迈向对神经网络的现象学理解：数据

Samuel Tovey斯图加特大学计算物理研究所，德国斯图加特 电子邮箱：
stovey@icp.uni-stuttgart.de

Sven Krippendorf
慕尼黑大学天文台和阿诺德·索末菲理论物理中心，物理学院，德国慕尼黑

电子邮箱： sven.krippendorf@physik.uni-muenchen.de

Konstantin Nikolaou
德国斯图加特，斯图加特大学计算物理研究所

Christian Holm
德国斯图加特，斯图加特大学计算物理研究所

摘要。 建立在集体变量基础上的神经网络（NN）理论，将为科学家提供工具，帮助他们更好地理解学习过程的每个阶段。在这项工作中，我们引入了两个这样的变量：熵，以及基于传入模型的训练数据构建的经验神经切线核（NTK）的迹。我们在这些变量的背景下对神经网络的性能进行了实证分析，发现初始熵、神经切线核的迹与模型训练完成后的泛化能力之间存在相关性。随后，我们将这一框架应用于神经网络训练的最优数据选择问题。为此，我们采用随机网络蒸馏（RND）来选择训练数据，并将其与数据随机选择进行比较。结果表明，RND 不仅能选择出性能优于随机选择的数据集，而且与 RND 数据集相关联的集体变量也大于随机选取的数据集所对应的集体变量。这项研究的结果为这一现象学框架指导神经网络训练数据的选择奠定了坚实基础。

关键词：神经切线核、以数据为中心的人工智能、随机网络蒸馏、神经网络的统计物理学、学习理论

arXiv:2305.00995v1 [cs.LG] 1 May 2023

1. 引言

神经网络（NN）是应对日益增多的数据驱动型挑战的有力工具。神经网络训练本质上是在一个极其庞大的参数空间中进行模型拟合的问题（某些情况下该空间甚至是无限的；Rasmussen 和 Williams（2005））。在有限宽度机制下，神经网络是强大的特征学习装置；在无限宽度机制下，则是由回归驱动的函数通用逼近器（Hornik 等（1989））。这些方法在日常技术应用（包括语音识别、定制广告和医学）以及众多科学领域中都取得了巨大成功。尽管理论方法在理解机器学习底层过程方面不断取得进展，但目前仍缺少一种简单、受物理学启发的现象学框架，用于理解神经网络训练。换言之，我们需要一种模型，能够独立于训练和部署中涉及的微观变量来描述学习过程，并且最好以经过充分研究的物理学原理为依据。这些变量包括由层数、层宽度以及神经网络所用传播算法决定的模型复杂度；用于训练模型的数据集，例如数据集的大小或其对问题空间的覆盖范围；以及用于最小化所选损失函数并训练神经网络的算法，例如优化器，甚至损失函数本身。在这项工作中，我们根据经验神经切线核（NTK）的初始状态来分析神经网络的性能（见 2.1）。在此使用 NTK 是顺理成章的，因为它包含有关训练动力学的重要信息，涉及神经网络本身及其训练所用的数据。采用这种方法，我们希望从 NTK 中找到一组可普遍计算的变量，用以分析神经网络在不同数据集和架构下的行为。已有多次观察发现，NTK 的谱是稀疏的（即主要由一个特征值及较小的特征值构成，其中绝大多数为 0），因此似乎可以将 NTK 中的信息压缩为少数几个集体变量；在此情形下，这些变量是 NTK 的迹，以及根据其特征值计算出的熵。这样的框架应有助于优化训练过程。为此，我们将确定这些变量与训练性能（例如泛化误差）的关系，并据此优化神经网络训练。具体而言，本文将该框架应用于神经网络训练中的数据选择问题。也就是说，我们考察随机网络蒸馏（RND）这一方法：它构建的数据集具有比随机选取的数据集更大的集体变量，从而带来更好的泛化效果。新的观察结果包括：NTK 的初始熵和迹之间存在相关性

一种神经网络，既具备出色的模型性能，也能让我们深入了解 RND 为何如此高效。本文所呈现的结果为未来研究指明了清晰方向：在物理启发的集体变量这一基础上，构建关于机器学习训练的现象学理论。

1.1. 相关工作

过去十年中，对神经切线核的研究呈爆炸式增长。因此，多个研究团队已经在与本文工作有一定关联的方向上取得了有希望的进展。早在1990年代，应用于神经网络的核方法就已被提出，当时研究人员首次发现了无限宽度神经网络与高斯过程之间的关系 Neal（1995）（GP）。此后，研究重点转向了神经网络的另一种核表示，即神经切线核。本文发展的理论建立在经验神经切线核之上，即针对固定数据集上的有限规模神经网络计算得到的神经切线核矩阵。关于神经切线核的研究最早出现在 Jacot 等人于2018年发表的论文中，其中神经切线核被引入作为理解神经网络在训练期间的动态，以及更好地刻画其极限的一种手段。此后，神经切线核被用作大量研究神经网络演化工作的起点。在特征值谱分析方向上，Gur-Ari 等人（2018）描述了训练期间海森矩阵特征向量的分裂，以及这一过程如何影响梯度下降。他们的研究表明，梯度会收敛到由海森矩阵的一组特征向量张成的小子空间中，而该子空间的维度由问题的复杂度决定，例如数据集中的类别数量。在2021年的论文中，Ortiz-Jim´enez 等人（2021）扩展了 Ortiz-Jimenez 等人（2020）的工作，进一步讨论了神经各向异性方向（NAD）的概念，以及如何利用它们解释训练数据为何是最优的。他们发现，无论是否经过线性化，神经网络都会以类似的方式利用神经各向异性方向对复杂度进行排序。此外，他们还借鉴了核理论的基础，特别是学习问题的复杂度受为该任务所选择的核范数约束这一观点。这可以归结为：学习模型的目标是拟合核的特征函数。将这一点应用于神经网络时，他们发现，神经网络难以学习具有较小相关特征值的特征函数。尽管这些研究旨在刻画静态形式下的神经切线核，但此前已有一些工作试图理解该核在训练期间的演化。在2022年的论文中，Krippendorf 和 Spannowsky（2022）展示了宇宙膨胀与训练过程中神经切线核迹演化之间的二重性。从数学上看，这涉及将神经网络演化重新写成神经切线核特征值的函数，这一表述在

这项工作旨在凸显集体变量的作用。上述示例都与神经网络直接相关。然而，大多数神经网络理论都以核理论为基础，因为核理论经过了深入研究，并且能够得到精确解。核回归领域早已确立，使用最大熵核可以为拟合模型提供最佳基础。Tsuda 和 Noble 在 2004年 首次描述了这类核的概念，并证明扩散核是通过最大化数据集的冯·诺依曼熵构建的。

本研究旨在将此处介绍的先前研究拓展至有限规模神经网络，并重点关注数据选择过程。本文其余部分安排如下：下一节将阐述理解集体变量所需的理论背景。随后，我们将介绍两项实验并讨论实验结果。第一项实验旨在理解集体变量在模型训练中的作用；第二项实验则尝试利用这些集体变量解释为何 RND 数据选择优于随机选择的数据集。最后，我们将展望这一框架，并讨论未来的研究工作。

2. 预备知识

本文涉及信息论、机器学习和物理学中的若干重要概念。本节将逐一介绍并解释这些概念，从而说明使用集体变量的动机。

2.1. 神经切线核

神经切线核（NTK）于2018年受到广泛关注。当时，研究人员开始发表论文，展示随机初始化、过参数化稠密神经网络的解析结果（Jacot 等人，2018年；Lee等人，2020年）。这项研究表明，在无限宽度极限下，神经网络的演化可视为线性算子，并为训练动力学提供了数学层面的洞见。此后，这一理论已得到扩展，适用于大多数神经网络架构；目前，研究人员正致力于更深入地理解这一机制下的学习机制（Yang，2019年）。给定数据集 $\mathcal{D}$，神经切线核记为 Θ，是由内积构成的格拉姆矩阵（Horn和Johnson，1990年）

$$\Theta_{ij} = \sum_k \frac{\partial}{\partial \theta_k} f(x_i, \{\theta\}) \cdot \frac{\partial}{\partial \theta_k} f(x_j, \{\theta\}) \ , \tag{1}$$

其中 Θ_{ij} 是 NTK矩阵中的一个元素， f 是一个输出维度为一的神经网络， $x_i \in \mathcal{D}$ 是一个数据点， $\{\theta\}$ 是网络的参数f。NTK矩阵中的各个元素提供了相关信息，说明一个点的表示 x_i 在参数发生变化时将如何相对于另一个点演变 x_j ，也就是说，它是神经网络在训练集数据上的表示相对于网络当前状态所形成的梯度向量之间的内积。NTK矩阵在神经网络训练中的作用，最适合通过无限宽度极限下神经网络对单个数据点的表示所满足的连续时间更新方程来描述；在梯度下降下，该方程可写为

$$\dot{f}(x_i) = -\sum_{x_j} \Theta_{i,j} \frac{\partial \mathcal{L}(f(x_i, \{\theta\}), y_i)}{\partial f_{x_j, \{\theta\}}} \ , \tag{2}$$

其中， $\mathcal{L}$ 是损失函数， y_i 是训练数据集中与输入点 x_i 对应的第i^{th} 个元素的标签（Lee 等人，2020年）。

2.2. 神经切线核谱

在无限宽度情形下，基于数据集构建的神经切线核在整个训练过程中保持不变，因此可用作逐步推进模型更新的算子（Jacot 等人，2018年）。然而，在有限网络情形下（例如科学和工业领域广泛部署的网络），情况并非如此：随着模型训练，神经切线核会不断演化。在本研究中，训练前神经切线核的状态成为一种测量工具，用于理解模型将如何泛化。因此，必须具备恰当的工具来讨论并量化这一矩阵。本研究所用工具可追溯至随机矩阵理论、信息论和物理学，其起点是熵。香农熵 S^{Sh}描述随机变量所包含的信息量（Shannon，1948年），可写为

$$S^{Sh} = -\sum_i p(x_i) \ln p_i(x_i), \tag{3}$$

其中， $p(x_i)$ 是随机变量 x_i实现的概率，ln 表示对数。在他的原始研究中，以及如今信息论中仍然常见的做法，是选择以 2 为底的对数，因为二进制数的取值范围有限。本文中，由于变量取连续值，我们使用自然对数。冯·诺伊曼熵起源于量子力学领域，当时引入密度矩阵作为研究复合系统的工具（Neumann，1927）。随机矩阵 S^{VN} 的冯·诺伊曼熵可以类似于香农熵的方式表述为

$$S^{VN} = -tr(\rho \ln \rho), \tag{4}$$

其中 ρ 是一个单位迹矩阵。不过，通常用 ρ 的特征值（记作 λ_i）来计算熵会更方便:

$$S^{VN} = -\sum_i \lambda_i \ln \lambda_i, \tag{5}$$

在这种情况下，它可以被视为香农熵在随机矩阵上的一种恰当推广。

在统计学中的协方差矩阵或量子力学中的密度矩阵的语境下，冯·诺伊曼熵可用于衡量系统各状态之间的相关性（Demarie，2018年；Tsuda 和 Noble，2004年）。

在这项工作中，NTK矩阵充当核矩阵，用于比较训练数据集各点梯度之间的相似性。Krippendorf 和 Spannowsky（2022年）的研究清楚地展现了这些梯度对模型更新的影响：他们将神经网络的连续时间演化推导为 NTK矩阵归一化特征值的函数，形式如下

$$\dot{\tilde{f}}(\mathcal{D}) = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)\mathcal{L}'(\mathcal{D}), \tag{6}$$

其中，$\tilde{f}$ 是经过基变换后的神经网络。公式6以一种特定方式描述神经网络模型更新，使冯·诺伊曼熵有望成为理解数据集质量的有用工具。具体而言，较高的冯·诺伊曼熵意味着更新步骤更加多样，因此模型或许训练得更充分。这种熵是本研究用于预测模型性能的第一个集体变量。需要指出的是，NTK矩阵不要求具有单位迹，因此在计算熵时，特征值会按其总和进行缩放。此外，公式6突显了特征值大小的作用，它们将充当缩放因子，促使模型沿指定方向采取更大的更新步长。我们将这种缩放作为基于NTK构建的第二个集体变量，具体而言，我们使用它的迹，

$$Tr(\Theta) = \sum \lambda_i \approx \lambda_{max}, \tag{7}$$

事实证明，在经验上，它可以由其最大特征值很好地近似。我们注意到，这两个变量在训练过程中的变化都反映了其偏离恒定神经切线核的程度，这一点 Krippendorf 和 Spannowsky（2022）曾进行过部分研究。

2.3. 用于数据选择的随机网络蒸馏

RND 最早出现在 Burda 等人于 2018年发表的一篇论文中，其中该方法被引入为深度强化学习问题中的一种环境探索方法。这一概念源于这样一种想法：随机初始化的神经网络所具有的随机性，能够在数据池的高维表示空间中充分区分独特的数据点。采用这种方法，RND架构似乎能够在一个数据样本中识别出独特的数据点，从而使最小数据集

可用于神经网络训练。此应用的目标与核心集方法 Feldman(2020) 的性质相似，不过它利用模型

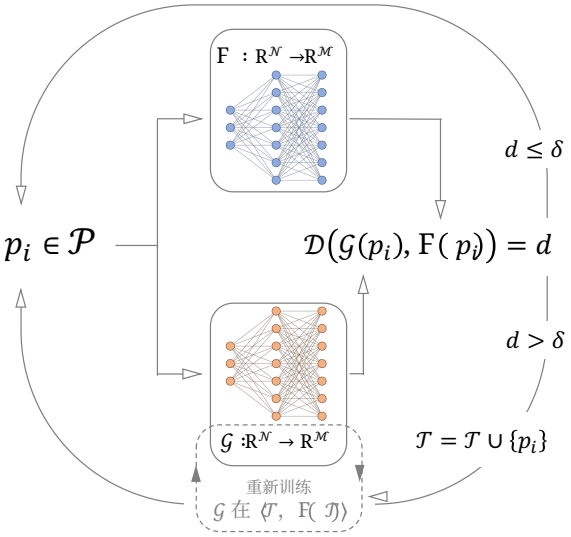

图 1. RND 的工作流程。将一个数据点 p_i 输入目标网络 $\mathcal{F}$ 和预测器网络 $\mathcal{G}$，以构建表示 $\mathcal{F}(p_i)$ 和 $\mathcal{G}(p_i)$。随后，使用度量 $\mathcal{D}(\mathcal{F}(p_i)$ $\mathcal{G}(p_i))$ 计算距离 d。如果 $d > \delta$，则将该点 p_i 添加到目标集 $\mathcal{T}$，并在完整集合 $\mathcal{T}$ 上重新训练预测器模型。如果 $d \leq \delta$，则假定 $\mathcal{T}$ 中已存在一个相似的数据点，因此将其丢弃。在我们的记号中，$\langle \mathcal{T}, \mathcal{F}(\mathcal{T}) \rangle$ 表示定义域为 $\mathcal{T}$、像为 $\mathcal{F}(\mathcal{T})$ 的函数集。

本身以无监督的方式提供有关训练数据独特性的信息。图 1 以图示方式概述了 RND 如何从数据池中筛选出点并纳入目标集。该方法通过随机初始化两个神经网络来实现，本文将它们称为目标网络 $\mathcal{F}: \mathcal{R}^{\mathcal{N}} \to \mathcal{R}^{\mathcal{M}}$ 和预测器网络 $\mathcal{G}: \mathcal{R}^{\mathcal{N}} \to \mathcal{R}^{\mathcal{M}}$；本研究中两者具有相同的架构。在数据选择过程中，目标网络将保持未训练状态，而预测器网络则会反复重新训练，以学习目标网络生成的表示。从理论上讲，这意味着预测器网络与目标网络之间的误差可以用来衡量某个点是否已被观察过。为了更好地理解这一过程，我们将其用数学形式加以表述，并逐一讨论相关步骤。考虑一个由点 p_i 构成的集合 $\mathcal{P}$，其中 $i \in \mathcal{I}$ 对 $\mathcal{P}$‡ 进行索引。现在考虑一个理论上的目标集 $\mathcal{T} \subset \mathcal{P}$，它由点 t_i 构成，并满足每个点与其中所有其他点的距离均达到最大，且误差在某个容差 δ 范围内。在每次重新训练时，网络 $\mathcal{G}$ 都使用 $\mathcal{T}$ 中的元素和目标值 $\mathcal{F}(t_i \in \mathcal{T})$ 进行训练。这样，预测器网络就能有效记住它在 $\mathcal{T}$ 中已经见过的点，从而区分出 $\mathcal{P}$ 中那些与已见点并不相似的点

‡ 索引 $\mathcal{P}$ 也就是说，对于每个 $i \in \mathcal{I}$，恰好存在一个点 $p_i \in \mathcal{P}$。

算法 1使用 RND 进行数据选择

输入： 数据池 $\mathcal{P}$，目标大小 S

当 $|\mathcal{T}| \leq S$ **时做**。

$D = \{d_i : d_i = \mathcal{D}(\mathcal{F}(p_i)\ G(p_i))\ \forall\ i \in \mathcal{P}\}$

$p_{\text{chosen}} = \{p_i : d_i \in D = \max(D)\}$

$\mathcal{T} = \mathcal{T} \cup p_{\text{chosen}}$

在 $\langle \mathcal{T}, \mathcal{F}$（$\mathcal{T}$）上重新训练 $\mathcal{G} \rangle$

结束 while 循环

$\mathcal{T}$。在使用 RND 进行数据选择时，$\mathcal{T}$ 的大小设为 S，并以贪心方式为目标集选取点：也就是说，在点云中的所有数据点上计算目标与预测器之间的距离，并选择距离最大的点。算法 1 概述了用于数据选择的 RND 算法。在本研究中，使用表示之间的均方差作为距离度量。

一般而言，RND 是一种高度复杂的数据选择方法。虽然该方法可用于构建包含数百个点的数据集，但若规模更大，就需要进行近似处理并进一步改进算法。这种优化有待后续研究，因此本文构建了规模较小的数据集，以便更好地理解它们对训练的影响。

2.4. ZnNL

本研究中使用的所有算法和工作流程均已编写成一个名为 ZnNL§ 的 Python 软件包。ZnNL 提供了一个框架，可灵活地对任意数据执行 RND，并使用本文讨论的集体变量分析所选数据。神经切线核的计算由 neural-tangents 库负责（Novak 等（2020，2022）、Hron 等（2020）、Sohl-Dickstein 等（2020）、Han 等（2022）），部分神经网络训练则由 Flax 负责（Heek 等（2020））。ZnNL 基于 Jax 框架构建（Bradbury 等（2018）），目前兼容基于 Jax 的模型。

3. 实验与结果

为了测试集体变量的有效性，我们进行了两项实验。第一项研究集体变量与模型性能之间的相关性。第二项实验展示了 RND 优于随机选取的数据集，随后使用我们的集体变量解释这一性能表现。

3.1. 所研究的数据

为确保研究全面，我们选取了涵盖分类和回归机器学习任务的若干数据集进行实验。为了进一步展示 RND 作为训练集生成工具在实际场景中的应用，我们选择了其中两个数据总体较为稀缺的问题，因此构建一个小型且具有代表性的数据集至关重要。表1介绍了所选的各个数据集，包括机器学习任务类型（分类或回归）、可用数据量以及测试集中使用的数据量。

3.2. 熵、迹与模型性能

在第一个实验中，我们考察了集体变量与模型性能之间的相关性。为此，我们使用 MNIST 和燃料数据集训练了架构固定、但初始化和训练数据各不相同的神经网络。除了更改数据集外，我们还针对 MNIST 分类测试了稠密模型架构和卷积模型架构。实验详情总结于表2。每次实验中，都会随机生成一个数据集大小，并随机初始化神经网络参数，以对熵空间和迹空间进行采样。随后，在训练过程开始时（即第一次反向传播之前）计算神经切线核的迹和熵。下文讨论的是采用标准 LeCun 初始化程序进行初始化的模型，LeCun等人（2012）。此外，我们也对采用神经切线核初始化的 Novak等人（2020）模型进行了相同研究，结果见附录 A.2。图2展示了该实验的结果，其中集体变量分别与最小测试损失以及彼此进行对照。在第一行中，颜色表示训练过程中达到的最小测试损失。其余各行中，颜色表示训练所用数据集的大小，颜色越深，数据集越小。

前两行展示了 NTK迹与矩阵初始熵之间的关系图。第一行的图根据模型达到的最小测试损失着色，第二行则展示数据集大小。对于稠密模型，我们可以看到一个损失曲面的形成，其中熵和迹都会影响训练期间取得的性能。在这些情况下，似乎需要结合熵与迹，才能在模型训练中实现最佳性能。对于卷积网络，这一趋势则不那么明显。虽然迹增大似乎有助于提升模型性能，但熵并未呈现出如此明确的趋势。

§ ZnNL 可在 https://github.com/zincware/ZnNL 找到

分析迹相对于.....的图表

表1. 列出实验所选问题的表格。对于 MNIST，在实验开始前，从总计 60000 个数据点中随机选取了 10000 个。

数据集	可用数据	测试数据	问题类型	特征	来源
MNIST	10000	500	分类	28x28x1	Lecun等人（1998）
燃油效率	398	120	回归	8	Quinlan（1993年）
步态数据	48	10	分类	328	G¨um¨u¸s¸c¨u（2019年）
混凝土数据	103	10	回归	10	Yeh（2007年）

表2。 所用参数 用于研究熵和 NTK迹与模型训练之间关系。网络架构命名 re
定义见表A1。再 隐藏层之间使用了LU激活函数，训练过程中使用ADAM优化器。

数据集名称	架构	# 样本数	最大准确率	最小测试损失
Fuel全连接网络	$\left(\mathcal{D}^{128}, \mathcal{D}^{128}, \mathcal{D}^{1}\right)$	7075	N/A	0.051
MNIST Dense	$\left(\mathcal{D}^{128}, \mathcal{D}^{128}, \mathcal{D}^{10}\right)$	5480	95.000	0.015
MNIST卷积网络	$\left(\mathcal{C}^{32}_{2\times2}, \mathcal{AP}^{4\times4}_{2\times2}, \mathcal{C}^{64}_{2\times2}, \mathcal{AP}^{4\times4}_{2\times2}, \mathcal{D}^{128}, \mathcal{D}^{10}\right)$	3082	99.000	0.008

模型训练期间最小测试损失的关系图时，可以发现不同数据集和架构之间存在一种有趣的相似性。具体而言，图形呈现出类似凸包的形状：随着初始迹增大，测试损失逐渐降低。结果表明，较大的初始神经切线核迹能够带来泛化能力更强的模型，这一点可由其较低的测试损失体现出来。值得注意的是，这一趋势在不同数据集、架构和初始化方式下均会出现。除这种简单关系外，似乎还存在一种约束效应。具体而言，迹值较低时，模型虽能实现较低的测试损失，但其取值范围通常较大；而迹值较高时，取值分布则会变窄。

将注意力转向熵图后，两者之间的关系就不那么明晰了。在稠密模型中，可以发现类似的趋势：熵较大的数据集对应较低的测试损失。不过，这一现象的作用机制有所不同。对于 MNIST，较大的熵似乎与较低的测试损失呈线性关系；而对于燃料数据集，这种关系虽也存在，但稍弱一些。两者共同之处在于，都存在迹与熵图中讨论的约束机制。似乎无论数据集大小如何，起始熵较大的数据集在训练后，其最小测试损失的取值范围都会更小。对于卷积模型，这一趋势并不存在，这表明至少对于所测试的架构而言，起始熵并非模型性能的指标。需要注意的是，起始熵和迹的取值取决于具体问题和所选架构。本研究固定了架构，因此未研究这些参数的影响，相关问题留待未来工作。

所有图中，数据集大小均呈现出一定程度的分带现象。不过值得注意的是，这些带之间存在混合：集体变量值较高的较小数据集，其测试损失与较大数据集的测试损失相近。这种混合表明，造成这些结果的是集体变量本身，而不仅仅是数据集大小。

除了图表之外，我们还计算了若干变量之间的皮尔逊相关系数，并在图3中展示了相关矩阵。为呈现集体变量与模型其他属性之间的相关性，我们在构建这些矩阵时纳入了额外指标。图表中讨论的趋势在此以数值形式呈现，与我们的结论相符。这些矩阵还展示了集体变量与训练损失之间的关系。在这些情况下，集体变量值越大，训练期间的训练损失似乎也越大。一种可能的解释是，熵值和迹值较大意味着要在数据集中拟合更多模式，因此会出现较高的训练损失和较低的测试损失。

这些结果凸显了在不同数据集大小下进行标准机器学习训练时，集体变量与模型性能之间的相关性。为了进一步探究该模型，了解固定数据集大小时熵的变化如何影响模型性能十分重要。为此，研究人员利用这些集体变量考察了数据选择方法的有效性。

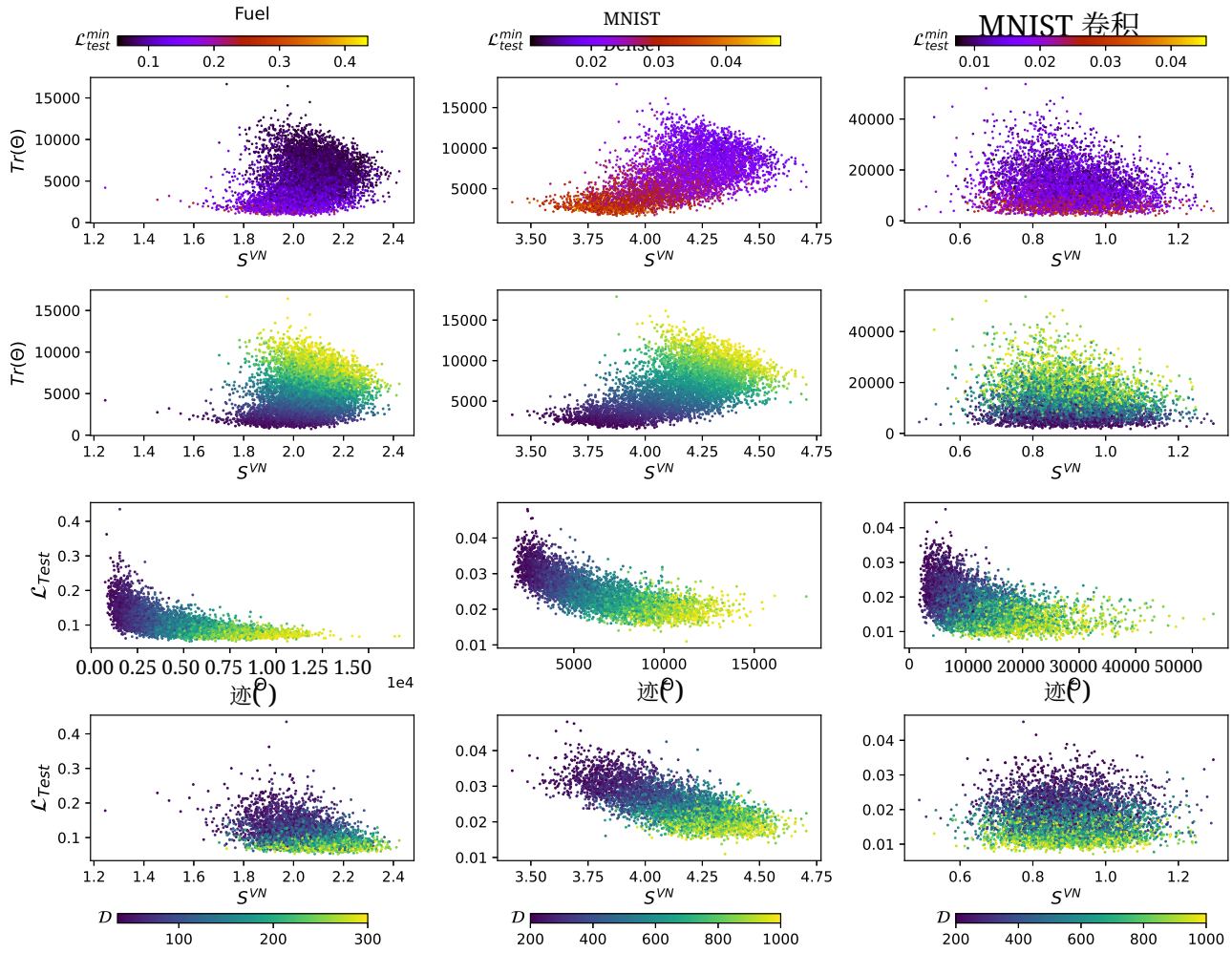

图2. 展示熵、神经切线核迹与最小测试损失之间关系的图。第一行中的颜色表示训练期间达到的最小测试损失，颜色越深表示损失越小。其余各行中的颜色表示神经网络训练所用数据集的大小，颜色越深表示数据集越小。

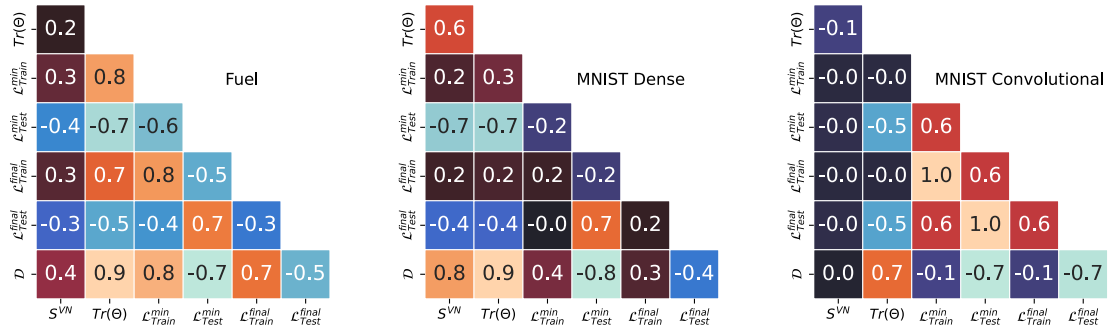

图3. 相关　　模型训练中若干变量的相关矩阵。颜色对应方框中的数字。

3.3. 随机网络蒸馏

目前的结果表明，熵、神经切线核迹与模型性能之间存在关联，

其余实验将围绕测试和解读 RND 的性能展开，以评估其作为一种数据选择方法、用于筛选训练数据的效果。

在 RND 研究过程中，所有实验均采用集成方法，即测试执行 20 次，并对结果取平均值，以构建有意义的统计学数据。通过这种方式，网络的随机初始化以及由于随机选择导致的数据集变化都得到了考虑。在所有图表中，误差条显示标准误差，即$\epsilon = \sigma/\sqrt{N}$ 其中 σ 是样本的标准差， N 是样本数量。

在实验的第一部分中，我们通过构建不同大小的数据集，并将其最小测试损失和最终测试损失与使用随机选择构建的数据集进行比较，来评估 RND 的有效性。图 4 展示了该实验的结果。首先考察最小测试损失，可以看到，使用 RND 生成的数据集始终优于使用随机数据选择构建的数据集。这适用于所有数据集大小、问题和机器学习任务，表明 RND 是构建最优数据集的合适工具。我们还比较了图 4 中的最终测试损失图，以确定模型是否存在过拟合方面的影响。比较结果表明，RND 选出的数据集不仅能带来更低的最小损失，而且似乎也较不易过拟合。

这些结果表明，RND 能够用尽可能少的点生成一个覆盖问题域的数据集，从而实现较低的最小损失。此外，RND 数据集中过拟合现象显著减少，这表明训练所用的数据覆盖了问题空间中更多样化的区域，避免了相似元素。

在实验的下一部分中，计算每种数据集大小对应的初始冯·诺伊曼熵和 NTK矩阵的迹，并考察随机选取的数据集与通过 RND 选取的数据集之间的差异。图 5 以图表形式展示了这项研究的结果，分别绘制了不同问题集的初始熵和迹随数据集大小变化的情况。这些图表清楚地表明，在每种情况下，RND 选取的数据集的初始熵和／或 NTK迹都高于随机选取的数据集。

要理解这些变量如何影响训练，必须更深入地考察NTK矩阵。NTK矩阵的元素描述了在当前参数状态下，神经网络（NN）嵌入空间中数据集各点的表示所形成的梯度向量之间的相似性。考虑由从数据集中选取的两个点构成的NTK。如果为这两个点计算出的梯度向量方向一致，那么它们的内积就会很大

这表明它们在网络进行参数更新时会以相似的方式变化。在这种情况下，NTK的熵较低，因为只需其中一个点就足以解释这种变化。如果这两个点几乎相互垂直，它们的内积就会很小，熵则会很高，因为此时NTK矩阵呈现为一个由对角线元素主导的核矩阵。这意味着在参数更新过程中，这两个点会以不同的方式对学习产生贡献。Krippendorf和Spannowsky（2022）的研究可以进一步阐释这些结论：他们将模型更新写成公式6的形式。由此可见，沿数据中某个特定特征模态的更新步长，将按其对应特征值的大小进行缩放 λ。因此，较大的特征值会使该模态上的梯度步长更大，并有望带来更好的训练效果。这一结果表明，应最大化迹，以聚焦于占主导地位的特征模态，从而更好地训练模型。然而，最大化熵相当于增加系统中占主导地位的特征模态数量，进而重新分配特征值。因此，要获得理想的模型性能，就应在系统中占主导地位的模态数量（以熵表示）与各模态的缩放因子（以迹表示）之间取得平衡。

这里已经表明，通过 RND 选取的数据集通常会产生这样的数据集：相对于某种神经网络架构，其迹和熵中的一项或两项都大于随机选取的类似数据集。虽然这些变量似乎与模型性能相关，但目前尚不清楚如何最好地厘清它们各自在模型更新中的作用，因此还需要进一步研究。此外，本文尚未探讨架构在集体变量缩放中的作用。这仍有待未来研究。

4. 结论

本研究通过考察在训练数据上计算得到的相关神经切线核矩阵的谱和熵，研究了有限宽度神经网络的性能。研究表明，初始熵和神经切线核矩阵的迹，与训练后通过测试损失衡量的模型泛化能力之间存在相关性。这些集体变量使我们能够量化不同数据选择方法对测试性能的影响。我们的结果支持了此前关于模型如何学习数据集模态的研究，具体而言，即特征值越大，训练效果越

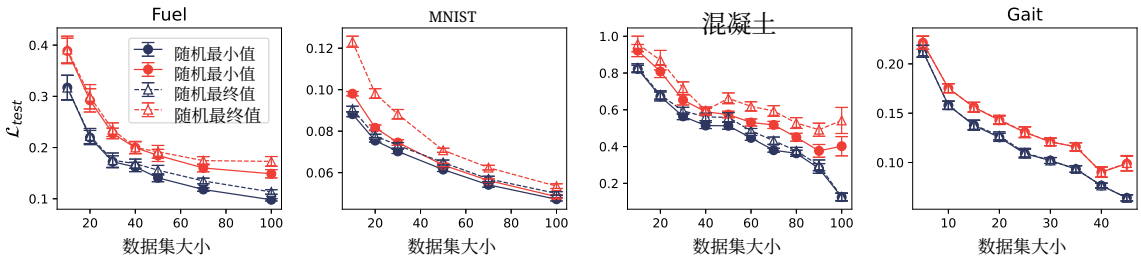

图 4. 使用 RND 和随机方法选取的数据集训练的模型，在多种数据集大小下的最小测试损失和最终测试损失。误差条的大小表示对模型进行集成操作的结果；每种数据集大小下的实验均重复进行了 20 次。

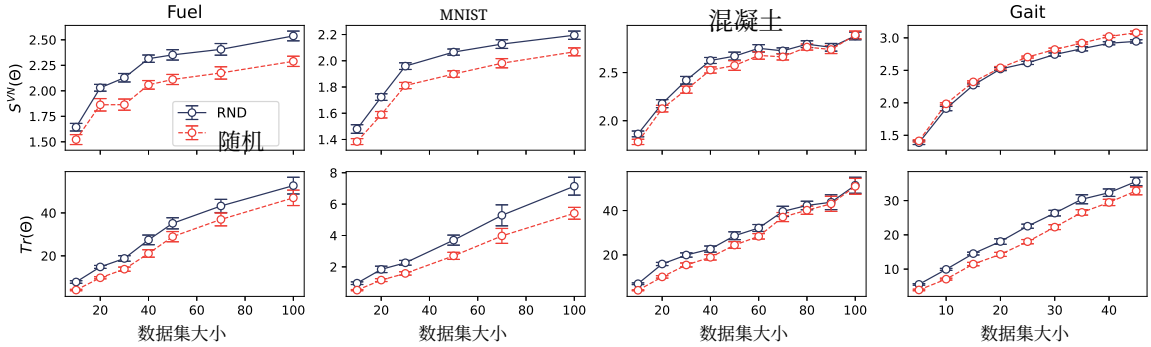

图 5. 初始冯·诺依曼熵，以及在不同大小的数据集上构建的 NTK 矩阵的迹，这些数据集由随机方法和 RND 方法生成。误差条的大小对应于正文中所述的集成。

好。该框架已用于理解 RND 作为一种数据选择方法的作用。研究表明，RND 在涵盖回归和分类任务、采用不同架构的多个数据集上均有效。为解释这一性能，研究发现，与随机选择相比，RND 选出的数据集具有更高的初始熵和/或更大的神经切线核迹。本研究是迈向以熵和神经切线核迹等集体变量构建通用现象学理论、理解机器学习训练过程的一步。未来的工作将进一步厘清熵和迹在神经网络训练其他方面所起的作用，包括架构和优化器构造，并更深入地理解它们在训练过程中的演变。随着神经网络日益复杂，基于物理启发的集体变量构建的框架，能为理解这些复杂模型的内部运作提供难得的可解释视角。本文的工作是深入理解这一框架的第一步，并有望为构建完整理论奠定基础。

致谢

C.H 和 S.T 感谢德国研究联合会（德国研究联合会—schaft DFG）在德国卓越战略 EXC 2075-390740016 项目下提供的资金支持；S.T 获得了巴登-符腾堡州的 LGF 奖学金。C.H 和 S.T 感谢德国研究联合会（德国研究联合会 DFG）在优先计划 SPP 2363 项目下提供的资金支持。

参考文献

J. Bradbury、R. Frostig、P. Hawkins、M. J. Johnson、C. Leary、D. Maclaurin、G. Necula、A. Paszke、J. VanderPlas、S. Wanderman-Milne 和 Q. Zhang。JAX: Python+NumPy 程序的可组合变换，2018年。Y. Burda、H. Edwards、A. Storkey 和 O. Klimov。通过随机网络蒸馏进行探索，2018年。T. F. Demarie。高斯态纠缠熵的教学导论。 *European Journalof Physics*，39(3):035302，2018年3月。doi: 10.1088/ 1361-6404/aaaad0。D. Feldman。核心集：最新综述。 *WIREs Data Mining and Knowledge Discovery*，10(1):e1335，2020年。doi: https://doi.org/10.1002/widm.1335。G. Gur-Ari、D. A. Roberts 和 E. Dyer。梯度下降发生在一个微小子空间中，2018年。A. G¨um¨u¸s¸c¨u。利用特征选择算法改进基于可穿戴步态分析传感器的人体分类。 *Fırat Üniversitesi Mühendislik Bilimleri*

Dergisi，31:463–471，2019年9月。doi: 10.35234/fumbd. 554789。I. Han、A. Zandieh、J. Lee、R. Novak、L. Xiao 和 A. Karbasi。适用于一般激活函数的快速神经核嵌入。载于 *Advances in Neural Information Processing Systems* 2022 ，。J. Heek、A. Levskaya、A. Oliver、M. Ritter、B. Rondepierre、A. Steiner 和 M. van Zee。Flax：用于 JAX 的神经网络库及生态系统，2020年。R. A. Horn 和 C. R. Johnson。 *Matrix Analysis*。剑桥大学出版社，1990年。ISBN 0521386322。K. Hornik、M. Stinchcombe 和 H. White。多层前馈网络是通用逼近器。*Neural Networks*，2(5):359–366，1989年。ISSN 0893-6080。doi: https://doi.org/10.1016/0893-6080(89)90020-8。J. Hron、Y. Bahri、J. Sohl-Dickstein 和 R. Novak。无限注意力：深度注意力网络的神经网络高斯过程和神经切线核。载于 *International Conference on Machine Learning*，2020年。A. Jacot、F. Gabriel 和 C. Hongler。神经切线核：神经网络中的收敛与泛化，2018年。S. Krippendorf 和 M. Spannowsky。连接神经网络与宇宙学动力学的一种对偶性。*MachineLearning: Science and Technology*，3(3):035011，2022年8月。doi: 10.1088/2632-2153/ac87e9。URL https://dx.doi.org/10.1088/2632-2153/ac87e9。Y. Lecun、L. Bottou、Y. Bengio 和 P. Haffner。应用于文档识别的基于梯度的学习。*Proceedings of the IEEE*，86(11):2278–2324，1998年。doi: 10.1109/5.726791。Y. A. LeCun、L. Bottou、G. B. Orr 和 K.-R. M¨uller。*Efficient BackProp*，第9–48页。施普林格柏林海德堡，柏林，海德堡，2012年。ISBN 978-3- 642-35289-8。doi: 10.1007/978-3-642-35289-8 3。URL https://doi.org/10.1007/978-3-642-35289-8_3。J. Lee、L. Xiao、S. S. Schoenholz、Y. Bahri、R. Novak、J. Sohl-Dickstein 和 J. Pennington。任意深度的宽神经网络在梯度下降下演化为线性模型。*Journal of Statistical Mechanics: Theory and Experiment* 2020 12 124002， 2020 ，( ): dec。doi: 10.1088/1742-5468/abc62b。R. M. Neal。神经网络的贝叶斯学习。1995年。J. v. Neumann。量子力学系综的热力学。*Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen Mathematisch- Physikalische Klasse* 1927 273 291, 1927 ，，：–。URL http: //eudml.org/doc/59231。R. Novak、L. Xiao、J. Hron、J. Lee、A. A. Alemi、J. Sohl-Dickstein 和 S. S. Schoenholz。神经切线：在 Python 中快速、简便地构建无限神经网络。载于 *International Conference on Learning Representations*，2020年。R. Novak、J. Sohl-Dickstein 和 S. S. Schoenholz。快速有限宽度神经切线核。载于 *International Conference on Machine Learning* 2022 ，。G. Ortiz-Jimenez、A. Modas、S.-M. Moosavi-Dezfooli 和 P. Frossard。《神经各向异性方向》，2020年。G. Ortiz-Jim´enez、S.-M. Moosavi-Dezfooli 和 P. Frossard。《线性化神经网络究竟能说明什么关于泛化的问题？》，2021年。J. R. Quinlan。结合基于实例的学习与基于模型的学习。载于 *Proceedings of the Tenth International Conference on International Conference on Machine Learning*，ICML'93，第 236–243 页，旧金山，加利福尼亚州，美国，1993年。Morgan Kaufmann 出版公司。ISBN 1558603077。C. E. Rasmussen 和 C. K. I. Williams。*Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* 2005 。麻省理工学院出版社，。ISBN 026218253X。C. E. Shannon。《通信的数学理论》。*The Bell System Technical Journal* 27 3 379 423, ，( ): –1948 10 1002 1538 7305 1948 01338。doi: . /j. - . .tb .x。J. Sohl-Dickstein 、 R. Novak 、 S. S. Schoenholz 和 J. Lee。关于采用标准参数化的神经网络的无限宽度极限，2020年。K. Tsuda 和 W. S. Noble。通过最大化熵从生物网络中学习核。《生物信息学》，20 (suppl 1):i326–i333，08 2004年。ISSN 1367-4803。doi: 10.1093/bioinformatics/bth906。G. Yang。具有权重共享的宽神经网络的缩放极限：高斯过程行为、梯度独立性与神经切线核推导，2019年。I.-C. Yeh。使用二阶回归和人工神经网络对混凝土坍落扩展度进行建模。*Cement and Concrete Com-posites*，29(6):474–480，2007年。ISSN 0958-9465。doi: https://doi.org/10.1016/j.cemconcomp.2007. 02.001。URL https://www.sciencedirect.com/science/article/pii/S0958946507000261。

附录 A. 附录

附录分为两个部分：第一部分讨论 RND 实验中所用神经网络的参数，第二部分展示针对神经切线核初始化网络的相关性实验。

附录 *A.1. RND*

在 RND 实验中，研究了三个涵盖回归和分类问题的数据集，以确定该方法能否优于随机选择数据。表 A1 概述了本研究使用的网络架构。

附录 *A.2.* 熵、神经切线核参数化的迹关系

为了理解初始化对所引入的集体变量的作用，还使用 NTK 初始化神经网络进行了第 3.2 节所述的研究，Novak 等人（2020年）。图 A1 和 A2 详细展示了这项研究的结果。

该图和相关矩阵与采用 LeCun初始化构造的结果有着很强的相似性，这表明至少对于这两种方案，模型的初始化对这些关系影响不大。值得注意的是迹的尺度，它的取值比 LeCun初始化中的小两个数量级。

表A1. 用于比较RND所选数据与随机选择数据的研究中的架构。 $\mathcal{D}^n$ 表示一个具有 n 个维度的全连接层， $\mathcal{C}^n_{l\times k}$ 表示一个具有 n 个输出通道、滤波器形状为 $l \times k$的卷积层。窗口形状为 $n \times m$、步幅为 $l \times k$ 的平均池化表示为 $\mathcal{AP}^{n\times m}_{l\times k}$ 。每个模型均使用ADAM优化器进行训练，并使用神经切线核初始化器进行初始化。

数据集名称	架构
Fuel	$\left(\mathcal{D}^{32}, \mathcal{D}^{32}, \mathcal{D}^{32}, \mathcal{D}^{1}\right)$
MNIST	$\left(\mathcal{C}^{32}_{3\times3}, \mathcal{AP}^{2\times2}_{2\times2}, \mathcal{C}^{32}_{2\times2}, \mathcal{AP}^{2\times2}_{2\times2}, \mathcal{D}^{128}, \mathcal{D}^{10}\right)$
Gait	$\left(\mathcal{D}^{32}, \mathcal{D}^{32}, \mathcal{D}^{16}\right)$
混凝土	$\left(\mathcal{D}^{32}, \mathcal{D}^{32}, \mathcal{D}^{32}, \mathcal{D}^{3}\right)$

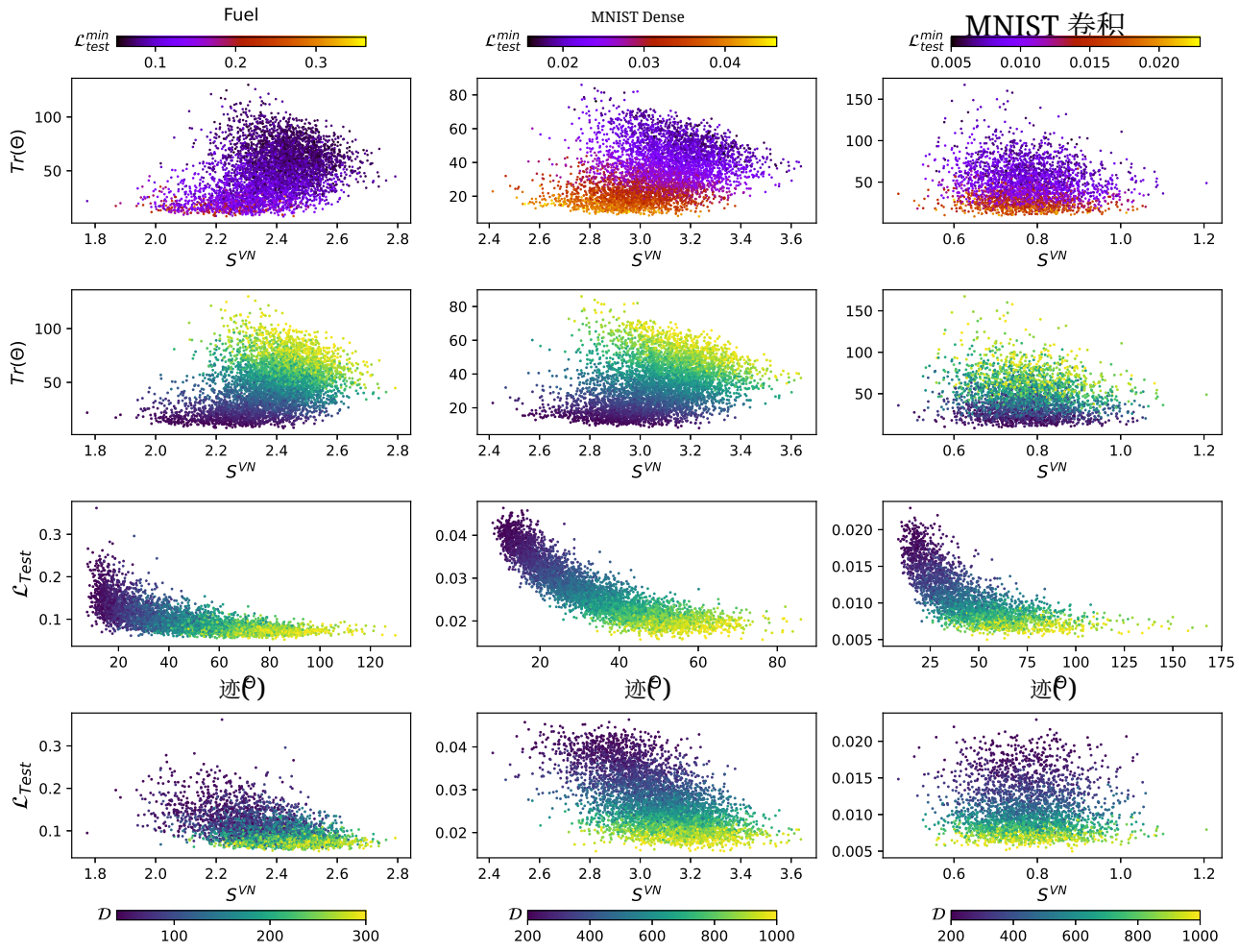

图 A1. 展示熵、NTK迹与NTK初始化神经网络的最小测试损失之间关系的图。第一行中的颜色对应训练期间的最小测试损失，其余各行中的颜色则对应数据集大小。在所有情况下，颜色越深，表示数值越小。

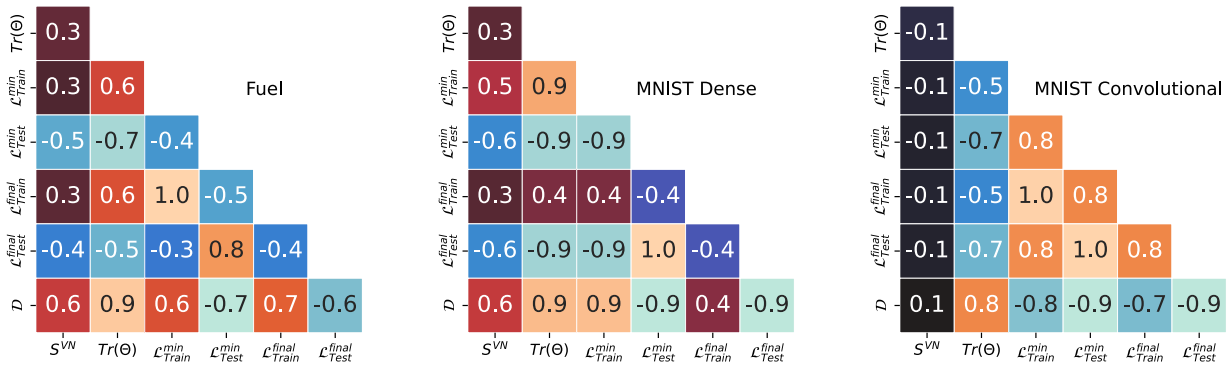

图 A2. 采用神经切线核 初始化进行模型训练时若干变量的相关矩阵。颜色对应方框中的数字。