arXiv:2301.01651v2 [cs.LG] 9 Jan 2023

低精度数值格式下随机梯度下降的收敛性

Matteo Cacciola[1] **Antonio Frangioni**[2] **MasoudAsgharian**[3] **Alireza Ghaffari**[4] **Vahid Partovi Nia**[41]
蒙特利尔理工学院 [2]比萨大学 [3]麦吉尔大学 [4]华为诺亚方舟实验室

摘要

深度学习模型在视觉、文本和语音处理等几乎所有人工智能任务中占据主导地位。随机梯度下降（SGD）是训练此类模型的主要工具，通常采用单精度浮点数值格式进行计算。由于实数计算的误差可以忽略不计，单精度随机梯度下降的收敛通常与实数计算的理论结果一致。然而，采用低精度数值格式进行计算时，数值误差会增大。因此，有充分理由研究适用于低精度计算的随机梯度下降收敛性。我们对随机梯度下降算法进行了确定性和随机性分析，并得到了能够揭示数值格式影响的界。这些界可为理解在各种约束导致难以进行高精度计算时随机梯度下降的收敛性如何受到影响提供指引。

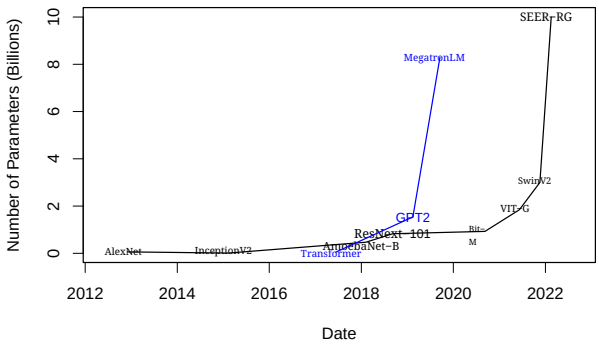


图1：图像分类（黑色）和语言模型（蓝色）深度学习模型规模随时间呈爆炸式增长的趋势。

1 引言

深度学习模型在不同机器学习任务中的成功，使其成为几乎所有视觉、文本和语音处理任务的事实标准。图 1 展示了深度学习模型的规模，表明模型规模呈指数级增长，因此高效计算的需求日益迫切。训练深度学习模型时常用的一种技术是随机梯度下降，但随机梯度下降在低精度数值格式下的理论表现却鲜有研究。尽管关于实数的研究成果层出不穷（例如，参见 Polyak (1967)、Schmidt 等人 (2011)、Ram 等人 (2009)），但近期才开始研究随机梯度下降在低精度数值格式下的表现。根据

精度的不同，损失景观可能发生显著变化。例如，图 2 展示了 ResNet-18 在单精度和低精度数值格式下的损失景观。受图 2 启发，我们对低精度数值格式下进行计算时，随机梯度下降在拟凸函数上的表现开展了形式化研究。

我们注意到，前向传播和反向传播中的数值误差都可能影响算法的收敛行为。可以想见，随着精度降低，数值误差可能会增大。为了理解数值格式对随机梯度下降收敛的影响，需要针对预先确定的精度对随机梯度下降算法进行仔细分析。我们对归一化随机梯度下降算法进行了确定性和随机分析，并得出了明确揭示精度（即数值格式）影响的界。这些界可以为以下问题提供指导：当各种限制使得高精度计算难以实现时，随机梯度下降的收敛会受到怎样的影响，以及在不损害其收敛性的前提下，精度可以降低到何种程度。

我们的实验是在 MNIST数据集上针对逻辑回归进行的。实验结果证实，在数值误差相对较小时，低精度随机梯度下降设置下的损失轨迹至少存在一个极限点，其损失值接近最小值，详见定理 4.1 和定理4.3。

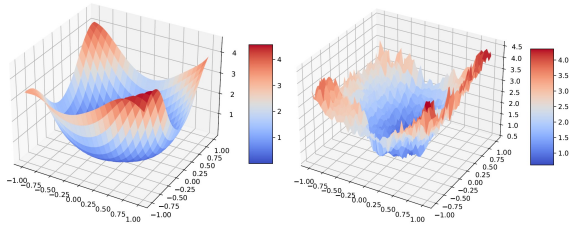


图 2：单精度（左）和低精度数值格式（右）下 ResNet-18 的损失景观。

本文结构如下。第 2 节回顾深度学习模型低精度训练的相关文献，并介绍随机梯度下降理论分析的一些常见背景。第 3 节介绍分析拟凸损失函数所需的一些预备符号和定义，以及浮点数格式。第 4 节给出主要理论结果。第 5 节提供一些实验结果，以支持我们的理论结论。第 6 节总结全文。

2 相关工作

近年来，深度学习模型在计算机视觉、语音和自然语言处理（NLP）等各类机器学习任务中取得了最先进的性能。自 AlexNet 问世以来，ImageNet 分类模型的规模已激增至 $200\times$，而语言模型的规模每年都在变得 $10\times$ 更大。深度学习模型的近期发展趋势表明，更大的模型（如 Transformer（Vaswani 等人，2017年））及其变体（如 GPT-2（Radford 等人，2019年）、MegatronLM（Shoeybi 等人，2019年）和 Brown 等人（2020年）提出的模型）更容易在不同的下游任务中实现泛化。此外，大型语言模型的示例已列于图 1（蓝线）中，显示其参数数量随时间不断增长。视觉模型中也出现了类似趋势，尤其是在视觉 Transformer（Zhai 等人，2022年；Goyal 等人，2022年）兴起之后，它们在多项任务上的表现超过了卷积神经网络（Mahajan 等人，2018年），见图 1（黑线）。尽管这类大型模型在准确率方面具有优势，但其训练和推理阶段的计算成本很高。此外，这些模型的高计算复杂度会导致能耗和内存使用量增加，使得模型训练和部署变得困难，有时甚至无法实现。因此，降低大型深度学习模型的计算复杂度至关重要。

另一方面，也有一些工作尝试手动重新设计更小的模型，使其达到与大型模型相近的准确率；而大型模型通常需要更复杂的训练。在图像分类领域，MobileNet（Howard 等人，2017年）等小型模型的准确率与 ResNet（He 等人，2016年）相近；在语言模型领域，DistilBERT（Sanh 等人，2019年）的表现也接近 BERT。与此同时，也有一些工作尝试自动设计模型，例如 Liu 等人（2018年）和 Zoph 等人（2018年）。其他方法则保留基线模型的架构，同时改动计算过程，例如利用稀疏估计压缩大型模型（Luo 等人，2017年；Ramakrishnan 等人，2020年；Furuya 等人，2022年），或通过采用低精度数值格式来简化计算（Jacob 等人，2018年；Wu 等人，2020年）。一些研究人员甚至进一步探索，将权重存储为二进制数并将激活值降至二进制数（Hubara 等人，2016年），或三进制数（Li 等人，2021年）。

使用单精度浮点数训练大型模型需要大量计算资源。因此，英伟达、谷歌、Transcent 和华为等硬件制造商开始支持低精度数值格式的硬件，例如 Bfloat、float16 和 int8。近期，研究人员尝试将单精度计算映射到更低位宽，参见 Zhang 等人（2020）；Zhao 等人（2021）；Ghaffari 等人（2022）。

大多数关于随机梯度下降的文献都假设损失函数是凸的。我们通过考虑拟凸类损失函数来放宽这一假设，该类函数将凸函数作为特例。拟凸优化领域最早的研究之一是 Kiwiel 和 Murty（1996），其中证明了梯度下降算法会收敛到一个驻点。随后，Kiwiel（2001）去除了可微性假设，并利用拟次梯度证明了收敛结果。对于扰动随机梯度下降，Hu 等人（2015）处理了拟次梯度计算中的有界偏差扰动。在后续工作中，Hu 等人（2016）分析了随机情形。近期，Zhang 等人（2022）研究了针对强凸损失函数的低精度随机梯度下降，并使用了朗之万动力学。相比之下，我们的工作有两点不同：（i）我们假设拟凸性；（ii）我们的设置在随机梯度下降中加入噪声，从而允许对噪声及其分布采用较宽松的假设。

3 预备知识

我们先介绍一些关于拟凸性和后文所用浮点数格式的预备符号。

3.1 拟凸性

定义 3.1。 函数 $f\colon \mathrm{IR}^d \to \mathrm{IR}$ 称为拟凸函数，如果 $\forall a \in \mathrm{IR}$，$f^{-1}[(-\infty, a)] = \{\mathbf{w} \in \mathrm{IR}^d \mid f(\mathbf{w}) \in (-\infty, a)\} = \mathcal{S}_{f,a}$ 是凸的。

定义 3.2. 给定拟凸函数 $f\colon \mathrm{IR}^d \to \mathrm{IR}$，$f$ 在 $\mathbf{w} \in \mathrm{IR}^d$ 处的拟次梯度定义为 $\bar{\partial}^* f(\mathbf{w}) = \{g \in \mathrm{IR}^d \mid \langle g, \mathbf{w}' - \mathbf{w} \rangle \leq 0, \quad \forall \mathbf{w}' \in \mathcal{S}_{f,\ f(\mathbf{w})}\}$

下文中，函数 f 在集合 $\mathbf{\Omega}$ 上的最优值和最优集合分别记为 f^* 和 $\mathbf{\Omega}^*$，即 $f^* = \inf_{\mathbf{w}\in\mathbf{\Omega}} f(\mathbf{w})$ 和 $\mathbf{\Omega}^* = \operatorname{argmin}_{\mathbf{w}\in\mathbf{\Omega}} f(\mathbf{w})$。

定义 3.3. 设 $p > 0$ 且 $L > 0$。 $f : \mathbb{R}^d \to \mathbb{R}$ 被称为满足阶数为 p、常数为 L 的 Hölder 条件，如果

$$f(\mathbf{w}) - f^* \leq L[\operatorname{dist}(\mathbf{w}, \mathbf{\Omega}^*)]^p$$

其中 $\operatorname{dist}(\mathbf{w}, \mathbf{w}') = \min_{\mathbf{w}'} \|\mathbf{w} - \mathbf{w}'\|$，其中 $\|\cdot\|$ 表示欧几里得范数。

3.2 浮点数

基数为 $\beta \in \mathbb{N}$、精度为 $t \in \mathbb{N}$、指数范围为 $[e_{\min}, e_{\max}] \subset \mathbb{Z}$ 的系统定义了一个浮点系统 $\mathbb{F}$，其中元素 $y \in \mathbb{F}$ 可表示为

$$y = \pm m \times \beta^{e-t}, \tag{1}$$

其中 $m \in \mathbb{Z}$、 $0 \leq m < \beta^t$ 和 $e \in [e_{\min}, e_{\max}]$。

对于一个 $x \in [\beta^{e_{\min}-t}, \beta^{e_{\max}}(1-\beta^{-t})]$，，令浮点投影函数为 $\mathrm{fl}(x) := \mathrm{P}_{\mathbb{F}}(x)$，那么对于 x， $y \in \mathbb{F}$，基本运算 $\mathrm{op} \in \{+, -, \times, /\}$ 的舍入误差为

$$\mathrm{fl}(x \operatorname{op} y) = (x \operatorname{op} y)(1 + \delta_0), \tag{2}$$

其中误差的界为 $\delta_0 \leq \beta^{1-t}$。

在尝试计算 $g \in \partial f(w)$ 次梯度时，对于一个$w \in \mathbb{F}$，误差的界为

$$\mathrm{fl}(g) = g(1 + \delta_1),$$

其中， $|\delta_1| \leq c\delta_0$ 和 $c > 0$ 取决于计算该次梯度所需的运算次数。浮点数中的一次次梯度下降迭代（在 $\mathbb{R}^d$ 中）为:

$$\mathbf{w}_{k+1} = \{\mathbf{w}_k - \eta[\mathbf{g}_k(1 + \boldsymbol{\delta}_{1k})]\}(1 + \boldsymbol{\delta}_{2k}),$$

其中我们假设 $\eta \in \mathbb{F}$， $\|\boldsymbol{\delta}_{1k}\| \leq c\delta_0\sqrt{d}$ 且 $\|\boldsymbol{\delta}_{2k}\| \leq \delta_0\sqrt{d}$，其中 d 是 $\boldsymbol{\delta}$ 的维度。我们可以用绝对误差的形式重新表述它

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta(\mathbf{g}_k + \mathbf{r}_k) + \mathbf{s}_k,$$

并且，如果次梯度 $\mathbf{g}_k$ 和 $\mathbf{w}_{k+1}$ 的范数有界，那么 $\|\mathbf{r}_k\| \leq R$ 和 $\|\mathbf{s}_k\| \leq S$ 也有界。注意，误差的无穷范数有界

$$\|\delta_{1k}\|_\infty \leq c\delta_0, \qquad \|\delta_{2k}\|_\infty \leq \delta_0,$$

so

$$R^2 \leq c\delta_0 \|\mathbf{g}_k\|, \quad S^2 \leq \delta_0 \|\mathbf{w}_k\|.$$

4 主要结果

虽然我们的研究主要受训练深度学习模型和浮点误差的启发，但这些方法也可以应用于其他领域。

4.1 确定性分析

令 $\mathrm{P}_\Omega(\cdot)$ 表示 Ω 上的投影算子。我们首先针对求和中的误差，对 Hu 等人（2015）的结果进行调整和改进

定理*4.1*。设 $f: \mathbb{R}^d \to \mathbb{R}$ 为满足阶数为 p、常数为 L 的 *Hölder* 条件的拟凸函数。令 $w_{k+1} = P_\Omega(w_k - \eta \hat{g}_k + s_k)$，其中 $\Omega \subset \mathbb{R}^d$ 为紧集， c 为 Ω 的直径，且 $\hat{g}_k = g_k / \|g_k\| + r_k$，其中 $g_k \in \partial^* f(w_k)$， $\|r_k\| < R < 1$， $\|s_k\| \leq S$。则

$$\liminf_{k\to\infty} f(\mathbf{w}_k) \leq f^* + L\Gamma^p(c),$$

其中

$$\Gamma(c) = \frac{\eta}{2}\left[1 + \left(R + \frac{S}{\eta}\right)^2\right] \bigvee \left[\frac{\eta}{2}\left\{1 - \left(R + \frac{S}{\eta}\right)^2\right\} + c\left(R + \frac{S}{\eta}\right)\right].$$

证明见附录。

注： 与 Hu 等人（2015）不同，减小 η 并不总能减小误差界，因此我们可以通过对 η 求最小值来推导其最优值。

定义

$$\eta_1 = \frac{S}{\sqrt{1+R^2}}, \quad \eta_2 = \sqrt{\frac{S(c-2S)}{1-R^2}}, \quad \eta_3 = \frac{c-S}{R}$$

推论 4.1.1. 使定理 *4.1* 中误差界最小的步长 η 的最优选择至少在以下 *3* 个点之一处取得： $\{\eta_1, \eta_2, \eta_3\}$。

下一个结果给出了前一结果的有限迭代版本。迭代次数 K 和起始点 $\mathbf{w}_0$ 的影响清楚地体现在 $\min_{k<K} f(\mathbf{w}_k)$ 的界中。

定理 4.2。设 $f: \mathbb{R}^d \to \mathbb{R}$ 是一个满足阶数为 p、常数为 L 的 *Hölder* 条件的拟凸函数。设 $\mathbf{w}'_{k+1} = \mathbf{w}_k - \eta \hat{\mathbf{g}}_k$ 以及 $\mathbf{w}_{k+1} = \mathbf{w}'_{k+1} + \mathbf{s}$，其中 $\hat{\mathbf{g}}_k = \mathbf{g}_k / \|\mathbf{g}_k\| + \bar{\mathbf{r}}_k$，其中 $\mathbf{g}_k \in \partial^* f(\mathbf{w}_k)$， $\|\mathbf{r}_k\| < R$， $\|\mathbf{s}_k\| \leq S$.那么，

$$\min_{k<K} f(\mathbf{w}_k) \leq f^* + L\left[\Gamma(c_0) + \frac{c_0}{2\eta K}\right]^p,$$

with $c_0 = \|\mathbf{w}_0 - \mathbf{w}^*\|$.

证明见附录。

4.2 随机分析

这里，我们给出定理 4.1 的随机对应版本。该定理对误差的前两阶矩仅要求温和条件。我们首先定义随机拟次梯度的概念。

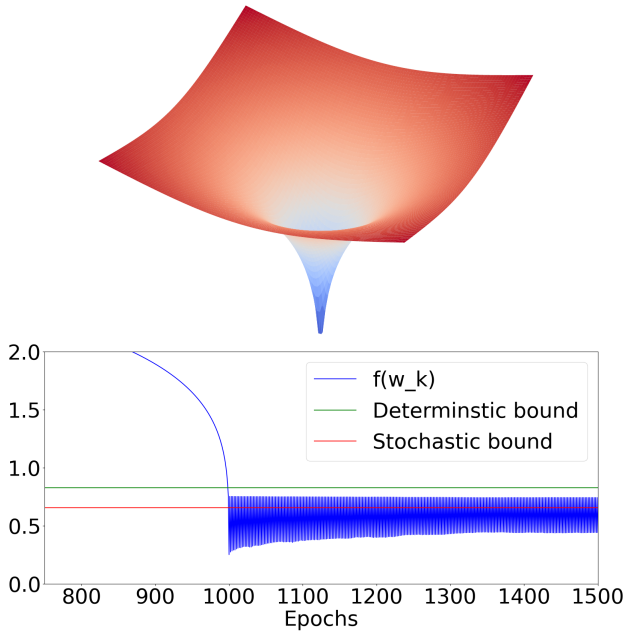


图3：一个二维拟凸函数（上图），以及随机梯度下降轨迹图，证实随机界和确定性界均成立（下图）。

定义4.1. 设$\mathbf{w}$ 和$\mathbf{w}'$ 是定义在概率空间$(\mathcal{W},\mathcal{F},\mathcal{P})$上的 d 维随机向量，且 $f:\mathrm{IR}^d \to \mathrm{IR}$是一个可测拟凸函数。那么$\mathbf{g}(\mathbf{w})$称为 f 在$\mathbf{w}$ 处的单位噪声拟次梯度，如果 $\|\mathbf{g}(\mathbf{w})\| \overset{a.s.}{=} 1$ 且 $\mathcal{P}\left\{\mathcal{S}_{f,f(\mathbf{w})} \cap \mathcal{A}_{\mathbf{w}} \neq \emptyset\right\} = 0$, 其中 $\mathcal{A}_{\mathbf{w}} = \{\mathbf{w}' : \langle \mathbf{g}(\mathbf{w}), \mathbf{w}' - \mathbf{w}\rangle > 0\}$.

因此，受 Hu 等人（2016年）研究结果的启发，我们证明以下定理，该定理同时考虑了梯度中的随机性和计算数值误差。

定理*4.3*。设 $f:IR^d \to IR$ 是满足 $H\ddot{o}lder$ 条件（阶数为 p，常数为 L）的连续拟凸函数。设 $w_{k+1} = P_\Omega[w_k - \eta(\hat{g}_k + r_k) + s_k]$，其中 Ω 为凸闭集，$\hat{g}_k$ 是 f 在 w_k 处的单位含噪拟次梯度，r_k 为独立同分布随机向量，满足 $IE\{r_k\} = 0$ 和 $IE\{\|r_k\|^2\} = d\sigma_r^2$；同样，$s_k$ 也为独立同分布随机向量，满足 $IE\{s_k\} = 0$ 和 $IE\{\|s_k\|^2\} = d\sigma_s^2$。进一步假设 $\hat{g}$、s 和 r 不相关。那么，

$$\liminf_{k\to\infty} f(\mathbf{w}_k) \le f^* + L\left[\frac{\eta}{2}(1+d\sigma_r^2) + \frac{d\sigma_s^2}{2\eta}\right]^p \quad a.s.$$

证明见附录。

类似于推论 4.1.1，可以推导出最优步长。

推论4.3.1。使定理*4.3*中的误差界最小的最优步长 η 为 $\eta^* = \sqrt{\frac{d\sigma_s^2}{d\sigma_r^2+1}}$.

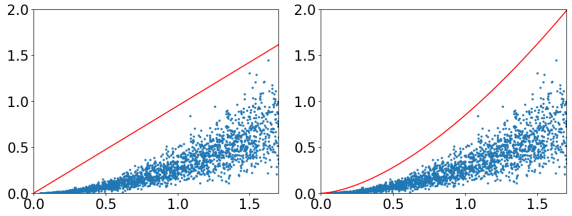


图 4：根据在距离最优点不同距离处计算的损失函数，手动拟合赫尔德参数 p 和 L。左侧面板为使用 $p=1$ 和 $L=0.95$进行的线性拟合。右侧面板展示了使用 $p=1.6$ 和 $L=0.85$进行的拟合。我们的实验使用了 $p=1.6$ 和$L=0.85$。

注：</style}显然，当 d 很大时，最优步长为$\frac{\sigma_s}{\sigma_r}$。
—

5 实验

我们进行了两类实验：一个简单的拟凸函数实验，以及一个基于 MNIST数据集的逻辑回归实验。

5.1 简单拟凸函数

为了评估定理中得到的界，我们从一个简单的拟凸函数入手，该函数严格满足赫尔德条件。我们选择了 $f(\mathbf{w}) = 3\|\mathbf{w}\|^{0.2}$，其中$\mathbf{w} \in \mathrm{IR}^{40}$。在此例中，赫尔德条件的参数为 $p=0.2$ 和 $L=3$。我们分别在梯度和权重更新中加入噪声，记为 $\|\mathbf{r}_k\|$ 和$\|\mathbf{s}_k\|$。这些噪声服从均匀分布$\mathbf{r}_{ki} \sim U(-B_r, B_r)$，和 $s_{ki} \sim U(-B_s, B_s)$。图 3 展示了此次实验的随机界和确定性界。请注意，理论界在随机和确定性两种情况下均成立。

5.2 逻辑回归

这里，我们展示了在 MNIST数据集前两个主成分上进行逻辑回归的实验结果。在这项实验中，我们需要估计损失函数的赫尔德条件参数，以便计算这些界。为此，我们根据损失函数在距最优点不同距离处的取值，手动拟合赫尔德参数 p 和L，见图 4。在我们的实验中，最优点$\mathbf{w}^*$ 通过单精度浮点梯度下降（GD）方法求得，并用作计算界参数 f^* 和 c的参考。

梯度的计算涉及由乘法器和累加器计算的内积。累加器会产生数值误差，其大小与尾数位数有关。

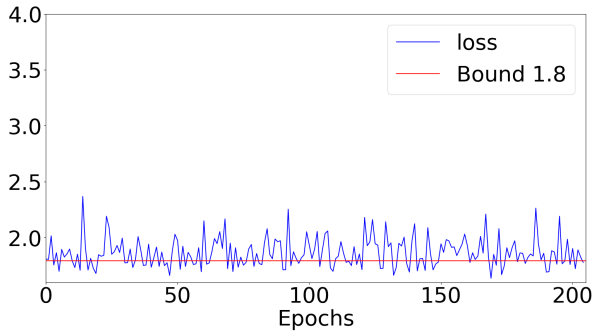


图5：使用单精度随机梯度下降和固定学习率训练的逻辑回归。

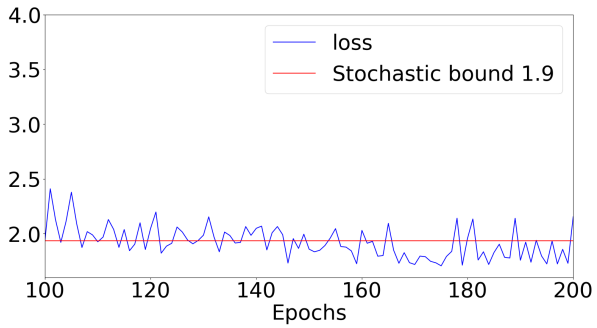


图6：使用累加器大小为15的Bfloat随机梯度下降训练的逻辑回归（$\mathbf{s}_k \neq 0$和$\mathbf{r}_k \neq 0$）。

我们使用 Bfloat 数字格式和较小的累加器大小测试了逻辑回归设置。另请注意，根据定理4.3，计算界需要用到 $d\sigma_s^2$ 和 $d\sigma_r^2$ 的值。因此，在我们的实验中，我们使用这些参数的经验值来计算界。还要注意，对于这些实验，我们没有绘制确定性界，因为它们过于保守。

为了评估赫尔德条件的参数 p和 L（如图4所示估算），我们使用单精度随机梯度下降来确认这些界是否成立。图5显示，损失轨迹（蓝线）的极限点位于凸损失函数最优点附近。图6显示了在权重更新和梯度计算均使用Bfloat数值格式时的损失轨迹。请注意，Bfloat具有8位指数和7位尾数，近期已用于训练深度学习模型。计算使用15位累加器尾数进行。图7显示了权重更新采用单精度、仅梯度计算采用Bfloat数值格式时的损失轨迹。在该实验中，随机界在数值上与单精度随机梯度下降相同，这表明，与梯度的精度相比，权重更新的精度更为重要。

减小累加器尾数位数会直接影响

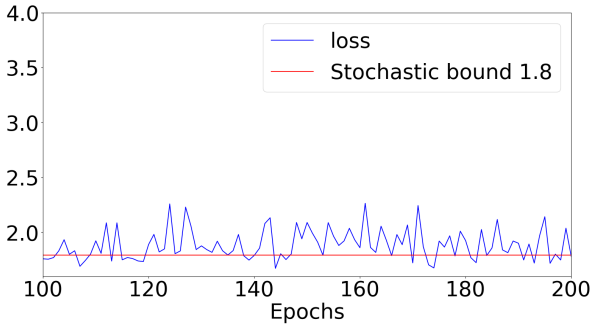


图7：使用累加器大小为 15 的 Bfloat 梯度进行训练，并采用单精度权重更新的逻辑回归（$\mathbf{s}_k = 0$和$\mathbf{r}_k \neq 0$）。

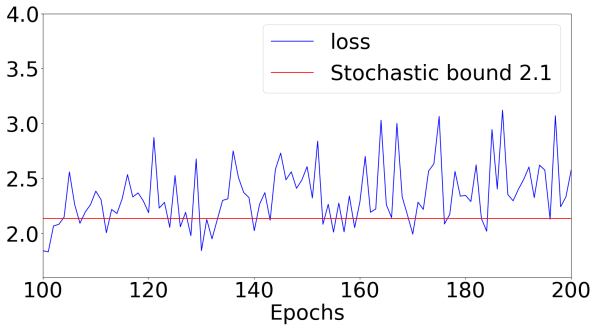


图 8：使用累加器大小为 10 的 Bfloat 梯度进行训练，并采用单精度权重更新的逻辑回归（$\mathbf{s}_k = 0$和$\mathbf{r}_k \neq 0$）。

随机梯度下降的收敛。在 10 位累加器大小的情况下，图 8 显示随机界增大。在本实验中，损失轨迹在最优点附近振荡得更加明显。这表明，累加器大小在减少低精度随机梯度下降计算中的数值误差方面发挥着重要作用，从而改善随机梯度下降的收敛。

6 结论

我们研究了拟凸损失函数下低精度浮点随机梯度下降的收敛性，并将一些现有的凸损失函数确定性界和随机界加以扩展。在理论框架中，我们考虑了权重更新和梯度计算中的数值误差。我们还根据理论结果推导出了最优步长。此外，我们通过实验展示了数值误差对权重更新和梯度计算的影响。实验表明，累加器尾数位数对于减少数值误差、提升随机梯度下降的收敛性起着关键作用。尽管我们在逻辑回归上的实验结果令人鼓舞，但将实验扩展到更复杂的模型，是一个值得探索的未来方向。

参考文献

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、GirishSastry、Amanda Askell 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020年。

Takashi Furuya、Kazuma Suetake、Koichi Taniguchi、Hi- royuki Kusumoto、Ryuji Saiin 和 Tomohiro Daimon。循环神经网络的谱剪枝。载于人工智能与统计国际会议，第 3458–3482 页。PMLR，2022年。

Alireza Ghaffari、Marzieh S Tahaei、Mohammadreza Tayaranian、Masoud Asgharian 和 Vahid Partovi Nia。整数运算是否足以用于深度学习训练？*arXiv* 预印本 *arXiv:2207.08822*，2022年。Priya Goyal、Quentin Duval、Isaac Seessel、Mathilde Caron、Mannat Singh、Ishan Misra、Levent Sagun、Ar- mand Joulin 和 Piotr Bojanowski。未经筛选且未经监督的图像预训练可使视觉模型更加稳健、公平。*arXiv* 预印本 *arXiv:2202.08360*，2022年。何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 770–778 页，2016年。Andrew G Howard、Menglong Zhu、Bo Chen、Dmitry Kalenichenko、Weijun Wang、Tobias Weyand、Marco Andreetto 和 Hartwig Adam。MobileNets：面向移动视觉应用的高效卷积神经网络。*arXiv* 预印本 *arXiv:1704.04861*，2017年。Yaohua Hu、Xiaoqi Yang 和 Chee-Khian Sim。拟凸优化问题的非精确次梯度方法。《欧洲运筹学杂志》，240(2)：315–327，2015年。Yaohua Hu、Carisa Yu 和 Chong Li。拟凸优化问题的随机次梯度方法。《非线性与凸分析杂志》，17：711–724，01 2016年。Itay Hubara、Matthieu Courbariaux、Daniel Soudry、Ran El-Yaniv 和 Yoshua Bengio。二值化神经网络。神经信息处理系统进展，29，2016年。Benoit Jacob、Skirmantas Kligys、Bo Chen、Menglong Zhu、Matthew Tang、Andrew Howard、Hartwig Adam 和 Dmitry Kalenichenko。用于高效纯整数运算推理的神经网络量化与训练。载于*IEEE*计算机视觉与模式识别会议论文集，第 2704–2713 页，2018年。Krzysztof Kiwiel 和 Katta Murty。最速下降法在最小化拟凸函数。《优化理论与应用杂志》，89，1996年4月。doi: 10.1007/BF02192649。Krzysztof C. Kiwiel。用于拟凸最小化的次梯度方法的收敛性与效率。《数学规划》，90:1–25，2001年。Xinlin Li、Bang Liu、Yaoliang Yu、Wulong Liu、Chunjing Xu 和 Vahid Partovi Nia。S3：用于有效训练低比特移位网络的符号稀疏移位重参数化。《神经信息处理系统进展》，34:14555–14566，2021年。Hanxiao Liu、Karen Simonyan 和 Yiming Yang。DARTS：可微架构搜索。*arXiv*预印本 *arXiv:1806.09055*，2018年。Jian-Hao Luo、Jianxin Wu 和 Weiyao Lin。ThiNet：一种用于深度神经网络压缩的滤波器级剪枝方法。发表于*IEEE*国际计算机视觉会议论文集，第 5058–5066 页，2017年。Dhruv Mahajan、Ross Girshick、Vignesh Ramanathan、何恺明、Manohar Paluri、Yixuan Li、Ashwin Bharambe 和 Laurens Van Der Maaten。探索弱监督预训练的极限。发表于欧洲计算机视觉会议（*ECCV*）论文集，第 181–196 页，2018年。Boris Polyak。求解极值问题的一种通用方法。《苏联数学·文献》，8，1967年1月。Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。语言模型是无监督多任务学习器。*OpenAI* 博客，1(8): 9，2019年。S. Sundhar Ram、A. Nedi´c 和 V. V. Veeravalli。用于凸优化的增量随机次梯度算法。《SIAM优化杂志》，20(2):691–717，2009年。doi: 10.1137/080726380。URL https: // doi.org/10.1137/080726380。Ramchalam Kinattinkara Ramakrishnan、Eyyub Sari 和 Vahid Partovi Nia。用于剪枝卷积网络和循环网络的可微掩码。发表于*2020*年第*17*届计算机与机器人视觉会议（*CRV*），第 222–229 页。IEEE，2020年。Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。DistilBERT：BERT的蒸馏版本，更小、更快、更便宜、更轻量。*arXiv* 预印本 *arXiv:1910.01108*，2019年。Mark Schmidt、Nicolas Roux 和 Francis Bach。用于凸优化的非精确近端梯度方法的收敛速率。《神经信息处理系统进展》，24，2011年。Mohammad Shoeybi、Mostofa Patwary、Raul Puri、Patrick LeGresley、Jared Casper 和 Bryan Catanzaro。Megatron-LM：使用模型并行训练数十亿参数的语言模型。*arXiv*预印本 *arXiv:1909.08053*，2019年。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。你所需要的只是注意力。《神经信息处理系统进展》，30，2017年。Hao Wu、Patrick Judd、Xiaojie Zhang、Mikhail Isaev 和 Paulius Micikevicius。用于深度学习推理的整数化：原理与实证评估。*arXiv*预印本 *arXiv:2004.09602*，2020年。Xiaohua Zhai、Alexander Kolesnikov、Neil Houlsby 和 Lucas Beyer。扩展视觉 Transformer。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 12104–12113 页，2022年。Ruqi Zhang、Andrew Gordon Wilson 和 Christopher De Sa。低精度随机梯度朗之万动力学。发表于国际机器学习会议，第 26624–26644 页。PMLR，2022年。Xishan Zhang、Shaoli Liu、Rui Zhang、Chang Liu、Di Huang、Shiyi Zhou、Jiaming Guo、Qi Guo、Zidong Du、Tian Zhi 等。定点反向传播训练。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 2330–2338 页，2020年。Kang Zhao、Sida Huang、Pan Pan、Yinghan Li、Yingya Zhang、Zhenyu Gu 和 Yinghui Xu。用于训练卷积神经网络的分布自适应 int8 量化。发表于*AAAI*人工智能会议论文集，第 35 卷，第 3483–3491 页，2021年。Barret Zoph、Vijay Vasudevan、Jonathon Shlens 和 Quoc V Le。学习可迁移架构以实现可扩展图像识别。发表于*IEEE*计算机视觉与模式识别会议论文集，第 8697–8710 页，2018年。

补充材料

7 证明

设 $\mathbf{w} \in \mathrm{IR}^d$ 且 $r \in \mathrm{IR}^+$。定义 $\mathbb{B}(\mathbf{w}\ r) = \left\{\mathbf{w}' \in \mathrm{IR}^d | \ \|\mathbf{w}' - \mathbf{w}\| \leq r\right\}$。已知 f 是拟凸的并满足 Holder 条件，且 $\boldsymbol{\Omega}$ 是一个紧集，由此可得 $\boldsymbol{\Omega}^* \neq \emptyset$ 且 $f^* > -\infty$。我们先介绍证明定理 4.1 所需的两个技术引理。

引理 7.1。 设 $\mathbf{g} \in \mathrm{IR}^d$ 是一个 d 维向量，且 R、C、B、$\eta \in \mathrm{IR}^+$ 为实常数，满足 $\|\mathbf{g}\| = 1$、 $R < 1$ 和 $C > B$。定义 $F(\mathbf{w}，\ \mathbf{r}) = \eta\|\mathbf{g} + \mathbf{r}\|^2 - 2\langle \mathbf{w}，\ \mathbf{g} + \mathbf{r}\rangle$，其中 $\mathbf{r}$、$\mathbf{w}$ $\in \mathrm{IR}^d$。考虑以下优化问题：

$$F^* = \max_{\substack{\mathbf{w},\mathbf{r}\in\mathbb{R}^d \\ \text{满足以下条件:}}} F(\mathbf{w}, \mathbf{r}), \tag{3}$$

$$\begin{aligned} &\langle \mathbf{g}，\ \mathbf{w}\rangle \geq B, \\ &\|\mathbf{r}\|^2 \leq R^2, \\ &\|\mathbf{w}\|^2 \leq C^2. \end{aligned} \tag{4}$$

Then

$$F^* \leq \max\left\{\eta\left(1 + R^2\right) + 2R\sqrt{\eta^2 + C^2 - 2\eta B} - 2B, \eta\left(R^2 - 1\right)\right\}.$$

推论 *7.0.1*。 *</style}*引理 *7.1* 中的上界在 $d \geq 2$时可达到。

引理 7.2。 （*Kiwiel*（*2001*）的引理 *6*）设 $f(\cdot)$ 满足 IR^d 上的 *H¨older* 条件，其参数为 p，常数为 L。若 $\mathbb{B}(\mathbf{w}^*，\ M) \subset \mathcal{S}_{f,\ f(\mathbf{w})}$，则 $\langle \mathbf{w} - \mathbf{w}^*，\ \mathbf{g}/\|\mathbf{g}\|\rangle \geq M，\quad \forall \mathbf{g} \in \partial^* f(\mathbf{w})$。

定理 4.1 采用此类结果的标准论证，通过反证法证明。为简单起见，我们在假设 $S = 0$ 的条件下证明该定理。一般情形的证明，即 $S \neq 0$，只需注意到$\mathbf{w}_{k+1} = \mathrm{P}_\Omega\left[\mathbf{w}_k - \eta\left(\mathbf{g}\| \ {}^{k}\mathbf{g}_k\| + \mathbf{r}_k + \frac{\mathbf{s}_k}{\eta}\right)\right]$ 并将 R 替换为 $R + \frac{S}{\eta}$，即可类似地得到。鉴于 $\overline{\eta}$ 假设为固定值，这一变化不会造成任何困难。

7.1 定理 4.1 的证明

证明。鉴于 c 是$\mathbf{\Omega}$的直径，当 $\Gamma(c) \geq c$ 时，所需结果显然成立。因此，假设 $\Gamma(c) < c$。设 $\exists\, \delta > 0,\ k > 0\ s.t.\ f(\mathbf{w}_k) > f^* + L(\Gamma(c) + \delta)^p \quad \forall\, k > k$且$\mathbf{w}^* \in \mathbf{\Omega}^*$。鉴于 $f(\cdot)$ 满足 H¨older 条件，¯对所有 $k > k$

$$f(\mathbf{w}) - f^* \leq L\,\mathrm{dist}(\mathbf{w}, \mathbf{\Omega}^*)^p \leq L(\Gamma(c) + \delta)^p < f(\mathbf{w}_k) - f^*, \qquad \forall\, \mathbf{w} \in \mathbb{B}(\mathbf{w}^*, \Gamma(c) + \delta,) \quad ,$$

¯这意味着 $\mathbb{B}(\mathbf{w}^*, \Gamma(c) + \delta) \subset \mathcal{S}_{f, f(\mathbf{w}_k)}\ \forall\, k > k_{\circ}$ 利用引理 7.2

$$\langle \frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}, \mathbf{w}_k - \mathbf{w}^* \rangle \geq \Gamma(c) + \delta, \quad \forall\, k > \bar{k},$$

因此，可以对 $B = \Gamma(c) + \delta,\ C = c,\ \mathbf{g} = \mathbf{g}\|^{k}\mathbf{g}_k\|,\ \mathbf{w} = \mathbf{w}^* - \mathbf{w}_k,\ \mathbf{r} = \mathbf{r}_k,\ R = R$ 以及 $\eta = \eta$ 使用引理 7.1。这就得到

$$\langle \mathbf{w}_k - \mathbf{w}^*, \mathbf{g}_k \rangle + \langle \mathbf{w}_k - \mathbf{w}^*, \mathbf{r}_k \rangle - \frac{\eta}{2} \| \frac{\mathbf{g}_k}{\|\mathbf{g}_k\|} + \mathbf{r}_k \|^2 \geq \min \left\{ \Gamma(c) + \delta - \frac{\eta}{2}\left(1 + R^2\right) - R\sqrt{\eta^2 + c^2 - 2\eta\left[\Gamma(c) + \delta\right]}, \frac{\eta}{2}\left(1 - R^2\right) \right\}. \tag{5}$$

可以证明，当 δ 足够小时

$$\Gamma(c) + \delta - \frac{\eta}{2}\left(1 + R^2\right) - R\sqrt{\eta^2 + c^2 - 2\eta\left[\Gamma(c) + \delta\right]} < \frac{\eta}{2}\left(1 - R^2\right).$$

鉴于 $\Gamma(c)$ 是两项中的最大值，我们分别证明这两项都满足该不等式，因此它们的最大值也满足该不等式。若 $\Gamma(c) = \frac{\eta}{2}(1 - R^2)$，则

$$\Gamma(c) + \delta - \frac{\eta}{2}(1 + R^2) - R\sqrt{\eta^2 + c^2 - 2\eta\left[\Gamma(c) + \delta\right]} = \delta - R\sqrt{\eta^2 + c^2 - 2\eta\left[\Gamma(c) + \delta\right]} \leq 0 \leq \frac{\eta}{2}\left(1 - R^2\right), \tag{6}$$

其中，第一个不等式是因为我们可以任意取小 δ，第二个不等式则由 $R < 1$ 得出。否则，$\Gamma(c) = \frac{\eta(1-R^2)}{2} + Rc$，因此

$$\begin{aligned}
\left[\Gamma(c) + \delta - \frac{\eta}{2}\left(1 + R^2\right)\right]^2 &= \frac{\eta^2}{4}\left(1 + R^2\right)^2 + \left[\Gamma(c) + \delta\right]^2 - \eta\left[\Gamma(c) + \delta\right]\left(1 + R^2\right) \\
&= \frac{\eta^2}{4}\left(1 + R^2\right)^2 + \left[\frac{\eta\left(1 - R^2\right)}{2} + Rc + \delta\right]^2 - \eta\left[\frac{\eta\left(1 - R^2\right)}{2} + R\delta + \delta\right]\left(1 + R^2\right) \\
&= \frac{\eta^2}{4}\left(1 + R^2\right)^2 + \frac{\eta^2\left(1 - R^2\right)^2}{4} + R^2c^2 + \eta Rc\left(1 - R^2\right) - \frac{\eta^2\left(1 - R^4\right)}{2} \\
&\quad - \eta Rc\left(1 + R^2\right) + \delta^2 + 2\delta Rc + \delta\eta\left(1 - R^2\right) - \delta\eta\left(1 + R^2\right) \\
&= \frac{\eta^2}{2}\left(1 + R^4\right) + R^2c^2 - 2\eta cR^3 - \frac{\eta^2\left(1 - R^4\right)}{2} + \delta^2 + 2\delta Rc - 2\delta\eta R^2 \\
&= R^4\eta^2 + R^2c^2 - 2\eta cR^3 + \delta^2 + 2\delta Rc - 2\delta\eta R^2 \\
&= R^2\left(c^2 + \eta^2R^2 - 2\eta Rc - 2\eta\delta\right) + \delta^2 + 2\delta Rc \\
&= R^2\left[\eta^2 + c^2 - \eta^2\left(1 - R^2\right) - 2\eta Rc - 2\eta\delta\right] + \delta^2 + 2\delta Rc \\
&= R^2\left\{\eta^2 + c^2 - 2\eta\left[\Gamma(c) + \delta\right]\right\} + \delta^2 + 2\delta Rc.
\end{aligned}$$

对第一项和最后一项开平方

$$\Gamma(c) + \delta - \frac{\eta}{2}\left(1 + R^2\right) = \sqrt{R^2\left\{\eta^2 + c^2 - 2\eta\left[\Gamma(c) + \delta\right]\right\} + \delta^2 + 2\delta Rc}. \tag{7}$$

那么，式 (5) 中 min 运算符的第一项为

$$\begin{aligned}
\Gamma(c) + \delta - \frac{\eta}{2}\left(1 + R^2\right) - R\sqrt{\eta^2 + c^2 - 2\eta\left[\Gamma(c) + \delta\right]} &= \sqrt{R^2\left\{\eta^2 + c^2 - 2\eta\left[\Gamma(c) + \delta\right]\right\} + \delta^2 + 2\delta Rc} \\
&\quad - R\sqrt{\eta^2 + c^2 - 2\eta\left[\Gamma(c) + \delta\right]} \xrightarrow{\delta \to 0} 0.
\end{aligned} \tag{8}$$

因此，如果 δ 足够小

$$\Gamma(c)+\delta-\frac{\eta}{2}\left(1+R^2\right)-R\sqrt{\eta^2+c^2-2\eta\left[\Gamma(c)+\delta\right]}\leq\frac{\eta}{2}\left(1-R^2\right). \tag{9}$$

如果 δ 足够小，则由式 (5)、(6) 和 (9) 可得

$$\langle\mathbf{w}_k-\mathbf{w}^*,\mathbf{g}_k\rangle+\langle\mathbf{w}_k-\mathbf{w}^*,\mathbf{r}_k\rangle-\frac{\eta}{2}\|\frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}+\mathbf{r}_k\|^2\geq\Gamma(c)+\delta-\frac{\eta}{2}\left(1+R^2\right)-R\sqrt{\eta^2+c^2-2\eta\left[\Gamma(c)+\delta\right]}. \tag{10}$$

由推导式 (7) 时的论证可知，如果 $\Gamma(c)\geq\frac{\eta}{(1-R^2)}2\quad+Rc,$

$$\Gamma(c)+\delta-\frac{\eta}{2}(1+R^2)\geq\sqrt{R^2\left\{\eta^2+c^2-2\eta\left[\Gamma(c)+\delta\right]\right\}+\delta^2+2\delta Rc}. \tag{11}$$

注意到 $\Gamma(c)>\frac{\eta}{2}$ 后，我们得到 $\eta^2+c^2-2\eta(\Gamma(c)+\delta)<c^2$。由此可得

$$\begin{aligned}R^2\left\{\eta^2+c^2-2\eta\left[\Gamma(c)+\delta\right]\right\}+\delta^2+2\delta Rc&=\left\{R\sqrt{\eta^2+c^2-2\eta\left[\Gamma(c)+\delta\right]}+\delta\right\}^2\\&\quad-2\delta R\sqrt{\eta^2+c^2-2\eta\left[\Gamma(c)+\delta\right]}+2R\delta c\\&\geq\left\{R\sqrt{\eta^2+c^2-2\eta\left[\Gamma(c)+\delta\right]}+\delta\right\}^2.\end{aligned} \tag{12}$$

由式 (10)、(11) 和 (12) 可知

$$\langle\mathbf{w}_k-\mathbf{w}^*,\mathbf{g}_k\rangle+\langle\mathbf{w}_k-\mathbf{w}^*,\mathbf{r}_k\rangle-\frac{\eta}{2}\|\frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}+\mathbf{r}_k\|^2\geq\delta,$$

从而

$$\|\mathbf{w}_{k+1}-\mathbf{w}^*\|\leq\|\mathbf{w}_k-\mathbf{w}^*\|^2-2\eta\left(\langle\mathbf{w}_k-\mathbf{w}^*,\mathbf{g}_k\rangle+\langle\mathbf{w}_k-\mathbf{w}^*,\mathbf{r}_k\rangle-\frac{\eta}{2}\|\frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}+\mathbf{r}_k\|^2\right)\leq\|\mathbf{w}_k-\mathbf{w}^*\|^2-2\eta\delta,$$

其中 w 我们在第一个不等式中使用了 $\mathbf{w}_{k+1}$ 的定义以及投影映射 $\mathrm{P}_\Omega(\cdot)$ 的性质 y.

Thus

$$\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2\leq\|\mathbf{w}_k-\mathbf{w}^*\|-2\eta\delta\leq\ldots\leq\|\mathbf{w}_0-\mathbf{w}^*\|-2\eta(k-\bar{k}+1)\delta,\quad\forall k>\bar{k}. \tag{13}$$

注意到上界在 k 趋于 $+\infty$ 时趋于 $-\infty$，即可得到所需结果，这将导致矛盾。

■

7.1.1 引理 7.1 背后的直觉

通过仔细审视定理 4.1 的证明以及 Hu 等人，2015 的定理 3.1 的证明，我们注意到证明的关键在于

$$\begin{aligned}\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2\leq\|\mathbf{w}_k-t\hat{\mathbf{g}}_k-\mathbf{w}^*\|^2&=\|\mathbf{w}_k-\mathbf{w}^*\|^2-2\eta\langle\mathbf{w}_k-\mathbf{w}^*,\frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}\rangle-2\eta\langle\mathbf{w}_k-\mathbf{w}^*,\mathbf{r}_k\rangle+\eta^2\|\frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}+\mathbf{r}_k\|^2\\&=\|\mathbf{w}_k-\mathbf{w}^*\|^2-2\eta\left(B+\langle\mathbf{w}_k-\mathbf{w}^*,\mathbf{r}_k\rangle-\frac{\eta}{2}\|\frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}+\mathbf{r}_k\|^2\right),\end{aligned}$$

其中 B 应满足 $\exists\delta'>0\mid\left(B+\langle\mathbf{w}_k-\mathbf{w}^*,\ \mathbf{r}_k\rangle-\frac{\eta}{2}\|\mathbf{g}\|\ \overset{k}{\mathbf{g}_k}\|+\mathbf{r}_k\|^2\right)>\delta'$。最终的界将显著取决于 B。更具体地说，B 值越大，得到的界也越大，因而越差。

在 Hu 等人（2015年）的研究中，作者使用

$$|\mathbf{w}_k-\mathbf{w}^*\|^2-2\eta\left(B+\langle\mathbf{w}_k-\mathbf{w}^*,\mathbf{r}_k\rangle-\frac{\eta}{2}\|\frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}+\mathbf{r}_k\|^2\right)\leq\|\mathbf{w}_k-\mathbf{w}^*\|^2-2\eta\left(B-Rc-\frac{\eta}{2}(1+R)^2\right), \tag{14}$$

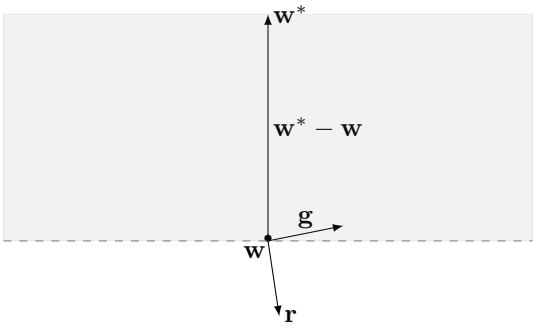


图 9：给定w 和w*，g 需要与w* −w 具有正的标量积（灰色区域）。因此，r 不可能同时与w* −w 方向相反，并且与 g 共线。

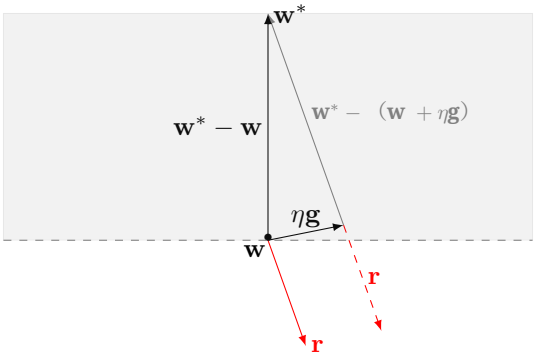


图10： 给定w、w* 和g，最坏情况似乎是 −r = w* − (w + ηg)，将其缩放为范数 R .

然后选择 $B = Rc + \frac{\eta}{2}(1+R)^2 + \delta$。为得到式（14），Hu 等人（2015年）使用

$$\langle \mathbf{w}_k - \mathbf{w}^*, \mathbf{r}_k \rangle \geq -\|\mathbf{w}_k - \mathbf{w}^*\|\|\mathbf{r}_k\|, \tag{15}$$

$$\|\frac{\mathbf{g}_k}{\|\mathbf{g}_k\|} + \mathbf{r}_k\| \leq \|\frac{\mathbf{g}_k}{\|\mathbf{g}_k\|}\| + \|\mathbf{r}_k\|. \tag{16}$$

当$\mathbf{r}_k = -\alpha(\mathbf{w}_k - \mathbf{w}^*)$和$\mathbf{r}_k = \beta \mathbf{g}_k$ 分别满足 $\alpha,\ \beta > 0$时，式 (15) 和 (16) 中的等号成立。这进而意味着 $\langle \mathbf{g}_k, \mathbf{w}_k - \mathbf{w}^* \rangle = \langle \mathbf{r}_k/\beta, -\mathbf{r}_k/\alpha \rangle = -\frac{1}{\alpha\beta}\|\mathbf{r}_k\|^2 < 0$，这与$\mathbf{g}_k \in \partial^* f(\mathbf{w}_k)$的拟次微分定义相矛盾。这意味着式 (15) 和 (16) 中的等号不可能同时成立。因此，我们使用引理 7.1 而非式 (14) 来证明定理 4.1。图 9 展示了我们论证背后的直观解释。

图10表明，当$\mathbf{r}_k$ 与$\mathbf{w}_k + \eta \frac{\mathbf{g}_k}{\|\mathbf{g}_k\|} - \mathbf{w}^*$共线时，会出现最坏情况。证明见第7.1.2节。

7.1.2 引理 7.1 的证明

粗略地说，我们的思路是利用 KKT 条件，证明最优解位于可行集的边界上。

证明。 目标函数的梯度为

$$\nabla F(\cdot) = (-2(\mathbf{g} + \mathbf{r}), 2\eta(\mathbf{g} + \mathbf{r}) - 2\mathbf{w}).$$

目标函数是凸函数，因此至少在可行域边界上的某一点达到最优解。我们

回顾一下，若存在 $\exists \mathbf{r},\mathbf{w} \in \mathrm{IR}^d$，使得问题 3 满足弱斯莱特条件

$$\begin{aligned}\langle \mathbf{g}, \mathbf{w}\rangle &\geq B,\\ \|\mathbf{r}\|^2 &< R^2,\\ \|\mathbf{w}\|^2 &< C^2.\end{aligned}$$

这些条件由 $\mathbf{r} = 0$和 $\mathbf{w} = B\mathbf{g}$满足，因此我们可以利用 KKT 方程，得到问题 (3) 的最优解所必须满足的条件。根据关于梯度分量$\mathbf{r}$的 KKT 条件

$$\frac{\partial F}{\partial \mathbf{r}} = \lambda \frac{\partial(\|\mathbf{r}\|^2 - R^2)}{\partial \mathbf{r}} \Leftrightarrow 2\eta(\mathbf{g}+\mathbf{r}) - 2\mathbf{w} = \lambda 2\mathbf{r}$$

与 $\lambda \in \mathrm{IR}^+$。

若 $\eta = \lambda$，则 $\mathbf{w} = \eta\mathbf{g}$，因此目标函数可简化为 $\eta + \eta\|\mathbf{r}\|^2 - 2\eta$。此时，任何满足$\|\mathbf{r}\| = R$ 的 $\mathbf{r}$ 都能取得最优值，即 $\eta(R^2 - 1)$。否则，

$$\mathbf{r} = \frac{\mathbf{w} - \eta\mathbf{g}}{\eta - \lambda}. \tag{17}$$

目标函数可写为

$$F(\mathbf{r}, \mathbf{w}) = \eta + 2\langle \eta\mathbf{g} - \mathbf{w}, \mathbf{r}\rangle + \eta\|\mathbf{r}\|^2 - 2\langle \mathbf{w}, \mathbf{g}\rangle.$$

Let ($\mathbf{r}^*$,$\mathbf{w}^*$) 是问题 (3) 的一个最优解。我们通过反证法证明以下事实 :

a) $\langle \mathbf{r}^*, \eta\mathbf{g} - \mathbf{w}^*\rangle \geq 0.$

假设 $\langle \mathbf{r}^*, \eta\mathbf{g} - \mathbf{w}^*\rangle < 0$。考虑 $\tilde{\mathbf{r}} = -\mathbf{r}^*$，解 ($\tilde{\mathbf{r}}$,$\mathbf{w}^*$) 是可行的，因为 $\|\tilde{\mathbf{r}}\| = \|\mathbf{r}^*\| \leq R$。此外，$\langle \eta\mathbf{g} - \mathbf{w}^*, \tilde{\mathbf{r}}\rangle > 0 > \langle \eta\mathbf{g} - \mathbf{w}^*, \mathbf{r}^*\rangle$，由此得到

$$F(\mathbf{r}^*, \mathbf{w}^*) - F(\tilde{\mathbf{r}}, \mathbf{w}^*) = \langle \eta\mathbf{g} - \mathbf{w}^*, \mathbf{r}^*\rangle - \langle \eta\mathbf{g} - \mathbf{w}^*, \tilde{\mathbf{r}}\rangle < 0.$$

最后一个不等式违背了$\mathbf{r}^*$的最优性，因此命题得证。

b) $\|\mathbf{r}^*\| = R.$

假设 $\|\mathbf{r}^*\| < R$。考虑 $\tilde{\mathbf{r}} = R\frac{\mathbf{r}^*}{\|\mathbf{r}^*\|}$，利用 (a) 点可得

$$\langle \eta\mathbf{g} - \mathbf{w}^*, \tilde{\mathbf{r}}\rangle - \langle \eta\mathbf{g} - \mathbf{w}^*, r^*\rangle = \left(\frac{R}{\|\mathbf{r}^*\|} - 1\right)\langle \eta\mathbf{g} - \mathbf{w}^*, \mathbf{r}^*\rangle \geq 0.$$

此外，$R = \|\tilde{\mathbf{r}}\| > \|\mathbf{r}^*\|$，因此

$$F(\mathbf{r}^*, \mathbf{w}^*) - F(\tilde{\mathbf{r}}, \mathbf{w}^*) = 2\left[\left(\frac{R}{\|\mathbf{r}^*\|} - 1\right)\langle \eta\mathbf{g} - \mathbf{w}^*, \mathbf{r}^*\rangle\right] + \eta(\|\mathbf{r}^*\| - R) < 0.$$

最后一个不等式违背了$\mathbf{r}^*$ 的最优性，因此结论成立。

c) $\langle \mathbf{w}^*, \mathbf{g}\rangle = B.$

假设 $\langle \mathbf{w}^*, \mathbf{g}\rangle > B$。那么 $\tilde{\mathbf{w}} = \mathbf{w}^* - (\langle \mathbf{w}^*, \mathbf{g}\rangle - B)\mathbf{g}$，($\mathbf{r}^*$, $\tilde{\mathbf{w}}$) 是一个可行点，因为

$$\|\tilde{\mathbf{w}}\|^2 = \|\mathbf{w}^*\|^2 + (\langle \mathbf{w}^*, \mathbf{g}\rangle - B)^2 - 2\left[\langle \mathbf{w}^*, \mathbf{g}\rangle - B\right]\langle \mathbf{w}^*, \mathbf{g}\rangle < \|\mathbf{w}^*\|^2$$

且 $\langle \tilde{\mathbf{w}}, \mathbf{g}\rangle = B$。

$$\begin{aligned}F(\mathbf{w}^*, \mathbf{r}^*) - F(\tilde{\mathbf{w}}, \mathbf{r}^*) =& 2\langle \tilde{\mathbf{w}} - \mathbf{w}^*, \mathbf{g} + \mathbf{r}^*\rangle = -2\langle [\langle \mathbf{w}^*, \mathbf{g}\rangle - B]\,\mathbf{g}, \mathbf{g} + \mathbf{r}^*\rangle\\ =& -2(\langle \mathbf{w}^*, \mathbf{g}\rangle - B) - 2\langle [\langle \mathbf{w}^*, \mathbf{g}\rangle - B]\mathbf{g}, \mathbf{r}^*\rangle\\ \leq& -2[\langle \mathbf{w}^*, \mathbf{g}\rangle - B](1 - \|\mathbf{g}\|\|\mathbf{r}^*\|) < 0.\end{aligned}$$

最后一个不等式与$\mathbf{w}^*$的最优性相矛盾，由此完成了这部分证明。

由式（17）、（a）和（b）可得

$$\mathbf{r}^* = \alpha(\eta\mathbf{g} - \mathbf{w}^*) \text{ with } \alpha = \frac{R}{\|\eta\mathbf{g} - \mathbf{w}^*\|}. \tag{18}$$

所需结果由 (18)、(4) 得出,

$$\begin{aligned} F(\mathbf{w}^*, \mathbf{r}^*) &= \eta(1 + R^2) + 2R\sqrt{\eta^2 + \|\mathbf{w}\|^2 - 2\eta B} - 2B \\ &\leq \eta(1 + R^2) + 2R\sqrt{\eta^2 + c^2 - 2\eta B} - 2B. \end{aligned}$$

■

现在给出推论 7.0.1 的证明

证明。假设 $d \geq 2$ 且 $\|\mathbf{w}^*\| < C$。考虑满足 $\langle \mathbf{g}, \mathbf{y}\rangle = 0$ 且 $\langle \mathbf{w}^*, \mathbf{y}\rangle \geq 0$ 的 $\mathbf{y}$。由于 $d \geq 2$，不失一般性，我们可以假设 $\|\mathbf{y}\| = 1$。此外，定义 $\tilde{\mathbf{w}} = \mathbf{w}^* + a\mathbf{y}$，其中 $a \in \mathbb{R}$，则 $\exists\, a > 0 \; s.t. \; \|\tilde{\mathbf{w}}\| = C$。点 $(\mathbf{r}^*, \tilde{\mathbf{w}})$ 是一个可行点，因为 $\langle \tilde{\mathbf{w}}, \mathbf{g}\rangle = \langle \mathbf{w}^*, \mathbf{g}\rangle = B$。

我们考虑两种情况:

- $\mathbf{g}$ 与 $\mathbf{w}^*$ 不共线。

在这种情况下，$\exists \mathbf{y}$ 使得 $\langle \mathbf{w}^*, \mathbf{y}\rangle > 0$。方程（18）得出

$$\begin{aligned} F(\mathbf{w}^*, \mathbf{r}^*) - F(\tilde{\mathbf{w}}, \mathbf{r}^*) &= 2\langle \tilde{\mathbf{w}} - \mathbf{w}^*, \mathbf{g} + \mathbf{r}^*\rangle \\ &= 2\langle a\mathbf{y}, \mathbf{g} + \mathbf{r}^*\rangle \\ &= 2\langle a\mathbf{y}, \mathbf{r}^*\rangle \\ &= 2\langle a\mathbf{y}, \alpha[\eta\mathbf{g} - \mathbf{w}^*]\rangle \\ &= -2a\alpha\langle \mathbf{y}, \mathbf{w}^*\rangle \\ &< 0. \end{aligned}$$

最后一个方程与 $\mathbf{w}^*$ 的最优性矛盾，因此

$$\|\mathbf{w}^*\| = C. \tag{19}$$

- $\mathbf{g}$ 与 $\mathbf{w}^*$ 共线。

在这种情况下，(c)意味着 $\mathbf{w}^* = B\mathbf{g}$。考虑 $\tilde{\mathbf{r}} = \beta[(\eta - B)\mathbf{g} - a\mathbf{y}]$, , 则

$$\begin{aligned} F(\mathbf{w}^*, \mathbf{r}^*) - F(\tilde{\mathbf{w}}, \tilde{\mathbf{r}}) =& 2\left[\langle \eta\mathbf{g} - \mathbf{w}^*, \mathbf{r}^*\rangle - \langle \eta\mathbf{g} - \tilde{\mathbf{w}}, \tilde{\mathbf{r}}\rangle - \langle \mathbf{g}, \mathbf{w}^* - \tilde{\mathbf{w}}\rangle\right] \\ =& 2\left\{\langle (\eta - B)\mathbf{g}, \operatorname{sgn}(\eta - B)R\mathbf{g}\rangle - \langle (\eta - B)\mathbf{g} - a\mathbf{y}, \beta\left[(\eta - B)\mathbf{g} - a\mathbf{y}\right]\rangle\right\} \\ =& 2\left\{\langle \operatorname{sgn}(\eta - B)(\eta - B)\mathbf{g}, R\mathbf{g}\rangle - \beta\langle (\eta - B)\mathbf{g} - a\mathbf{y}, (\eta - B)\mathbf{g} - a\mathbf{y}\rangle\right\} \\ =& 2\left\{|\eta - B|R - \beta\left[(\eta - B)^2 + a^2\right]\right\} \\ =& 2\left\{|\eta - B|R - R\sqrt{(\eta - B)^2 + a^2}\right\} \\ =& R\sqrt{(\eta - B)^2 + a^2}\left(\frac{|\eta - B|}{\sqrt{(\eta - B)^2 + a^2}} - 1\right) \\ <& 0, \end{aligned}$$

其中，第二个等式由 $\mathbf{w}^* = B\mathbf{g}$、方程（18）和 $\langle \mathbf{g},\mathbf{y}\rangle = 0$ 得出，第四个和第五个等式则由 $\langle \mathbf{g},\mathbf{y}\rangle = 0$ 推出。最后一个等式违背了 $(\mathbf{w}^*, \mathbf{r}^*)$ 的最优性，因此（19）必须成立。于是可得所需结果,

$$F(\mathbf{w}^*, \mathbf{r}^*) = \eta(1 + R^2) + 2R\sqrt{\eta^2 + c^2 - 2\eta B} - 2B.$$

■

7.2 推论 4.1.1

证明。 最小化定理 4.1 中的界，等价于最小化

$$\max\left\{\frac{\eta}{2}\left[1+\left(R+\frac{S}{\eta}\right)^2\right],\frac{\eta}{2}\left[1-\left(R+\frac{S}{\eta}\right)^2\right]+c\left[R+\frac{S}{\eta}\right]\right\}.$$

定义函数 $G(\eta)$ 如下

$$\begin{aligned}G(\eta):=&\max\left\{\frac{\eta}{2}(1+R^2+\frac{S^2}{\eta^2}+\frac{2SR}{\eta}),\frac{\eta(1-R^2-\frac{S^2}{\eta^2}-\frac{2RS}{\eta})}{2}+Rc+\frac{Sc}{\eta}\right\}\\=&\max\left\{\frac{\eta}{2}(1+R^2)+\frac{S^2}{2\eta}+SR,\frac{\eta}{2}(1-R^2)-\frac{S^2}{2\eta}-RS+Rc+\frac{Sc}{\eta}\right\}.\end{aligned}$$

注意

$$G(\eta)=\begin{cases}\frac{\eta}{2}(1+R^2)+\frac{S^2}{2\eta}+SR & \text{if } t\geq\frac{c-S}{R}=\eta_3,\\ \frac{\eta}{2}(1-R^2)-\frac{S^2}{2\eta}-RS+Rc+\frac{Sc}{\eta} & \text{otherwise.}\end{cases}$$

第一部分的最小值在 $\eta_1=\frac{S}{\sqrt{1+R^2}}$ 处取得，而第二部分的最小值在 $\eta_2=\sqrt{S(c-2S)\ 1-R^2}$ 处取得。由于 $G(\eta)$ 是两个凸函数的最大值，因此它在集合 $\{\eta_1.\eta_2,\eta_3\}$ 中取得全局最小值。

■

7.2.1 定理 4.2 的证明

如果$\mathbf{w}_{k+1}=\mathrm{P}_\Omega(\mathbf{w}_k-\eta\hat{}\mathbf{g}_k+\mathbf{s}_k)$。 则该定理可由公式（13）立即推出。然而，要证明定理 4.2，关键在于 $\|\mathbf{w}_{k+1}-\mathbf{w}^*\|\leq\|\mathbf{w}_k-\mathbf{w}^*\|,\quad\forall 0<k<\tilde{k}$其中 $\tilde{k}=\inf\{k|\,f(\mathbf{w}_k)>f^*+L\,[\Gamma(C_0)]^p\}$。因此，只需证明最后这个不等式，即可完成定理 4.2 的证明。

证明。 固定 $K>0$。假设 $\delta>0$且 $f(\mathbf{w}_k)>f^*+L(\Gamma(C_0)+\delta)^p,\quad\forall\,k\leq K.$通过归纳法可以证明

$$\|\mathbf{w}_k-\mathbf{w}^*\|\leq C_0-2\eta\delta k,\quad\forall\,k\leq K.$$

情形 $k=0$是显然的。

假设 $\|\mathbf{w}_j-\mathbf{w}^*\|\leq C_0-2\eta\delta j\leq C_0,\quad\forall j\leq k$，则可以结合$c=C_0$重复定理 4.1 的证明过程。这是因为，在定理 4.1 的$\mathbf{w}_{k+1}$ 定义中，投影算子仅在使用引理 7.1 时用于界定 $\|\mathbf{w}_k-\mathbf{w}^*\|$ 的范数。采用导出公式（13）的类似论证，并利用归纳假设，可得

$$\|\mathbf{w}_{k+1}-\mathbf{w}^*\|^2\leq\|\mathbf{w}_k-\mathbf{w}^*\|-2\eta\delta\leq C_0-2\eta\delta(k+1),$$

若 $\delta\geq\frac{C_0}{2\eta K}$，则上述不等式会导致矛盾，因为 $f(\mathbf{w}_K)>f^*$。因此 $\exists\,k\leq K\ s.t.\ f(\mathbf{w}_k)\leq f^*+L(\Gamma(C_0)+\delta)^p$。证明至此完成。■

7.2.2 推论4.3.1的证明

证明。 定义 $G(\eta)=\eta(d\sigma_r^2+1)+\frac{d\sigma_s^2}{\eta}$。函数 $G(\cdot)$ 在 $\mathrm{IR}^+\setminus\{0\}$ 上是凸且连续可微的。因此， $G(\cdot)$ 的最小值在其导数的根处取得。显然，我们只关心正根。 $G(\cdot)$ 的导数为

$$\frac{\partial G(\cdot)}{\partial\eta}=d\sigma_r^2+1-\frac{d\sigma_s^2}{\eta^2},$$

具有正根 $\eta^*=\sqrt{\frac{d\sigma_s^2}{d\sigma_r^2+1}}$。

■

7.3 定理 4.3 的证明

证明。 令 $\mathcal{F}_n = \sigma\{\mathbf{w}_0, \mathbf{w}_1, \ldots\ldots, \mathbf{w}_n\}$ 为由 $\{\mathbf{w}_0, \mathbf{w}_1, \ldots\ldots, \mathbf{w}_n\}$生成的 σ-代数。假设 $\mathbf{w}_n \notin \mathbf{\Omega}^*$，由$\mathbf{w}_{n+1}$的定义可知

$$\begin{aligned}\|\mathbf{w}_{n+1} - \mathbf{w}^*\|^2 &\leq \|\mathbf{w}_n - \eta\hat{\mathbf{g}}(\mathbf{w}_n) - \eta\mathbf{r}_n - \mathbf{s}_n - \mathbf{w}^*\|^2 \\ &\leq \|\mathbf{w}_n - \mathbf{w}^*\|^2 - 2\eta\langle\hat{\mathbf{g}}(\mathbf{w}_n), \mathbf{w}_n - \mathbf{w}^*\rangle - 2\eta\langle\mathbf{w}_n - \mathbf{w}^*, \mathbf{r}_n\rangle \\ &\quad - 2\langle\mathbf{w}_n - \mathbf{w}^*, \mathbf{s}_n\rangle + \eta^2\|\hat{\mathbf{g}}(\mathbf{w}_n) + \mathbf{r}_n + \frac{\mathbf{s}_n}{\eta}\|^2.\end{aligned}$$

对 $\mathcal{F}_n$取条件期望

$$\begin{aligned}\mathbb{E}\left\{\|\mathbf{w}_{n+1} - \mathbf{w}^*\|^2|\mathcal{F}_n\right\} &\leq \|\mathbf{w}_n - \mathbf{w}^*\|^2 - 2\eta\mathbb{E}\left\{\langle\hat{\mathbf{g}}(\mathbf{w}_n), \mathbf{w}_n - \mathbf{w}^*\rangle|\mathcal{F}_n\right\} - 2\eta\mathbb{E}\left\{\langle\mathbf{w}_n - \mathbf{w}^*, \mathbf{r}_n\rangle|\mathcal{F}_n\right\} \\ &\quad - 2\mathbb{E}\left\{\langle\mathbf{w}_n - \mathbf{w}^*, \mathbf{s}_n\rangle|\mathcal{F}_n\right\} + \eta^2\mathbb{E}\left\{\|\hat{\mathbf{g}}(\mathbf{w}_n) + \mathbf{r}_n + \frac{\mathbf{s}_n}{\eta}\|^2|\mathcal{F}_n\right\} \\ &\leq \|\mathbf{w}_n - \mathbf{w}^*\|^2 - 2\eta\left(\frac{f(\mathbf{w}) - f^*}{L}\right)^{\frac{1}{p}} - 2\eta\langle\mathbf{w}_n - \mathbf{w}^*, \mathbb{E}\left\{\mathbf{r}_n\right\}\rangle \\ &\quad - 2\langle\mathbf{w}_n - x^*, \mathbb{E}\left\{\mathbf{s}_n\right\}\rangle + \eta^2\left(1 + \mathbb{E}\left\{\|\mathbf{r}_n\|^2\right\} + 2\mathbb{E}\left\{\langle\hat{\mathbf{g}}(\mathbf{w}_n), \mathbf{r}_n\rangle|\mathcal{F}_n\right\}\right) \\ &\quad + \mathbb{E}\left\{\|\mathbf{s}_n\|^2\right\} + 2\eta\mathbb{E}\left\{\langle\hat{\mathbf{g}}(\mathbf{w}_n), \mathbf{s}_n\rangle|\mathcal{F}_n\right\} + 2\eta\mathbb{E}\left\{\langle\mathbf{s}_n, \mathbf{r}_n\rangle\right\} \\ &\leq \|\mathbf{w}_n - \mathbf{w}^*\|^2 - 2\eta\left(\frac{f(\mathbf{w}) - f^*}{L}\right)^{\frac{1}{p}} + \eta^2\left(1 + d\sigma_r^2\right) + d\sigma_s^2, \qquad (20)\end{aligned}$$

其中，第二个不等式的推导用到了引理 7.3。第三个不等式由$\mathbf{g(w}_n\mathbf{)}, \mathbf{r}_n, \mathbf{s}_n$的正交性以及 $\mathrm{IE}\ \{\mathbf{r}_n\} = \mathrm{IE}\{\mathbf{s}_n\} = 0$这一事实得出。现在考虑 $\delta > 0$和

$$\mathbf{\Omega}_\delta = \mathbf{\Omega} \cap \mathcal{S}_{f,a},$$

其中 $a = f^* + L\left[\frac{\eta}{2}\left(1 + d\sigma_r^2\right) + \frac{d\sigma_s^2}{2\eta} + \delta\right]^p$。由于 f连续，$\exists \mathbf{y} \in \mathrm{IR}^d | \mathbf{y}_\delta \in \mathbf{\Omega}$ 且 $f\mathbf{(y}_\delta\mathbf{)} = f^* + L(\delta)^p$。定义过程

$$\tilde{\mathbf{w}}_{k+1} = \begin{cases}\mathsf{P}_\Omega\left[\eta\mathbf{w}_k - \eta\hat{\mathbf{g}}\left(\tilde{\mathbf{w}}_k\right)\right], & \text{if } \tilde{\mathbf{w}}_k \notin \mathbf{\Omega}_\delta, \\ \mathbf{y}_\delta & \text{otherwise.}\end{cases}$$

我们证明 $\tilde{\mathbf{w}}_k \notin \mathbf{\Omega}_\delta,\ \forall k > K,\ K \in \mathbb{N}$ 会导致矛盾。不失一般性，我们考虑 $K = 0$。假设对于任意 k，$\tilde{\mathbf{w}}_k \notin \mathbf{\Omega}_\delta$，并令 $\hat{\mathcal{F}}_k = \{\tilde{\mathbf{w}}_0, \tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_k\}$。由于 $f(\tilde{\mathbf{w}}_k) \geq f^* + L\left[\frac{\eta}{2}\left(1 + d\sigma_r^2\right) + \frac{d\sigma_s^2}{2\eta} + \delta\right]^p$，根据式 (20) 可得 $\forall \mathbf{w}^* \in \mathbf{\Omega}^*$,

$$\begin{aligned}\mathbb{E}\left\{\|\tilde{\mathbf{w}}_{k+1} - \mathbf{w}^*\|^2|\,\hat{\mathcal{F}}_k\right\} &\leq \|\tilde{\mathbf{w}}_k - \mathbf{w}^*\|^2 - 2\eta\left[\left(\frac{f(\tilde{\mathbf{w}}_k) - f^*}{L}\right)^{\frac{1}{p}} - \frac{t}{2}(1 + d\sigma_r^2) - \frac{d\sigma_s^2}{2\eta}\right] \\ &\leq \|\tilde{\mathbf{w}}_k - \mathbf{w}^*\|^2 - 2\eta\delta.\end{aligned}$$

因此，定理 7.4 推出 $\sum_{k=0}^{+\infty} 2\eta\delta < +\infty$。这是一个矛盾，因此证明完毕。 ■

引理 7.3。 （*Hu* 等人（*2016*）的引理 *2.4*）。设为 $\mathbf{w} \notin \mathbf{\Omega}^*$，且 $\hat{\mathbf{g}}(\mathbf{w})$ 是 f 在 $\mathbf{w}$ 处的单位噪声拟次梯度。那么，$\forall \mathbf{w}^* \in \mathbf{\Omega}^*$ 满足：对于任意 $\mathcal{F} = \sigma\{\mathbf{w}_1, ..., \mathbf{w}_k\}$，其中 $\mathbf{w}_i$是随机变量 $\forall i = 1、..、k$，有

$$\mathbb{E}\left\{\langle\hat{g}(\mathbf{w}), \mathbf{w}^* - \mathbf{w}\rangle|\mathcal{F}\right\} \geq \left(\frac{f(\mathbf{w}) - f^*}{L}\right)^{\frac{1}{p}}.$$

引理 7.4。 （*Hu* 等人（*2016*）的引理 *2.5*）。设 $\{Y_k\}$、$\{Z_k\}$ 和 $\{W_k\}$ 为三个非负随机变量序列，设 $\{\mathcal{F}_k\}$ 为一个滤过。假设对于每个 k，以下条件均得到满足：

a) Y_k、Z_k 和 W_k 是 $\mathcal{F}_k$ 中随机变量的函数；

b) $\mathbb{E}\left\{Y_{k+1}|\mathcal{F}_k\right\} \leq Y_k - Z_k + W_k$;

c) $\sum_{k=0}^{\infty} W_k < +\infty$.

那么 $\sum_{k=0}^{\infty} Z_k < +\infty$，序列 $\{Y_k\}$ 几乎必然收敛到非负随机变量 Y。

8 实验设置详情

我们使用归一化梯度下降（NGD）算法，对第5节中给出的确定性函数 $f(\mathbf{w}) = \|\mathbf{w}\|^{0.2}$ 进行实验。最大训练轮次设为 1500。我们使用了不同的 C_0、 B_r、 B_s 和学习率。人为添加的误差在每个坐标上均服从均匀分布。因此，定理4.3中所需的方差为 $\sigma_r^2 = \frac{B_r^2}{3}$ 和 $\sigma_s^2 = \frac{B_s^2}{3}$。计算定理 4.1 中的界时，使用了 $c = C_0$。

8.1 测试最优学习率

我们固定 B_s、 B_r 的值，并尝试不同的 η 取值，以获得推论4.3.1 给出的最优选择 η^*。对于每个测试的 η 值，实验重复进行 10 次。最后，在每个 epoch，绘制所有具有相同 η 的实验中观察到的最大损失函数值。

$B_r = B_s = 0.1$ 的实验结果如图11和图12所示。使用推论4.3.1 建议的值 $\eta = 0.0348$ 时，观察到的损失轨迹（蓝线）处于最低水平。我们的定理正确预测了降低 η 的值有时并无益于收敛，见图12。

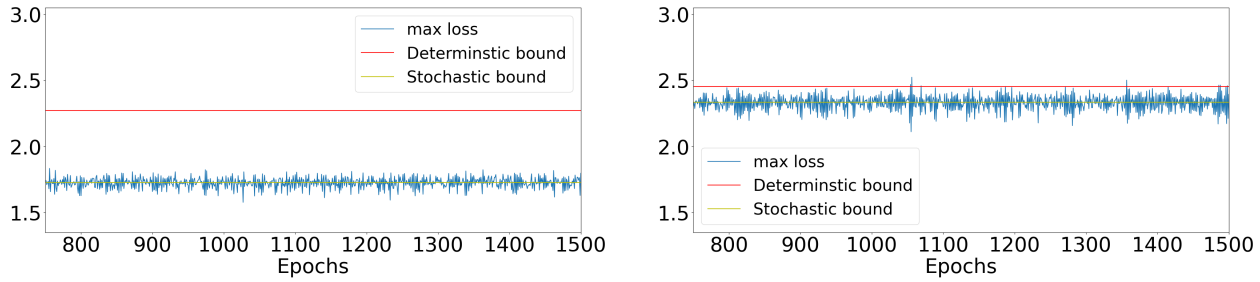


图11： $\eta = 0.1$ （左图）和 $\eta = 0.5$ （右图）的实验结果

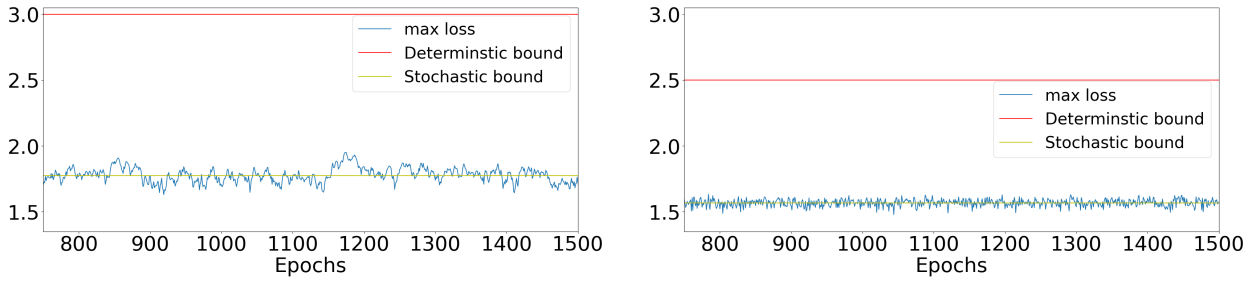


图12：使用 $\eta = 0.01$ （左图）和 $\eta = 0.0348$ （右图）的结果；降低 η 的值会导致界更差、收敛效果更差

8.2 原始MNIST上的实验结果

本节报告在原始MNIST数据集上获得的结果。与正文中的实验不同，这里未使用主成分分析来缩减输入的规模。

图13表明，损失轨迹（蓝线）在凸损失函数最优点附近存在一个极限点。图14展示了权重更新采用单精度、而仅梯度计算使用 Bfloat数值格式时的损失轨迹。在此实验中，随机界在数值上与单精度随机梯度下降相同，这证实了图7中已经观察到的结果。

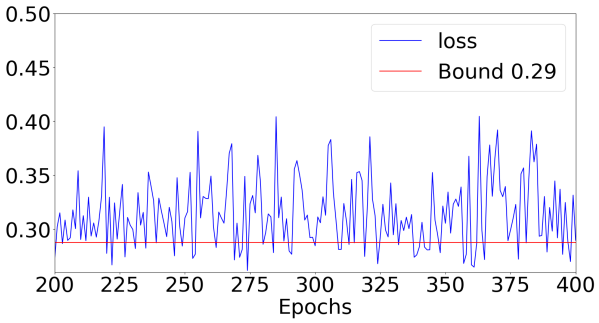


图13：使用单精度随机梯度下降和固定学习率训练的逻辑回归。

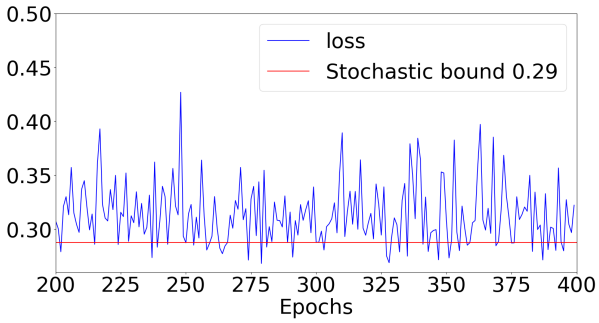
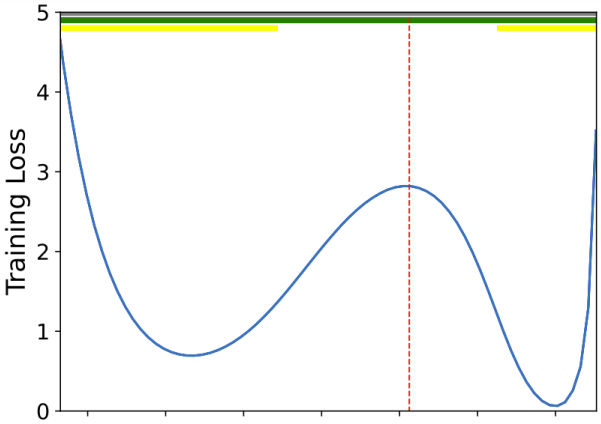


图14：使用 Bfloat 梯度（累加器大小为 50）和单精度权重更新（$\mathbf{s}_k = 0$ 和 $\mathbf{r}_k \neq 0$）训练的逻辑回归。

9 关于拟凸假设的更多动机

为进一步说明拟凸假设的合理性，我们展示了在 CIFAR10 数据集上不使用跳跃连接的 ResNet-56 损失景观投影的二维和三维视图，见图15。图中标出了凸区域和拟凸区域。拟凸区域比凸区域更大。这意味着我们的理论适用于损失函数更大的范围。



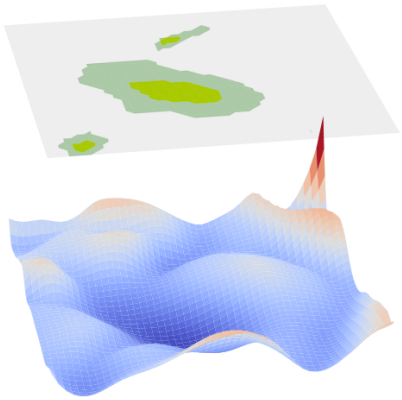


图15：拟凸区域（绿色）比凸区域（黄色）更大).