Heavy Ball 动量加速随机梯度下降中的隐式正则化

Avrajit Ghosh*[1]、He Lyu*[1]、Xitong Zhang[1], 和 Rongrong Wang[1]

[1]密歇根州立大学计算数学、科学与工程系

arXiv:2302.00849v1 [cs.LG] 2 Feb 2023

摘要

众所周知，梯度下降（GD）中的有限步长（h）会隐式地将解正则化至更平坦的极小值。一个自然的问题是*"动量参数 β 是否会在 Heavy-ball（H.B）动量加速梯度下降（GD+M）中发挥隐式正则化作用？"*为回答这一问题，我们首先证明，离散的 H.B 动量更新（GD+M）遵循一条由修正损失诱导的连续轨迹，该修正损失由原始损失和一个隐式正则化项组成。随后，我们证明，（GD+M）的隐式正则化项比（GD）的强 $(\frac{1+\beta}{1-\beta})$ 倍，从而解释了为什么（GD+M）比（GD）具有更好的泛化性能和更高的测试准确率。此外，我们将分析扩展至带动量的随机梯度下降（SGD+M），并以逐点意义刻画（SGD+M）更新的连续轨迹。我们通过一系列实验探究（SGD+M）和（GD+M）中的隐式正则化，并验证了我们的理论。

1 引言

近年来，深度神经网络（神经网络）在众多任务中取得了巨大的实证成功，应用范围涵盖计算机视觉、自然语言处理、自动驾驶、医学影像、天文学和物理学（Bengio & LeCun，2007；Hinton 等人，2006；Goodfellow 等人，2016）。大多数深度学习问题本质上都是在求解过参数化、大规模的非凸优化问题。近年来，一个引起广泛关注的神经网络之谜是：神经网络为何具有如此出色的泛化能力。事实上，即使模型极度过参数化，神经网络也很少表现出过拟合的迹象（Neyshabur，2017）。迄今为止，这方面的研究已成功揭示了多种隐式正则化形式；当使用梯度下降或随机梯度下降算法进行训练时，这些形式可能有助于实现良好的泛化，包括范数惩罚（Soudry 等人，2018）、隐式梯度正则化（Barrett & Dherin，2020），以及通过噪声注入实现的隐式海森正则化（Orvieto 等人，2022a；b）。

相比之下，动量加速梯度下降方法家族包括 Polyak 重球动量（Polyak，1964）、Nesterov 动量（Sutskever 等，2013）、RMSProp（Tieleman 等，2012）和 Adam（Kingma 与 Ba，2014）。尽管这些方法是 SGD 的有力替代方案，且收敛速度更快，但就隐式正则化而言，我们对它们的理解仍远远不够。本文分析重球动量加速 SGD（SGD+M）算法中的隐式梯度正则化，旨在进一步从理论上揭示动量如何影响 SGD 的泛化性能，以及它为何往往会带来方差缩减效应，且该效应的强度会随动量参数增大而增强。

2 相关文献

已有大量研究表明，基于梯度的优化会隐式地使解偏向复杂度较低的模型，从而有助于提升泛化能力。例如，在过参数化的

*表示贡献相同

二次模型中，从接近于零的初始化开始进行梯度下降，会隐式地使解偏向核范数较小的模型（Arora et al., 2019; Gunasekar et al., 2017; Razin & Cohen, 2020）；在最小二乘回归问题中，从 0 初始猜测出发的梯度下降解会偏向具有最小 ℓ_2 范数的解（Soudry et al., 2018; Neyshabur et al., 2014; Ji & Telgarsky, 2019; Poggio et al., 2020）。类似地，对于具有可分数据的线性分类问题，梯度下降的解会偏向最大间隔（即具有最小 ℓ_2 范数的）解（Soudry et al.,2018）。然而，在(Vardi &Shamir,2021)中，作者指出，在简单情形下证明的这些基于范数的正则化结果未必适用于非线性神经网络。

针对所有非线性模型（包括神经网络）发现的首个通用梯度下降隐式正则化方法是隐式梯度正则化（IGR）（Barrett & Dherin，2020）。研究表明，梯度下降（GD）中的学习率会惩罚损失函数梯度的二阶矩，从而促使模型找到更平坦的最优解。更平坦的最优解通常能带来更高的测试准确率，并且对参数扰动更加稳健（Barrett & Dherin，2020）。

随机梯度下降（SGD）也被发现具有隐式梯度正则化（Smith 等人，2021；Li 等人（2019）），这是其泛化性能良好的原因之一（但可能并非唯一原因）。人们认为，SGD 也受益于其随机性，这种随机性可能起到噪声注入的作用，从而提升性能。事实上，研究表明（Wu 等人，2020），通过向梯度注入噪声，全批量梯度下降能够达到使用小批量的 SGD 的性能。除了向梯度注入噪声，人们还发现许多其他噪声注入方式对模型参数具有隐式正则化效果，包括向模型空间注入噪声（Orvieto 等人，2022b）以及向网络激活值注入噪声（Camuto 等人，2020）。然而，这些不同类型的正则化如何协同影响泛化，目前仍不甚明了。

动量加速梯度下降的泛化效应鲜少受到研究。Li 等人（2019）分析了 SGD+M 的轨迹，发现该轨迹可以由某些伊藤随机微分方程的解弱逼近，这暗示了 (SGD+M) 中存在 IGR。然而，IGR 的显式公式及其与泛化的关系仍然未知。最近，(Wang et al., 2021) 基于可分数据上的线性分类问题分析了动量 (GD+M) 中的隐式正则化，并表明 (GD+M) 会收敛到 L_2 最大间隔解。尽管这是首次为基于动量的方法提出的隐式正则化形式之一，但它未能揭示非线性神经网络中动量隐式正则化的本质。

最近，(Jelassi 和 Li，2022) 表明，在某些特殊设置下（即使用双层网络的简单二分类问题，且部分输入特征远弱于其余特征），(GD+M) 能够提升网络的泛化能力；但尚不清楚从这一特殊设置中获得的见解在多大程度上适用于实际的神经网络模型。

据我们所知，尚无先前研究针对一般的非线性神经网络推导出（SGD+M）的隐式正则化。

3 梯度下降的隐式梯度正则化及其与泛化的关系

我们简要回顾针对梯度下降定义的隐式梯度正则化（Barrett 与 Dherin，2020），我们的分析将以此为基础。设 $E(\mathbf{x})$为定义在神经网络参数空间$\mathbf{x} \in \mathbb{R}^p$ 上的损失函数。梯度下降迭代会沿当前迭代点处损失函数梯度的反方向迈出一个离散步长（h）

$$\mathbf{x}^{k+1} = \mathbf{x}^k - h\nabla E(\mathbf{x}^k). \tag{1}$$

当步长趋于无穷小（$h \to 0$）时，梯度下降的轨迹收敛于一阶常微分方程的轨迹

$$\mathbf{x}'(t) = -\nabla E(\mathbf{x}(t)) \tag{2}$$

该方程称为梯度流。但当步长 h 有限（尽管很小）时，梯度下降的更新步骤会偏离梯度流的路径，而更接近修正流的路径：

$$\mathbf{x}'(t) = -\nabla \hat{E}(\mathbf{x}(t)), \qquad \text{where } \hat{E}(\mathbf{x}) = E(\mathbf{x}) + \frac{h}{4}\|\nabla E(\mathbf{x})\|^2. \tag{3}$$

通过所谓的经典后向分析，已有研究（Barrett 与 Dherin，2020）表明，当梯度下降和两个梯度流 2 与 3 均从同一点 $\mathbf{x}^k$ 出发时，下一次梯度更新 $\mathbf{x}^{k+1}$与原始梯度流 (2) 在下一个时间点 $t_{k+1} = h + t_k$ 处的取值相距 $O(h^2)$，但与修正流 (3) 在 t_{k+1} 处的取值相距 $O(h^3)$。因此，在局部范围内，修正流比原始流更贴近梯度下降轨迹。要讨论梯度下降的全局行为，我们需要以下关于两条轨迹之间接近程度的定义。

定义 3.1 （$O(h^\alpha)$-强意义下的接近）**.** 固定某个 $T > 0$，若离散梯度下降更新 $\mathbf{x}^k$ 与连续流 $\tilde{\mathbf{x}}(t_k)$ 在强意义下是 $O(h^\alpha)$-接近的*，则称：

$$\max_{k \in \mathcal{K}} \|\mathbf{x}^k - \tilde{\mathbf{x}}(t_k)\|_2 \leq c h^\alpha, \quad \mathcal{K} = \left\{1, \ldots \left\lfloor \frac{T}{h} \right\rfloor \right\},$$

其中 $t_k = hk$，且 c 是一个与 h 和 k 无关的常数

定义 3.1 量化了离散轨迹与连续轨迹之间的全局接近性。通过累积局部误差，可以证明从相同位置出发时，$\mathbf{x}^0$, 原始梯度流方程 2 与 $O(h)$梯度下降轨迹的接近程度，而修正流 3 与其的接近程度为$O(h^2)$。基于这一观察，作者将修正流中的 $O(h)$ 项 $\frac{h}{4}\|\nabla E(\mathbf{x})\|^2$ 定义为 IGR 项，并指出它能引导解趋向非凸景观中更平坦的极小值。

为了说明为什么最小化 $\|\nabla E(\mathbf{x})\|^2$ 是合理的，以及为什么它会促使极小化解趋于平坦（原论文似乎没有对此作出解释），我们借鉴了 Foret 等人（2020）的一种论证。该论证最初是为另一种目的提出的。

根据 PAC-贝叶斯分析（Neyshabur，2017），在满足某些技术条件时，可以得到神经网络的一个简化泛化界（Foret 等人，2020）：

$$\mathcal{L}_{\mathcal{D}}(\mathbf{x}) \leq \max_{\|\epsilon\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\mathbf{x} + \epsilon) + \hat{h}\left(\frac{\|\mathbf{x}\|^2}{\rho}\right), \text{ for any } \rho > 0,$$

其中，$\mathcal{L}_{\mathcal{D}}$ 是总体损失（即泛化误差），$\mathcal{L}_{\mathcal{S}}$ 是经验损失/训练损失，而 $h : \mathbb{R}_+ \to \mathbb{R}_+$ 是某个严格递增函数。可以尝试最小化这个上界，以降低泛化误差 $\mathcal{L}_{\mathcal{D}}$。通过在训练过程中启用权重衰减惩罚，可以控制上界中的项$\hat{h}\left(\frac{\|\mathbf{x}\|^2}{\rho}\right)$，而上界中的第一项通常写成

$$\max_{\|\epsilon\| \leq \rho} \mathcal{L}_{\mathcal{S}}(\mathbf{x} + \epsilon) = \underbrace{\max_{\|\epsilon\| \leq \rho} (\mathcal{L}_{\mathcal{S}}(\mathbf{x} + \epsilon) - \mathcal{L}_{\mathcal{S}}(\mathbf{x}))}_{\text{sharpness}} + \underbrace{\mathcal{L}_{\mathcal{S}}(\mathbf{x})}_{\text{training loss}}$$

它由训练损失和一个称为尖锐度的额外项组成，最小化尖锐度有助于泛化。由于直接最小化尖锐度很困难，因此可以采用以下一阶泰勒近似

$$\max_{\|\epsilon\| \leq \rho} (\mathcal{L}_{\mathcal{S}}(\mathbf{x} + \epsilon) - \mathcal{L}_{\mathcal{S}}(\mathbf{x})) \approx \max_{\|\epsilon\| \leq \rho} \epsilon^T \nabla \mathcal{L}_{\mathcal{S}}(\mathbf{x}) = \rho \|\nabla \mathcal{L}_{\mathcal{S}}(\mathbf{x})\|.$$

用我们的记号，$\nabla \mathcal{L}_{\mathcal{S}}(\mathbf{x})$ 是 $\nabla E(\mathbf{x})$，因此尖锐度近似正比于 $\|\nabla E\|$，后者是 IGR 项的平方根。

*Li 等人（2019）研究的弱意义近似只要求样本过程$\mathbf{x}$ 和 $\tilde{\mathbf{x}}$ 的分布接近，而我们的强意义近似则要求$\mathbf{x}$ 和 $\tilde{\mathbf{x}}$ 的每个实例都彼此接近。通过后者得到的强意义 IGR 适用于具有任意固定随机批次序列和任意固定初始化的训练；而前者只能刻画对许多不同训练（每次训练都使用随机批次序列和初始化）取期望得到的平均轨迹。

4 重球法加速梯度下降的隐式梯度正则化

研究动量更新中的 IGR 时，主要的数学挑战在于，我们现在需要进行全局误差分析，而不是局部反向分析，因为动量更新利用了整个更新历史。

此前，Kovachki 和 Stuart（2021）通过研究重球动量与阻尼二阶哈密顿动力学之间的关系，对重球动量的隐式梯度正则化进行了分析。

$$m\mathbf{x}'(t) + \gamma\mathbf{x}''(t) + \nabla E(\mathbf{x}(t)) = 0$$

该动力学通常被认为是动量更新所对应的底层常微分方程。然而，在一般步长假设下，只证明了动量更新与该常微分方程轨迹之间存在$O(h)$级的接近程度*，这还不够，因为隐式正则化项 $\frac{h}{4}\|\nabla E\|^2$ 本身的阶数为 $O(h)$。此外，这种方法难以应用于随机情形。

本文规避了对二阶常微分方程的使用（因为它只能给出 $O(h)$级的接近程度），并直接构造了一条连续路径，使其对梯度下降和随机梯度下降的动量更新均达到 $O(h^2)$级的接近程度。具体而言，我们将动量更新与一阶分段常微分方程联系起来，证明该常微分方程存在一条分段可微轨迹，与动量更新的接近程度为 $O(h^2)$，随后利用该轨迹研究隐式梯度正则化。详细论证见附录。这里给出基于（重球法）动量的梯度下降法（IGR-M）之隐式正则化的最终数学公式。

定理 4.1。*(IGR-M)*：设全批量梯度 E 对应的损失函数光滑且四次可微，则 *(*梯度下降+动量*)* 更新

$$\mathbf{x}^{k+1} = \mathbf{x}^k - h\nabla E(\mathbf{x}^k) + \beta(\mathbf{x}^k - \mathbf{x}^{k-1}) \quad \forall k = 1, 2, ..., n$$

在 $O(h^2)$ 意义下，接近分段一阶常微分方程所描述的连续轨迹流

$$\widetilde{\mathbf{x}}'(t) = -\frac{1-\beta^{k+1}}{1-\beta}\nabla E(\widetilde{\mathbf{x}}(t)) - \frac{h\gamma(1+\beta)}{2(1-\beta)^3}\nabla^2 E(\widetilde{\mathbf{x}}(t))\nabla E(\widetilde{\mathbf{x}}(t)), t \in [t_k, t_{k+1}] \tag{4}$$

其中 $t_k = kh$和

$$\gamma = (1-\beta^{2k+2}) - 4(k+1)\beta^{k+1}\frac{(1-\beta)}{(1+\beta)}.$$

*Sinc*当 k 增大时，*e* β^k 会迅速衰减至 *0*；对于足够大的迭代次数 k，方程 *4* 可简化为：

$$\widetilde{\mathbf{x}}'(t) = -\frac{1}{1-\beta}\nabla\hat{E}(\widetilde{\mathbf{x}}(t)), \quad t \in [0, T] \tag{5}$$

由修正损失 $\hat{E}(\widetilde{\mathbf{x}}(t)) := E(\widetilde{\mathbf{x}}(t)) + \frac{h(1+\beta)}{4(1-\beta)^2}\|\nabla E(\widetilde{\mathbf{x}}(t))\|_2^2$驱动。更具体地说，对于固定时间 T，存在一个常数 C，使得对于任意学习率 $0 < h < T$，我们有

$$\|\widetilde{\mathbf{x}}(t_k) - \mathbf{x}^k\|_2 \leq Ch^2,\ t_k = kh, \quad k = 1, 2, ..., \lfloor\frac{T}{h}\rfloor. \tag{6}$$

比较由式 5 描述的连续轨迹与不含动量的式 3 所描述的轨迹，我们立即得到几个重要观察：

注 4.1。忽略 5 和 3 中的 $O(h)$项，我们可以看到，动量轨迹受到的力是 GD 的 $\frac{1}{1-\beta}$ 倍。因此，在使用相同学习率的情况下，预计 (GD+M) 的收敛速度将是 GD 的 $\frac{1}{1-\beta}$ 倍。或者，学习率为 h 的 (GD+M) 与学习率为 $\frac{h}{1-\beta}$的 GD 具有大致相同的收敛速度。从现在起，我们将 $\frac{h}{1-\beta}$ 称为 (GD+M) 的有效学习率。

*在对学习率施加极其严格的条件下，证明了 $O(h^2)$ 的接近性；这些条件排除了这样一个值得关注的情形：在该情形下，动量在收敛速度和稳定性方面优于普通梯度下降

注释 4.2。从 IGR 的角度来看，我们可以看到，加入动量会增强 IGR 的强度（IGR 项前的系数从 (GD) 中的 $\frac{h}{4}$ 增加到 (GD+M) 中的 $\frac{h}{4}\frac{1+\beta}{(1-\beta)^2}$。即使令两种方法的有效学习率相等（即 $\frac{h_{GD+M}}{1-\beta}=h_{GD}$），(GD+M) 中的隐式正则化仍然强 $\frac{1+\beta}{1-\beta}$ 倍。

注 4.3. (GD+M) 的隐式梯度正则化在 $\beta=0$时简化为 GD 的隐式梯度正则化。

增加一个动量参数 β后，隐式正则化项的强度会增加 $\frac{1+\beta}{1-\beta}$ 倍。因此，随着动量参数 β 增大，正则化强度也会增加，从而使其下降过程更倾向于选择较平坦的轨迹。

4.1 二维线性模型中的 IGR-M

我们首先在一个非常简单的设置中展示 IGR-M 的影响：该设置旨在最小化一个含有两个可学习参数的损失函数，即 $(\hat{w}_1,\hat{w}_2) = \arg\min_{w_1,w_2} E(w_1,w_2)$ where $E(w_1,w_2) = \frac{1}{2}(y-w_1w_2x)^2$。这里 x, y, w_1, w_2 均为标量，该模型模拟了一个简单的标量线性双层网络。对于给定的标量 (x,y)，$E(w_1,w_2)$ 的全局极小值是曲线$w_1w_2=\frac{y}{x}$上的所有点。参数的连续梯度流为 $w_1'(t) = -\frac{\partial E}{\partial w_1}$ 和 $w_2'(t) = -\frac{\partial E}{\partial w_2}$。

这个问题的隐式梯度正则化为 $\frac{h_{GD}}{4}\|\nabla E\|_2^2 = \frac{h_{GD}}{4}(w_1^2+w_2^2)E$，它会对轨迹进行正则化，使其在所有位于 $w_1w_2=\frac{y}{x}$ 上的全局极小值中，找到取值更小的极小值（$w_1^2+w_2^2$）（朝向原点）。我们特意选择了与 (Barrett & Dherin，2020) 相同的实验，以比较 (GD) 和 (GD+M) 的隐式正则化效果。

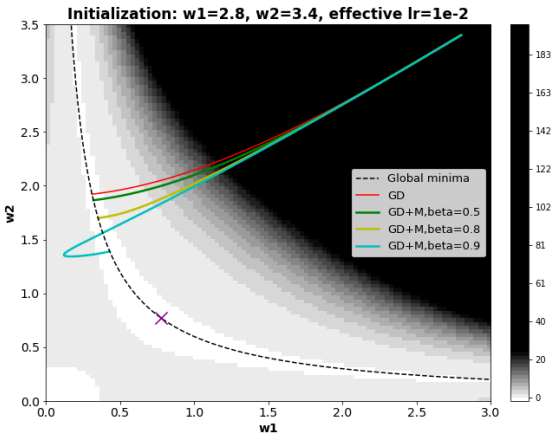


图 1：在相同的有效学习率 $\frac{h}{(1-\beta)}$ 下，(GD+M) 的隐式正则化强于 (GD)。随着 β 增大，最优解似乎会找到范数更小的解。这证实了注释 4.2：隐式正则化的强度会随 β 增大。背景颜色表示梯度范数的大小，即，$\|\nabla E\|_2^2$

为了公平比较（GD）和（GD+M），我们将有效学习率设为相同，即 $h_{GD}=\frac{h_{GD+M}}{1-\beta}$，如注 4.1 所述。采用相同的初始化 $(w_1^0,w_2^0)=(2.8,3.4)$，我们探索并追踪四条轨迹的路径，其中（h_{GD+M}, β）分别为 $(5\times10^{-3},0.5)$、$(2\times10^{-3},0.8)$、$(10^{-3},0.9)$ 和 $(10^{-2},0.0)$。在这四种情况下，有效学习率均相同，即 $\frac{h_{GD+M}}{(1-\beta)}=h_{GD}=10^{-2}$。我们有如下观察：a）对于所有四条轨迹，收敛后的权重参数 $w^*=(w_1^*,w_2^*)$ 都位于全局极小值曲线上。b）随着 β 的值增大，收敛解的 ℓ_2 范数减小（或隐式偏置的强度增大式正则化），即 $\|w^*_{(10^{-3},0.9)}\|_2<\|w^*_{(2\times10^{-3},0.8)}\|_2<\|w^*_{(5\times10^{-3},0.5)}\|_2<\|w^*_{(10^{-2},0.0)}\|_2$。这一观察支持注释 4.2：即使考虑有效学习率，隐式正则化的强度也会随着 β 增大。

5 带动量的随机梯度下降中的隐式正则化

在随机梯度下降中，全批量梯度由其抽样得到的无偏估计量取代。假设损失函数 $E(\mathbf{x})$具有以下形式

$$E(\mathbf{x})=\frac{1}{M}\sum_{j=1}^{M}E_{(j)}(\mathbf{x}) \tag{7}$$

其中 $E_{(j)}$ 是 j^{th} 小批次损失。在第 k^{th} 次迭代中，我们随机选取一个小批量，其损失记为 E_k，并据此更新参数。重球加速随机梯度下降的迭代形式为

$$\begin{cases}\mathbf{x}^{k+1}=\mathbf{x}^k-h\nabla E_k(\mathbf{x}^k)+\beta(\mathbf{x}^k-\mathbf{x}^{k-1}) \quad k=1,2,...,n\\ \mathbf{x}^1=\mathbf{x}^0-h\nabla E_0(\mathbf{x}^0)\\ \mathbf{x}^0=\mathbf{x}^{-1}=\mathbf{0}\end{cases} \tag{8}$$

对于每次迭代 k，更新由当前小批量损失函数 E_k 驱动。其连续近似为

$$\mathbf{x}'(t)=-\nabla E_k(\mathbf{x}(t))+\frac{\beta}{h}(\mathbf{x}(t_k)-\mathbf{x}(t_{k-1})) \quad \text{for } t_k<t<t_{k+1} \tag{9}$$

在时间 $[t_k,t_{k+1}]$ 内。因此，$\mathbf{x}$ 的轨迹是连续的，但分段可微；显然，在从一个批次切换到另一个批次的过渡点 t_k，左右两侧的导数并不相等：$\mathbf{x}'(t_k^+)\neq\mathbf{x}'(t_k^-)$。因此，我们预期隐式正则化项在不同区间上的导数也不连续。下面给出随机情形下 IGR-M 的数学公式。

定理 5.1. *[IGR-M 随机版本（IGRM$_s$）]设每个小批次 E_k 的损失函数均为四次可微，则重球动量更新 8 与具有修正损失的梯度流轨迹 $O(h^2)$ 接近，*

$$\begin{aligned}&\widetilde{\mathbf{x}}'(t)=-\nabla\hat{E}_k(\widetilde{\mathbf{x}}(t)) \quad \textit{for } t_k\le t<t_{k+1}\\ &\textit{where,}\quad \hat{E}_k(\widetilde{\mathbf{x}})=\underbrace{G_k(\widetilde{\mathbf{x}})}_{force}+\underbrace{\frac{h}{4}(\|\nabla G_k(\widetilde{\mathbf{x}})\|_2^2+2\sum_{r=0}^{k-1}\beta^{k-r}\|\nabla G_r(\widetilde{\mathbf{x}})\|_2^2)}_{IGRM_s}\end{aligned} \tag{10}$$

其中 $G_k(\tilde{\mathbf{x}}(t))=\sum_{r=0}^{k}\beta^{k-r}E_r(\tilde{\mathbf{x}}(t))$。具体而言，存在一个常数 C，使得

$$\|\widetilde{\mathbf{x}}(t_k)-\mathbf{x}^k\|_2\le Ch^2,\ k=1,2,...,n.$$

该定理告诉我们，动量对应的 IGR 表现为 ℓ_2 范数形式，∇G_k 它是历史梯度 ∇E_k、$k=0,...,n$ 按 β 的幂加权求和，并在当前位置 $\tilde{\mathbf{x}}(t)$ 处取值。此外，由于相邻批次之间发生突变，随机 IGR-M 在不同区间上有不同的表达式。下面列出关于 (SGD+M) 算法的一些进一步说明。

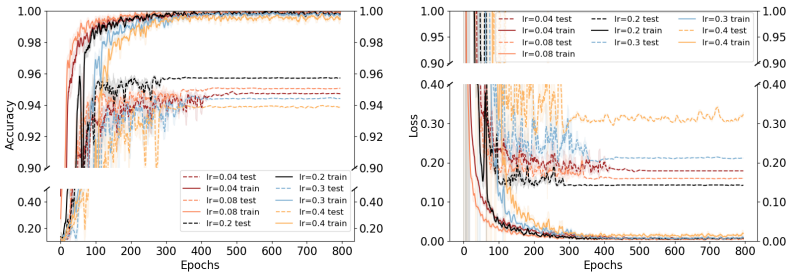
注释 5.1。连续轨迹的流由驱动力 $-\nabla G_k(\mathbf{x}(t))$ 和隐式正则化项的负梯度 $IGRM_s(\tilde{\mathbf{x}})=\frac{h}{4}(\|\nabla G_k(\tilde{\mathbf{x}})\|_2^2+2\sum_{r=0}^{k-1}\beta^{k-r}\|\nabla G_r(\tilde{\mathbf{x}})\|_2^2$ 所支配，该隐式正则化项同时取决于学习率 h 和动量 β。**注释 5.2**。当 $\beta=0$时，10 化简为

$$\widetilde{\mathbf{x}}'(t)=-\nabla\left(\underbrace{E_k(\widetilde{\mathbf{x}}(t)}_{force}+\underbrace{\frac{h}{4}\|\nabla E_k(\widetilde{\mathbf{x}}(t))\|_2^2)}_{IGR}\right),\quad t\in[t_k,t_{k+1}], \tag{11}$$

这是随机梯度下降的修正损失函数。

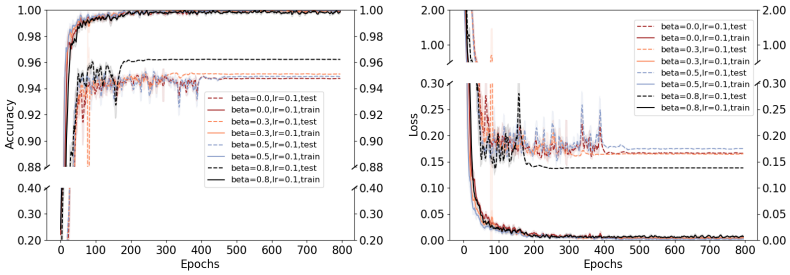
备注 5.3. 对批次的随机选择取期望，我们得到 $\mathbb{E}(\text{IGRM}_s)(\mathbf{x})=\frac{h(1+\beta)}{4(1-\beta)^3}\|\nabla E\|^2+\frac{h}{4(1-\beta)^2}F$，其中 $F:=\mathbb{E}\|E_n-E\|^2$(附录 定理 3.1)。相比之下，随机梯度下降中的 IGR 项在取期望后为 $\frac{h}{4}(\|\nabla E\|^2+F)$（Smith 等人，2021），弱得多。即使调整了学习率，(SGD+M) 中的 IGR 仍比随机梯度下降强约 $1/(1-\beta)$ 倍。

备注 5.4（方差缩减）。我们注意到，动量具有方差缩减效应。具体而言，假设采用有效学习率（注 4.1），使得 (SGD) 和 (SGD+M) 中的力项具有相同的期望，随后我们可以比较它们的方差。设 ∇E_k 在固定点 $\mathbf{x}$ 处的协方差矩阵为 $C:=\mathbb{E}(\nabla E_k(\mathbf{x})-\nabla E(\mathbf{x}))(\nabla E_k(\mathbf{x})-\nabla E(\mathbf{x}))^T$。其中，$\nabla E(\mathbf{x})$ 表示全批量梯度。那么，驱动 (SGD) 的力 $-E_k$ 的协方差矩阵恰好为 C，而驱动 (SGD+M) 的力 $-G_k$ 的协方差矩阵仅为 $\frac{(1-\beta)}{1+\beta}C$（附录定理 4.1），约小 $\frac{(1-\beta)}{(1+\beta)}$倍。



(a) 准确率 (b) 损失函数

图 2：（梯度下降）：使用全批量梯度下降对 MNIST 数据集上的 ResNet-18 进行分类的结果。图中展示了有限学习率 h 所带来的隐式正则化效果。随着学习率增大，测试准确率不断提升，直至 $h = 0.2$。当 $h = 0.2$ 时，测试准确率达到最高值 95.72%



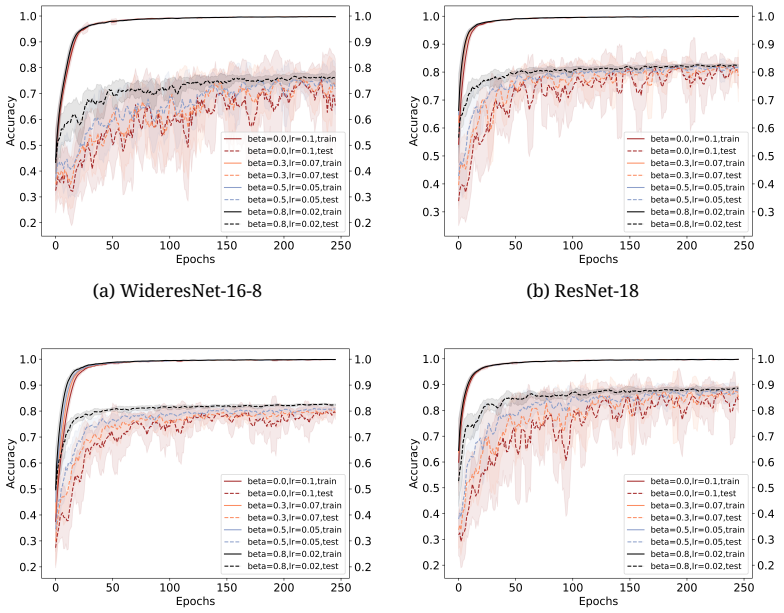
(a) 准确率 (b) 损失函数

图 3：（GD+M）：采用不同动量参数 β 进行全批量梯度下降训练的 ResNet-18 在 MNIST 数据集上的分类结果。据报告，最佳测试准确率为 96.22%

备注 5.5。许多文献都观察并证实，较大的 SGD 迭代方差（由较小的批量大小或较大的学习率导致（Smith 和 Le，2017；Li 等，2017））会增强泛化能力。小批量梯度中较大的方差会提高 SGD 逃离不良局部极小值的逃逸效率（Ibayashi 和 Imaizumi，2022）[有关详细讨论，请参见附录第 7 节]，，从而增强泛化能力。因此，(SGD+M) 的方差缩减效应会损害泛化能力。另一方面，(SGD+M) 具有更强的隐式梯度正则化（备注 5.3），并且相比 (SGD)，(SGD+M) 对使用较大的有效学习率更加稳定（例如参见（Cohen 等，2021）），这些因素往往有助于其泛化。这解释了为什么在实践中我们通常观察到，(SGD+M) 仅比 (SGD) 略好一些。

6 数值实验

我们的第一个实验是比较全批量（GD）与（GD+M）。对于海森矩阵的谱范数有界于 L 的线性最小二乘问题，众所周知（例如，Cohen 等，2021），只要满足 $h \leq \frac{2+2\beta}{L}$，（GD+M）就是稳定的；而只要满足 $h \leq \frac{2}{L}$，GD 就是稳定的。这意味着，（GD）可达到的最大有效学习率为 $\frac{2}{L}$，而（GD+M）可达到的最大有效学习率则可高达 $\frac{2(1+\beta)}{L(1-\beta)}$。由于更大的有效学习率意味着更强的隐式梯度正则化，（GD+M）显然得益于其较大的稳定区域。为验证这一点，我们使用 ResNet-18 对均匀子采样的 MNIST 数据集进行分类，该数据集包含 1000 张训练图像。除学习率衰减和批归一化外，所有外部正则化方案均已关闭。我们进行了以下实验：**1**）使用 $\beta = 0$（即 GD）进行全批量梯度下降，测试不同的学习率 h，并记录最佳测试准确率（见图 2），结果为 95.72%。**2**）使用动量进行全批量梯度下降（GD+M），测试不同的 β，同时固定步长 $h = 0.1$，并记录最佳测试准确率（见图 3），结果为 96.22%。我们的观察是，GD 的最佳性能（在所有学习率中）不如（GD



(a) WideresNet-16-8 (b) ResNet-18

(c) ResNet-50 (d) DenseNet-121

图 4：在 CIFAR-10 数据集上，采用不同网络架构及不同的（h，β）组合进行分类的结果，其中所选组合使有效学习率 $h_{(1-\beta)}$ 保持不变。在所有实验中，~~均~~关闭外部正则化方法，如权重衰减、学习率调度器、丢弃法和标签平滑（批归一化除外）。结果是对采用不同初始化的 3 个随机种子取平均所得。(SGD+M) 的测试准确率会随着 β 增大而提高，优于 (SGD)，这证实了备注 5.4；其测试准确率的方差更小，这证实了备注 5.3。

.

（在所有 β 中）的最佳性能。这一观察无法用已知的稳定性边缘理论 * 来解释，但我们的（GD+M）隐式正则化理论可以很好地解释这一现象，因为加入动量会增强隐式梯度正则化的强度。

为研究隐式正则化（SGD+M）的影响，我们在图像分类任务上进行了一系列实验。我们训练了四种知名且广受欢迎的网络架构，即 DenseNet（Iandola 等人，2014）、ResNet-18、ResNet-50（He 等人，2016）和 WideResNet（Zagoruyko 和 Komodakis，2016），以对 CIFAR-10 和 CIFAR-100 数据集中的图像进行分类。我们想了解分别使用（SGD）和（SGD+M）训练这些网络后，它们在测试数据集上的泛化能力如何。为了单独观察动量参数 β 和学习率 h 对泛化的影响，我们关闭了所有外部正则化方法，例如丢弃法、权重衰减和标签平滑。在所有实验中，我们将批量大小固定为 640。

在第一个实验中，我们展示了 (GD+M) 的稳定区域比 (GD) 更大，因此可以使用更大的有效学习率。(SGD+M) 和 (SGD) 也得出了相同的结论。然而，在这里我们想要说明，即使在两种算法都稳定的区域内，(SGD+M) 仍不只是 (SGD) 的缩放版本。为此，我们选取较小的学习率，以确保两种算法都稳定，并使 (SGD+M) 的有效学习率 $h_{(1-\beta)}$ 与 (SGD) 的学习率相同（两者均为 0.1）。从表1（另见图 4）中可以观察到，最高测试准确率几乎总是在 β 的最大取值处达到。这一观察结果

*“稳定性边缘”（EOS）是 Cohen 等人（2021）提出的一种现象：在使用全批量梯度下降进行网络训练时，尖锐度 $||\nabla^2 E||_2$ 会逐渐增大，直至达到阈值 $\frac{2}{h}$，随后在该值附近徘徊。对于 GD+M，尖锐度会在一个较大的值 $^{2}{}_{(1+\beta)\,h}$ 附近徘徊。由于尖锐度越大通常意味着泛化能力越差，因此 EOS 理论预测，加入动量会损害泛化能力。

	CIFAR-10				CIFAR-100			
β /h	DN-121	RN-18	RN-50	WRN-16-8	DN-121	RN-18	RN-50	WRN-16-8
0.0/0.10	84.0±5.0	79.7±5.6	79.3±2.5	65.3±18.1	60.4±4.9	53.1±0.6	47.4±2.1	38.6±3.8
0.3/0.07	85.1±5.4	78.7±9.3	80.0±1.5	72.5±7.4	60.0±8.6	52.7±1.0	48.9±2.2	37.0±6.1
0.5/0.05	87.6±1.2	81.5±0.9	80.7±0.7	71.8±9.8	63.2±2.4	53.3±1.1	**50.3±1.0**	39.4±3.9
0.8/0.02	**88.6±0.7**	**82.4±0.4**	**82.4±0.7**	**75.4±2.8**	**64.7±0.8**	**54.3±0.6**	49.6±0.7	**40.6±1.2**

表1：CIFAR-10 和 CIFAR-100 在不同动量 β 和学习率 h 下的测试准确率，但有效学习率 $\frac{h}{(1-\beta)} = 0.1$ 相同。不同模型的最佳表现已突出显示。测试准确率的均值和标准差是根据最后 5 个训练轮次以及 3 次随机种子初始化计算得出的。DN-121 表示 Densenet-121，RN-18 和 RN-50 分别表示 Resnet-18 和 Resnet-50，WRN-16-8 表示深度为 16、宽度因子为 8 的 WideResnet。

cons与注5.3一致，在注5.3中我们证明了 (SGD+M) 中的隐式正则化更强
于 (SGD)，即使调整学习率后也是如此。

表1中的测试准确率标准差是在最后5个训练轮次和三个随机种子初始化的情况下计算得出的。标准差越低，表示测试准确率曲线越平滑，也意味着在一个训练轮次区间内测试准确率的波动越小。我们观察到，β取值最高时，标准差最低。因此，β 越高，方差缩减效应越明显，这一观察与备注 5.4 一致。

7 隐式梯度正则化与噪声注入的综合效应

尽管 IGR 项 $\|\nabla E\|^2$ 与尖锐度密切相关（第3节），但当$\mathbf{x}^k$ 趋近局部极小值点时，它会变得非常微弱且无关紧要，因为 $\nabla E \to 0$。然而，我们发现，如果进行噪声注入，情况就会有所不同；噪声注入有助于 IGR 项即使在局部极小值附近也能保持作用。更具体地说，正如先前文献（Orvieto 等人，2022b；Camuto 等人，2020）所研究的那样，在梯度下降的每次迭代中注入噪声所得到的算法通常称为扰动梯度下降（PGD），其本质上是在最小化一个平均目标函数

$$R(\mathbf{x}) := \mathbb{E}_{\eta\sim N(0,\sigma^2\mathbf{I})}E(\mathbf{x}+\eta).$$

当 σ取较小值时，我们可以将 $R(\mathbf{x})$展开为

$$R(\mathbf{x}) = E(\mathbf{x}) + \mathbb{E}_\eta \eta^T \nabla E(\mathbf{x}) + \frac{1}{2}\mathbb{E}_\eta \eta^T \nabla^2 E(\mathbf{x})\eta + O(\sigma^3) = E(\mathbf{x}) + \frac{1}{2}\sigma^2 Tr(\nabla^2 E(\mathbf{x})) + O(\sigma^3),$$

其中，Tr 表示迹算子。因此，最小化 $R(\mathbf{x})$会对 E的海森矩阵的迹进行正则化。使用随机梯度下降类型的更新来最小化 $R(\mathbf{x})$时，迭代过程为

$$\mathbf{x}^{k+1} = \mathbf{x}^k - h\nabla E(\mathbf{x}^k + \eta_k), \text{ where } \eta_k \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$

这被称为一种扰动梯度下降（PGD）形式。由于学习率有限，更新将遵循带有 IGR 项的修正流，在这种情况下，该项为

$$E(\mathbf{x}+\eta_k) + \frac{h}{4}\|\nabla E(\mathbf{x}+\eta_k)\|^2.$$

取期望后，修正损失为

$$\begin{aligned}
&\mathbb{E}_{\eta_k}\left[E(\mathbf{x}+\eta_k) + \frac{h}{4}\|\nabla E(\mathbf{x}+\eta_k)\|^2)\right] \\
=&\mathbb{E}_{\eta_k}\left[E(\mathbf{x}) + \eta_k^T\nabla E(\mathbf{x}) + \frac{1}{2}(\eta_k)^T\nabla^2 E(\mathbf{x})\eta_k\right] \\
&+ \frac{h}{4}\mathbb{E}_{\eta_k}\|\nabla E(\mathbf{x}) + \nabla^2 E(\mathbf{x})\eta_k + \nabla^3 E(\mathbf{x})[\eta_k,\eta_k]\|^2) + O(\sigma^3) \\
=&E(\mathbf{x}) + \sigma^2 Tr(\nabla^2 E(\mathbf{x})) + \frac{h}{4}\left[\sigma^2\|\nabla^2 E(\mathbf{x})\|_F^2 + \mathbb{E}_{\eta_k}\|\nabla E(\mathbf{x}) + \nabla^3 E(\mathbf{x})[\eta_k,\eta_k]\|^2\right] + O(\sigma^3).
\end{aligned}$$

我们可以看到，现在出现了一个来自隐式梯度正则化的海森正则化项 $\frac{h}{4}\sigma^2\|\nabla^2 E(\mathbf{x})\|_F^2$，即使在局部极小值点附近也不会消失，其强度与学习率成正比。我们预计，加入动量后，这个新的正则化项会变得更强，因为动量会增强隐式梯度正则化的作用（注释 4.2、5.3）。这一观察表明，隐式梯度正则化和噪声注入这两种不同类型的隐式正则化，在协同使用时或许能够相互增强。

8 结论

本研究从隐式正则化（IR）的视角探讨了动量驱动的梯度下降方法的泛化，并提供了理论分析和实验验证。我们考察了（SGD）与（SGD+M）之间的异同，发现采用合适参数的（SGD+M）在几乎所有设置下都优于（SGD）。此外，我们还发现，除了动量之外，噪声注入也可能增强隐式梯度正则化（IGR），这是我们未来希望进一步探索的课题。

参考文献

Sanjeev Arora、Nadav Cohen、Wei Hu 和 Yuping Luo。《深度矩阵分解中的隐式正则化》。神经信息处理系统进展，32，2019。

David GT Barrett 和 Benoit Dherin。隐式梯度正则化。arXiv 预印本 arXiv:2009.11162，2020。Yoshua Bengio 和 Yann LeCun。面向人工智能扩展学习算法。载于《大规模核机器》。麻省理工学院出版社，2007。Alexander Camuto、Matthew Willetts、Umut Simsekli、Stephen J Roberts 和 Chris C Holmes。高斯噪声注入中的显式正则化。载于《神经信息处理系统进展》，33：16603–16614，2020。Jeremy M Cohen、Simran Kaur、Yuanzhi Li、J Zico Kolter 和 Ameet Talwalkar。神经网络上的梯度下降通常发生在稳定性边缘。arXiv 预印本 arXiv:2103.00065，2021。Pierre Foret、Ariel Kleiner、Hossein Mobahi 和 Behnam Neyshabur。用于高效提升泛化能力的尖锐度感知最小化。arXiv 预印本 arXiv:2010.01412，2020。Ian Goodfellow、Yoshua Bengio、Aaron Courville 和 Yoshua Bengio。《深度学习》，第1卷。麻省理工学院出版社，2016。Suriya Gunasekar、Blake E Woodworth、Srinadh Bhojanapalli、Behnam Neyshabur 和 Nati Srebro。矩阵分解中的隐式正则化。载于《神经信息处理系统进展》，30，2017。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于《电气电子工程师学会计算机视觉与模式识别会议论文集》，第770–778页，2016。Geoffrey E. Hinton、Simon Osindero 和 Yee Whye Teh。一种快速学习深度信念网络的算法。《神经计算》，18：1527–1554，2006。Forrest Iandola、Matt Moskewicz、Sergey Karayev、Ross Girshick、Trevor Darrell 和 Kurt Keutzer。DenseNet：实现高效的卷积网络描述符金字塔。arXiv 预印本 arXiv:1404.1869，2014。Hikaru Ibayashi 和 Masaaki I maizumi。逃逸问题的拟势理论：定量尖锐度对随机梯度下降逃离局部极小值的影响，2022。网址 https://openreview.net/forum?id=vLz0e9S-iF3。

Samy Jelassi 和 Yuanzhi Li。探究动量如何提升深度学习的泛化能力。载于《国际机器学习会议》，第9965–10040页。PMLR，2022。Ziwei Ji 和 Matus Telgarsky。不可分数据上梯度下降的隐式偏差。载于《学习理论会议》，第1772–1798页。PMLR，2019。Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980，2014。Nikola B Kovachki 和 Andrew M Stuart。动量方法的连续时间分析。《机器学习研究期刊》，22(17)：1–40，2021。Qianxiao Li、Cheng Tai 和 E Weinan。随机修正方程与自适应随机梯度算法。载于《国际机器学习会议》，第2101–2110页。PMLR，2017。Qianxiao Li、Cheng Tai 和 Weinan E。随机修正方程与随机梯度算法的动力学 I：数学基础。《机器学习研究期刊》，20(40)：1–47，2019。网址 http://jmlr.org/papers/v20/17-526.html。Behnam Neyshabur。深度学习中的隐式正则化。arXiv 预印本 arXiv:1709.01953，2017。Behnam Neyshabur、Ryota Tomioka 和 Nathan Srebro。探寻真正的归纳偏置：论隐式正则化在深度学习中的作用。arXiv 预印本 arXiv:1412.6614，2014。Antonio Orvieto、Hans Kersting、Frank Proske、Francis Bach 和 Aurelien Lucchi。通过反相关噪声注入提升泛化能力。arXiv 预印本 arXiv:2202.02831，2022a。Antonio Orvieto、Anant Raj、Hans Kersting 和 Francis Bach。通过噪声注入对过参数化模型进行显式正则化。arXiv 预印本 arXiv:2206.04613，2022b。Tomaso Poggio、Andrzej Banburski 和 Qianli Liao。深度网络中的理论问题。《美国国家科学院院刊》，117(48)：30039–30045，2020。Boris T Polyak。加速迭代方法收敛的一些方法。《苏联计算数学与数学物理》，4(5)：1–17，1964。Noam Razin 和 Nadav Cohen。深度学习中的隐式正则化或许无法用范数解释。《神经信息处理系统进展》，33：21174–21187，2020。Samuel L Smith 和 Quoc V Le。关于泛化与随机梯度下降的贝叶斯视角。arXiv 预印本 arXiv:1710.06451，2017。Samuel L Smith、Benoit Dherin、David GT Barrett 和 Soham De。论随机梯度下降中隐式正则化的起源。arXiv 预印本 arXiv:2101.12176，2021。Daniel Soudry、Elad Hoffer、Mor Shpigel Nacson、Suriya Gunasekar 和 Nathan Srebro。可分数据上梯度下降的隐式偏差。《机器学习研究期刊》，19(1)：2822–2878，2018。Ilya Sutskever、James Martens、George Dahl 和 Geoffrey Hinton。初始化与动量在深度学习中的重要性。载于《国际机器学习会议》，第1139–1147页。PMLR，2013。Tijmen Tieleman、Geoffrey Hinton 等。第6.5讲——RMSprop：用近期梯度幅值的运行平均值对梯度进行缩放。COURSERA：《机器学习神经网络》，4(2)：26–31，2012。Gal Vardi 和 Ohad Shamir。平方损失下 ReLU 网络中的隐式正则化。载于《学习理论会议》，第4224–4258页。PMLR，2021。Bohan Wang、Qi Meng、Huishuai Zhang、Ruoyu Sun、Wei Chen、Zhi-Ming Ma 和 Tie-Yan Liu。动量会改变可分数据上的隐式正则化吗？，2021。网址 https://arxiv.org/abs/2110.03891。

Jingfeng Wu、Wenqing Hu、Haoyi Xiong、JunHuan、Vladimir Braverman 和 Zhanxing Zhu。关于能够像随机梯度下降一样实现泛化的噪声梯度下降。载于《国际机器学习会议》，第 10367–10376 页。PMLR，2020。

Sergey Zagoruyko 和 Nikos Komodakis。《更宽的残差网络》。arXiv 预印本 arXiv:1605.07146，2016。

arXiv:2302.00849v1 [cs.LG] 2 Feb 2023

论文《动量随机梯度下降中的隐式正则化》的补充材料

本附录按以下方式组织。在第1节中，我们通过证明以下两部分来证明主论文中关于 (SGD+M) 的定理 5.1：

- 在定理 1.1 中，我们证明了连续分段可微轨迹与重球动量随机梯度下降在每次迭代中的局部误差为 $O(h^3)$。该定理依赖于引理 1.3，后者表明高阶导数有界。

- 随后，在定理 1.4 中，我们证明了连续轨迹与重球动量更新之间的全局误差为 $O(h^2)$。

在第2节中，我们通过证明当小批次损失 E_k 是全批量损失 E 时，推论 2.1（附录）中的每条分段可微轨迹都属于定理 5.1（正文）的特例，从而证明正文中关于（梯度下降+M）的定理 4.1。最后，我们在第4节和第3节中分别给出了注5.2和注5.3的证明。

1 定理5.1的证明

在整个附录中，$\|\cdot\|$ 表示向量和矩阵的 ℓ_2 范数，而对于三维或四维张量，则表示弗罗贝尼乌斯范数。

定理 1.1. */局部误差界/* 设每个小批量 E_n 的损失函数光滑且具有足够的（*4* 次）可微性，并且其零阶至四阶导数有界，则在每次迭代中，重球法动量随机梯度下降更新

$$\begin{cases} \mathbf{x}^{n+1} = \mathbf{x}^n - h\nabla E_n(\mathbf{x}^n) + \beta(\mathbf{x}^n - \mathbf{x}^{n-1}), \quad n = 1, 2, ..., N \\ \mathbf{x}^1 = \mathbf{x}^0 - h\nabla E_0(\mathbf{x}^0) \\ \mathbf{x}^0 = \mathbf{x}^{-1} = \mathbf{0} \end{cases} \tag{1.1}$$

在更新 n^{th} 小批量时，与以下修正常微分方程的流局部 $O(h^3)$接近

$$\widetilde{\mathbf{x}}'(t) = -\nabla G_n(\widetilde{\mathbf{x}}(t)) - A_n(\widetilde{\mathbf{x}}(t)), \qquad for\ t_n \le t < t_{n+1} \tag{1.2}$$

其中 $t_n = nh$。具体而言，每次迭代都满足以下等式

$$\widetilde{\mathbf{x}}(t_{n+1}) = \widetilde{\mathbf{x}}(t_n) - h\nabla E_n(\widetilde{\mathbf{x}}(t_n)) + \beta(\widetilde{\mathbf{x}}(t_n) - \widetilde{\mathbf{x}}(t_{n-1})) + O(h^3). \tag{1.3}$$

Here

$$G_n(\widetilde{\mathbf{x}}) = \sum_{k=0}^{n} \beta^{n-k} E_k(\widetilde{\mathbf{x}}), A_n(\widetilde{\mathbf{x}}) = \frac{h}{2}\sum_{k=0}^{n} \beta^{n-k} C_k(\widetilde{\mathbf{x}}),$$

并且 $C_k(\tilde{\mathbf{x}}) = \nabla^2 G_k(\tilde{\mathbf{x}})\nabla G_k(\tilde{\mathbf{x}}) + \beta\nabla^2 G_{k-1}(\tilde{\mathbf{x}})\nabla G_{k-1}(\tilde{\mathbf{x}}))$ 初始条件为 $C_0(\mathbf{x}) = \nabla^2 G_0(\tilde{\mathbf{x}})\nabla G_0(\tilde{\mathbf{x}})$。

证明。 在继续证明之前，我们指出，连续轨迹 $\tilde{\mathbf{x}}(t)$ 在整个定义域 $[0, T]$ 上可微，但在网格点 $t_n, n = 0, 1, ..., N$, $N = \lfloor \frac{T}{h} \rfloor$ 处除外。这是因为在区间$t_n \le t < t_{n+1}$内，连续轨迹表示为 $\tilde{\mathbf{x}}'(t) = -\nabla G_n(\tilde{\mathbf{x}}(t)) - A_n(\tilde{\mathbf{x}}(t))$，而在区间 $t_{n-1} \le t < t_n$内，轨迹定义为 $\tilde{\mathbf{x}}'(t) = -\nabla G_{n-1}(\tilde{\mathbf{x}}(t)) - A_{n-1}(\tilde{\mathbf{x}}(t))$。我们注意到以下几点：

- 在任意边界点 t_n 处，轨迹的右侧导数与左侧导数不相等，即 $\tilde{\mathbf{x}}'(t_n^+) \neq \tilde{\mathbf{x}}'(t_n^-)$ $n = 0, 1, ..., N$。这里定义 $\tilde{\mathbf{x}}'(t_0^-) = \mathbf{0}$。
- 根据引理 1.3，轨迹 $\tilde{\mathbf{x}}(t)$ 的一阶至三阶导数的范数 $t \in [0, T]$可以由与 h 无关的常数进行界定。因此，根据 1.2，我们可以计算 t_k（任意边界点 $\forall k = 1$，2，.....，n）处的左右导数，如下所示：

1. $\tilde{\mathbf{x}}'(t_n^+) = -\nabla G_n(\tilde{\mathbf{x}}(t_n)) - A_n(\tilde{\mathbf{x}}(t_n))$，因为它属于轨迹 $t_n \leq t < t_{n+1}$2. $\tilde{\mathbf{x}}''(t_n^+) = \nabla^2 G_n(\tilde{\mathbf{x}}(t_n))\nabla G_n(\tilde{\mathbf{x}}(t_n)) + O(h)$ 3. $\tilde{\mathbf{x}}'(t_n^-) = -\nabla G_{n-1}(\tilde{\mathbf{x}}(t_n)) - A_{n-1}(\tilde{\mathbf{x}}(t_n))$，因为它属于轨迹 $t_{n-1} \leq t < t_n$4. $\tilde{\mathbf{x}}''(t_n^-) = \nabla^2 G_{n-1}(\tilde{\mathbf{x}}(t_n))\nabla G_{n-1}(\tilde{\mathbf{x}}(t_n)) + O(h)$

由于 $\|\tilde{\mathbf{x}}'''(t)\|$ 由一个常数（引理 1.3）界定，因此在点 t_n 处对轨迹 $\tilde{\mathbf{x}}(t)$ 两侧进行泰勒展开：

$$\widetilde{\mathbf{x}}(t_{n+1}) = \widetilde{\mathbf{x}}(t_n) + \widetilde{\mathbf{x}}'(t_n^+)h + \frac{h^2}{2}\widetilde{\mathbf{x}}''(t_n^+) + O(h^3), \tag{1.4}$$

$$\widetilde{\mathbf{x}}(t_n) = \widetilde{\mathbf{x}}(t_{n-1}) + h\widetilde{\mathbf{x}}'(t_n^-) - \frac{h^2}{2}\widetilde{\mathbf{x}}''(t_n^-) + O(h^3). \tag{1.5}$$

回想一下，该定理的主要目标是确定连续轨迹 $\tilde{\mathbf{x}}(t)$ 对 H.B 动量更新方程 1.1 的满足程度。因此，我们将 $\tilde{\mathbf{x}}(t_n)$ 代入 1.1 并考察由此产生的误差，这种误差在数值常微分方程文献中也称为局部截断误差（LTE）。计算如下

$$\begin{aligned}
& \widetilde{\mathbf{x}}(t_{n+1}) - \widetilde{\mathbf{x}}(t_n) - \beta(\widetilde{\mathbf{x}}(t_n) - \widetilde{\mathbf{x}}(t_{n-1})) \\
= \; & h\widetilde{\mathbf{x}}'(t_n^+) + \frac{h^2}{2}\widetilde{\mathbf{x}}''(t_n^+) - \beta(h\widetilde{\mathbf{x}}'(t_n^-) - \frac{h^2}{2}\widetilde{\mathbf{x}}''(t_n^-)) + O(h^3) \\
= \; & -h\nabla G_n(\widetilde{\mathbf{x}}(t_n)) - hA_n(\widetilde{\mathbf{x}}(t_n)) + \frac{h^2}{2}\nabla^2 G_n(\widetilde{\mathbf{x}}(t_n))\nabla G_n(\widetilde{\mathbf{x}}(t_n)) + \beta h\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)) + \beta h A_{n-1}(\widetilde{\mathbf{x}}(t_n)) \\
& + \beta\frac{h^2}{2}\nabla^2 G_{n-1}(\widetilde{\mathbf{x}}(t_n))\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)) + O(h^3) \\
= \; & -h\underbrace{(\nabla G_n(\widetilde{\mathbf{x}}(t_n)) - \beta\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)))}_{\nabla E_n(\widetilde{\mathbf{x}}(t_n))} - h\underbrace{(A_n(\widetilde{\mathbf{x}}(t_n)) - \beta A_{n-1}(\widetilde{\mathbf{x}}(t_n)))}_{\frac{h}{2}C_n} \\
& + \frac{h^2}{2}\underbrace{(\nabla^2 G_n(\widetilde{\mathbf{x}}(t_n))\nabla G_n(\widetilde{\mathbf{x}}(t_n)) + \beta\nabla^2 G_{n-1}(\widetilde{\mathbf{x}}(t_n))\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n)))}_{C_n(\text{by definition})} + O(h^3) \\
= \; & -h\nabla E_n(\widetilde{\mathbf{x}}(t_n)) - h\frac{h}{2}C_n(\widetilde{\mathbf{x}}(t_n)) + \frac{h^2}{2}C_n(\widetilde{\mathbf{x}}(t_n)) + O(h^3) \\
= \; & -h\nabla E_n(\widetilde{\mathbf{x}}(t_n)) + O(h^3).
\end{aligned}$$

该方程利用以下两个易于验证的恒等式进行了简化：

$$\nabla G_n(\widetilde{\mathbf{x}}(t_n)) - \beta\nabla G_{n-1}(\widetilde{\mathbf{x}}(t_n))) = E_n(\widetilde{\mathbf{x}}(t_n))$$

and

$$A_n(\widetilde{\mathbf{x}}(t_n)) - \beta A_{n-1}(\widetilde{\mathbf{x}}(t_n)) = \frac{h}{2}C_n.$$

因此，我们证明了连续轨迹满足离散 H.B. 动量更新，误差阶为 $O(h^3)$。

□

Co推论 1.2。 定理 *1.1* 中的分段常微分方程 (1.1) 也可以等价地写为 :

$$\begin{aligned} &\widetilde{\mathbf{x}}'(t) = -\nabla \hat{E}_n(\widetilde{\mathbf{x}}(t)) \quad \textit{for } t_n \le t < t_{n+1}, \\ &\textit{where,} \quad \hat{E}_n(\widetilde{\mathbf{x}}) = G_n(\widetilde{\mathbf{x}}) + \frac{h}{4}(\|\nabla G_n(\widetilde{\mathbf{x}})\|^2 + 2\sum_{r=0}^{n-1}\beta^{n-r}\|\nabla G_r(\widetilde{\mathbf{x}})\|^2), \end{aligned} \tag{1.6}$$

这正是我们在正文中陈述定理 *5.1* 时所用的内容。

证明。 根据定理 1.1 中常微分方程的定义，我们可以看出

$$\begin{aligned} A_n(\widetilde{\mathbf{x}}(t)) &= \frac{h}{2}\sum_{k=1}^{n}\beta^{n-k}\left(\nabla^2 G_k(\widetilde{\mathbf{x}})\nabla G_k(\widetilde{\mathbf{x}}) + \beta\nabla^2 G_{k-1}(\widetilde{\mathbf{x}})\nabla G_{k-1}(\widetilde{\mathbf{x}})\right) + \frac{h}{2}\beta^n\nabla^2 G_0(\widetilde{\mathbf{x}})\nabla G_k(\widetilde{\mathbf{x}}) \\ &= \frac{h}{2}\sum_{k=0}^{n}\beta^{n-k}\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\nabla G_k(\widetilde{\mathbf{x}}(t)) + \frac{h}{2}\sum_{k=1}^{n}\beta^{n-k+1}\nabla^2 G_{k-1}(\widetilde{\mathbf{x}}(t))\nabla G_{k-1}(\widetilde{\mathbf{x}}(t)) \\ &= \frac{h}{2}\nabla^2 G_n(\widetilde{\mathbf{x}}(t))\nabla G_n(\widetilde{\mathbf{x}}(t)) + h\sum_{k=0}^{n-1}\beta^{n-k}\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\nabla G_k(\widetilde{\mathbf{x}}(t)). \end{aligned}$$

因此，我们可以将 (1.2) 改写为

$$\widetilde{\mathbf{x}}'(t) = -\nabla G_n(\widetilde{\mathbf{x}}(t)) - \frac{h}{4}\nabla\left(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2 + 2\sum_{k=0}^{n-1}\beta^{n-k}\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2\right) := \nabla\hat{E}_n(\widetilde{\mathbf{x}}(t)), \tag{1.7}$$

$\hat{E}_n(\widetilde{\mathbf{x}}) = G_n(\widetilde{\mathbf{x}}) + \frac{h}{4}\left(\|\nabla G_n(\widetilde{\mathbf{x}})\|^2 + 2\sum_{k=0}^{n-1}\beta^{n-k}\|\nabla G_k(\widetilde{\mathbf{x}})\|^2\right)$。

□

引理 1.3。 在定理 *1.1* 的假设下，设 $\widetilde{\mathbf{x}}(t)$ 如(1.2)中所定义，则 $\widetilde{\mathbf{x}}$ 关于时间的一阶至三阶导数有界，即存在常数 c_1，c_2，c_3 ，使得$\|\widetilde{\mathbf{x}}'(t)\| \le c_1$, $\|\widetilde{\mathbf{x}}''(t)\| \le c_2$, $\|\widetilde{\mathbf{x}}'''(t)\| \le c_3$，对所有 $t \in [0,T]$ 均成立。

证明。 尽管连续轨迹 $\widetilde{\mathbf{x}}(t)$ 是分段定义的，且步长为 h，我们希望得到其导数的一个与步长无关的常数上界，从而能够忠实地截断泰勒展开 (1.4)，使误差为 $O(h^3)$。根据定理 1.1 的假设，对于 $0 \le \alpha \le 4$, $\|\nabla^{(\alpha)}E_k(\mathbf{x})\|$ 有界；也就是说，存在某个常数 c_0，使得对所有 $1 \le n \le N$,

$$\sup\|\nabla^{(\alpha)}E_n(\mathbf{x})\| \le c_0, \quad 0 \le \alpha \le 4,$$

其中 $\|\cdot\|$ 表示张量的弗罗贝尼乌斯范数。

那么，对于任意 k、$\mathbf{x}$ 和 $0 \le \alpha \le 4$，我们可以立即给出 G_k 和 C_k 的导数上界，

$$\|\nabla^{(\alpha)}G_k(\widetilde{\mathbf{x}})\| = \left\|\sum_{k=0}^{n}\beta^{n-k}\nabla^{(\alpha)}E_k(\widetilde{\mathbf{x}})\right\| \le \sum_{k=0}^{n}\beta^{n-k}\|\nabla^{(\alpha)}E_k(\widetilde{\mathbf{x}})\| \le \frac{c_0}{1-\beta}, \tag{1.8}$$

$$\|C_k(\widetilde{\mathbf{x}})\| \le \|\nabla^2 G_k(\widetilde{\mathbf{x}})\|\|\nabla G_k(\widetilde{\mathbf{x}})\| + \beta\|\nabla^2 G_{k-1}(\widetilde{\mathbf{x}})\|\|\nabla G_{k-1}(\widetilde{\mathbf{x}})\| \le \frac{1+\beta}{(1-\beta)^2}c_0^2. \tag{1.9}$$

现在，利用 (1.8) 并 由 (1.9) 以及 h （即 $h \le T$）的有界性，我们可以证明存在某个常数 c_1,

$$\begin{aligned} \|\widetilde{\mathbf{x}}'(t)\| &\le \|G_n(\widetilde{\mathbf{x}}(t))\| + \frac{h}{2}\left\|\sum_{k=0}^{n}\beta^{n-k}C_k(\widetilde{\mathbf{x}})\right\| \\ &\le \sum_{k=0}^{n}\beta^{n-k}\max_{\widetilde{\mathbf{x}},k}\|E_k(\widetilde{\mathbf{x}})\| + \frac{h}{2}\sum_{k=0}^{n}\beta^{n-k}\max_{\widetilde{\mathbf{x}},k}\|C_k(\widetilde{\mathbf{x}})\| \\ &\le \frac{c_0}{(1-\beta)^2} + \frac{c_0^2 h(1+\beta)}{2(1-\beta)^3} := c_1. \end{aligned} \tag{1.10}$$

接下来我们证明 $\|\tilde{\mathbf{x}}''(t)\|$ 一致有界，

$$\begin{aligned}\widetilde{\mathbf{x}}''(t) = &-\sum_{k=0}^{n}\beta^{n-k}\nabla^2 E_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\\ &-\frac{h}{2}\sum_{k=0}^{n}\beta^{n-k}\underbrace{\left(\nabla^3 G_k(\widetilde{\mathbf{x}}(t))[\widetilde{\mathbf{x}}'(t)]\nabla G_k(\widetilde{x}(t)) + \nabla^2 G_k(\widetilde{\mathbf{x}}(t))\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\right)}_{(I)}\\ &-\frac{h}{2}\beta\sum_{k=0}^{n}\beta^{n-k}\underbrace{\left(\nabla^3 G_{k-1}(\widetilde{\mathbf{x}}(t))[\widetilde{\mathbf{x}}'(t)]\nabla G_{k-1}(\widetilde{\mathbf{x}}(t)) + \nabla^2 G_{k-1}(\widetilde{\mathbf{x}}(t))\nabla^2 G_{k-1}(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\right)}_{(II)}.\end{aligned}$$

因此，我们有：

$$\|\widetilde{\mathbf{x}}''(t)\| \le \|\sum_{k=0}^{n}\beta^{n-k}\nabla^2 E_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\| + \frac{h}{2(1-\beta)}\|(I)\| + \frac{h\beta}{2(1-\beta)}\|(II)\|. \tag{1.11}$$

分别考察 $\|(I)\|$ 和 $\|(II)\|$，可得

$$\begin{aligned}\|(I)\| &\le \|\nabla^3 G_k(\widetilde{\mathbf{x}}(t))[\widetilde{\mathbf{x}}'(t)]\nabla G_k(\widetilde{\mathbf{x}}(t))\| + \|\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\|\\ &\le \|\nabla^3 G_k(\widetilde{\mathbf{x}}(t))\|\|\widetilde{\mathbf{x}}'(t)\|\|\nabla G_k(\widetilde{\mathbf{x}}(t))\| + \|\nabla^2 G_k(\widetilde{\mathbf{x}}(t))\|^2\|\widetilde{\mathbf{x}}'(t)\| \le 2c_0^2 c_1.\end{aligned}$$

这里在最后一个不等式中，我们用到了 $\|\tilde{\mathbf{x}}'\| \le c_1$，如式 (1.10) 所示。同理，$\|(II)\| \le 2c_0^2 c_1$。将这些不等式代入式 1.11，对于某个常数 c_2，我们有：

$$\|\widetilde{\mathbf{x}}''(t)\| \le \|\sum_{k=0}^{n}\beta^{n-k}\nabla^2 E_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}'(t)\| + \frac{h}{2(1-\beta)}\|(I)\| + \frac{h\beta}{2(1-\beta)}\|(II)\| \le c_2.$$

最后，我们对三阶导数进行估计

$$\begin{aligned}\widetilde{\mathbf{x}}'''(t) = &-\sum_{k=0}^{n}\beta^{n-k}\nabla^3 E_k(\widetilde{\mathbf{x}}(t))[\widetilde{\mathbf{x}}'(t)]\widetilde{\mathbf{x}}'(t) - \sum_{k=0}^{n}\beta^{n-k}\nabla^2 E_k(\widetilde{\mathbf{x}}(t))\widetilde{\mathbf{x}}''(t)\\ &-\frac{h}{2}\sum_{k=0}^{n}\beta^{n-k}\frac{d(I)}{dt} - \frac{h}{2}\beta\sum_{k=0}^{n}\beta^{n-k}\frac{d(II)}{dt}.\end{aligned}$$

根据这个表达式估计 $\tilde{\mathbf{x}}'''(t)$上的范数是显而易见的。

□

定理 1.4 (全局误差界)。 设 $\tilde{\mathbf{x}}(t)$ 为 (1.2) 的解，并假设定理 *1.1* 的条件成立。则全局误差 $\|\mathbf{e}^n\| = \|\tilde{\mathbf{x}}(t_n) - \mathbf{x}^n\|$ 的阶为 $O(h^2)$，其中 $\tilde{\mathbf{x}}(t_n)$ 是边界点处式 (1.2) 所示常微分方程的解，而$\mathbf{x}^n$ 是离散的 *Heavy Ball* 动量更新。

证明。 在定理 1.1 中，我们已经证明，分段常微分方程 $\tilde{\mathbf{x}}(t_n)$ 的解满足

$$\widetilde{\mathbf{x}}(t_{n+1}) = \widetilde{\mathbf{x}}(t_n) - h\nabla E_n(\widetilde{\mathbf{x}}(t_n)) + \beta(\widetilde{\mathbf{x}}(t_n) - \widetilde{\mathbf{x}}(t_{n-1})) + O(h^3), \tag{1.12}$$

根据定义，离散的 Heavy Ball 动量更新满足

$$\mathbf{x}^{n+1} = \mathbf{x}^n - h\nabla E_k(\mathbf{x}^n) + \beta(\mathbf{x}^n - \mathbf{x}^{n-1}). \tag{1.13}$$

令误差为 n^{th} 更新记作 $\mathbf{e}_n = \tilde{\mathbf{x}}(t_n) - \mathbf{x}_n$，那么将这两个更新相减，得到 :

$$\mathbf{e}_{n+1} = \mathbf{e}_n + \beta(\mathbf{e}_n - \mathbf{e}_{n-1}) - h\underbrace{(\nabla E_n(\mathbf{x}^n) - \nabla E_n(\widetilde{\mathbf{x}}(t_n)))}_{M} + O(h^3). \tag{1.14}$$

根据定理 1.1 中的假设，存在某个常数 c_1 使得 $\|\nabla^2 E_n\|_\infty := \max_{\mathbf{x}} \|\nabla^2 E_n(\mathbf{x})\| \leq c_1$。那么 M 可以界定为

$$\|M\| = \|\nabla E_n(\mathbf{x}^n) - \nabla E_n(\widetilde{\mathbf{x}}(t_n))\| \leq \|\nabla^2 E_n\|_\infty \|\mathbf{e}_n\| \leq c_1 \|\mathbf{e}_n\|. \tag{1.15}$$

现在对 $\mathbf{e}_{n+1} - \mathbf{e}_n$ 取范数，并对 (1.14) 右侧应用三角不等式，可得存在某个常数 c:

$$\|\mathbf{e}_{n+1} - \mathbf{e}_n\| \leq \beta \|\mathbf{e}_n - \mathbf{e}_{n-1}\| + h\|M\|_2 + ch^3 \leq \beta \|\mathbf{e}_n - \mathbf{e}_{n-1}\| + hc_1 \|\mathbf{e}_n\| + ch^3. \tag{1.16}$$

定义三个量 $d_1 = \frac{c}{c_1}$、$d_2 = \frac{2c_1}{1-\beta}$ 和 $d_3 = \frac{2c}{1-\beta}$ 后，我们用数学归纳法证明以下命题

$$\|\mathbf{e}_n\| \leq d_1 \mathbf{e}^{d_2 hn} h^2, \quad \|\mathbf{e}_{n+1} - \mathbf{e}_n\| \leq d_3 \mathbf{e}^{d_2 hn} h^3, \quad n \geq 0. \tag{1.17}$$

我们首先证明基础情形。当 $n = 0$ 时，根据定义有 $\|\mathbf{e}_0\| = \|\widetilde{\mathbf{x}}(0) - \mathbf{x}^0\| = 0$。由 (1.14) 可得，$\|\mathbf{e}_1 - \mathbf{e}_0\| \leq ch^3 < d_3 h^3$, 因此归纳基础成立。

假设命题 (1.17) 对 $(n-1)$ 成立，即 $\|\mathbf{e}_{n-1}\| \leq d_1 \mathbf{e}^{d_2 h_{(n-1)}} h^2$, 且 $\|\mathbf{e}_n - \mathbf{e}_{n-1}\| \leq d_3 \mathbf{e}^{d_2 h_{(n-1)}} h^3$，那么在第 n 种情形下,

$$\begin{aligned}
\|\mathbf{e}_n\| &\leq \|\mathbf{e}_{n-1}\| + \|\mathbf{e}_n - \mathbf{e}_{n-1}\| \\
&\leq d_1 \mathbf{e}^{d_2 h(n-1)} h^2 + d_3 \mathbf{e}^{d_2 h(n-1)} h^3 \\
&= d_1 (1 + \frac{d_3 h}{d_1}) \mathbf{e}^{d_2 h(n-1)} h^2 = d_1 (1 + d_2 h) \mathbf{e}^{d_2 h(n-1)} h^2 \\
&\leq d_1 \mathbf{e}^{d_2 hn} h^2
\end{aligned}$$

并由 (1.16) 可得,

$$\begin{aligned}
\|\mathbf{e}_{n+1} - \mathbf{e}_n\| &\leq \beta d_3 \mathbf{e}^{d_2 h(n-1)} h^3 + hc_1 d_1 \mathbf{e}^{d_2 hn} h^2 + ch^3 \\
&\leq d_3 \underbrace{(\frac{d_1 c_1}{d_3} + \beta + \frac{c}{d_3})}_{=1} \mathbf{e}^{d_2 hn} h^3 = d_3 \mathbf{e}^{d_2 hn} h^3.
\end{aligned}$$

因此，我们已经证明，命题 (1.17) 对 n^{th} 情形也成立。

□

2 定理 4.1（IGR-M）的证明

推论 2.1。设全批量梯度下降的损失函数 E 光滑且四次可微，则梯度下降动量更新

$$\mathbf{x}^{n+1} = \mathbf{x}^n - h\nabla E(\mathbf{x}^n) + \beta(\mathbf{x}^n - \mathbf{x}^{n-1}), \qquad n = 1, 2, ..., N$$

与分段常微分方程的流 $O(h^2)$ 相近

$$\widetilde{\mathbf{x}}'(t) = -\frac{1}{1-\beta} \nabla \hat{E}_n(\widetilde{\mathbf{x}}(t)), \quad t \in [t_n, t_{n+1}], \tag{2.1}$$

其中，修正损失函数定义为

$$\hat{E}_n(\widetilde{\mathbf{x}}(t)) = (1 - \beta^{n+1}) \nabla E(\mathbf{x}(t)) + \frac{h}{2(1-\beta)} \left[\frac{(1+\beta)}{(1-\beta)} (1 - \beta^{2n+2}) - 4(n+1)\beta^{n+1} \right] \nabla^2 E(\mathbf{x}(t)) \nabla E(\mathbf{x}(t)).$$

对于后续区间，当 $k \gg \log h \log \beta$（我们只排除了极少数初始区间，因为这个下界随 $h \to 0$ 增长得非常缓慢（呈对数增长））时，修正损失函数和常微分方程都不再依赖于 k，也就是说，后续的梯度下降+动量更新 $\mathbf{x}^n$ 与 $\widetilde{\mathbf{x}}(t_n)$ 相近，后者是下列方程的解

$$\widetilde{\mathbf{x}}'(t) = -\frac{1}{1-\beta} \nabla \hat{E}(\widetilde{\mathbf{x}}(t)),$$

与 $\hat{E}(\widetilde{\mathbf{x}}(t)) = E(\widetilde{\mathbf{x}}(t)) + \frac{(1+\beta)h}{4(1-\beta)^2} \|\nabla E(\widetilde{\mathbf{x}}(t))\|^2$。

证明。 推论 (2.1) 是定理 1.1 的一个特例，将$E_n = E$ 设为所有 $1 \leq n \leq N$即可直接得到该结果。

□

3 注记5.2的证明

定理 *3.1.* 对随机批次的抽取取期望后，*(SGD+M)* 的隐式梯度正则化方法为$\mathbb{E}(IGRM_s) = \frac{h}{(1+\beta)}\frac{}{4(1-\beta)^3}\|\nabla E(\tilde{} x(t))\|^2 + \frac{h}{4}\frac{}{(1-\beta)^2}\mathbb{E}(\|\nabla E_j(\tilde{} x(t)) - \nabla E(\tilde{} x(t))\|^2)$。对批次抽取取期望后，*(SGD+M)* 的隐式梯度正则化大于 *(SGD)* 的期望隐式梯度正则化。

证明。为避免符号混淆，我们用 $\mathbb{E}$ 表示期望，用 E 表示损失函数。这里，算子 $\mathbb{E}$ 表示对所有随机批次的抽取取期望。回顾式 (1.7)，更新 n^{th} 小批量时的隐式正则化项形式如下:

$$
\begin{aligned}
IGRM_s &= \frac{h}{4}\left(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2 + 2\sum_{k=0}^{n-1}\beta^{n-k}\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2\right), \\
\mathbb{E}(IGRM_s) &= \frac{h}{4}\left(\mathbb{E}(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2) + 2\sum_{k=0}^{n-1}\beta^{n-k}\mathbb{E}(\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2)\right).
\end{aligned} \tag{3.1}
$$

这里，由于期望 $\mathbb{E}$ 是对随机批次的抽取取的，我们首先推导 $\mathbb{E}(\|\nabla G_n(\tilde{}\mathbf{x})\|^2)$，如下所示

$$
\begin{aligned}
&\mathbb{E}\left(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2\right) \\
&= \mathbb{E}\left(\left(\sum_{k=0}^{n}\beta^{n-k}\nabla E_k(\widetilde{\mathbf{x}}(t))\right)^T\left(\sum_{r=0}^{n}\beta^{n-r}\nabla E_k(\widetilde{\mathbf{x}}(t))\right)\right) \\
&= \sum_{i=0}^{n}\beta^{2n-2i}\mathbb{E}\left(\|\nabla E_i(\widetilde{\mathbf{x}}(t))\|^2\right) + \sum_{k=0}^{n}\sum_{r=0,k\neq r}^{n}\beta^{2n-r-k}\mathbb{E}\left(\nabla E_k(\widetilde{\mathbf{x}}(t))^T\nabla E_r(\widetilde{\mathbf{x}}(t))\right) \\
&= \sum_{i=0}^{n}\beta^{2n-2i}\mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i))\|^2) + \underbrace{\sum_{k=0}^{n}\sum_{r=0,k\neq r}^{n}\beta^{2n-r-k}\mathbb{E}\left(\nabla E_k(\widetilde{\mathbf{x}}(t_k))^T\nabla E_r(\widetilde{\mathbf{x}}(t_r))\right)}_{III} + O(h).
\end{aligned} \tag{3.2}
$$

我们通过分别将变量 $\tilde{}\mathbf{x}(t)$ 替换为 $\tilde{}\mathbf{x}(t_i)$、$\tilde{}\mathbf{x}(t_k)$ 和 $\tilde{}\mathbf{x}(t_r)$，得到最后一步。注意，将 $\tilde{}\mathbf{x}(t)$ 改为 $\tilde{}\mathbf{x}(t_i)$ 会引入泰勒级数产生的 $O(h)$ 误差。该 $O(h)$ 误差会与正则化项前面的系数 $\frac{h}{4}$ 相乘。这个总体 $O(h^2)$ 误差不会影响正则化项，因为它是 $O(h)$ 阶的。

注意，随机选取 i^{th} 小批量 E_i 与 $\tilde{}\mathbf{x}(t_i)$ 相互独立，因此我们计算 III 为

$$
III = \sum_{k=0}^{n}\sum_{r=0,k\neq r}^{n}\beta^{2n-r-k}\nabla E(\widetilde{\mathbf{x}}(t_k))^T\nabla E(\widetilde{\mathbf{x}}(t_r)).
$$

回顾一下，全批量梯度损失定义为 $\nabla E(\tilde{}\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M}\nabla E_{(j)}(\tilde{}\mathbf{x})$，其中 $E_{(j)}$ 是主论文式（7）中的第 j 个小批量。式（ 3.2 ）中第一项的各加数变为

$$
\mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i))\|^2) = \|\nabla E(\widetilde{\mathbf{x}}(t_i))\|^2 + \mathbb{E}\|\nabla E_i(\widetilde{\mathbf{x}}(t_i)) - \nabla E(\widetilde{\mathbf{x}}(t_i))\|^2. \tag{3.3}
$$

我们可以如下计算 $\mathbb{E}(\|\nabla G_n(\tilde{\mathbf{x}}(t))\|^2)$

$$
\begin{aligned}
&\mathbb{E}(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2) \\
&= \sum_{i=0}^{n} \beta^{2n-2i} \mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i))\|^2) + \sum_{k=0}^{n} \sum_{r=0,k\neq r}^{n} \beta^{2n-r-k} \nabla E(\widetilde{\mathbf{x}}(t_k))^T \nabla E(\widetilde{\mathbf{x}}(t_r)) + O(h) \\
&= \sum_{i=0}^{n} \beta^{2n-2i} \left(\|\nabla E(\widetilde{\mathbf{x}}(t_i))\|^2 + \mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i)) - \nabla E(\widetilde{\mathbf{x}}(t_i))\|^2) \right) \\
&\quad + \sum_{k=0}^{n} \sum_{r=0,k\neq r}^{n} \beta^{2n-r-k} \nabla E(\widetilde{\mathbf{x}}(t_k))^T \nabla E(\widetilde{\mathbf{x}}(t_r)) + O(h) \\
&= \sum_{k=0}^{n} \sum_{r=0}^{n} \beta^{2n-r-k} \nabla E(\widetilde{\mathbf{x}}(t_k))^T \nabla E(\widetilde{\mathbf{x}}(t_r)) + \sum_{i=0}^{n} \beta^{2n-2i} \mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t_i)) - \nabla E(\widetilde{\mathbf{x}}(t_i))\|^2) + O(h) \\
&= \left\| \sum_{k=0}^{n} \beta^{n-k} \nabla E(\widetilde{\mathbf{x}}(t)) \right\|^2 + \sum_{i=0}^{n} \beta^{2n-2i} \mathbb{E}\left(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2 \right) + O(h). \\
&= \left(\frac{1-\beta^{n+1}}{1-\beta} \right)^2 \|\nabla E(\widetilde{\mathbf{x}}(t))\|^2 + \frac{1-\beta^{2n+2}}{1-\beta^2} \mathbb{E}\left(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2 \right) + O(h).
\end{aligned}
$$

与之前类似，我们将倒数第二行写成这样，因为将 $\tilde{\mathbf{x}}(t_k)$ 改为 $\tilde{\mathbf{x}}(t)$ 只会引入 $O(h)$ 误差。

同样，对于任意这样的 k，我们有

$$
\mathbb{E}(\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2) = \left(\frac{1-\beta^{k+1}}{1-\beta} \right)^2 \|\nabla E(\widetilde{\mathbf{x}}(t))\|^2 + \frac{1-\beta^{2k+2}}{1-\beta^2} \mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2) + O(h).
$$

将 $\mathbb{E}(\|\nabla G_n(\tilde{\mathbf{x}}(t))\|^2)$ 和 $\mathbb{E}(\|\nabla G_k(\tilde{\mathbf{x}}(t))\|^2)$ 的表达式代入 3.1，得到：

$$
\begin{aligned}
&\mathbb{E}(IGRM_s) && (3.4)\\
&\approx \frac{h}{4} \left(\mathbb{E}(\|\nabla G_n(\widetilde{\mathbf{x}}(t))\|^2) + 2 \sum_{k=0}^{n-1} \beta^{n-k} \mathbb{E}(\|\nabla G_k(\widetilde{\mathbf{x}}(t))\|^2) \right) && (3.5)\\
&= \frac{h}{4} \|\nabla E(\widetilde{\mathbf{x}}(t))\|^2 \left(\left(\frac{1-\beta^{n+1}}{1-\beta} \right)^2 + 2 \sum_{k=0}^{n-1} \beta^{n-k} \left(\frac{1-\beta^{k+1}}{1-\beta} \right)^2 \right) && (3.6)\\
&\quad + \frac{h}{4} \mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2) \left(\frac{1-\beta^{2n+2}}{1-\beta^2} + 2 \sum_{k=0}^{n-1} \beta^{n-k} \frac{1-\beta^{2k+2}}{1-\beta^2} \right) + O(h^2) && (3.7)
\end{aligned}
$$

对于较大的迭代次数 n，(3.7) 可化简为

$$
\mathbb{E}(IGRM_s) = \frac{h(1+\beta)}{4(1-\beta)^3} \|\nabla E(\widetilde{\mathbf{x}}(t))\|^2 + \frac{h}{4(1-\beta)^2} \mathbb{E}(\|\nabla E_i(\widetilde{\mathbf{x}}(t)) - \nabla E(\widetilde{\mathbf{x}}(t))\|^2).
$$

□

4 备注 5.3 的证明

定理 4.1. 设$\mathbf{C}$ 为第 k^{th} 次迭代时 *(SGD)* 驱动力的协方差矩阵，即 $cov(\nabla E_k(\mathbf{x})) = \mathbf{C} \in \mathbb{R}^{p\times p}$ 那么，*(SGD+M)*（采用调整后的学习率）的驱动力协方差矩阵为 $cov((1-\beta)\nabla G_k(\mathbf{x})) = \frac{1-\beta}{1+\beta}\mathbf{C}$。这里，随机向量 E_k 和 G_k 均在固定点$\mathbf{x}$处取值。

证明. 假设随机梯度 $\nabla E_k(\mathbf{x})$ 取自一个元素独立同分布的分布，其均值为全批量梯度 $\nabla E(\mathbf{x})$，协方差矩阵为$\mathbf{C}$。那么根据定义，

$$\mathbb{E}((1-\beta)\nabla G_k(\mathbf{x})) = (1-\beta)\sum_{i=0}^{k}\beta^{k-i}\mathbb{E}(\nabla E_i(\mathbf{x})) = (1-\beta^{k+1})\nabla E(\mathbf{x}) \approx \nabla E(\mathbf{x}).$$

根据$\mathbf{C}$ 的定义，我们有

$$\mathbf{C} = \mathbb{E}((\nabla E_k(\mathbf{x}) - \nabla E(\mathbf{x}))(\nabla E_k(\mathbf{x}) - \nabla E(\mathbf{x}))^T) = \mathbb{E}(\nabla E_k(\mathbf{x})\nabla E_k(\mathbf{x})^T) - \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T. \tag{4.1}$$

(SGD+M) 的协方差矩阵为:

$$\begin{aligned}
&cov((1-\beta)G_k(\mathbf{x})) \\
&\quad= \mathbb{E}(((1-\beta)\nabla G_k(\mathbf{x}) - \mathbb{E}((1-\beta)\nabla G_k(\mathbf{x})))((1-\beta)\nabla G_k(\mathbf{x}) - \mathbb{E}((1-\beta)\nabla G_k(\mathbf{x}))^T) \\
&\quad= \mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T)(1-\beta)^2 - \mathbb{E}((1-\beta)\nabla G_k(\mathbf{x}))\mathbb{E}((1-\beta)\nabla G_k(\mathbf{x}))^T \\
&\quad\approx \mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T)(1-\beta)^2 - \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T.
\end{aligned} \tag{4.2}$$

现在按如下方式计算 $\mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T)$:

$$\begin{aligned}
&\mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T) \\
&\quad= \mathbb{E}((\sum_{k=0}^{n}\beta^{n-k}\nabla E_k(\mathbf{x}))(\sum_{k=0}^{n}\beta^{n-k}\nabla E_k(\mathbf{x}))^T) \\
&\quad= \sum_{p=0}^{n}\beta^{2n-2p}\underbrace{\mathbb{E}(\nabla E_p(\mathbf{x})\nabla E_p(\mathbf{x})^T)}_{\mathbf{C}+\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T \text{from 4.1}} + \sum_{i=0}^{n}\sum_{j=0,j\neq i}^{n}\beta^{2n-i-j}\underbrace{\mathbb{E}(\nabla E_i(\mathbf{x})\nabla E_j(\mathbf{x})^T)}_{\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T} \\
&\quad= (\mathbf{C}+\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T)\frac{(1-\beta^{2k+2})}{(1-\beta^2)} + \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T[(\frac{1-\beta^{k+1}}{1-\beta})^2 - \frac{1-\beta^{2k+2}}{1-\beta^2}] \\
&\quad= \frac{1-\beta^{2k+2}}{1-\beta^2}\mathbf{C} + (\frac{1-\beta^{k+1}}{1-\beta})^2\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T.
\end{aligned} \tag{4.3}$$

将 4.3 代入 4.2，得到:

$$\begin{aligned}
&cov((1-\beta)G_k(\mathbf{x})) \\
&\quad\approx \mathbb{E}(\nabla G_k(\mathbf{x})\nabla G_k(\mathbf{x})^T)(1-\beta)^2 - \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T \\
&\quad= \left(\frac{1-\beta^{2k+2}}{1-\beta^2}\mathbf{C} + (\frac{1-\beta^{k+1}}{1-\beta})^2\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T\right)(1-\beta)^2 - \nabla E(\mathbf{x})\nabla E(\mathbf{x})^T \\
&\quad= \frac{(1-\beta^{2k+2})(1-\beta)^2}{1-\beta^2}\mathbf{C} - \beta^{k+1}(2-\beta^{k+1})\nabla E(\mathbf{x})\nabla E(\mathbf{x})^T.
\end{aligned}$$

当迭代次数足够多时 k，它可简化为:

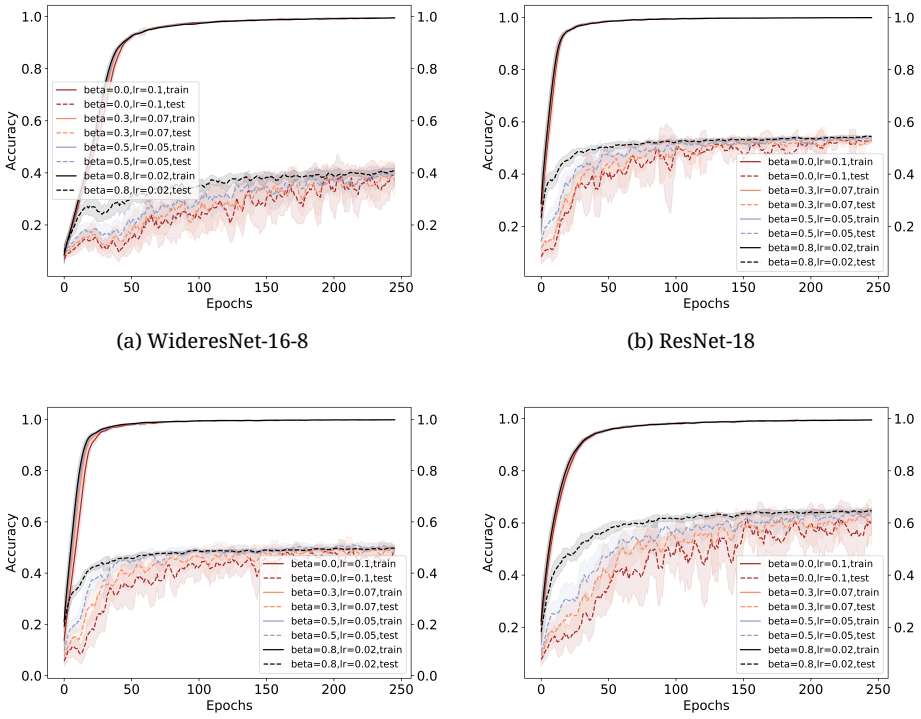
$$cov((1-\beta)G_k(\mathbf{x})) = \frac{(1-\beta)^2}{1-\beta^2}\mathbf{C} = \frac{1-\beta}{1+\beta}\mathbf{C}.$$

□

5 其他实验

我们将在此处的附录中介绍 CIFAR-100 分类的结果。最终测试准确率见正文中的表1。

5.1 CIFAR-100 分类结果



(a) WideresNet-16-8

(b) ResNet-18

(c) ResNet-50

(d) DenseNet-121

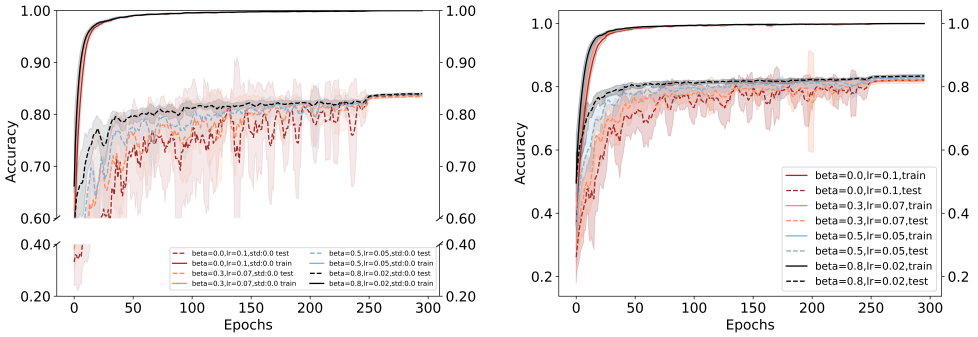
图 1： 在 CIFAR-100 数据集上，采用不同网络架构及不同的 $(h,\ \beta)$ 组合进行分类的结果，其中各组合经过选择，使有效学习率 $^{h}\ _{(1-\beta)}$ 保持不变。在所有实验中，均未启用权重衰减、学习率调度器、丢弃法、标签平滑等外部正则化方法（批归一化除外）。结果取自 3 个具有不同初始化的随机种子对应实验的平均值

.

5.2 学习率调度器的影响

学习率调度器是训练分类网络时的常用方法，因此，探究 IGR 和 IGR-M 在调度器下的效果十分重要。在实验中，我们训练 ResNet-18 和 ResNet-50 网络对 CIFAR-10 数据集进行分类，以比较学习率调度器作用下 (随机梯度下降) 和 (随机梯度下降+M) 的性能。

我们观察到，与之前比较 (SGD) 和 (SGD+M) 的实验一样，随着 β 增大，测试准确率也会提高。我们将这一现象归因于动量相比普通随机梯度下降具有更强的隐式正则化作用。然而，经过调度器调整后，学习率降低了 10 倍。这减弱了 SGD 和 SGD+M 的隐式正则化项的作用， $IGR \propto h$ 也是如此。不过，根据经验观察（图 2），(SGD) 和 (SGD+M) 的测试准确率差异在接近收敛时仍然存在，但可能不像初始迭代阶段那么显著。我们认为，这是因为在较早的迭代阶段，(SGD+M) 显著更强的隐式梯度正则化会引导其轨迹穿过比 (SGD) 更平坦的子流形。这种影响足够显著，以至于即使在调度器启动后（同样



(a) ResNet-18 (b) ResNet-50

图 2：在训练轮次 =250 启用阶梯学习率调度器（1 10）时，使用不同的 β、但相同的有效学习率 $h_{(1-\beta)}$ 对 CIFAR-10 进行分类的结果。

接近收敛时），(SGD+M) 的测试准确率仍略高于 (SGD)。

6 在稳定区域内，根据经典收敛分析，(梯度下降+M) 的收敛速率比 (梯度下降) 高 $^{1}{}_{(1-\beta)}$

本节将证明：当采用学习率 h 的 (梯度下降) 和采用有效学习率 $h_{(1-\beta)}$ 的 (梯度下降+M) 均处于 (梯度下降) 的稳定区域内时，(梯度下降 +M) 的收敛速率比 (梯度下降) 高 $^{1}{}_{(1-\beta)}$ 。

我们在局部二次曲面上考察 (梯度下降) 和 (梯度下降+M) 的经典收敛性。对于标准二次函数，最小化问题为 $\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - b^T\mathbf{x} + c$，其中 $\mathbf{A}$ 是特征值属于 $[\mu, L]$ 的半正定矩阵。通过简单的变量替换，可将最小化问题化为 $\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\Sigma\mathbf{x}$，其中 Σ 的对角线上包含 A 的特征值。因此，$\nabla f(\mathbf{x}) = \Sigma\mathbf{x}$，而 $\nabla^2 f(\mathbf{x}) = \Sigma$。此外，目标函数的条件数记为 $\kappa = \frac{L}{\mu}$。

对于重球法，迭代过程如下：

$$\mathbf{x}^{k+1} = \mathbf{x}^k - h\nabla f(\mathbf{x}^k) + \beta(\mathbf{x}^k - \mathbf{x}^{k-1}) \tag{6.1}$$

在局部二次情形下，迭代过程大致如下：

$$\mathbf{x}^{k+1} = \mathbf{x}^k - h\Sigma\mathbf{x} + \beta(\mathbf{x}^k - \mathbf{x}^{k-1}) = ((1+\beta)\mathbf{I} - h\Sigma)\mathbf{x}^k - \beta\mathbf{x}^{k-1} \tag{6.2}$$

稍作整理后，可写成：

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix} = \begin{bmatrix} (1+\beta)\mathbf{I} - h\Sigma & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{x}^{k-1} \end{bmatrix} \tag{6.3}$$

Denoting $\mathbf{y}^k = \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix}$ and $\mathbf{T} = \begin{bmatrix} (1+\beta)\mathbf{I} - h\Sigma & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$, the norm of $\|\mathbf{y}^k\|_2$ is derived as follows:

$$\|\mathbf{y}^k\| = \|\mathbf{T}\mathbf{y}^{k-1}\| = \|\mathbf{T}^k\mathbf{y}^0\| \leq \|\mathbf{T}^k\|_2\|\mathbf{y}^0\| \leq (\rho(\mathbf{T}))^k\kappa(V)\|\mathbf{y}^0\| \tag{6.4}$$

其中，$\mathbf{T} = \begin{bmatrix} \mathbf{T}_0 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_0 & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{T}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{T}_n \end{bmatrix}$ 是一个 2×2 矩阵，$j = 1, 2..n$。令 r_j 表示每个分块矩阵 $\mathbf{T}j$ 的特征值，则 $\rho(\mathbf{T})$ 是 $\mathbf{T}$ 的谱半径，而 $\mathbf{T}$ 具有特征分解 $\mathbf{T} = VDV^{-1}$，$\kappa(V)$ 是 V 的条件数。$\mathbf{T}$ 与分块对角矩阵置换相似，

其中，$\mathbf{T}_j = \begin{bmatrix} 1+\beta-\alpha\lambda_j & -\beta \\ 1 & 0 \end{bmatrix}$ 是一个 2×2 矩阵，$j = 1, 2..n$。令 r_j 表示每个分块矩阵 $\mathbf{T}j$ 的特征值，则满足

$$r_j = \begin{cases} \frac{1}{2}((1+\beta-\alpha\lambda_j) \pm \sqrt{(1+\beta-h\lambda_j)^2 - 4\beta}), & \text{如果 } (1+\beta-h\lambda_j)^2 - 4\beta = \Delta_j > 0 \\ \frac{1}{2}((1+\beta-\alpha\lambda_j) \pm i\sqrt{|\Delta_j|}, & \text{否则} \end{cases}$$

其中 $i = \sqrt{-1}$。由于由于 $\mathbf{T}$ 具有块矩阵结构，收敛因子 $\rho(\mathbf{T})$ 由以下矩阵中最大的特征值决定所有块矩阵 $\mathbf{T}j$，即 $\rho(\mathbf{T}) = \max_j r_j = \max r_1, \; r_n$。

现在，根据 4 个条件 $\Delta_j \leq 0 \equiv \beta \geq (1 - \sqrt{h\lambda_j})$、$\Delta_j > 0 \equiv \beta \leq (1 - \sqrt{h\lambda_j})$、$|1 - \sqrt{h\mu}| < |1 - \sqrt{hL}|$ 和 $|1 - \sqrt{h\mu}| > |1 - \sqrt{hL}|$，我们有四种子情况来确定 $\rho(\mathbf{T})$：

1. 若 $0 < h \leq (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ 且 $\beta \geq (1 - \sqrt{h\mu})^2$
2. 若 $0 < h \leq (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ 且 $\beta < (1 - \sqrt{h\mu})^2$
3. $h > (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ 且 $\beta \geq (\sqrt{hL} - 1)^2$
4. $h > (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2$ 且 $\beta < (\sqrt{hL} - 1)^2$

对于较小的 h 和固定的 β，满足条件 2，且有效学习率位于梯度下降的稳定区域内。在这一特定条件 (2) 下，我们有 $\Delta_1 > 0$，因此谱半径 $\rho(\mathbf{T})$ 变为（取较大的 r_j）：

$$\rho^{(GD+M)} = \frac{1}{2}(1+\beta-h\mu+\sqrt{(1+\beta-h\mu)^2-4\beta}) \quad \text{[considering the larger term]} \tag{6.5}$$

$$= \frac{1}{2}(1+\beta-h\mu+\sqrt{(1-\beta)^2-2h\mu(1+\beta)+h^2\mu^2}) \tag{6.6}$$

$$= \frac{1}{2}(1+\beta-h\mu+(1-\beta)(\underbrace{\sqrt{1-\frac{2h\mu(1+\beta)+h^2\mu^2}{(1-\beta)^2}}}_{1-\frac{1}{2}\frac{2h\mu(1+\beta)}{(1-\beta)^2}+O(h^2)}-1)+(1-\beta)) \tag{6.7}$$

$$\approx \frac{1}{2}(1+\beta-h\mu-\frac{h\mu(1+\beta)}{(1-\beta)}+(1-\beta)) \text{[小 } h \text{ 近似]} \tag{6.8}$$

$$= 1 - \frac{h\mu}{(1-\beta)} \tag{6.9}$$

类似地，对于以学习率 h 最小化局部二次函数的（梯度下降），采用经典收敛方法，我们有 $\|\mathbf{x}^k\| \leq \rho_{\tilde{h}}^k \|\mathbf{x}^0\|$ 其中 $\rho_{\tilde{h}} = \max(|1-\tilde{h}\mu|, |1-\tilde{h}L|)$。因此，对于足够小的 h，即（$0 < \tilde{h} \leq \frac{2}{L+\mu}$），梯度下降的收敛速度为：

$$\rho^{GD} = 1 - \tilde{h}\mu \tag{6.10}$$

代入 $\tilde{h} = \frac{h}{(1-\beta)}$，可得 $\rho^{(GD+M)} \approx \rho^{(GD)}$。这意味着，如果梯度下降使用大 $\frac{1}{(1-\beta)}$ 倍的学习率，其收敛速度就能与（GD+M）相匹配。

换言之，在（GD）和（GD+M）采用相同学习率（设为 h）的情况下，（GD+M）的收敛速度是（GD）的 $(1-\beta)^{-1}$ 倍，即，$(1-\beta)\rho^{(GD)}$。

7 v 的作用 在寻找更优极小值中的作用

深度神经网络的损失函数通常具有高度非凸性，包含许多局部极小值。一个优秀的优化器应当能够逃离局部且较差（即尖锐）的极小值，从而收敛到较好的平坦极小值。在随机梯度下降中，小批量梯度可以被视为全批量梯度的带噪版本：$\nabla E_i(x) = \nabla E(x) + \eta_i$。因此，当优化器陷入一个具有较差局部极小值的谷底时，带噪梯度中的随机性 $\nabla E_i(x)$ 提供了逃离 该谷底（其中包含较差局部极小值）的可能性。最近，Ibayashi and Imaizumi（2021）从数学上形式化了这一直觉。在他们的定理2中，作者证明了随机梯度下降的逃逸效率（平均逃逸时间的倒数）为

$\propto \exp(-\frac{B}{h}\Delta E \lambda_{max}^{-\frac{1}{2}})$，其中 B、 h、 ΔE 和 λ_{max} 分别表示批量大小、学习率、极小值的深度以及海森矩阵的最大特征值。简言之，较小的批量大小 (B) 和较大的学习率对于逃离较差的局部极小值至关重要。

8 带非线性（Sigmoid）激活函数的 IGR-Min二维模型

除手稿第 4.1 节中的线性情形外，我们现在还考虑一个带有 Sigmoid激活函数的二维非线性模型，以探究 IGR-M 的影响。损失函数 E 使用两个可学习参数（w_1， w_2）进行最小化，但中间加入了一个 sigmoid 层。优化问题如下：

$$(\hat{w_1}, \hat{w_2}) = \arg\min_{w_1,w_2} \frac{1}{2}(y - w_1\sigma(w_2 x))^2 \equiv \arg\min_{w_1,w_2} \frac{1}{2}\left(y - \frac{w_1}{1+e^{-w_2 x}}\right)^2 := E(w_1, w_2)$$

其中 σ 是 Sigmoid激活函数。在这种情况下，梯度的范数表达式如下：

$$\|\nabla E\|^2 = \left|\frac{\partial E}{\partial w_1}\right|^2 + \left|\frac{\partial E}{\partial w_2}\right|^2 = \left(\frac{1}{(1+e^{-w_2 x})^2} + \frac{w_1^2 x^2 e^{-2w_2 x}}{(1+e^{-w_2 x})^4}\right)\left(y - \frac{w_1}{1+e^{-w_2 x}}\right)^2.$$

黑色虚线曲线表示由方程 $w_2 = -\frac{\log(\frac{w_1}{y}-1)}{x}$ 给出的全局极小值。不同于线性情形（其中 IGR 与权重的范数 w_1 和 w_2 成正比），这里的 IGR $\|\nabla E\|^2$ 具有更复杂的等值线（图 3）。因此，为了帮助理解 IGR-M 的作用，我们绘制了两条参考曲线，其中一条是深蓝色曲线，表示原始损失函数的梯度流，其表达式为

$$\mathbf{x}'(t) = -\nabla E(\mathbf{x}(t)).$$

其中$\mathbf{x} = [w_1, w_2]^T$。另一条是实心黑色曲线，表示隐式正则化项的梯度流 $\|\nabla E\|^2$，其表达式为：

$$\mathbf{x}'(t) = -\nabla \|\nabla E(\mathbf{x}(t))\|_2^2.$$

隐式梯度正则化更强的方法，其轨迹会更接近黑色实线。因此，我们绘制了（梯度下降）（$\beta = 0$）和（梯度下降+M）在 $\beta =$ 下的轨迹

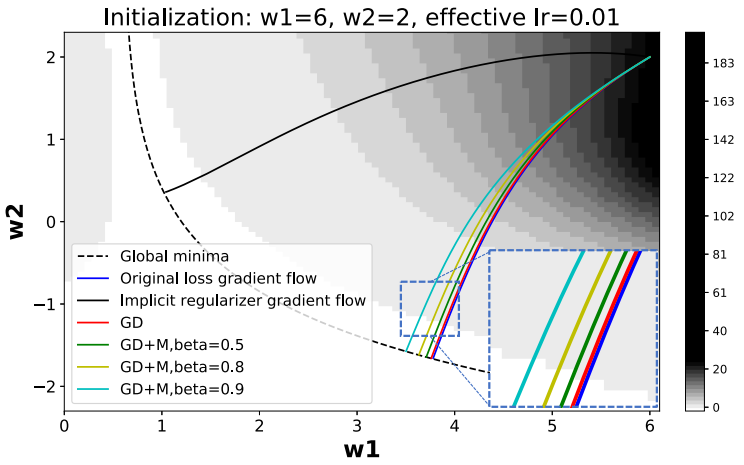


图 3：不同 β 下（GD）和（GD+M）的轨迹，但有效学习率相同 $h_{(1-\beta)}$。随着 β 增大，轨迹逐渐接近隐式正则化项的梯度流（实心黑线），从而支持我们的理论。背景颜色表示 $\|\nabla E\|_2^2$ 的大小

0.5，0.8，以及 0.9，采用 使用相同的初始化（$w_1 = 6, w_2 = 2$）。有效学习率保持不变

对于所有轨迹，其学习率都等于梯度下降的学习率，即 $\frac{h}{1-\beta} = 0.01$。我们可以看到，与梯度下降（GD）相比，（GD+M）的轨迹更接近隐式正则化项对应的梯度流（黑色实线）。更具体地说，我们观察到所有轨迹都收敛到全局极小值曲线。不过，β 越大，轨迹就越接近最小化 $\|\nabla E\|^2$ 的梯度流（黑色实线）。这一观察与我们的定理一致：修正损失是原始损失函数 E 与隐式正则化项 $\|\nabla E\|^2$ 的加权组合，而更大的 β 会赋予正则化项 $\|\nabla E\|^2$ 更大的权重，因此使其更接近黑色实线。

9 未来方向

隐式梯度正则化虽然是研究泛化的有力工具，但它依赖于低阶泰勒近似，而这种近似在学习率较小时效果良好。此外，当前分析基于固定的 β，同时令 $h \to 0$。在实践中，为了更好地指导超参数的选择，针对 h 和 β（$h \to 0$，$\beta \to 1$）均具有渐近性质的界可能更有帮助。这留待未来工作。

参考文献

Sanjeev Arora、Nadav Cohen、Wei Hu 和 Yuping Luo。深度矩阵分解中的隐式正则化。《神经信息处理系统进展》，第 32 卷，2019。

David GT Barrett 和 Benoit Dherin。隐式梯度正则化。arXiv 预印本 arXiv:2009.11162，2020。

Yoshua Bengio 和 Yann LeCun。面向人工智能扩展学习算法。载于《大规模核机器》。麻省理工学院出版社，2007。

Alexander Camuto、Matthew Willetts、Umut Simsekli、Stephen J Roberts 和 Chris C Holmes。《高斯噪声注入中的显式正则化》。《神经信息处理系统进展》，33：16603–16614，2020。

Jeremy M Cohen、Simran Kaur、Yuanzhi Li、J Zico Kolter 和 Ameet Talwalkar。《神经网络上的梯度下降通常发生在稳定性边缘》。arXiv 预印本 arXiv:2103.00065，2021。

Pierre Foret、Ariel Kleiner、Hossein Mobahi 和 Behnam Neyshabur。《通过尖锐度感知最小化高效提升泛化能力》。arXiv 预印本 arXiv:2010.01412，2020。

Ian Goodfellow、Yoshua Bengio、Aaron Courville 和 Yoshua Bengio。《深度学习》，第 1 卷。麻省理工学院出版社，2016。

Suriya Gunasekar、Blake E Woodworth、Srinadh Bhojanapalli、Behnam Neyshabur 和 Nati Srebro。矩阵分解中的隐式正则化。《神经信息处理系统进展》，30，2017。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于《电气电子工程师学会计算机视觉与模式识别会议论文集》，第 770–778 页，2016。Geoffrey E. Hinton、Simon Osindero 和 Yee Whye Teh。深度信念网络的快速学习算法。《神经计算》，18：1527–1554，2006。Forrest Iandola、Matt Moskewicz、Sergey Karayev、Ross Girshick、Trevor Darrell 和 Kurt Keutzer。DenseNet：实现高效的卷积网络描述符金字塔。arXiv 预印本 arXiv:1404.1869，2014。Samy Jelassi 和 Yuanzhi Li。探究动量如何改善深度学习的泛化能力。载于《国际机器学习会议》，第 9965–10040 页。PMLR，2022。Ziwei Ji 和 Matus Telgarsky。梯度下降在不可分数据上的隐式偏差。载于《学习理论会议》，第 1772–1798 页。PMLR，2019。

Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980，2014。Nikola B Kovachki 和 Andrew M Stuart。动量方法的连续时间分析。《机器学习研究期刊》，22(17)：1–40，2021。Qianxiao Li、Cheng Tai 和 E Weinan。随机修正方程与自适应随机梯度算法。载于《国际机器学习会议》，第 2101–2110 页。PMLR，2017。Behnam Neyshabur。深度学习中的隐式正则化。arXiv 预印本 arXiv:1709.01953，2017。Behnam Neyshabur、Ryota Tomioka 和 Nathan Srebro。探寻真正的归纳偏置：论隐式正则化在深度学习中的作用。arXiv 预印本 arXiv:1412.6614，2014。Antonio Orvieto、Hans Kersting、Frank Proske、Francis Bach 和 Aurelien Lucchi。通过反相关噪声注入改善泛化能力。arXiv 预印本 arXiv:2202.02831，2022a。Antonio Orvieto、Anant Raj、Hans Kersting 和 Francis Bach。通过噪声注入对对参数化模型进行显式正则化。arXiv 预印本 arXiv:2206.04613，2022b。Tomaso Poggio、Andrzej Banburski 和 Qianli Liao。深度网络中的理论问题。《美国国家科学院院刊》，117(48)：30039–30045，2020。Boris T Polyak。加速迭代方法收敛的一些方法。《苏联计算数学与数学物理》，4(5)：1–17，1964。Noam Razin 和 Nadav Cohen。深度学习中的隐式正则化可能无法用范数解释。《神经信息处理系统进展》，33：21174–21187，2020。Samuel L Smith 和 Quoc V Le。关于泛化与随机梯度下降的贝叶斯视角。arXiv 预印本 arXiv:1710.06451，2017。Samuel L Smith、Benoit Dherin、David GT Barrett 和 Soham De。随机梯度下降中隐式正则化的起源。arXiv 预印本 arXiv:2101.12176，2021。Daniel Soudry、Elad Hoffer、Mor Shpigel Nacson、Suriya Gunasekar 和 Nathan Srebro。梯度下降在可分数据上的隐式偏差。《机器学习研究期刊》，19(1)：2822–2878，2018。Ilya Sutskever、James Martens、George Dahl 和 Geoffrey Hinton。初始化和动量在深度学习中的重要性。载于《国际机器学习会议》，第 1139–1147 页。PMLR，2013。Tijmen Tieleman、Geoffrey Hinton 等。第 6.5 讲——RMSprop：用近期梯度幅值的运行平均值对梯度进行缩放。COURSERA：《机器学习神经网络》，4(2)：26–31，2012。Gal Vardi 和 Ohad Shamir。平方损失下 ReLU 网络中的隐式正则化。载于《学习理论会议》，第 4224–4258 页。PMLR，2021。Bohan Wang、Qi Meng、Huishuai Zhang、Ruoyu Sun、Wei Chen、Zhi-Ming Ma 和 Tie-Yan Liu。动量会改变可分数据上的隐式正则化吗？，2021。URL https://arxiv.org/abs/ 2110.03891。Jingfeng Wu、Wenqing Hu、Haoyi Xiong、Jun Huan、Vladimir Braverman 和 Zhanxing Zhu。论能够像随机梯度下降一样实现泛化的噪声梯度下降。载于《国际机器学习会议》，第 10367–10376 页。PMLR，2020。

Sergey Zagoruyko 和 Nikos Komodakis。宽残差网络。arXiv 预印本 arXiv:1605.07146，2016。