arXiv:2611.30880v1 [q-fin.CP] 23 Nov 2026

线性神经元随机梯度下降的临界步长：精确解及输入谱的作用

Sebastien Valeyre *†

摘要

我们研究教师–学生设置下、具有任意协方差高斯输入的单线性神经元的在线随机梯度下降。 N 权重的均方误差遵循一个维度为 N的封闭线性递推，我们对其进行了精确求解。我们证明，当且仅当步长低于临界值 λ_c时，动力学才会收敛；该临界值由一个涉及输入协方差全部特征值的显式标量方程给出。我们还得到了每个权重的稳态误差，该误差在 λ_c处发散。该条件与最小均方滤波器的经典均方稳定性条件一致；我们的推导简单且自洽。随后我们证明， λ_c 由输入协方差的有效秩 Ω_1 控制，而不是由参数数量1/控制；在重缩放单位下为（$\Omega_1 + 2$） $\leq \lambda_c < 1/\Omega_1$ ：仅对于各向同性输入，临界步长才与 N 成反比；对于幂律谱，临界步长衰减得更慢，或趋于饱和。最后，我们将该解推广到大小为 B的小批量：临界步长随 B 线性增长，并在 B 超过有效秩后趋于饱和。

关键词： 随机梯度下降；学习率；线性神经元；教师–学生模型；最小均方算法；均方稳定性；有效秩；小批量

1 引言

现代神经网络通过随机梯度下降（SGD）及其变体进行训练，而步长（或称学习率）的选择是最重要的实践决策之一[1, 2]。步长过小，学习就会很慢；步长过大，动力学则会变得不稳定。随着模型的参数数量 N 增长了几个数量级[3, 4], ，一个自然的问题是，最大可用步长如何依赖于 N。

本文在能够提出这一问题的最简单设定中给出精确解：一个单线性神经元接收 N 个输入，并通过在线方式学习由具有相同架构的教师神经元生成的样本。教师–学生框架在学习的统计力学中由来已久 [5–7], ，在线梯度下降曾在该框架下得到详细分析 [8–10]。在信号处理中，同一模型被称为 Widrow 和 Hoff 的最小均方（LMS）自适应滤波器 [11]。其均方收敛性已得到确立
对于高斯输入， [12–14]，如今已是教科书中的内容 [15, 16]。在机器学习文献中，最小二乘回归的 SGD 已得到广泛研究[17–20], ，而最大步长与输入协方差的迹有关。

这三个群体使用不同的语言，其中一个群体的研究结果并不总是为其他群体所知。特别是，当问题以参数数量的形式提出时，输入谱所起的作用往往会被忽视：基于梯度的 Lipschitz

*Machina Capital，21 rue de la Banque，75002 巴黎，法国，svaleyre@machinacap.com
†Valeyre Research，114 rue d'Antibes，06400 戛纳，法国。电子邮件：sv@valeyreresearch.com

常数的界仅涉及输入协方差的最大特征值 [1, 21]，这对于确定性梯度下降是合适的，但对于其随机版本则并非如此。我们的目的是用教师–学生模型的语言，给出一种简短、初等且自洽的推导，并明确展示其对 N 以及谱的依赖关系。我们的贡献如下。

1. 我们证明，权重的均方误差遵循一个维度为 N 的封闭线性系统，其矩阵由一个对角矩阵与一个秩一矩阵相加构成（第3节）。无需使用闭合近似。2. 我们证明，当且仅当步长小于临界值 λ_c 时，动力学才会收敛； λ_c 是一个显式标量方程的唯一根（定理1）。这一条件与高斯输入下 LMS 滤波器的均方稳定性条件一致[13, 14]；这里给出的证明只需几行。3. 我们以闭式形式给出各个权重的稳态误差，以及收敛率（第5节和第6节）。4. 我们证明， λ_c 在一个较窄的区间内由输入协方差的有效秩决定，而非由 N 决定（命题 2 ）。这一标度关系 $\lambda_c \propto 1/N$ 仅适用于各向同性输入；对于幂律谱， λ_c 随 N 衰减得更慢，或趋于常数（第7节）。5. 我们将解推广至批量大小为 B 的小批量，并在各向同性情形下以闭式形式得到临界步长： $\lambda_c(B) = B/(N + B + 1)$。该结果体现了实践中采用的随批量大小线性增长的关系 [22, 23] 及其饱和现象（第8节）。

前面三项使用不同的方法重新审视了经典结果；后面两项明确指出了一些据我们所知，统计物理学文献中关于学习的相关表述尚未明确提出的结论。模型的局限性将在第 9 节讨论。

2 模型

教师是一个具有 N固定权重 $\bar{p}_1$、……、 $\bar{p}_N$的线性神经元，学生则是一个权重为 $p_1(t)$……、 $p_N(t)$的线性神经元，这些权重取决于迭代次数 t。在每次迭代中，二者都接收相同的随机输入 $\varepsilon_1(t)$……、 $\varepsilon_N(t)$，而教师的输出会受到噪声 $\varepsilon_0(t)$的干扰。学生的误差为

$$R(t) = \varepsilon_0(t) + \sum_{i=1}^{N} \big(p_i(t) - \bar{p}_i\big)\, \varepsilon_i(t). \tag{1}$$

学生通过对瞬时平方误差进行梯度下降训练，步长˜为常数 $\lambda > 0$:

$$p_i(t+1) - p_i(t) = -\tilde{\lambda}\, \frac{\partial R^2(t)}{\partial p_i(t)} = -2\tilde{\lambda}\, \varepsilon_i(t)\, R(t) \qquad (i = 1, \ldots, N). \tag{2}$$

我们假设 $\varepsilon_0(t)$ $\varepsilon_1(t)$……、 $\varepsilon_N(t)$是相互独立的高斯变量，均值为零，标准差为 σ_0、 $\tilde{\sigma}_1$、……、 $\tilde{\sigma}_N$，并且在每次迭代中独立抽取（在线学习：每个样本只使用一次）。因此，权重 $p_i(t)$仅取决于时刻 $t' < t$的输入，并且与 $\varepsilon_j(t)$相互独立。不失一般性，假设这些标准差按升序排列且严格为正，

$$\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \ldots \geq \tilde{\sigma}_N > 0, \tag{3}$$

因为输入方差为零的权重不会发生变化。

注释 1. 独立输入这一假设并不构成限制。如果输入服从协方差矩阵为 C 的联合高斯分布，就可以将输入和权重旋转到 C 的特征基中：式 (2) 保持形式不变，旋转后的输入相互独立，且总权重误差保持不变。此时，$\tilde{\sigma}_i^2$ 是 C 的特征值。

通过最大方差 $\sigma^2 = \tilde{\sigma}_1^2$ 对方差和步长进行重新缩放会很方便：

$$\sigma_i = \tilde{\sigma}_i/\sigma \in (0,1], \qquad \lambda = \tilde{\lambda}\,\sigma^2, \qquad \nu = \sigma_0^2/\sigma^2. \tag{4}$$

我们通过每个权重的均方误差及其加权和来刻画学生模型的收敛，

$$u_i(t) = \langle \big(p_i(t) - \bar{p}_i\big)^2 \rangle, \qquad F_n(t) = \sum_{i=1}^{N} \sigma_i^{2n}\, u_i(t), \tag{5}$$

其中 $\langle \cdots \rangle$ 表示对输入和噪声取期望。总权重误差为 $F_0(t)$，学生模型的超额预测误差为 $\langle R^2(t)\rangle - \sigma_0^2 = \sigma^2 F_1(t)$。我们还定义

$$\Omega_n = \sum_{i=1}^{N} \sigma_i^{2n}, \qquad 1 \le \Omega_n \le N. \tag{6}$$

量 $\Omega_1 = \mathrm{Tr}\, C/\|C\|$ 是输入协方差的有效秩：对于各向同性输入，它等于 N；当少数几个方向占主导时，它可能远小于 N。

3 权重误差的精确动力学

Pro位置 1. 权重的均方误差满足封闭的线性递推 n

$$u_i(t+1) = \rho_i\, u_i(t) + 4\lambda^2\sigma_i^2\big[F_1(t) + \nu\big], \qquad \rho_i = 1 - 4\lambda\sigma_i^2 + 8\lambda^2\sigma_i^4. \tag{7}$$

证明。令 $d_i(t) = p_i(t) - \bar{p}_i$。由式（2）可得，

$$u_i(t+1) = u_i(t) - 4\tilde{\lambda}\,\langle d_i(t)\,\varepsilon_i(t)\,R(t)\rangle + 4\tilde{\lambda}^2\,\langle \varepsilon_i^2(t)\,R^2(t)\rangle. \tag{8}$$

由于 $d_j(t)$ 与 $\varepsilon_k(t)$ 及 $\langle \varepsilon_i(t)\varepsilon_j(t)\rangle = \tilde{\sigma}_i^2\delta_{i,\,j}$ 相互独立，第二项为

$$\langle d_i\,\varepsilon_i\,R\rangle = \sum_{j=1}^{N} \langle d_i d_j\rangle\,\langle \varepsilon_i\varepsilon_j\rangle = \tilde{\sigma}_i^2\,u_i(t). \tag{9}$$

对于第三项，关于 ε_0 或 ε_j 为奇次的部分均为零，因此

$$\langle \varepsilon_i^2 R^2\rangle = \sigma_0^2\tilde{\sigma}_i^2 + \sum_{j=1}^{N} \langle \varepsilon_i^2\varepsilon_j^2\rangle\, u_j(t). \tag{10}$$

对于高斯变量，$\langle \varepsilon_i^2\varepsilon_j^2\rangle = \tilde{\sigma}_i^2\tilde{\sigma}_j^2 + 2\tilde{\sigma}_i^4\delta_{i,\,j}$，由此可得

$$\langle \varepsilon_i^2 R^2\rangle = \sigma_0^2\tilde{\sigma}_i^2 + 2\tilde{\sigma}_i^4\,u_i(t) + \tilde{\sigma}_i^2\sum_{j=1}^{N} \tilde{\sigma}_j^2\,u_j(t). \tag{11}$$

结合这些结果并采用重缩放（4），即可得到式（7）。 □

式（7）的两个特点值得说明。首先，交叉相关 $\langle d_i d_j \rangle$ 与 $i \neq j$ 并未出现： N 均方误差构成了一个封闭系统。其次，权重仅通过超额预测误差 $F_1(t)$相互耦合，也就是说，只通过一个标量耦合。用矩阵形式表示，令 $\mathbf{u}(t) = (u_1(t) \dots, u_N(t))^\dagger$ 且 $\mathbf{s} = (\sigma_1^2, \dots, \sigma_N^2)^\dagger$，则有

$$\mathbf{u}(t+1) = M\,\mathbf{u}(t) + 4\lambda^2 \nu\,\mathbf{s}, \qquad M = D + 4\lambda^2\,\mathbf{s}\,\mathbf{s}^\dagger, \qquad D = \mathrm{diag}(\rho_1, \dots, \rho_N). \tag{12}$$

矩阵 M 是对称的，并且是一个对角矩阵与一个秩一矩阵之和。由于 $\rho_i = (1 - 2\lambda\sigma_i^2)^2 + 4\lambda^2\sigma_i^4 > 0$，它也是正定的，且所有元素均严格为正。当 $I - M$ 可逆时，方程 (12) 的解为

$$\mathbf{u}(t) = \mathbf{u}(\infty) + M^t\big[\mathbf{u}(0) - \mathbf{u}(\infty)\big], \qquad \mathbf{u}(\infty) = 4\lambda^2\nu\,(I - M)^{-1}\mathbf{s}. \tag{13}$$

将方程 (7) 乘以 σ_i^{2n}并对 i 求和，可得关于矩 $F_n(t)$ 的层级方程组，详见附录 A。

注 2。 平均误差满足 $\langle d_i(t+1)\rangle = (1 - 2\lambda\sigma_i^2)\langle d_i(t)\rangle$ ，并且对于任意 $\lambda < 1$均呈指数衰减。因此，均值收敛是比均方收敛弱得多的要求，后者将在下文进行研究。

4 临界步长

对于 $0 \le \lambda < 1/2$，我们定义

$$\Sigma_n(\lambda) = \sum_{i=1}^{N} \frac{\sigma_i^{2n}}{1 - 2\lambda\sigma_i^2}. \tag{14}$$

函数 $\lambda\,\Sigma_1(\lambda)$ 在 $(0,\ 1/2)$ 上连续且严格递增，在 $\lambda = 0$ 处为零，并且当 $\lambda \to 1/2$ 时发散（因为 $\sigma_1 = 1$）。因此，存在唯一的 $\lambda_c \in (0,\ 1/2)$ 使得

$$\lambda_c\,\Sigma_1(\lambda_c) = 1. \tag{15}$$

定理1。 （i）如果 $\lambda < \lambda_c$，则 M 的所有特征值都位于 （0，1） 内，并且对于任意初始条件，$\mathbf{u}$（t）指数收敛于 $\mathbf{u}$ （∞）。（ii）如果 $\lambda > \lambda_c$，则 M 的最大特征值超过 1，且总误差 F_0（t）呈指数发散，除非处于平凡情形 $\nu = 0$ 且 $\mathbf{u}$（0） $= 0$。

证明。由于 M 是对称正定的，其特征值均为实数且为正；当且仅当 $I - M = (I - D) - 4\lambda^2\mathbf{s}\mathbf{s}^\dagger$ 为正定矩阵时，它们才全部小于 1。如果对某个 i 有 $\rho_i \ge 1$，则 $I - M$ 的第 i 个对角元素为负，因而 $I - M$ 不是正定矩阵。否则， $A = I - D$ 为正定矩阵，并且当且仅当 $4\lambda^2\mathbf{s}^\dagger A^{-1}\mathbf{s} < 1$ 时， $A - 4\lambda^2\mathbf{s}\mathbf{s}^\dagger$ 为正定矩阵（事实上， $A^{-1/2}(A - 4\lambda^2\mathbf{s}\mathbf{s}^\dagger)A^{-1/2}$ 是单位矩阵减去一个秩一矩阵，其特征值为 1 和 $1 - 4\lambda^2\mathbf{s}^\dagger A^{-1}\mathbf{s}$）。现在 $1 - \rho_i = 4\lambda\sigma_i^2(1 - 2\lambda\sigma_i^2)$，因此对于所有 i， $\rho_i < 1$ 当且仅当 $\lambda < 1/2$，并且

$$4\lambda^2\,\mathbf{s}^\dagger(I - D)^{-1}\mathbf{s} = \sum_{i=1}^{N} \frac{4\lambda^2\sigma_i^4}{4\lambda\sigma_i^2(1 - 2\lambda\sigma_i^2)} = \lambda\,\Sigma_1(\lambda). \tag{16}$$

因此，当且仅当 $\lambda < 1/2$ 和 $\lambda\Sigma_1(\lambda) < 1$时， M 的特征值才全都小于 1；也就是说，当且仅当 $\lambda < \lambda_c$时成立。由式（13）可得陈述 (i)。对于 (ii)，我们首先证明，当 $\lambda > \lambda_c$时， M 的最大特征值 $\mu_{\max}$ 大于 1。若 $\lambda \ge 1/2$，则 $\rho_1 \ge 1$和$\mu_{\max} \ge M_{1,1} = \rho_1 + 4\lambda^2 > 1$。若 $\lambda_c < \lambda < 1/2$，则 $a = \mathbf{s}^\dagger(I - D)^{-1}\mathbf{s}$ 满足$4\lambda^2 a = \lambda\Sigma_1(\lambda) > 1$，而向量 $\mathbf{x} = (I - D)^{-1}\mathbf{s}$ 给出 $\mathbf{x}^\dagger(I - M)\mathbf{x} = a\,(1 - 4\lambda^2 a) < 0$，因此 $I - M$ 有一个负特征值。由于 M 的元素严格为正，与 $\mu_{\max}$ 对应的特征向量 $\mathbf{v}$ 可以取为分量严格为正（佩龙–弗罗贝尼乌斯定理）。将式（12）投影到 $\mathbf{v}$ 上，得到 $\mathbf{v}^\dagger\mathbf{u}(t+1) = \mu_{\max}\,\mathbf{v}^\dagger\mathbf{u}(t) + 4\lambda^2\nu\,\mathbf{v}^\dagger\mathbf{s}$。由于 $u_i(t) \ge 0$，只要 $\nu > 0$ 或 $\mathbf{u}(0) \neq 0$，量 $\mathbf{v}^\dagger\mathbf{u}(t)$ 就为正，并且随着 $\mu_{\max}^t$ 增长。 □

当 $\lambda = \lambda_c$ 时，最大特征值等于 1；若 $\nu > 0$，误差会随时间线性增长。换算为原始单位，临界步长为 $\tilde{\lambda}_c = \lambda_c / \tilde{\sigma}_1^2$。

推论1 （各向同性输入）。 如果所有方差都相等， $\sigma_i = 1$，那么 $\Sigma_1(\lambda) = N/(1-2\lambda)$ 并且

$$\lambda_c = \frac{1}{N+2}. \tag{17}$$

因此，在各向同性情形下，对于较大的 N，临界步长与参数数量成反比。下一个结果表明，一般而言，相关量不是 N，而是有效秩 Ω_1。

命题2 （界）。 对于任意一组方差，

$$\frac{1}{\Omega_1 + 2} \;\le\; \lambda_c \;\le\; \frac{1}{\Omega_1 + 2\sigma_N^2} \;<\; \frac{1}{\Omega_1}, \tag{18}$$

当且仅当输入是各向同性输入时，前两个不等式中的等号成立。

证明。 由于 $\sigma_N^2 \le \sigma_i^2 \le 1$，有 $(1 - 2\lambda\sigma_N^2)^{-1} \le (1 - 2\lambda\sigma_i^2)^{-1} \le (1 - 2\lambda)^{-1}$，因此 $\lambda\Omega_1/(1-2\lambda\sigma_N^2) \le \lambda\Sigma_1(\lambda) \le \lambda\Omega_1/(1-2\lambda)$。令 $\lambda = \lambda_c$ 并使用公式（15）即可得到公式（18）。□

由于 $\Omega_1 \le N$，公式（18）意味着 $\lambda_c \ge 1/(N+2)$：在所有具有给定最大方差的输入分布中，各向同性分布的约束最强。用原始单位表示，公式（18）为

$$\frac{1}{\operatorname{Tr} C + 2\|C\|} \;\le\; \tilde{\lambda}_c \;<\; \frac{1}{\operatorname{Tr} C}, \tag{19}$$

其中，$\operatorname{Tr} C = \sum_i \tilde{\sigma}_i^2$ 是输入的均方范数，$\|C\| = \tilde{\sigma}_1^2$ 是其最大方差。~这两个界的相对差异至多为 $2/\Omega_1$，因此当有效秩较大时，$\lambda_c \simeq 1/\operatorname{Tr} C$。相比之下，平均误差上的确定性梯度下降 $\langle R^2 \rangle$ 在~时对 $\lambda < 1/\|C\|$是稳定的；这一条件只涉及最大特征值 [21]。因此，随机动力学所受的约束比确定性动力学更强，约强 Ω_1倍。

与LMS滤波器的关系。 公式（2）是LMS递推 $\mathbf{w}(t+1) = \mathbf{w}(t) + \mu\, e(t)\, \mathbf{x}(t)$，其中 $\mu = 2\tilde{\lambda}$。记输入协方差的特征值为 $a_i = \tilde{\sigma}_i^2$，条件 $\lambda < \lambda_c$写作

$$\mu < \frac{1}{a_1} \qquad \text{and} \qquad \sum_{i=1}^{N} \frac{\mu\, a_i}{2 - 2\mu\, a_i} < 1, \tag{20}$$

这正是针对实高斯输入的 LMS滤波器所得到的均方稳定性条件 [13, 14], ，而公式 (19) 中的上界是自适应滤波文献中的实用规则 $\mu < 2/\operatorname{Tr} C$ [12, 15]。本文的推导从教师–学生表述中重新得到了这些结果。

注 3（非高斯输入）。 公式 (11) 中的项 $2\tilde{\sigma}_i^4$ 来源于高斯分布的四阶矩。如果输入彼此独立且具有相同的峰度 $\kappa = \langle \varepsilon_i^4 \rangle / \tilde{\sigma}_i^4$，则将其替换为 $(\kappa - 1)\tilde{\sigma}_i^4$。在各向同性情形下，可得到 $\lambda_c = 1/(N + \kappa - 1)$；当 $\kappa = 3$ 时，该结果退化为公式 (17)，而对于二值输入 ($\kappa = 1$)，则给出 $\lambda_c = 1/N$。因此，对主导行为 $1/N$ 的修正是输入分布尾部的特征。

5 稳态误差

对于 $\lambda < \lambda_c$，在公式 (7) 中令 $u_i(t+1) = u_i(t) = u_i(\infty)$，可得 $u_i(\infty) = \lambda\,[F_1(\infty)+\nu]/(1-2\lambda\sigma_i^2)$。乘以 σ_i^2 并对 i 求和，即可确定 $F_1(\infty)$，由此

$$u_i(\infty) = \frac{\nu\,\lambda}{\big(1-2\lambda\sigma_i^2\big)\big(1-\lambda\,\Sigma_1(\lambda)\big)}, \qquad F_n(\infty) = \frac{\nu\,\lambda\,\Sigma_n(\lambda)}{1-\lambda\,\Sigma_1(\lambda)}. \tag{21}$$

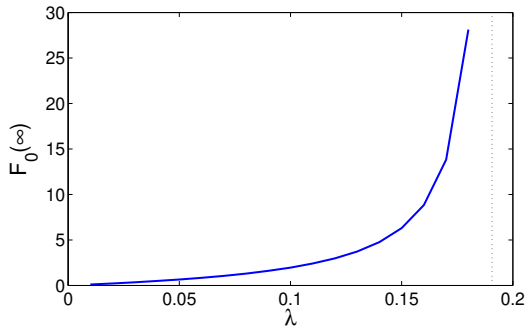
第二个关系对任意实数 n 都成立，尤其适用于总权重误差（$n=0$）和超额预测误差（$n=1$）。稳态超额预测误差与噪声方差之比，在自适应滤波文献中称为失调量 [12], ，其为

$$\frac{\sigma^2 F_1(\infty)}{\sigma_0^2} = \frac{\lambda\,\Sigma_1(\lambda)}{1-\lambda\,\Sigma_1(\lambda)}. \tag{22}$$

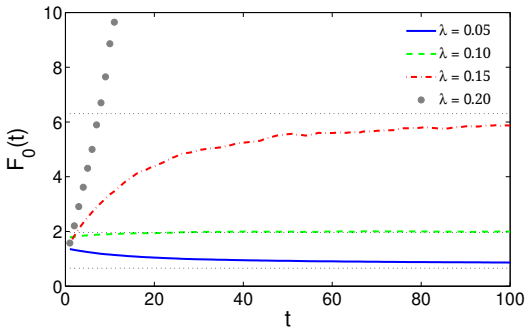
由此可得出三个结论。(a) 在无噪声时（$\nu=0$），学生能够准确地恢复教师。(b) 存在噪声时，稳态误差与噪声方差成正比；在步长较小时，还与步长成正比：$u_i(\infty) \simeq \nu\lambda$，即在原始单位下为 $u_i(\infty) \simeq \tilde{\lambda}\,\sigma_0^2$，且对所有权重都相同。(c) 当步长趋近于临界值时，所有稳态误差都会发散：$(\lambda_c-\lambda)^{-1}$；λ_c 不仅是稳定性的极限，也是学生准确性丧失的点。

利用单位矩阵恒等式 $2\lambda\,\Sigma_1(\lambda) = \Sigma_0(\lambda) - N$，总权重误差也可以写成$F_0(\infty) = 2\nu\lambda\,\Sigma_0(\lambda)/[N+2-\Sigma_0(\lambda)]$, ，而稳定性条件为 $\Sigma_0(\lambda) < N+2$。由于$\Sigma_0(\lambda)$ 随 λ 增大，$F_0(\infty)$ 是步长的增函数。在各向同性情形下，$F_0(\infty) = \nu\lambda N/[1-\lambda(N+2)]$。

图1将这些预测与在 $N=10$条件下对式（2）进行的蒙特卡洛模拟结果进行了比较，其中标准差呈线性分布，$\sigma_i = 1-(i-1)/N$，此时 $\Omega_1 = 3.85$ 和$\lambda_c \simeq 0.1906$，而 $1/(N+2) \simeq 0.083$。



(a) 稳态权重误差 $F_0(\infty)$ 随重标度步长 λ 的变化：式（21）（实线）和对式（2）的蒙特卡洛模拟结果（圆圈）。竖直虚线表示 $\lambda_c \simeq 0.1906$。



(b) 从 $u_i(0) = 1$ 开始的权重误差 $F_0(t)$，对应四种步长：式（13）（实线）和对 4×10^4 次实现取平均的模拟结果（符号）。当 $\lambda = 0.20 > \lambda_c$ 时，动力学发散。

图1：具有 $N=10$ 个输入的单线性神经元，$\sigma_i = 1-(i-1)/N$和 $\nu=1$

6 收敛率

根据式（13），在长时间尺度下，趋向稳态的过程由 M 的最大特征值 $\mu_{\max}$ 决定：$F_n(t) - F_n(\infty) \propto \mu_{\max}^t$。由于 M 是对角矩阵的秩一扰动，其不同于 ρ_i 的特征值是世俗方程 [24], 的根

并且 $\mu_{\max}$ 是大于 $\max_i \rho_i$ 的唯一解

$$4\lambda^2 \sum_{i=1}^{N} \frac{\sigma_i^4}{\mu - \rho_i} = 1. \tag{23}$$

将 $\mu = 1$ 代入左侧，可得 $\lambda\,\Sigma_1(\lambda)$，与定理1一致。在各向同性情形下，

$$\mu_{\max} = 1 - 4\lambda + 4\lambda^2(N+2), \tag{24}$$

其在 $\lambda = \lambda_c/2 = 1/(2(N+2))$ 时达到最小值，其中 $\mu_{\max} = 1 - 1/(N+2)$：最快收敛所需的迭代次数约为 N，即约需 N 个样本，并且在临界步长的一半处实现。

在一般情形下，对所有 i 都有 $\mu_{\max} > \rho_i$，意味着 $1 - \mu_{\max} < 4\lambda\sigma_N^2$；并且由于 $\lambda < \lambda_c < 1/\Omega_1$，弛豫时间 $\tau = 1/(1-\mu_{\max})$ 满足

$$\tau > \frac{\Omega_1}{4\,\sigma_N^2} = \frac{\operatorname{Tr} C}{4\,\tilde{\sigma}_N^2}. \tag{25}$$

与低方差输入相关联的权重学习得较慢，而步长无法增大以作补偿，因为它受所有方差之和的限制。较小的步长会降低稳态误差（21），但会延长弛豫时间，因此，对于有限次迭代，存在一个最优步长，如图1(b)所示，其中 $\lambda = 0.05$ 和 $\lambda = 0.10$ 的曲线相交。

7 对输入谱的依赖

命题2表明，临界步长随参数数量变化的标度完全由有效秩 Ω_1 随 N 的增长决定。为说明这一点，考虑幂律谱，

$$\sigma_i^2 = i^{-\alpha} \qquad (i = 1, \ldots, N), \tag{26}$$

它是描述自然数据 [25] 以及宽神经网络核 [26] 的常见方式。此时 $\Omega_1 = \sum_{i=1}^{N} i^{-\alpha}$，公式（18）给出了大 N 时的三种渐近情形：

$$\lambda_c \simeq \begin{cases} (1-\alpha)\,N^{-(1-\alpha)} & (0 \le \alpha < 1), \\ 1/\ln N & (\alpha = 1), \\ \text{constant} & (\alpha > 1), \end{cases} \tag{27}$$

其中，该常数介于 $1/(\zeta(\alpha)+2)$ 和 $1/\zeta(\alpha)$ 之间，ζ 为黎曼ζ函数。只有在 $\alpha = 0$ 时，才能得到标度 $\lambda_c \propto 1/N$。例如，对于 $\alpha = 2$，公式（15）的数值解表明，对所有 $N \ge 100$，均有 $\lambda_c \simeq 0.306$；而当 $\alpha = 1$ 和$N = 1000$ 时，则有 $\lambda_c \simeq 0.125$（图2(a)）。换言之，加入方差较小的输入不会减小临界步长。不过，它确实会增大弛豫时间；根据公式（25），该时间大于 $\Omega_1 N^{\alpha}/4$。

8 小批量

实际上，梯度是在由 B 独立样本组成的小批量上取平均的：

$$p_i(t+1) - p_i(t) = -\frac{2\tilde{\lambda}}{B} \sum_{b=1}^{B} \varepsilon_i^{(b)}(t)\, R^{(b)}(t). \tag{28}$$

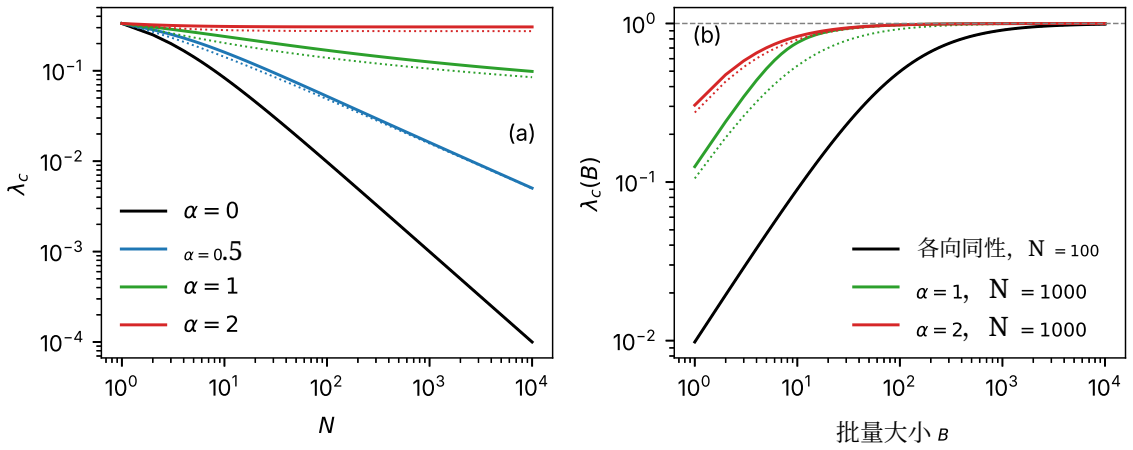


图2：（a）幂律谱 $\sigma_i^2 = i^{-\alpha}$（实线）下，临界步长 λ_c 随输入数量 N 的变化，以及公式（18）的下界 $1/(\Omega_1 + 2)$（虚线）；当 $\alpha = 0$ 时，两者重合。（b）临界步长 $\lambda_c(B)$ 随批量大小 B 的变化（实线），以及公式（31）的下界 $B/(\Omega_1 + B + 1)$（虚线）；在各向同性情形下，该下界是精确的。水平虚线表示确定性梯度下降的极限 $\lambda = 1$。

第3节的推导可以轻松调整。具有相同样本的 B 项与公式（11）中的贡献相同，而涉及两个不同样本的 $B(B-1)$ 项各自贡献 $\tilde{\sigma}_i^4\, u_i(t)$，因为对于给定的权重，$\langle \varepsilon_i^{(b)} R^{(b)} \rangle = \tilde{\sigma}_i^2\, d_i$。由此可得

$$u_i(t+1) = \rho_i^{(B)}\, u_i(t) + \frac{4\lambda^2}{B}\,\sigma_i^2 \big[F_1(t) + \nu\big], \qquad \rho_i^{(B)} = 1 - 4\lambda\sigma_i^2 + 4\Big(1 + \frac{1}{B}\Big)\lambda^2\sigma_i^4, \tag{29}$$

它的结构与公式（7）相同，并且在 $B = 1$ 时可化为公式（7）。可以验证，$\rho_i^{(B)} > 0$且 $1 - \rho_i^{(B)} = 4\lambda\sigma_i^2\,[1 - (1 + 1/B)\lambda\sigma_i^2]$，因此定理1的证明在以 $4\lambda^2/B$ 替换 $4\lambda^2$ 后仍然适用，并得出以下结果。

定理2。 当且仅当 $\lambda < \lambda_c(B)$ 时，动力学 *(28)* 在均方意义下收敛，其中$\lambda_c(B)$ 是$(0, B/(B+1))$ 中的唯一解，满足

$$\sum_{i=1}^{N} \frac{\lambda\,\sigma_i^2}{B - (B+1)\lambda\,\sigma_i^2} = 1. \tag{30}$$

它满足

$$\frac{B}{\Omega_1 + B + 1} \;\le\; \lambda_c(B) \;\le\; \frac{B}{\Omega_1 + (B+1)\sigma_N^2}, \tag{31}$$

而在各向同性情形下，

$$\lambda_c(B) = \frac{B}{N + B + 1}. \tag{32}$$

稳态误差的求法与第5节相同：$u_i(\infty) = \nu\lambda/\{[B - (B+1)\lambda\sigma_i^2]\,[1 - \lambda\Sigma_1^{(B)}(\lambda)]\}$，其中 $\lambda\Sigma_1^{(B)}(\lambda)$ 表示公式（30）左侧。

公式（31）和（32）描述了两种情况（图2(b)）。只要批量大小远小于有效秩，$B \ll \Omega_1$，临界步长就与批量大小成正比，$\lambda_c(B) \simeq B/\Omega_1$。这就是用于训练大型网络的线性缩放规则[22, 23]，，它在这里源于梯度噪声降低了 B倍这一事实。当$B \gg \Omega_1$，临界步长会饱和于 $\lambda_c = 1$，即 $\lambda_c = 1/\|C\|$，这就是稳定性极限

确定性梯度下降的稳定性极限。交叉点出现在批量大小与有效秩同一量级时，这与 [19]中对最小二乘回归小批量随机梯度下降的分析一致。

在各向同性情形下，最大特征值为 $\mu_{\max} = 1 - 4\lambda + 4\lambda^2(N + B + 1)/B$，当 $\lambda = \lambda_c(B)/2$时取最小值，其中 $1 - \mu_{\max} = B/(N + B + 1)$。每个样本的收敛率$(1 - \mu_{\max})/B = 1/(N + B + 1)$随 B递减：在该模型中，较大的批次允许采用较大的步长，但不会减少学习所需的样本数。

9 讨论与结论

我们精确求解了教师–学生设定中，采用随机梯度下降训练的线性神经元的在线学习动力学。权重的均方误差构成一个维度为 N 的封闭线性系统，其稳定性由一个步长的标量函数控制。当且仅当步长低于式（15）定义的临界值时，动力学才会收敛；在该临界值处，稳态误差发散。

核心结论涉及其对参数数量的依赖关系。对于各向同性输入，最大允许步长与 N 成反比这一说法是正确的，其中 $\lambda_c = 1/(N + 2)$。然而，一般而言，临界步长取决于输入协方差的有效秩 $\lambda_c \simeq 1/\mathrm{Tr}\, C$；当谱衰减足够快时，它便不再依赖于 N。大量弱激发方向所影响的是学习这些方向所需的时间。采用小批量时，临界步长会随批量大小线性增加，直至批量大小达到有效秩的量级。

其中若干结果以其他形式为人所知。稳定性条件和失调量与 LMS滤波器 [13–15], 中的结果一致；在对输入作出更一般假设的情况下，协方差迹和批量大小所起的作用已在最小二乘回归中得到确立 [17, 19, 20]。本文方法的优点在于简洁：高斯教师–学生模型给出了阈值、每个权重的稳态误差以及收敛率的闭式公式，从中可以直接看出它们随 N、谱和批量大小的变化规律。

该模型存在明显的局限性。单线性神经元等价于线性回归：没有隐藏层、没有特征学习，并且只有一个极小值。每个样本只使用一次，因此这些结果描述的是在线机制，而不是重复使用有限训练集的情形。对于深度非线性网络，损失曲率会在训练过程中发生变化，而据观察，梯度下降的运行状态接近稳定性边缘[27–29]；此时，学习率对宽度的依赖与网络的参数化 [30, 31]相关。这里给出的精确解可以作为研究这些问题的参考情形：在宽网络的行为相对于其参数呈线性模型的机制中，$\tilde{\sigma}_i^2$ 是相应核的特征值，而临界步长由其迹决定。

附录 A 矩层级体系

将方程（7）乘以 σ_i^{2n}并对 i 求和，得到

$$F_n(t+1) - F_n(t) = 4\lambda^2\,\nu\,\Omega_{n+1} - 4\lambda\big[F_{n+1}(t) - \lambda\,\Omega_{n+1}\,F_1(t) - 2\lambda\,F_{n+2}(t)\big], \tag{33}$$

一个无限层级体系，其中关于 F_n 的方程涉及 F_{n+1} 和 F_{n+2}。在各向同性情形下，所有 F_n 都相等，该层级体系化简为单个方程，

$$F_0(t+1) = \big[1 - 4\lambda + 4\lambda^2(N+2)\big]F_0(t) + 4\lambda^2\nu N, \tag{34}$$

由此可得公式（17）和（24）。在稳态下，公式（33）变为 $F_n(\infty) - 2\lambda F_{n+1}(\infty) = \lambda\Omega_n[\nu + F_1(\infty)],$，其解为 $F_n(\infty) = \lambda[\nu + F_1(\infty)]\sum_{k\ge 0}(2\lambda)^k\Omega_{n+k} = \lambda[\nu + F_1(\infty)]\Sigma_n(\lambda)$，并重新得到公式（21）。

然而，对于时变问题，层级体系并不是一个方便的起点。各阶矩并不相互独立：如果 σ_i 取 $K \leq N$ 个不同的值，那么 F_n 是 K 个量的线性组合，而 $F_0, \ldots, F_{K-1}$ 则决定了其他所有量（相应的矩阵属于范德蒙德类型）。因此，在任意阶数处截断层级体系并采用近似闭合关系，会导致病态矩阵或奇异矩阵。此外，传播该层级体系的无限矩阵具有与动力学无关的谱。事实上，公式（33）的齐次部分（不含 F_1 中的项）作用于形如 $F_n = z^n$ 的序列时，相当于乘以 $1 - 4\lambda z + 8\lambda^2 z^2$。物理序列是 $z = \sigma_i^2 \in (0, 1]$，的组合，其中该因子为 ρ_i；具有 $|z| = 1$ 和 $z \neq 1$ 的有界序列（例如 $z = e^{i\theta}$）也属于形式上的特征向量，而某些相应的特征值对于任意 $\lambda > 0$ 的模都大于1（例如 $1 + 4\lambda + 8\lambda^2$ 对于 $z = -1$），但它们并不对应任何一组权重误差。像第3节中那样使用 N 变量 $u_i(t)$，可以消除这些伪模态。

参考文献

[1] L. Bottou, F. E. Curtis, and J. Nocedal, 《大规模机器学习的优化方法》，SIAM Review**60**，223–311（2018年）。[2] Y. Bengio, 《深度架构基于梯度训练的实用建议》，载于*Neural Networks: Tricks of the Trade*，Springer，437–478（2012）。[3] M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. d'Ascoli, G. Biroli, C. Hongler, and M. Wyart, 《深度学习中泛化与参数数量的尺度描述》，J. Stat. Mech. 023401（2020年）。[4] C. H. Martin and M. W. Mahoney, 《神经网络模型中的传统自正则化与重尾自正则化》，arXiv:1901.08276（2019）。[5] H. S. Seung, H. Sompolinsky, and N. Tishby, 《从样例中学习的统计力学》，Phys. Rev. A **45**，6056–6091（1992）。[6] T. L. H. Watkin, A. Rau, and M. Biehl, 《规则学习的统计力学》，Rev. Mod. Phys.**65**，499–556（1993）。[7] A. Engel and C. Van den Broeck, 《统计学习力学》，剑桥大学出版社（2001）。[8] M. Biehl and H. Schwarze, 《通过在线梯度下降进行学习》，J. Phys. A **28**，643–656（1995）。[9] D. Saad and S. A. Solla, 《软委员会机中的在线学习》，Phys. Rev. E **52**，4225–4243（1995）。[10] J. Werfel, X. Xie, and H. S. Seung, 《线性前馈网络中随机梯度下降的学习曲线》，Neural Comput. **17**，2699–2718（2005）。[11]B. Widrow and M. E. Hoff, 《自适应开关电路》，IRE WESCON Convention Record，第4部分，96–104（1960）。[12] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, 《LMS自适应滤波器的平稳与非平稳学习特性》，Proc. IEEE **64**，1151–1162（1976）。

[13] L. L. Horowitz and K. D. Senne, 《复数LMS用于控制窄带自适应阵列的性能优势》, IEEE《声学、语音与信号处理汇刊》**29**，722–736（1981年）。[14] A. Feuer and E. Weinstein, 《不相关高斯数据下LMS滤波器的收敛分析》, IEEE《声学、语音与信号处理汇刊》**33**，222–230（1985年）。[15] S. Haykin, 《自适应滤波器理论》, 第4版, Prentice Hall（2002年）。[16] A. H. Sayed, 《自适应滤波基础》, Wiley（2003年）。[17] F. Bach and E. Moulines, 《非强凸光滑随机近似及其收敛率》 $O(1/n)$, Adv. Neural Inf. Process. Syst. **26** （2013年）。[18] A. Défossez and F. Bach, 《平均最小均方: 偏差-方差权衡与最优采样分布》, Proc. AISTATS（2015年）。[19] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford, 《最小二乘回归中随机梯度下降的并行化：小批量处理、平均与模型误设》, 《机器学习研究杂志》**18**，1–42（2018年）。[20] D. Zou, J. Wu, V. Braverman, Q. Gu, and S. M. Kakade, 《线性回归中恒定步长随机梯度下降的良性过拟合》, Proc. COLT（2021年）。[21] Y. LeCun, I. Kanter, and S. A. Solla, 《协方差矩阵的特征值：在神经网络学习中的应用》, 《物理评论快报》**66**，2396–2399（1991年）。[22]P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, 《准确的大批量随机梯度下降：1小时内训练ImageNet》, arXiv:1706.02677（2017年）。[23] S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, 《不要衰减学习率，而要增大批量大小》, Proc. ICLR（2018年）。[24] G. H. Golub, 《若干修正的矩阵特征值问题》, SIAM Review **15**，318–334（1973年）。[25] D. L. Ruderman and W. Bialek, 《自然图像的统计特性：尺度律之谜》, 《物理评论快报》**73**, 814–817（1994年）。[26] B. Bordelon, A. Canatar, and C. Pehlevan, 《核回归与宽神经网络中谱依赖的学习曲线》, Proc. ICML（2020年）。[27] J. M. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar, 《神经网络中的梯度下降通常发生在稳定性边缘》, Proc. I CLR（2021年）。[28] A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari, 《深度学习的大学习率阶段：弹射机制》, arXiv:2003.02218（2020年）。[29] L. Wu, C. Ma, and W. E, 《过参数化学习中随机梯度下降如何选择全局极小值：动力学稳定性视角》, Adv. Neural Inf. Process. Syst. **31** （2018年）。[30] A. Jacot, F. Gabriel, and C. Hongler, 《神经正切核：神经网络中的收敛与泛化》, Adv. Neural Inf. Process. Syst. **31** （2018年）。[31] G. Yang and E. J. Hu, 《张量程序 IV：无限宽神经网络中的特征学习》, Proc. ICML（2021年）。