tag{24}$$

若我们进一步假设 $\mathbb{P}(\bigcap_{n\in\mathbb{N}_0}\mathbb{A}_n) < \mathbb{P}(\mathbb{A}_0)$ $ab \geq C_L$，则对于每个 $0 < \varepsilon < m :=\min(a, \mu - a + b)$，存在常数 $C, \bar{\gamma} \geq 0$，使得若 $\sup_{n\in\mathbb{N}}\gamma_n \leq \bar{\gamma}$，则有

(i)
$$\max(\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)], \mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}}|V_n|^2]) \leq C\exp(-(m-\varepsilon)t_n)$$
对于所有 $n \in \mathbb{N}$，其中 $t_n = \sum_{i=1}^{n}\gamma_i$。

(ii) 令 $m' < m-\varepsilon$，并假设 $\sum_{i=0}^{\infty}\exp((m'-(m-\varepsilon))t_i) < \infty$。则在事件 $\mathbb{A}_\infty = \bigcap_{n\in\mathbb{N}_0}\mathbb{A}_n$ 上，$\exp(m't_n)f(X_n) \to 0$ 几乎必然成立。(iii) 过程 $(X_n)^{n\in\mathbb{N}_0}$ 在 $\mathbb{A}_\infty$上几乎必然收敛。

证明。**(i)**：首先注意到，对于所有 $x \in \mathbb{R}$, $1 + x \leq \exp(x)$ 和 $(1+x)^{-1} = e^{-x+o(x)}$。因此，对于每个 $\varepsilon' > 0$，存在一个 $\bar{\gamma}' > 0$，使得当 $\max_{i=1,\dots,n}\gamma_i \leq \bar{\gamma}'$ 时，我们有

$$\prod_{i=1}^{n}(1+a\gamma_i)^{-1} \leq \Big(\prod_{i=1}^{n}\exp((-a+\varepsilon')\gamma_i)\Big) = \exp((-a+\varepsilon')t_n).$$

此外，对于 $(\delta_n)_{n\in\mathbb{N}}$（见式 (20)），我们有 $\delta_n = 1 - \gamma_n(\mu - a + b) + o(\gamma_n)$，因此，对于所有 $\varepsilon'' > 0$，都存在一个 $\bar{\gamma}'' > 0$，使得若 $\max_{i=1,\dots,n}\gamma_i \leq \bar{\gamma}''$，则对于所有 $i = 1, \dots, n$，都有 $\delta_i \geq 0$，并且

$$\prod_{i=1}^{n}\delta_i \leq \exp(-(\mu - a + b - \varepsilon'')t_n).$$

注意，$\beta_n = -\gamma_n + o(\gamma_n)$、$\alpha_n - a\delta_n + 2\beta_n L = \gamma_n(a\mu - a^2 + ab - 2L) + o(\gamma_n)$，

$$\epsilon_n - \frac{b}{2}\delta_n = \gamma_n\Big(C_L - \frac{b}{2}(\mu + a - b)\Big) + o(\gamma_n) \text{ and } \tfrac{C_L}{2}\gamma^2 - \tfrac{b}{2}\gamma = -\tfrac{b}{2}\gamma + o(\gamma),$$

并利用假设 (24)，我们可以将 $\bar\gamma \le \min(\bar\gamma', \bar\gamma'')$ 选得足够小，使得当 $\sup_{n\in\mathbb{N}}\gamma_n \le \bar\gamma$ 时，上述所有项都严格为负。然后，利用命题 3.8 可得，对所有$n \in \mathbb{N}$

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \le \exp((-a + \varepsilon' + \varepsilon'')t_n) \Big(\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} f(X_0)] + \sum_{i=1}^{n} \gamma_i \exp(-(\mu - 2a + b)t_i)\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} E_0]\Big).$$

若 $a < \mu - a + b$，函数 $t \mapsto \exp(-(\mu - 2a + b)t)$ 单调递减，因此可得

$$\sum_{i=1}^{n} \gamma_i \exp(-(\mu - 2a + b)t_i) \le \int_0^{t_n} \exp(-(\mu - 2a + b)t)\, dt.$$

因此，

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \le \exp((-a + \varepsilon' + \varepsilon'')t_n)\Big(\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} f(X_0)] + \frac{\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} E_0]}{\mu - 2a + b}\Big).$$

对于 $a > \mu - a + b$，函数 $t \mapsto \exp(-(\mu - 2a + b)t)$ 单调递增，因此可得

$$\sum_{i=1}^{n} \gamma_i \exp(-(\mu - 2a + b)t_i) \le \int_0^{t_n} \exp(-(\mu - 2a + b)(t + \bar\gamma))\, dt.$$

因此，

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \le \exp((-(\mu - a + b) + \varepsilon' + \varepsilon'')t_n)\Big(\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} f(X_0)] + \frac{\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} E_0]}{2a - \mu - b}\big(\exp(-(\mu - 2a + b)\bar\gamma)\Big).$$

最后，对于 $a = \mu - a + b$，我们得到 $\sum_{i=1}^{n} \gamma_i \exp(-(\mu - 2a + b)t_i) = t_n$，因此，

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \le \exp((-a + \varepsilon' + \varepsilon'')t_n)\Big(\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} f(X_0)] + t_n\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} E_0]\Big).$$

注意，$\exp(-\varepsilon'''t)t \to 0$ 对所有 $\varepsilon''' > 0$ 均成立。因此，存在一个常数 $C' > 0$，使得

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \le \exp((-a + \varepsilon' + \varepsilon'' + \varepsilon''')t_n)\Big(\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} f(X_0)] + C'\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} E_0]\Big).$$

第一个断言的证明可通过取 $\varepsilon' = \varepsilon'' = \varepsilon''' = \frac{1}{3}\varepsilon$ 得出。

对于第二个断言，注意到由引理 2.1 可得 $|\nabla f(x)|^2 \le 2C_L f(x)$ 对所有 $x \in \mathbb{R}^d$ 均成立，并且利用 $f(X_n) \ge 0$，由柯西-施瓦茨不等式可得

$$\frac{b}{2}|V_n|^2 \le E_n - \langle \nabla f(X_n), V_n\rangle \le E_n + |\nabla f(X_n)|\,|V_n|,$$

其中 E_n 由式 (13) 定义。因此，利用杨氏不等式，

$$\frac{b}{2}\mathbb{E}\big[\mathbb{1}_{\mathbb{A}_{n-1}}|V_n|^2\big] \le \mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} E_n] + \frac{1}{b}\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}}|\nabla f(X_n)|^2] + \frac{b}{4}\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}}|V_n|^2] \tag{25}$$

并且，利用 $\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)]$ 的界以及 (22)，可得一个常数 $C \ge 0$，使得

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}}|V_n|^2]^{1/2} \le C\exp\Big(-\frac{1}{2}(m - \varepsilon)t_n\Big). \tag{26}$$

(ii)：令 $m' < m-\varepsilon$。对于 $\varepsilon' > 0$ 和 $n \in \mathbb{N}$，考虑集合 $\mathbb{B}_n = \mathbb{A}_\infty \cap \{\sup_{i \geq n} \exp(m' t_i) f(X_i) \geq \varepsilon'\}$。由马尔可夫不等式和 (i)，存在一个 $C > 0$，使得

$$\begin{aligned}\mathbb{P}(\mathbb{B}_n) &\leq \sum_{i=n}^{\infty} \mathbb{P}\big(\mathbb{A}_{i-1} \cap \big\{\exp(m' t_i) f(X_i) \geq \frac{\varepsilon'}{2}\big\}\big) \\ &\leq \sum_{i=n}^{\infty} \frac{2\exp(m' t_i)}{\varepsilon'} \mathbb{E}[\mathbb{1}_{\mathbb{A}_{i-1}} f(X_i)] \leq \frac{C}{\varepsilon'} \sum_{i=n}^{\infty} \exp((m' - (m-\varepsilon)) t_i) \overset{n\to\infty}{\longrightarrow} 0.\end{aligned}$$

With

$$\mathbb{P}\Big(\mathbb{A}_\infty \cap \big\{ \limsup_{n\to\infty} \exp(m' t_n) f(X_n) \geq \varepsilon' \big\}\Big) \leq \mathbb{P}\Big(\bigcap_{n\in\mathbb{N}} B_n\Big) = 0$$

我们得到 $\exp(m' t_n) f(X_n) \to 0$ 在 $\mathbb{A}_\infty$ 上几乎必然成立。

(iii)：我们考虑事件 $\mathbb{A}_\infty$，并对过程 $(X_n)_{n\in\mathbb{N}_0}$ 行进的距离进行界定。由于 $\varepsilon < m$ 映射 $t \mapsto \exp(-\frac{1}{2}(m-\varepsilon)t)$ 单调递减，因此，利用 (26) 可得

$$\begin{aligned}\mathbb{E}\Big[\mathbb{1}_{\mathbb{A}_\infty} \sum_{i=1}^{\infty} |X_i - X_{i-1}|\Big] &\leq \sum_{i=1}^{\infty} \gamma_i \mathbb{E}[\mathbb{1}_{\mathbb{A}_{i-1}} |V_i|^2]^{1/2} \leq C \sum_{i=1}^{\infty} \gamma_i \exp\Big(-\frac{1}{2}(m-\varepsilon) t_i\Big) \\ &\leq C \int_0^{\infty} \exp\Big(-\frac{1}{2}(m-\varepsilon)t\Big)\, dt < \infty,\end{aligned}$$

这意味着 $\sum_{i=1}^{\infty} |X_i - X_{i-1}|$ 在 $\mathbb{A}_\infty$ 上几乎必然有限。因此，$(X_n)_{n\in\mathbb{N}_0}$ 在 $\mathbb{A}_\infty$ 上几乎必然收敛。
□

引理 3.10. 对于所有 $\mu > 0$，都存在 $a, b > 0$，使得 $ab \geq C_L$，且满足 (24)。

证明。令 $\varepsilon, b > 0$，并选取 $a = b + 2\frac{C_L}{b} - \mu + \varepsilon$。注意，$a > 0$ 当且仅当 $\mu < b + 2\frac{C_L}{b} + \varepsilon$。此外，$\mu - a + b = 2(\mu - \frac{C_L}{b}) - \varepsilon$ 为正当且仅当 $\mu > \frac{C_L}{b} + \frac{\varepsilon}{2}$。现在，$C_L - \frac{b}{2}(\mu + a - b) = -\frac{b}{2}\varepsilon < 0$，并且我们有

$$a\mu - a^2 + ab - 2L = -2\mu^2 + 2\mu\Big(b + \frac{3C_L}{b} + \frac{3\varepsilon}{2}\Big) - 2C_L - 2L - 4\frac{C_L^2}{b^2} - 2\varepsilon\Big(\frac{b}{2} + 2\frac{C_L}{b} + \frac{\varepsilon}{2}\Big).$$

后一个项是关于 μ 的二次函数，在两个根之外为负。因此，$a\mu - a^2 + ab - 2L < 0$ 当且仅当 $\mu \notin [\mu_-^{\varepsilon,b}, \mu_+^{\varepsilon,b}]$，其中

$$\mu_\pm^{\varepsilon,b} = \frac{1}{2}\Big(b + \frac{3C_L}{b} + \frac{3\varepsilon}{2} \pm \sqrt{\Big(b + \frac{C_L}{b} + \frac{\varepsilon}{2}\Big)^2 - 4L}\Big). \tag{27}$$

注意，对于所有 $b > 0$，$(b + \frac{C_L}{b})^2 \geq 4C_L$，且 $C_L \geq L$，因此 $\mu_\pm^{\varepsilon,b}$ 有定义。此外，$\mu_-^{\varepsilon,b} > \frac{C_L}{b} + \frac{\varepsilon}{2}$，且 $\mu_+^{\varepsilon,b} < b + 2\frac{C_L}{b} + \varepsilon$。附加假设 $ab \geq C_L$ 成立当且仅当 $\mu \leq b + \frac{C_L}{b} + \varepsilon$。因此，对于给定的数对 (a,b)，满足 (24) 的摩擦参数集合 μ 等于 $(\frac{C_L}{b} + \frac{\varepsilon}{2}, \mu_-^{\varepsilon,b}) \cup (\mu_+^{\varepsilon,b}, b + 2\frac{C_L}{b} + \varepsilon)$；而对于给定的数对 (a,b)，同时满足 (24) 和 $ab \geq C_L$ 的摩擦参数集合 μ 包含区间 $(\frac{C_L}{b} + \frac{\varepsilon}{2}, \mu_-^{\varepsilon,b} \wedge (b + \frac{C_L}{b} + \varepsilon))$。对于所有 $b > 0$，后一个区间都非空，上下限关于 b 连续，且下限满足

$$\frac{C_L}{b} + \frac{\varepsilon}{2} \overset{b\to\infty}{\longrightarrow} \frac{\varepsilon}{2} \quad \text{and} \quad \frac{C_L}{b} + \frac{\varepsilon}{2} \overset{b\to 0}{\longrightarrow} \infty.$$

通过令 $\varepsilon \to 0$，我们证明了对于每个 $\mu > 0$，都存在一对 $(a,\ b)$，使得 (24) 得以满足且 $ab \geq C_L$。□

定理 *3.1* 的证明。 根据引理 3.10，存在参数 a，$b>0$，使得 $ab \geq C_L$ 且 (24) 得以满足。注意到，选择 $(\mathbb{A}_n)_{n\in\mathbb{N}} = (\{X_i \in \mathcal{D} \text{ 对所有 } i=0, \ldots\ldots, n\})_{n\in\mathbb{N}}$ 满足命题 3.9 的假设。现在，结论 (i)、(ii) 和 (iii) 由命题 3.9 得出。□

我们给出一个关于收敛速率大小的一般性表述，该表述仍依赖于技术参数 b。

引理 3.11。 设 L、b、$\varepsilon>0$ 和 $\sigma\geq 0$。设 $(\mathbb{A}_n)^{n\in\mathbb{N}_0}$ 为一个递减事件序列，使得对于所有 $n\in\mathbb{N}_0$、$\mathbb{A}_n \in \mathcal{F}_n$，在 $\mathbb{A}_n$ 上均有

$$|\nabla f(X_n)|^2 \geq 2Lf(X_n) \quad and \quad \mathbb{E}[|D_{n+1}|^2|\mathcal{F}_n] \leq \sigma f(X_n).$$

设 $\mu \in (\frac{C_L}{b}+\frac{\varepsilon}{2}, \mu^{\varepsilon}_{+,b}) \cup (\mu^{\varepsilon}_{-,b}, b+2\frac{C_L}{b}+\varepsilon)$，其中 $\mu^{\varepsilon}_{\pm,b}$ 由(27)定义。若 $\mathbb{P}(\bigcap_{n\in\mathbb{N}_0}\mathbb{A}_n) < \mathbb{P}(\mathbb{A}_0)$，进一步假设 $\mu \leq b+\frac{C_L}{b}+\varepsilon$。那么，存在 C、$\bar{\gamma}>0$，使得若$\sup_{n\in\mathbb{N}}\gamma_n \leq \bar{\gamma}$，则有

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \leq C\exp(-m(\varepsilon,b,\mu)t_n),$$

其中

$$m(\varepsilon,b,\mu) = \begin{cases} 2(\mu-\frac{C_L}{b}-\varepsilon), & if\ \mu < \frac{1}{3}(b+4\frac{C_L}{b}+2\varepsilon) \\ b+2\frac{C_L}{b}-\mu, & if\ \mu \geq \frac{1}{3}(b+4\frac{C_L}{b}+2\varepsilon). \end{cases}$$

证明。 令 $\varepsilon, b>0$ 以及 μ 如引理中的假设所述，并令 $a = b+2\frac{C_L}{b}-\mu+\varepsilon$。回顾引理 3.10 的证明，我们已证明这一参数选择满足 (24)，并且若$\mathbb{P}(\bigcap_{n\in\mathbb{N}_0}\mathbb{A}_n) < \mathbb{P}(\mathbb{A}_0)$，还满足 $ab \geq C_L$。因此，我们可以应用命题 3.9，并推得存在常数 $C, \bar{\gamma}>0$，使得若 $\sup_{n\in\mathbb{N}}\gamma_n \leq \bar{\gamma}$，则有

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}} f(X_n)] \leq C\exp(-m(\varepsilon,b,r)t_n),$$

其中

$$m(\varepsilon,b,\mu) = \min(a, \mu-a+b)-\varepsilon = \min\Big(b+2\frac{C_L}{b}-\mu, 2\Big(\mu-\frac{C_L}{b}-\varepsilon\Big)\Big).$$

最小值的计算是直接的。 □

定理 *3.2* 的证明。 我们在所有允许的参数 μ 和 b 上最大化引理 3.11 得出的收敛速率。首先，假设 $\kappa = \frac{C_L}{L} < \frac{9}{8}$。那么，对于所有充分小的 $\varepsilon>0$，我们有 $\kappa^{\varepsilon} := \big(\frac{\sqrt{C_L}}{\sqrt{L}}+\frac{\varepsilon}{4\sqrt{L}}\big)^2 < \frac{9}{8}$，因此 Set

$$b^{\varepsilon}_{\pm} := \frac{3}{\sqrt{8}}\sqrt{L}-\frac{\varepsilon}{4} \pm \sqrt{\Big(\frac{3}{\sqrt{8}}\sqrt{L}-\frac{\varepsilon}{4}\Big)^2 - C_L}$$

是良定义的。对于 $b\in(b^{\varepsilon}_{-}, b^{\varepsilon}_{+})$，我们有$\frac{1}{3}(b+4\frac{C_L}{b}+2\varepsilon) < \mu^{\varepsilon}_{-,b}$，因此利用引理 3.11 可得，对于 $\mu \in (\frac{C_L}{b}+\frac{\varepsilon}{2}, \frac{1}{3}(b+4\frac{C_L}{b}+2\varepsilon))$，收敛速率为 $m(\varepsilon, b, \mu) = 2\big(\mu-\frac{C_L}{b}-\varepsilon\big)$。注意，

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\frac{1}{3}\Big(b^{\varepsilon}_{+}-\varepsilon^2+4\frac{C_L}{b^{\varepsilon}_{+}-\varepsilon^2}+2\varepsilon\Big) = \frac{1}{4}+\frac{3\sqrt{L}}{4\sqrt{9L-8C_L}} > 0.$$

因此，当 ε 足够小时，我们有

$$\mu^*_1 := \frac{1}{3}\Big(b^0_{+}+4\frac{C_L}{b^0_{+}}\Big) = \frac{1}{\sqrt{8}}(5-\sqrt{9-8\kappa})\sqrt{L} < \frac{1}{3}\Big(b^{\varepsilon}_{+}-\varepsilon^2+4\frac{C_L}{b^{\varepsilon}_{+}-\varepsilon^2}+2\varepsilon\Big)$$

由连续性可知

$$m(\varepsilon, b^{\varepsilon}_{+}-\varepsilon^2, \mu^*_1) \xrightarrow{\varepsilon\to 0} \frac{2}{3}\Big(b^0_{+}+\frac{C_L}{b^0_{+}}\Big) = \frac{2}{3}\frac{(b^0_{+})^2+C_L}{b^0_{+}} = \sqrt{2L}.$$

此外，注意到 b_+^0 满足 $\frac{1}{3}(b_+^0+4\frac{C_L}{b_+^0}) < b_+^0 + \frac{C_L}{b_+^0}$，因为

$$\frac{C_L}{b_+^0} = \frac{1}{\sqrt{8}}(3\sqrt{L}-\sqrt{9L-8C_L}) < \frac{1}{\sqrt{2}}(3\sqrt{L}+\sqrt{9L-8C_L}) = 2b_+^0.$$

因此，$b_+^0(b_+^0+2\frac{C_L}{b_+^0}-\mu_1^*) > C_L$ 并且，当 $\varepsilon>0$ 足够小时，参数 $b=b_+^\varepsilon-\varepsilon^2$ 和 $a=b_+^\varepsilon-\varepsilon^2+2\frac{C_L}{b_+^\varepsilon-\varepsilon^2}-\mu_1^*+\varepsilon$ 满足 $ab\geq C_L$。

类似地，若 $b\in(b_-^\varepsilon, b_+^\varepsilon)$ 且 $\mu\in(\frac{1}{3}(b+4\frac{C_L}{b}+2\varepsilon)\ \mu^{\varepsilon,b}_-)$，则可得收敛速率 $m(\varepsilon,b,\mu) = b+2\frac{C_L}{b}-\mu$。注意，由于 $\kappa<\frac{9}{8}$，我们有

$$\frac{d}{d\varepsilon}\big|_{\varepsilon=0}\frac{1}{3}\Big(b_-^\varepsilon+\varepsilon^2+4\frac{C_L}{b_-^\varepsilon+\varepsilon^2}+2\varepsilon\Big) = \frac{1}{4}-\frac{3\sqrt{L}}{4\sqrt{9L-8C_L}} < 0.$$

因此，当 ε 足够小时，我们有

$$\mu_2^* := \frac{1}{3}\Big(b_-^0+4\frac{C_L}{b_-^0}\Big) = \frac{1}{\sqrt{8}}(5+\sqrt{9-8\kappa})\sqrt{L} > \frac{1}{3}\Big(b_-^\varepsilon+\varepsilon^2+4\frac{C_L}{b_-^\varepsilon+\varepsilon^2}+2\varepsilon\Big)$$

由连续性可得

$$m(\varepsilon, b_-^\varepsilon+\varepsilon^2, \mu_2^*) \xrightarrow{\varepsilon\to 0} b_-^0+2\frac{C_L}{b_-^0}-\mu^* = \frac{2}{3}\frac{(b_-^0)^2+C_L}{b_-^0} = \sqrt{2L}.$$

若 $L<C_L$, b_-^0 满足 $\frac{1}{3}(b_-^0+4\frac{C_L}{b^0}) < b_-^0+\frac{C_L}{b^0}$，因为

$$\frac{C_L}{b_-^0} = \frac{1}{\sqrt{8}}(3\sqrt{L}+\sqrt{9L-8C_L}) < \frac{1}{\sqrt{2}}(3\sqrt{L}-\sqrt{9L-8C_L}) = 2b_-^0.$$

因此，当 $\varepsilon>0$ 足够小时，$b=b_-^\varepsilon+\varepsilon^2$ 和 $a=b_-^\varepsilon+\varepsilon^2+2\frac{C_L}{b_-^\varepsilon+\varepsilon^2}-\mu_2^*+\varepsilon$ 满足 $ab\geq C_L$。如果 $L=C_L$，则有 $\mu_2^* = b_-^0+\frac{C_L}{b_-^0}$。利用 $\frac{d}{d\varepsilon}\big|_{\varepsilon=0}\Big(b_-^\varepsilon+\frac{C_L}{b_-^\varepsilon}+\varepsilon\Big) = \frac{1}{2}$，我们得到，当 $\varepsilon>0$ 足够小时，$\mu_2^* \leq b_-^\varepsilon+\varepsilon^2+\frac{C_L}{b_-^\varepsilon+\varepsilon^2}+\varepsilon$，因此，对于参数 $b=b_-^\varepsilon+\varepsilon^2$ 和 $a=b_-^\varepsilon+\varepsilon^2+2\frac{C_L}{b_-^\varepsilon+\varepsilon^2}-\mu_2^*+\varepsilon$，有 $ab\geq C_L$。最后注意到，不失一般性，只要 $\kappa<\frac{9}{8}$，我们就可以增大 Lipschitz 常数 C_L。因此，我们得到所有摩擦参数

$$\mu\in\Big[\frac{1}{\sqrt{8}}(5-\sqrt{9-8\kappa})\sqrt{L}, \frac{1}{\sqrt{8}}(5+\sqrt{9-8\kappa})\sqrt{L}\Big]\Big\backslash\Big\{\frac{5}{\sqrt{8}}\sqrt{L}\Big\}$$

给出 $\sqrt{2L}-\varepsilon$ 的最优收敛速率。情形 $\mu=\frac{5}{\sqrt{8}}\sqrt{L}$ 对应于 $\kappa=\frac{9}{8}$，将在下文讨论。

接下来，假设 $\kappa\geq\frac{9}{8}$，这意味着对所有 $\varepsilon>0$，都有 $\kappa^\varepsilon\geq\frac{9}{8}$。利用引理 3.11，我们得到收敛速率 $m(\varepsilon, b, \mu) = 2(\mu-\frac{C_L}{b}-\varepsilon)$，条件是 $\mu\in(\frac{C_L}{b}+\frac{\varepsilon}{2}, \mu^{\varepsilon,b}_-)$。首先，注意到如果 $\mu^{0,b}_-\in(\frac{C_L}{b}+\frac{\varepsilon}{2}, \mu^{\varepsilon,b}_-)$，则有

$$m(\varepsilon, b, \mu_-^{0,b}) = b+\frac{C_L}{b}-\sqrt{\Big(b+\frac{C_L}{b}\Big)^2-4L}-2\varepsilon.$$

此外，$m(\varepsilon, b, \mu^{0,b}_-) \xrightarrow{b\to 0} -2\varepsilon$ 以及 $m(\varepsilon, b, \mu^0_-) \xrightarrow{b\to\infty} -2\varepsilon$，并且

$$\frac{d}{db}m(\varepsilon, b, \mu_-^{0,b}) = \Big(1-\frac{C_L}{b^2}\Big)\Big(1-\frac{b+C_L/b}{\sqrt{(b+C_L/b)^2-4L}}\Big),$$

因此，对于所有 $b < \sqrt{C_L}$，$\frac{d}{db} m(\varepsilon, b, \mu^{0,b}_{-}) > 0$；对于所有 $b > \sqrt{C_L}$，$\frac{d}{db} m(\varepsilon, b, \mu^{0,b}_{-}) < 0$。因此，$m(\varepsilon, b, \mu^{0,b}_{-})$ 的最大值在 $b^* = \sqrt{C_L}$ 处取得，其中

$$\mu_3^* = \mu_-^{0,b^*} = 2\sqrt{C_L} - \sqrt{C_L - L} = (2\sqrt{\kappa} - \sqrt{\kappa - 1})\sqrt{L}$$

and

$$m(\varepsilon, b^*, \mu_3^*) = 2(\sqrt{C_L} - \sqrt{C_L - L} - \varepsilon) \xrightarrow{\varepsilon \to 0} 2(\sqrt{\kappa} - \sqrt{\kappa - 1})\sqrt{L}.$$

现在，考虑

$$\mu_-^{\varepsilon, b^*} = \frac{1}{2}\Big(4\sqrt{C_L} + \frac{3\varepsilon}{2} - \sqrt{(2\sqrt{C_L} + \frac{\varepsilon}{2})^2 - 4L}\Big)$$

作为 ε 的函数。注意，对于所有 $\varepsilon \geq 0$，我们有 $\frac{L}{C_L} \leq \frac{2}{9}\Big(2 + \frac{\varepsilon}{2\sqrt{C_L}}\Big)^2$，因此，

$$\frac{d}{d\varepsilon}\mu_-^{\varepsilon, b^*} = \frac{1}{2}\Big(\frac{3}{2} - \frac{1}{2}\Big((2\sqrt{C_L} + \frac{\varepsilon}{2})^2 - 4L\Big)^{-1/2}\big(2\sqrt{C_L} + \frac{\varepsilon}{2}\big)\Big) \geq 0$$

因此，对于足够小的 $\varepsilon > 0$，我们有 $\mu_3^* \in (\frac{C_L}{b^*} + \frac{\varepsilon}{2}, \mu_-^{\varepsilon, b^*})$。此外，注意存在 $\mu_3^* < 2\sqrt{C_L} = b^* + \frac{C_L}{b^*}$，使得参数 b^* 和 $a = b^* + 2\frac{C_L}{b^*} - \mu + \varepsilon$ 满足 $ab^* \geq C_L$。最后，对于 $\kappa = \frac{9}{8}$，我们得到 $2(\sqrt{\kappa} - \sqrt{\kappa - 1})\sqrt{L} = \sqrt{2L}$ 和 $\mu_3^* = \frac{5}{\sqrt{8}}\sqrt{L}$。□

推论 *3.3* 的证明。设 $r > 0$ 满足 $B_r(y) \subset \mathcal{D}$。那么，对于每个 $r_0 < r$，我们有 $\{X_0 \in B_{r_0(y)}\} \subset \mathbb{A}_0$。按照命题 3.9 和引理 3.10 选择 $\bar{\gamma}$，a，b，$\varepsilon > 0$，使得 $m = \min(a, \mu - a + b) > \varepsilon$。然后，命题 3.9 (i) 表明，存在常数 $C(r_0) \geq 0$，使得对于所有 $n \in \mathbb{N}$

$$\mathbb{E}[\mathbb{1}_{\mathbb{A}_{n-1}}|V_n|^2]^{1/2} \leq C(r_0)\exp\Big(-\frac{1}{2}(m - \varepsilon)t_n\Big).$$

注意到，根据引理 2.1，逆 PL 不等式 (9) 对所有 $x \in B_{r/2}(y)$ 均成立。因此，结合 (25)、(15) 和 (22)，常数 $C(r_0)$ 仅取决于 $\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} f(X_0)]$，$\mathbb{E}[\mathbb{1}_{\mathbb{A}_0} E_0]$以及 $\mathbb{E}[\mathbb{1}_{\mathbb{A}_0}|V_0|^2]$。现在，利用 $f(y) = 0$和 $\nabla f(y) = 0$这一事实，可得 $C(r_0) \to 0$，因为 $r_0 \to 0$。由马尔可夫不等式可得，对于 $r_0 < r$

$$\begin{aligned}\mathbb{P}(\mathbb{A}_\infty^c) &= \mathbb{P}\Big(\bigcup_{n=1}^{\infty} \mathbb{A}_n^c \cap \mathbb{A}_{n-1}\Big) \leq \mathbb{P}\Big(\sup_{n\in\mathbb{N}} \mathbb{1}_{\mathbb{A}_{n-1}} \sum_{i=1}^{n} \gamma_i |V_i| > r - r_0\Big) \\ &\leq \frac{1}{r - r_0}\sum_{i=1}^{\infty} \gamma_i \mathbb{E}[\mathbb{1}_{\mathbb{A}_{i-1}}|V_i|] \leq \frac{C(r_0)}{r - r_0}\int_0^\infty \exp\Big(-\frac{1}{2}(m - \varepsilon)t\Big)\, dt,\end{aligned}$$

并且，因此，$\mathbb{P}(\mathbb{A}_\infty^c) \to 0$，因为 $r_0 \to 0$。 □

4. Momentum stochastic gradient descent in continuous time

本节中，我们研究 (5) 中定义的扩散过程 $(X_t)_{t\geq 0}$。我们证明，如果摩擦参数相对于随机噪声的大小足够大，那么对于在开集 $\mathcal{D} \subset \mathbb{R}^d$ 中满足 PL 条件的目标函数 $f \in C^2$，$(f(X_t))_{t\geq 0}$ 几乎必然指数收敛。

定理 4.1。令 L、$\sigma > 0$ 和 $\mathcal{D} \subset \mathbb{R}^d$ 为一个开集。设 $T := \text{下确界}\{t \geq 0 : \ X_t \notin \mathcal{D}\}$，并假设对于所有 $x \in \mathcal{D}$

$$|\nabla f(x)|^2 \geq 2Lf(x) \quad \text{and} \quad \|\Sigma(x)\|_F^2 \leq \sigma f(x). \tag{28}$$

如果 $\mu > \frac{C_L \sigma}{4L}$，那么：

(i) 存在 C、$m>0$，使得对所有 $t\geq 0$

$$\mathbb{E}[\mathbb{1}_{\{T>t\}}f(X_t)]\leq C\exp(-mt).$$

（ii）对所有 $m'<m$，我们都有 $\exp(m't)f(X_t)\to 0$ 在事件 $\{T=\infty\}$ 上几乎必然成立。（iii）过程 $(X_t)_{t\geq 0}$ 在 $\{T=\infty\}$ 上几乎必然收敛。

为了推导收敛速率 m 的显式值，需要求解一个约束优化任务。优化任务的精确定式可见下文引理 4.8 的陈述（另见注 4.9）。接下来，我们给出摩擦参数 μ 的最优选择及相应收敛速率的估计。

定理 4.2。令 L，$\sigma>0$。定义 $C_L^*=C_L\vee\frac{9}{8}L$，假设 $0<\sigma<4\sqrt{\frac{L}{C_L^*}}$ 并选择

$$\mu=2\sqrt{C_L^*}-\sqrt{C_L^*-L+\frac{1}{4}\sqrt{C_L^*}\sigma}.$$

那么，在假设（28）下，存在一个 $C\geq 0$ 使得

$$\mathbb{E}[\mathbb{1}_{\{T>t\}}f(X_t)]\leq C\exp(-mt),$$

对所有 $t\geq 0$，其中

$$m=2\left(\sqrt{C_L^*}-\sqrt{C_L^*-L+\frac{1}{4}\sqrt{C_L^*}\sigma}\right).$$

注4.3。本注中，我们比较连续时间动量随机梯度下降(5)的收敛速率与随机梯度下降的连续时间对应过程的收敛速率，后者由以下随机微分方程给出

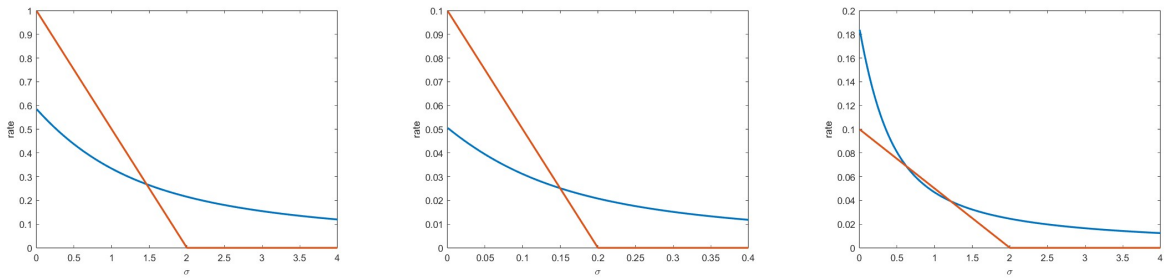
$$d\hat{X}_t=-\nabla f(\hat{X}_t)\,dt+\Sigma(\hat{X}_t)\,dW_t. \tag{29}$$

在非过参数化情形下，文献 [DK22a] 推导了随机微分方程 (29) 的收敛速率。遵循文献 [DK22a] 中的论证并使用定理 4.2 的假设，可以直接证明

$$\mathbb{E}[\mathbb{1}_{\{T>t\}}f(\hat{X}_t)]\leq C\exp(-m_{\mathrm{SGD}}t)$$

其收敛速率为 $m_{\mathrm{SGD}}=2L-\frac{1}{2}C_L\sigma$。此外，选择目标函数 $f(x)=\frac{L}{2}x^2$ 可知，如果 $\sigma>4\frac{L}{C_L}$，则 $\mathbb{E}[\mathbb{1}_{\{T>t\}}f(\hat{X}_t)]$ 不会收敛到零。相比之下，只要摩擦参数满足 $\mu>\frac{C_L\sigma}{4L}$，对于所有 $\sigma\geq 0$，MSGD 过程 (5) 都会指数收敛至临界点集合。(2) 的显式收敛速率由关于摩擦参数 μ 的优化任务的解给出，见引理 4.8 和注 4.9。下方图 4 中使用 *fminsearch* 在*Matlab*中对不同的 L，C_L 和 σ 数值求解了该优化任务。

我们观察到，在噪声较大或收敛到平坦极小值（即 L 较小）的情况下，连续时间动量随机梯度下降相比连续时间随机梯度下降收敛得更快；而较大的条件数 $\kappa=\frac{C_L}{L}$ 会在噪声较小时削弱这一效果。



(A) $L = 0.5$, $C_L = 1$ (B) $L = 0.005$, $C_L = 1$ (C) $L = 0.05$, $C_L = 0.1$

FIGURE 4. 动量随机梯度下降的收敛速率比较 m ： 连续时间下，依照定理 4.1 (i)，蓝色表示动量随机梯度下降，橙色表示随机梯度下降；收敛速率随噪声强度变化 σ (x轴)，对应不同的 L 和 C_L。

定理 4.1 和定理 4.2 是针对过程 $(X_t)_{t\geq 0}$ 在定义域 $\mathcal{D}$ 上的局部分析，其中 0 是唯一的临界水平，并且随着目标函数值趋近于其最小值，随机噪声会消失。[3] 我们可以利用定理 4.1 证明中的估计，证明轨迹的期望长度可以由一个常数界定，该常数随着初始速度、初始化时损失函数的梯度大小和值的增大而衰减。由此，我们可以界定集合 $\mathcal{D}$ 的退出概率。因此，如果我们从 $\mathcal{D}$ 中接近最优值的位置开始过程 $(X_t)_{t\geq 0}$ ，且初始速度 V_0 较小，则 $(X_t)_{t\geq 0}$ 以高概率永远不会触及 $\mathcal{D}$ 的边界，并会收敛到全局最小值。

推论 4.4。设 $y \in \mathcal{D}$ ，其中 $f(y) = 0$ 且 $\mu > \frac{C_L\sigma}{4L}$ 。那么，在满足定理 *4.1* 的假设时，对于每个 $\varepsilon > 0$ ，存在一个 $r_0 > 0$ ，使得如果 $X_0 \in B_{r_0}(y)$，则几乎必然有，且$\mathbb{E}[|V_0|^2] \leq r_0$ 我们有

$$\mathbb{P}(T < \infty) \leq \varepsilon.$$

注释 4.5. 注意， C_L 表示 ∇f的 Lipschitz 常数。 Σ 的 Lipschitz 常数的精确值不会出现在结果的表述中。我们可以在以下意义下放宽对 ∇f 和 Σ 的 Lipschitz 连续性假设。假设 ∇f 和Σ 仅在 $\mathcal{D}$上 Lipschitz 连续。那么，存在一个连续半鞅（X_t， V_t）$_{t\geq 0}$，使得式（5）在停时 $T = \inf\{t \geq 0 : X_t \notin \mathcal{D}\}$之前成立。现在，为了推出定理 4.1 和定理 4.2 的结论，只需假设 $\inf_{x\in\mathcal{D}} f(x) = 0$ ，且对所有 $x \in \mathcal{D}$,

$$|\nabla f(x)|^2 \leq 2C_L f(X_t),$$

其中 C_L 表示 ∇f 在 $\mathcal{D}$上的 Lipschitz 常数。引理 2.1 说明，后一个不等式如何由 ∇f 在更大定义域上的 Lipschitz 连续性推出。对于定理 3.1 (ii) 的表述，还需要额外假设 ∇f 和 Σ 在包含 $\mathcal{D}$的凸集上具有 Lipschitz 连续性。

我们开始证明本节的主要结果。下面的命题在对参数 μ、C_L、L 和 σ作出技术假设的条件下，给出了目标函数值期望的指数收敛。同样，证明基于随机李雅普诺夫函数 $(E_t)_{t\geq 0}$ ，其定义如下:

$$E_t = af(X_t) + \langle \nabla f(X_t), V_t\rangle + \frac{b}{2}|V_t|^2.$$

[3]在本节中，只需假设 $0 = \inf_{x\in\mathcal{D}} f(x)$。

注意，$(E_t)_{t\geq 0}$ 是一个连续的可积过程。如果 $(X_t)_{t\geq 0}$ 能够离开 $\mathcal{D}$，我们就必须确保李雅普诺夫函数在退出时刻非负；当$ab \geq C_L$时，这一条件成立，见引理 3.7。

命题 4.6. 设 $L, \sigma > 0$。设 T 是一个 $(\mathcal{F}_t)_{t\geq 0}$-停时，使得对所有 $t \geq 0$，在 $\{T > t\}$ 上

$$2Lf(X_t) \leq |\nabla f(X_t)|^2 \quad \text{and} \quad \|\Sigma(X_t)\|_F^2 \leq \sigma f(X_t). \tag{30}$$

此外，设 $a, b > 0$，并假设

$$\mu - a + b > 0\,,\ \frac{b}{2}\sigma - a^2 + a\mu + ab - 2L \leq 0 \text{ and } C_L - \frac{b}{2}(\mu + a - b) \leq 0. \tag{31}$$

若 $\mathbb{P}(T = \infty) < \mathbb{P}(T > 0)$，还需假设 $ab \geq C_L$。则：

(i) 存在一个常数 $C > 0$，使得

$$\max(\mathbb{E}[\mathbb{1}_{\{T>t\}} f(X_t)], \mathbb{E}[\mathbb{1}_{\{T>t\}} |V_t|^2]) \leq \begin{cases} C\exp(-mt), & \text{if } a \neq \mu - a + b, \\ C(1+t)\exp(-mt), & \text{if } a = \mu - a + b, \end{cases}$$

对所有 $t \geq 0$ 均成立，其中 $m = \min(a, \mu - a + b)$。

(ii) $(X_t)_{t\geq 0}$ 在 $\{T = \infty\}$ 上几乎必然收敛。

证明。(i)：首先，我们证明 $(\mathbb{E}[\mathbb{1}_{\{T>t\}} E_t])_{t\geq 0}$ 的指数收敛，随后证明这意味着 (i)。

由于 $(X_t)_{t\geq 0}$ 具有有界变差，根据伊藤公式可得

$$df(X_t) = \langle \nabla f(X_t), V_t\rangle dt \quad \text{and} \quad d|V_t|^2 = 2\langle V_t, dV_t\rangle + \|\Sigma(X_t)\|_F^2 dt,$$

根据伊藤乘积法则可得

$$d\langle \nabla f(X_t), V_t\rangle = \langle \nabla f(X_t), dV_t\rangle + \langle V_t, \operatorname{Hess} f(X_t) V_t\rangle dt.$$

因此，对于 $(E_t)_{t\geq 0} = (\mathbb{1}_{\{T>t\}} E_t)_{t\geq 0}$，我们有

$$\begin{aligned} d\tilde{E}_t =& \mathbb{1}_{\{T>t\}} \Big((a - \mu - b)\langle \nabla f(X_t), V_t\rangle - |\nabla f(X_t)|^2 - b\mu |V_t|^2 + \langle V_t, \operatorname{Hess} f(X_t) V_t\rangle \\ &+ \frac{b}{2}\|\Sigma(X_t)\|_F^2 \Big) dt + dM_t - d\xi_t, \end{aligned}$$

其中 $(M_t)_{t\geq 0}$ 表示 L^2-鞅

$$(M_t)_{t\geq 0} = \Big(\int_0^{T\wedge t} \langle \nabla f(X_u) + bV_u, \Sigma(X_u) dW_u\rangle \Big)_{t\geq 0},$$

且 $(\xi_t)_{t\geq 0}$ 表示如下定义的（几乎必然）非负递增过程

$$\xi_t := \begin{cases} 0, & \text{if } t < T \text{ or } T = 0, \\ E_T, & \text{otherwise.} \end{cases}$$

利用（30）以及 ∇f 的 Lipschitz 连续性，我们得到，对于所有 $0 \leq s < t$,

$$\begin{aligned} \tilde{E}_t - \tilde{E}_s \leq& \int_{s\wedge T}^{t\wedge T} (a - \mu - b)\langle \nabla f(X_u), V_u\rangle - |\nabla f(X_u)|^2 - (b\mu - C_L)|V_u|^2 \, du \\ &+ \int_{s\wedge T}^{t\wedge T} \frac{b}{2}\sigma f(X_u)\, du + M_t - M_s - (\xi_t - \xi_s). \end{aligned}$$

根据 $(E_t)_{t\geq 0}$ 的定义，对于所有 $u \geq 0$，我们有，

$$\langle \nabla f(X_u), V_u \rangle = E_u - af(X_u) - \frac{b}{2}|V_u|^2,$$

因此，利用 PL 不等式 (30)，我们得到

$$\begin{aligned}\tilde{E}_t - \tilde{E}_s \leq & \int_{s\wedge T}^{t\wedge T} (a-\mu-b)\tilde{E}_u - \Big(\frac{b}{2}\mu - C_L + \frac{b}{2}a - \frac{b^2}{2}\Big)|V_u|^2\, du \\ & + \int_{s\wedge T}^{t\wedge T} \Big(\frac{b}{2}\sigma - a^2 + a\mu + ab - 2L\Big) f(X_u)\, ds + M_t - M_s - (\xi_t - \xi_s).\end{aligned}$$

根据控制收敛定理 $(e_t)_{t\geq 0} := (\mathbb{E}[\mathbb{1}_{\{T>t\}} E_t])_{t\geq 0}$ 是下半连续的，因此利用 (31) 和 [MNPR20] 中的命题2.3，我们有 $e_t \leq e_0 \exp((a-\mu-b)t)$，对所有$t \geq 0$均成立。

接下来，我们利用 $(e_t)_{t\geq 0}$ 的估计来推导 $\varphi_t = \mathbb{E}[\mathbb{1}_{\{T>t\}} f(X_t)]$的收敛速率。回顾一下，

$$df(X_t) = \langle \nabla f(X_t), V_t \rangle\, dt = E_t dt - af(X_t)dt - \frac{b}{2}|V_t|^2 dt.$$

因此，$(\tilde{f}_t)_{t\geq 0} := (\mathbb{1}_{\{T>t\}} f(X_t))_{t\geq 0}$是一个非负过程，满足

$$d\tilde{f}_t = \mathbb{1}_{\{T>t\}}\big(E_t - af(X_t) - \frac{b}{2}|V_t|^2\big)dt - d\zeta_t,$$

其中，$(\zeta_t)_{t\geq 0}$ 是一个非负递增过程，定义为

$$\zeta_t := \begin{cases} 0, & \text{if } t < T \text{ or } T = 0 \\ f(X_T), & \text{otherwise.}\end{cases}$$

取期望后，我们注意到 $(\varphi_t)_{t\geq 0}$ 是下半连续的，并且对于所有 $0 \leq s < t$，都有

$$\varphi_t - \varphi_s \leq \mathbb{E}\Big[\int_s^t \mathbb{1}_{\{T<u\}}(E_u - af(X_u))\, du\Big] = \int_s^t (e_u - a\varphi_u)\, du.$$

在 [MNPR20] 中应用命题2.3，可得对于所有 $t \geq 0$,

$$\begin{aligned}\varphi_t &\leq \varphi_0 \exp(-at) + \int_0^t \exp(a(s-t)) e_s ds \\ &= \varphi_0 \exp(-at) + e_0 \exp(-at) \int_0^t \exp((2a-\mu-b)s) ds.\end{aligned}$$

$$\varphi_t \leq \varphi_0 \exp(-at) + e_0\Big(\frac{1}{2a-\mu-b}\big(\exp((a-\mu-b)t) - \exp(-at)\big)\Big)$$

反过来，对于 $2a-\mu-b=0$ ，我们得到

$$\varphi_t \leq (\varphi_0 + e_0 t)\exp(-at).$$

关于 $(\mathbb{E}[\mathbb{1}_{\{T>t\}}|V_t|^2])_{t\geq 0}$的收敛性，注意到由于 $f(X_t) \geq 0$

$$\frac{b}{2}|V_t|^2 \leq E_t - \langle \nabla f(X_t), V_t \rangle \leq E_t + |\nabla f(X_t)|\,|V_t|,$$

由此类比（25）可得

$$\frac{b}{4}\mathbb{E}[\mathbb{1}_{\{T>t\}}|V_t|^2] \leq e_t + \frac{1}{b}\varphi_t. \tag{32}$$

根据上述计算，存在一个常数 $C \geq 0$，使得

$$\mathbb{E}[\mathbb{1}_{\{T>t\}}|V_t|^2] \leq \begin{cases} C\exp(-mt), & \text{if } a \neq \mu - a + b, \\ C(1+t)\exp(-mt), & \text{if } a = \mu - a + b. \end{cases}$$

(ii)：利用 (i)，我们得到

$$\mathbb{E}\Big[\int_0^T |V_s|\,ds\Big] \leq \int_0^\infty \mathbb{E}[\mathbb{1}_{\{T>s\}}|V_s|^2]^{1/2}\,ds < \infty$$

使得 $\int_0^T |V_s|\,ds$ 几乎必然有限。由于，$|X_t - X_s| \leq \int_s^t |V_u|\,du$，对所有 $0 \leq s \leq t$，$(X_t)_{t\geq 0}$在 $\{T = \infty\}$上几乎必然收敛。□

下一个引理表明，如果摩擦相对于随机噪声的大小足够大，我们就可以找到参数 $a,\ b > 0$，使得命题 4.6 适用。

引理 4.7。设 $L, \sigma > 0$。那么，对所有 $\mu > \frac{C_L\sigma}{4L}$，都存在 $a, b > 0$，使得(31) 成立且$ab \geq C_L$。

证明。设 $b > 0$并选取 $a = b + 2\frac{C_L}{b} - \mu$。注意，$a > 0$当且仅当 $\mu < b + 2\frac{C_L}{b}$，且 $a - \mu - b < 0$当且仅当 $\mu > \frac{C_L}{b}$。现在，$\frac{b}{2}\mu - C_L + \frac{b}{2}a - \frac{b^2}{2} = 0$ 并且

$$\frac{b}{2}\sigma - a^2 + a\mu + ab - 2L = -2\mu^2 + \Big(2b + \frac{6C_L}{b}\Big)\mu - 4\frac{C_L^2}{b^2} - 2(C_L + L) + \frac{b}{2}\sigma.$$

后一个方程右侧是一个仅在根之间为正的二次函数

$$\mu_\pm^b = \frac{1}{2}\Big(b + \frac{3C_L}{b} \pm \sqrt{\Big(b + \frac{C_L}{b}\Big)^2 - 4L + b\sigma}\Big). \tag{33}$$

注意，对于 $b < \frac{4L}{\sigma}$，我们有 $\mu_-^b > \frac{C_L}{b}$ 和 $\mu_+^b < b + 2\frac{C_L}{b}$。此外，假设 $ab \geq C_L$ 当且仅当 $\mu \leq b + \frac{C_L}{b}$ 成立。因此，对于给定的数对 $(a,\ b)$，满足 (31) 的摩擦参数集合 μ 等于

$$\Big(\frac{C_L}{b}, \mu_-^b\Big] \cup \Big[\mu_+^b, b + 2\frac{C_L}{b}\Big)$$

此外，对于给定的数对 $(a,\ b)$，还满足 $ab \geq C_L$ 的摩擦参数集合 μ 包含于 $(\frac{C_L}{b},\ \mu_-^b \wedge (b + \frac{C_L}{b}))$。注意，对于所有 $0 < b < \frac{4L}{\sigma}$，后一个区间均非空，上界和下界关于 b 连续，且下界满足

$$\frac{C_L}{b} \xrightarrow{b\to 0} \infty \quad \text{and} \quad \frac{C_L}{b} \xrightarrow{b\to 4L/\sigma} \frac{C_L\sigma}{4L}.$$

因此，我们证明了对于每个 $\mu > \frac{C_L\sigma}{4L}$，都存在一个数对 $(a,\ b)$，使得 (31) 成立且 $ab \geq C_L$。□

现在我们可以证明定理 4.1。与离散时间中的对应结果相比，证明的第二部分更加复杂，因为我们无法立即使用博雷尔-坎特利引理。还需要额外一步，证明该过程与其在离散时刻取值之间的偏差不会太大。

定理 4.1 的证明。(i) 和 (iii)：显然，T 是一个满足命题 4.6 假设的 $(\mathcal{F}_t)_{t\geq 0}$-停时。由引理 4.7，存在参数 $a,\ b > 0$，使得 (31) 成立，并且

$ab \geq C_L$。若 $a \neq \mu - a + b$，则令 $m =$ 为 $\min(a, \mu - a + b)$；否则令其为 $m \in (a, \infty)$。那么，命题 4.6 意味着存在一个 $C > 0$，使得对所有 $t \geq 0$

$$\mathbb{E}[\mathbb{1}_{\{T>t\}} f(X_t)] \leq C \exp(-mt).$$

并且，$(X_t)_{t\geq 0}$ 在事件 $\{T = \infty\}$ 上几乎必然收敛。

(ii)：记 $C_L' := \|\nabla f\|_{\mathrm{Lip}(\mathbb{R}^d)} \vee \|\Sigma\|_{F,\ \mathrm{Lip}(\mathbb{R}^d)}$，其中 $\|\cdot\|_{F,\ \mathrm{Lip}(\mathbb{R}^d)}$ 是由 Frobenius 范数诱导的 Lipschitz 范数。令 $n \in \mathbb{N}_0$，并注意，对于 $t \in [n, n+1]$

$$\sup_{s\in[n,t]} \mathbb{1}_{\{T>s\}} |X_s - X_n|^2 \leq \int_{n\wedge T}^{t\wedge T} |V_u|^2 \, du \leq (t-n) \sup_{s\in[n,t]} \mathbb{1}_{\{T>s\}} |V_s|^2.$$

Now,

$$\begin{aligned}\mathbb{E}\Big[\sup_{s\in[n,t]} \mathbb{1}_{\{T>s\}} |V_s|^2\Big] \leq 4\Big(&\mathbb{E}[\mathbb{1}_{\{T>n\}} |V_n|^2] + \mu^2 \mathbb{E}\Big[\int_n^t \sup_{u\in[n,s]} \mathbb{1}_{\{T>u\}} |V_u|^2 \, ds\Big] \\ &+ \mathbb{E}\Big[\int_{n\wedge T}^{t\wedge T} |\nabla f(X_u)|^2 \, du\Big] + \mathbb{E}\Big[\sup_{s\in[n,t]} \Big|\int_{n\wedge T}^{s\wedge T} \Sigma(X_u) \, dW_u\Big|^2\Big]\Big).\end{aligned}$$

利用 ∇f 的 Lipschitz 连续性，可得

$$\mathbb{E}\Big[\int_{n\wedge T}^{t\wedge T} |\nabla f(X_u)|^2 \, du\Big] \leq 2\mathbb{E}[\mathbb{1}_{\{T>n\}} |\nabla f(X_n)|^2] + 2(C_L')^2 \mathbb{E}\Big[\int_n^t \sup_{s\in[n,u]} \mathbb{1}_{\{T>s\}} |V_s|^2 \, du\Big].$$

此外，利用 Doob 的 L^2-不等式和 Itˆo 等距性，

$$\begin{aligned}\mathbb{E}\Big[\sup_{s\in[n,t]} \Big|\int_{n\wedge T}^{s\wedge T} \Sigma(X_u) \, dW_u\Big|^2\Big] &\leq 4\mathbb{E}\Big[\int_{n\wedge T}^{t\wedge T} \|\Sigma(X_u)\|_F^2 \, du\Big] \\ &\leq 8\mathbb{E}[\mathbb{1}_{\{T>n\}} \|\Sigma(X_n)\|_F^2] + 8(C_L')^2 \mathbb{E}\Big[\int_n^t \sup_{s\in[n,u]} \mathbb{1}_{\{T>s\}} |V_s|^2 \, du\Big].\end{aligned}$$

因此，

$$\begin{aligned}\mathbb{E}\Big[\sup_{s\in[n,t]} \mathbb{1}_{\{T>s\}} |V_s|^2\Big] \leq 32\Big(&\mathbb{E}[\mathbb{1}_{\{T>n\}} (|V_n|^2 + |\nabla f(X_n)|^2 + \|\Sigma(X_n)\|_F^2)] \\ &+ (\mu^2 + 2(C_L')^2) \int_n^t \mathbb{E}\Big[\sup_{s\in[n,u]} \mathbb{1}_{\{T>s\}} |V_s|^2\Big] \, du\Big).\end{aligned}$$

因此，根据 Gronwall 不等式，存在一个常数 $C \geq 0$，使得对于所有 $n \in \mathbb{N}_0$ 和 $n \leq t \leq n+1$，我们有

$$\mathbb{E}\Big[\sup_{s\in[n,t]} \mathbb{1}_{\{T>s\}} |V_s|^2\Big] \leq C\, \mathbb{E}[\mathbb{1}_{\{T>n\}} (|V_n|^2 + |\nabla f(X_n)|^2 + \|\Sigma(X_n)\|_F^2)].$$

根据命题 4.6 (i)、引理 2.1 和 (30)，存在常数 $C, > 0$ 使得对于所有$n \in \mathbb{N}_0$

$$\mathbb{E}[\mathbb{1}_{\{T>n\}} (|V_n|^2 + |\nabla f(X_n)|^2 + \|\Sigma(X_n)\|_F^2)] \leq C \exp(-mn).$$

因此，根据 ∇f 的利普希茨连续性，

$$\begin{aligned}\mathbb{E}\Big[\sup_{s\in[n,n+1]}\mathbb{1}_{\{T>s\}}|f(X_s)-f(X_n)|\Big] &\le \mathbb{E}\Big[\int_{n\wedge T}^{(n+1)\wedge T}|\nabla f(X_u)|\,|V_u|\,du\Big]\\ &\le \mathbb{E}[\mathbb{1}_{\{T>n\}}|\nabla f(X_n)|^2]^{1/2}\mathbb{E}\Big[\sup_{s\in[n,n+1]}\mathbb{1}_{\{T>s\}}|V_s|^2\Big]^{1/2}\\ &\quad + C_L'\mathbb{E}\Big[\sup_{s\in[n,n+1]}\mathbb{1}_{\{T>s\}}|V_s|^2\Big]\\ &\le C\exp(-mn),\end{aligned}$$

对于常数 $C>0$。

接下来，我们证明该论断。对于 $m'<m$、$\varepsilon>0$ 和 $n\in\mathbb{N}$，考虑集合

$$\mathbb{B}_n=\{T=\infty\}\cap\Big\{\sup_{t\ge n}\exp(m't)f(X_t)\ge\varepsilon\Big\}.$$

我们使用马尔可夫不等式、∇f 的 Lipschitz 连续性以及 (i) 可得，对于某个 $C\ge 0$，有

$$\begin{aligned}\mathbb{P}(\mathbb{B}_n) &\le \sum_{i=n}^{\infty}\mathbb{P}(\{T\ge i\}\cap\{\exp(m'(i+1))f(X_i)\ge\varepsilon/4\})\\ &\quad+\sum_{i=n}^{\infty}\mathbb{P}\Big(\{T\ge i+1\}\cap\Big\{\sup_{t\in[i,i+1]}\exp(m'(i+1))(f(X_t)-f(X_i))\ge\varepsilon/4\Big\}\Big)\\ &\le \sum_{i=n}^{\infty}\exp(m'(i+1))\frac{4}{\varepsilon}\mathbb{E}[\mathbb{1}_{\{T\ge i\}}f(X_i)]\\ &\quad+\sum_{i=n}^{\infty}\exp(m'(i+1))\frac{4}{\varepsilon}\mathbb{E}\Big[\sup_{t\in[i,i+1]}\mathbb{1}_{\{T\ge t\}}|f(X_t)-f(X_i)|\Big]\\ &\le C\sum_{i=n}^{\infty}\exp((m'-m)i)\overset{n\to\infty}{\longrightarrow}0.\end{aligned}$$

因此，

$$\mathbb{P}\Big(\{T=\infty\}\cap\Big\{\limsup_{t\to\infty}\exp(mt)f(X_t)\ge\varepsilon\Big\}\Big)\le\mathbb{P}\Big(\bigcap_{n\in\mathbb{N}}\mathbb{B}_n\Big)=0$$

从而在 $\{T=\infty\}$ 上几乎必然有 $\exp(mt)f(X_t)\to 0$。 □

对于可容许摩擦参数 $\mu>\frac{C_L\sigma}{4L}$，我们利用命题 4.6 得到期望目标函数值关于技术参数 b 的收敛速率。为了获得最优收敛速率，我们对下述引理中定义的 $m(b,\mu)$ 在所有可容许的 b 取值上进行优化。

引理 4.8。设 $L,\sigma>0$。设 T 是一个 $(\mathcal{F}_t)_{t\ge0}$-停时，使得对所有 $t\ge0$

$$2Lf(X_t)\le|\nabla f(X_t)|^2\quad\text{and}\quad\|\Sigma(X_t)\|_F^2\le\sigma f(X_t),\quad\text{on }\{T>t\}.$$

令 $0<b<\frac{4L}{\sigma}$ 和 $\mu\in(\frac{C_L}{b},\mu_-^b]\cup[\mu_+^b,b+2\frac{C_L}{b})$，其中 $\mu_\pm^b$ 由(33)给出。若 $\mathbb{P}(T=\infty)<\mathbb{P}(T>0)$ 此外满足 $\mu\le b+\frac{C_L}{b}$，则对所有 $\varepsilon>0$，存在一个 $C>0$，使得对所有 $t\ge0$

$$\mathbb{E}[\mathbb{1}_{\{T>t\}}f(X_t)]\le C\exp(-m(b,\mu)t),\tag{34}$$

其中

$$m(b,\mu)=\begin{cases}2(\mu-\frac{C_L}{b}), & if\ \mu<\frac{1}{3}(b+4\frac{C_L}{b})\\ b+2\frac{C_L}{b}-\mu, & if\ \mu>\frac{1}{3}(b+4\frac{C_L}{b})\\ \frac{2}{3}(b+\frac{C_L}{b})-\varepsilon & if\ \mu=\frac{1}{3}(b+4\frac{C_L}{b}).\end{cases} \tag{35}$$

证明。设 b 和 μ 如假设中所述，并令 $a=b+2\frac{C_L}{b}-\mu$。那么，由命题 4.6 和引理 4.7 可得 $(\mathbb{E}[\mathbb{1}_{\{T>t\}}f(X_t)])_{t\geq 0}$ 以速率 $m(b,\ \mu)$ 指数收敛，其中当 $a\neq\mu-a+b$ 时，我们有

$$m(b,\mu)=\min(a,\mu-a+b)=\min\Big(b+2\frac{C_L}{b}-\mu,2\Big(\mu-\frac{C_L}{b}\Big)\Big)$$

否则，对于所有 $\varepsilon>0$，我们都可以选择 $m(b,\ \mu):=a-\varepsilon$。 □

为了推导任意固定摩擦参数 $\mu>\frac{C_L\sigma}{4L}$ 下的最优收敛速率，需要在所有可行参数 b中最大化 $m(b,\mu)$。我们将同时对 μ 和 b进行优化。

注 4.9。我们在可行的 b 和 μ取值范围内优化速率 $m(b,\ \mu)$。首先，固定 $b>0$，并注意到，在 $m(b,\ \mu)=2(\mu-\frac{C_L}{b})$ 的情况下，当 μ取可行范围内的最大值时，速率最优。若 $\Phi(b):=b^4+\frac{9}{8}\sigma b^3+\Big(2C_L-\frac{9}{2}L\Big)b^2+C_L^2>0$，则有 $\mu_-^b<\frac{1}{3}(b+4\frac{C_L}{b})$。因此，取 $\mu=\mu_-^b$，可得

$$m(b,\mu_-^b)=b+\frac{C_L}{b}-\sqrt{\Big(b+\frac{C_L}{b}\Big)^2-4L+b\sigma}.$$

如果 $\Phi(b)\leq 0$我们有 $\mu_-^b\geq\frac{1}{3}(b+4\frac{C_L}{b})$。因此，取 $\mu=\frac{1}{3}(b+4\frac{C_L}{b})$，对于任意 $\varepsilon>0$，

$$m\Big(b,\frac{1}{3}\Big(b+4\frac{C_L}{b}\Big)\Big)=\frac{2}{3}\Big(b+\frac{C_L}{b}\Big)-\varepsilon.$$

在 $m(b,\ \mu)=b+2\frac{C_L}{b}-\mu$的情形下，对于最小的可取 μ，该收敛速率是最优的。如果 $\Phi(b)\geq 0$，我们取 $\mu=\mu_+^b$，得到

$$m(b,\mu_+^b)=\frac{1}{2}\left(b+\frac{C_L}{b}-\sqrt{\Big(b+\frac{C_L}{b}\Big)^2-4L+b\sigma}\right).$$

如果 $\Phi(b)<0$，我们取 $\mu=\frac{1}{3}(b+4\frac{C_L}{b})$，对于任意 $\varepsilon>0$，得到

$$m\Big(b,\frac{1}{3}\Big(b+4\frac{C_L}{b}\Big)\Big)=\frac{2}{3}\Big(b+\frac{C_L}{b}\Big)-\varepsilon.$$

定理 4.2 的证明。 注意，$C_L^*\geq C_L$ 使得 C_L^* 也是 ∇f 的 Lipschitz 常数。根据上述计算，将 C_L 替换为C_L^*、$b=\sqrt{C_L^*}$、$a=3\sqrt{C_L^*}-\mu$ 和 $\mu=2\sqrt{C_L^*}-\sqrt{C_L^*-L+\frac{1}{4}\sqrt{C_L^*}\sigma}$后，满足命题 4.6 的假设。特别地，$ab\geq C_L^*$，因为 $\mu\leq b+\frac{C_L^*}{b}$；且 $\Phi(b)>0$，因为$\sigma>0$。因此，存在一个常数 $C>0$ 使得

$$\mathbb{E}[\mathbb{1}_{\{T>t\}}f(X_t)]\leq C\exp(-mt),$$

其中 $m=2(\mu-\frac{C_L^*}{b})=2\sqrt{C_L^*}-\sqrt{4C_L^*-4L+\sqrt{C_L^*}\sigma}$.

□

˜推论 *4.4* 的证明。取足够小的 r_0，使得 $B_{r_0}(y) \subset \mathcal{D}$。令 $r_1 < r_0$, $T'_{r_1} = \inf\{t \geq 0 : \int_0^t |V_s|\, ds > r_0 - r_1\}$，并注意到 $T'_{r_1} < T$ 几乎必然成立。因此，由 (32) 可得，

$$\mathbb{P}(T < \infty) \leq \mathbb{P}\Big(\int_0^T |V_s|\, ds \geq r_0 - r_1\Big) \leq \frac{1}{r_0 - r_1}\mathbb{E}\Big[\int_0^T |V_s|\, ds\Big]$$
$$\leq \frac{1}{r_0 - r_1} C(e_0, \varphi_0) \int_0^\infty \exp(-ms)\, ds,$$

对于一个 $m > 0$ 和一个常数 $C(e_0, \varphi_0)$，它仅依赖于 $e_0 = \mathbb{E}[1\!\!1_{\{T>0\}} E_0]$ 和 $\varphi_0 = \mathbb{E}[1\!\!1_{\{T>0\}} f(X_0)]$，并满足 $C(e_0, \varphi_0) \to 0$ 当 $(e_0, \varphi_0) \to 0$ 时。因此，对于每个 $\varepsilon > 0$，都存在一个 $r_0 > 0$，使得 $\mathbb{P}(T < \infty) \leq \varepsilon$。□

致谢。作者感谢 Vitalii Konarovskyi 仔细校对本文并提出许多宝贵建议。感谢 Benjamin Fehrman 在本项目初期进行的富有成效的讨论。作者获得了德国研究联合会（DFG，德国研究联合会）资助，项目为 SFB 1283/22021 – 317210226。BG 感谢马克斯·普朗克学会通过研究组“科学中的随机分析（SAiS）”提供的支持。

References

[ADR22a] J.-F. Aujol、C. Dossal 和 A. Rondepierre。拟强凸优化中重球法的收敛速率。 *SIAM J. Optim.*，32(3):1817–1842，2022年。[ADR22b] J.-F. Aujol、C. Dossal 和 A. Rondepierre。Łojasiewicz 性质下重球法的收敛速率。-，第 1–60 页，2022年。 *Math. Program.*AGV[V. Apidopoulos、N. Ginatta 和 S. Villa。在 Polyak–-Łojasiewicz 条件下，非凸优化中重球连续动力学的收敛速率。 *J. Global Optim.*，第 1–27 页，2022年。[Ben12] Y. Bengio。深层架构基于梯度的训练实用建议。载于 *Neural networks: Tricks of the trade: Second edition* 437 478 2012 ，第 – 页。

施普林格，柏林，海德堡，。[CEG07] A. Cabot、H. Engler 和 S. Gadat。关于具有渐近微小耗散的二阶微分方程的长期行为。 *Trans. Amer. Math. Soc.*，361，2007年 11 月。[Coo21] Y. Cooper。过参数化神经网络的全局极小值。 *SIAM J. Math. Data Sci.*，3(2):676–691，2021年。[DHS11] J. Duchi、E. Hazan 和 Y. Singer。用于在线学习和随机优化的自适应次梯度方法。 *J. Mach. Learn. Res.*，12(7)，2011年。[DK22a] S. Dereich 和 S. Kassing。随机微分方程的冷却：几乎必然收敛。*Stochastic Process. Appl.*，152:289–311，2022年。[DK22b] S. Dereich 和 S. Kassing。关于浅层 ReLU 网络的最小表示。 *Neural Networks*，148 121 128，2022年：–。[DK23] S. Dereich 和 S. Kassing。具有平均化的随机梯度下降在稳定流形上的中心极限定理。 *Electron. J. Probab.*，28:1–48，2023年。[DK24] S. Dereich 和 S. Kassing。Łojasiewicz 景观下随机梯度下降方案的收敛性。 *J. Mach. Learn.*，3(3):245–281，2024年。[DKP20] M. Danilova、A. Kulakova 和 B. Polyak。重球法的非单调行为。载于 *Difference Equations and Discrete Dynamical Systems with Applications: 24th ICDEA*、 *Dresden*、 *Germany*、 *May 21–25*，2018年 24， 第 213–230 页。施普林格，2020年。[EBB+21] M. Even、R. Berthier、F. Bach、N. Flammarion、H. Hendrikx、P. Gaillard、L. Massouli´e 和 A. Taylor。确定性与随机梯度下降以及 gossip 算法的连续化加速方法。*Neural Information Processing Systems*，34:28054–28066，2021年。[Fee19] P. Feehan。奇点消解与 Łojasiewicz 不等式的几何证明。，( ): –。 - *Geom.* FGJ[B. *Topol.* 23 7 3273 3313, 2019
Fehrman、B. Gess 和 A. Jentzen。非凸目标函数的随机梯度下降法的收敛速率。20] ，21(136):1–48，2020年。 *J. Mach. Learn. Res.*Gar[G. Garrigos。平方距离函数是 Polyak-Łojasiewicz 的，反之亦然。arXiv:2301.10332，2023年。

[GFJ15] E. Ghadimi、H. R. Feyzmahdavian 和 M. Johansson。凸优化中重球法的全局收敛性。载于 *European control conference*，第 310–315 页。IEEE，2015年。[GG22] S. Gadat 和 I. Gavra。非凸景观中随机自适应算法的渐近研究。*J. Mach. Learn. Res.*，23(228):1–54，2022年。[GGK22] B. Gess、R. S. Gvalani 和 V. Konarovskyi。守恒型随机偏微分方程作为随机梯度下降的涨落平均场极限。arXiv:2207.05705，2022年。[GKK24] B. Gess、S. Kassing 和 V. Konarovskyi。随机修正流、平均场极限与随机梯度下降的动力学。*J. Mach. Learn. Res.*，25(30):1–27，2024年。[GPS18] S. Gadat、F. Panloup 和 S. Saadane。随机重球法。*Electron. J. Stat.*，12(1):461–529，2018年。[GTD23] B. Goujaud、A. Taylor 和 A. Dieuleveut。可证明的重球法非加速性。arXiv:2307.11291，2023年。[HLZ19] W. Hu、C. J. Li 和 X. Zhou。关于非凸优化连续时间随机重球法的全局收敛性。载于 *2019 IEEE International Conference on Big Data*，第 94–104 页。IEEE，2019年。[KB15] D. P Kingma 和 J. Ba。Adam：一种随机优化方法。载于 *3rd International Conference on Learning Representations* 2015 ，。[KMN+16] N. S. Keskar、D. Mudigere、J. Nocedal、M. Smelyanskiy 和 P. T. P. Tang。关于深度学习的大批量训练：泛化差距与尖锐极小值。arXiv:1609.04836，2016年。[KNS16] H. Karimi、J. Nutini 和 M. Schmidt。Polyak-Łojasiewicz 条件下梯度法和近端梯度法的线性收敛。载于 *Joint European Conference on Machine Learning andKnowledge Discovery in Databases*，第 795–811 页。施普林格，2016年。[KR23] A. Khaled 和 P. Richt ´ *Transactions on Machine* arik。非凸 *Learning Research* 2023 世界中关于随机梯度下降的更好理论。，。[KSA23] I. A. Kuruzov、F. S. Stonyakin 和 M. S. Alkousa。具有 Polyak-Łojasiewicz 条件的优化问题的梯度型方法：提前停止与对不精确参数的自适应性。载于 *Advances in Optimization and Applications: 13th International Conference*、 *OPTIMA 2022*，第 18–32 页。施普林格，2023年。[Loj63] S. ´ ´ ´ *Les équations aux dérivées partielles* 117 87 89, 1963

Lojasiewicz。实解析子集的一个拓扑性质。，：–。[LP16] B. Lessard、L. 和 Recht 以及 A. Packard。通过积分二次约束分析和设计优化算法。 *SIAM J. Optim.*，26(1):57–95，2016年。[LR17] N. Loizou 和 P. Richt´arik。用于最小化泛化误差的线性收敛随机重球法。载于 *OPTML 2017*，2017年。[LR20] N. Loizou 和 P. Richt´arik。随机梯度、牛顿法、近端点法和子空间下降法中的动量与随机动量。 *Comput. Optim. Appl.*，77(3):653–710，2020年。[LTE19] Q. Li、C. Tai 和 W. E. 随机梯度算法的随机修正方程与动力学 I：数学基础。 *J. Mach. Learn. Res.*，20(1):1474–1520，2019年。[LY23] J. Liu 和 Y. Yuan。在没有有界梯度假设的情况下随机梯度法几乎必然规避鞍点。arXiv:2302.07862，2023年。[MNPR20] R. Matusik、A. Nowakowski、S. Plaskacz 和 A. Rogowski。微分包含的有限时间稳定性及其在神经网络中的应用。 *SIAM J. Control Optim.*，58(5):2854–2870，2020年。[Nes83] Y. Nesterov。具有收敛速率$O(1/k^2)$的无约束凸极小化问题的一种方法。载于 *Doklady an ussr*，第 269 卷，第 543–547 页，1983年。[Pol63] B. T. Polyak。泛函极小化的梯度法。 *U.S.S.R. Comput. Math. and* ，（）：–。[*Math. Phys.* 3 4 864 878, 1963
Pol64] B. T. Polyak。加速迭代法收敛的一些方法。 *U.S.S.R. Comput.* ，（）：–。[*Math. and Math. Phys.* 4 5 1 17, 1964
PS17] B. T. Polyak 和 P. Shcherbakov。李雅普诺夫函数：优化理论视角。 *IFAC-PapersOnLine* 50 1 7456 7461, 2017 ，（）：–。[

Qia99] N. Qian。梯度下降学习算法中的动量项。 *Neural Networks*，12(1):145–151，1999年。[RB24] Q. Rebjock 和 N. Boumal。快速收敛到非孤立极小值：关于C^2 函数的四个等价条件。 *Math. Program.*，2024年。[RM51] H. Robbins 和 S. Monro。一种随机逼近法。 *Ann. Math. Statistics*，第 400–407 页，1951年–。

[随机梯度下降21] O. Sebbouh、R. M. Gower 和 A. Defazio。随机梯度下降和随机重球法的几乎必然收敛速率。载于 *Conference on Learning Theory*，第 3935–3971 页。PMLR，2021年。[SMDH13] I. Sutskever、J. Martens、G. Dahl 和 G. Hinton。论初始化和动量在深度学习中的重要性。载于 *International Conference on Machine Learning*，第 1139–1147 页。PMLR，2013年。[VBS19] S. Vaswani、F. Bach 和 M. Schmidt。过参数化模型的随机梯度下降更快收敛及加速感知机。载于 *The 22nd international conference on artificial intelligenceand statistics*，第 1195–1204 页。PMLR，2019年。[WKM23] D. Wu、V. Kungurtsev 和 M. Mondelli。重球法的平均场分析：Dropout 稳定性、连通性与全局收敛。 *Transactions on Machine Learning Research*，2023年。[Woj23] S. Wojtowytsch。带有机器学习类型噪声的随机梯度下降，第一部分：离散时间分析。 *J. Nonlinear Sci.*，33(3):45，2023年。[Woj24] S. Wojtowytsch。带有机器学习类型噪声的随机梯度下降，第二部分：连续时间分析。 *J. Nonlinear Sci.*，34(1):16，2024年。[YFL23] P. Yue、C. Fang 和 Z. Lin。关于最小化 Polyak-

The Łojasiewicz 函数

Thirty Sixth Annual Conference on Learning Theory 2948 2968 2023
的下界。载于 ，第 – 页。PMLR，。

Benjamin Gess, Fakultät für Mathematik, Universität Bielefeld, Universitätsstrasse 25, 33615Bielefeld, Germany, Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103Leipzig, Germany*Email address*:bgess@math.uni-比勒费尔德.de

Sebastian Kassing, Fakultät für Mathematik, Universität Bielefeld, Universitätsstrasse 25,33615 Bielefeld, Germany*Email address*: skassing@math.uni-bielefeld.de