arXiv:2601.03244v1 [stat.ML] 6 Jan 2026

从含噪与不完整数据中进行自监督学习

Juli´an Tachella Mike Davies 法国国家科学研究中心、里昂高等师范学院 爱丁堡大学

2026年1月7日

摘要

科学与工程中的许多重要问题都涉及从含噪和／或不完整观测中推断信号，且观测过程已知。传统上，人们通过手工设计正则化方法（例如稀疏性、全变差）来解决这类问题，以获得有意义的估计值。近年来，数据驱动方法通常能提供更好的解决方案：它们直接利用真实信号及其对应观测的示例来学习求解器。然而，在许多现实应用中，获取用于训练的真实参考数据代价高昂，甚至根本不可能。自监督学习方法提供了一种颇具前景的替代方案，仅从测量数据中学习求解器，从而不再需要真实参考数据。本文全面总结了用于逆问题的各种自监督方法，尤其着重探讨其理论基础，并介绍了这些方法在成像逆问题中的实际应用。

目录

1 逆问题自监督学习简介 4

1.1 逆问题.5 1.2 从解析重建到机器学习.6 1.3 监督学习.7 1.4 自监督学习.9 1.5 本文不涉及的内容.11 1.6 文章结构.12

2 学习　　从含噪测量值中学习 14

2.1 从独立含噪对中学习.14 2.2 已知噪声分布.16 2.3 部分未知噪声分布.23 2.4 从含噪数据中学习生成模型.29 2.5 迈向一般逆问题.31 2.6 总结.31

3 学习　　g 来自不完整测量值 33

3.1 利用多个算子.33 3.2 利用变换的不变性.39 3.3 从不完整测量值中学习生成模型.44 3.4 模型识别理论.45 3.5 总结.48

4 有限数据集效应 50

4.1 使用自监督损失的留出法.50 4.2 损失及其梯度的方差.50 4.3 与监督学习的差距.53 4.4 微调与测试时自适应.55

5 扩展与开放问题 56

5.1 非线性逆问题.56 5.2 面向大规模自监督成像.57 5.3 鲁棒解与部分定义模型.57 5.4 超越 ℓ_2 损失.58 5.5 不确定性量化与生成式建模.59 5.6 样本复杂度.59

5.7 选择合适的自监督方法.60

A Noisier2Noise 与 R2R 的等价性 62 B 矩的识别 63 C 补充证明 65

第1章

逆问题的自监督学习简介

科学和工程领域中的许多重要问题，归根结底都是要从含噪声和／或不完整的观测中推断信号或图像，而测量过程通常是一个先验已知的物理系统。例如，这涵盖了传感和成像逆问题中广泛的应用，从利用计算显微术学习分子结构到天文成像。在医疗保健领域，通过计算机断层扫描（CT）、磁共振成像（MRI）和超声进行的医学成像，是疾病早期诊断的重要组成部分。时间序列和音频领域的应用则包括源分离、声学断层成像和盲反卷积。

过去，这类逆问题通常通过基于模型的方法求解。 [1],深度神经网络强大的表征学习能力，使研究人员得以开发出全新的先进数据驱动重建方法。这些方法利用大量真实数据进行训练，能够挖掘复杂的统计依赖关系，而先前的手工设计模型（如稀疏表示或全变分（TV）正则化）则无法捕捉这些关系。 [2], 这显著提高了图像重建所能达到的性能，例如在加速磁共振成像图像重建中， [3],与全变分正则化相比，峰值信噪比（PSNR）提高了 6 dB。

尽管这类方法取得了非凡的成功，但它们依赖大量真实训练数据，这构成了该技术的一项关键局限，限制了其在能够轻易获取真实数据的问题中的应用——而这类问题实际上早已被预先“解决”。在重要的科学、医疗和工程场景中，以及在复杂环境下运行的传感系统中，这一问题尤为突出，因为这些场景的真实数据十分稀缺，且预测准确率至关重要。因此，人们日益关注开发新型自监督学习方法，旨在无需直接获取真实数据即可学习重建结果。

本专著旨在对近年来出现的此类自监督学习技术进行系统而完整的介绍，并突出其与支撑这些方法的统计学和几何理论之间的联系。

1.1 逆问题

所有此类方法的主要 focus 是求解一个数学逆问题，以估计或重建所关注的信号或图像。尽管这些对象在现实中通常可定义为连续函数，但为了计算解，有必要将其表示为离散形式，例如通过适当的基函数展开 [1]。为避免犯下反问题犯罪 [4]，本文将重点讨论离散信号，并将其表示或近似为有限维向量 $\boldsymbol{x} \in \mathbb{R}^n$（或 $\boldsymbol{x} \in \mathbb{C}^n$）；该向量可通过对测量值 $\boldsymbol{y} \in \mathbb{R}^m$所对应的采集过程进行稳定的反演来估计，这一过程也称为前向算子 $\boldsymbol{A}$：$\mathbb{R}^n \mapsto \mathbb{R}^m$，我们假设其中已经完成了离散化过程:

$$\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x}) + \boldsymbol{\epsilon}. \tag{1.1}$$

这里，$\boldsymbol{\epsilon}$ 表示任何噪声或建模误差，并且应假设它可能与信号相关，例如泊松分布噪声的情况 [5]。

示例

我们可以通过几个理想化示例来说明式 (1.1) 中的前向模型，并将在全文中使用这些示例:

- **去噪** 是最简单的逆问题，其中前向算子是恒等映射，也就是 $\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{x}$，目标是从观测测量值中去除噪声。

- **图像修复** 是指恢复图像中的一组缺失像素，也就是$\boldsymbol{A}(\boldsymbol{x}) = \text{diag}\,(\boldsymbol{b})\ \boldsymbol{x}$，其中掩码为 $\boldsymbol{b} \in \{0,1\}^n$。

- **超分辨率** 通常被建模为一个逆问题 [6]，其中 $\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{S}\text{circ}(\boldsymbol{k})\boldsymbol{x}$，这里 $\text{circ}(\boldsymbol{k}) \in \mathbb{R}^{n\times n}$ 表示与核 $\boldsymbol{k} \in \mathbb{R}^n$ 进行卷积，而 $\boldsymbol{S} \in \mathbb{R}^{m\times n}$ 则是一个子采样操作。

- **加速磁共振成像** 可以表示为一个线性反问题 [3]。在单线圈设置中，采集过程可以建模为 $\boldsymbol{A}(\boldsymbol{x}) = \text{diag}\,(\boldsymbol{b})\ \boldsymbol{F}\boldsymbol{x}$，其中 $\boldsymbol{F} \in \mathbb{C}^{n\times n}$ 是二维离散傅里叶变换，$\boldsymbol{b} \in \{0,1\}^n$ 是加速掩模。

- **相位恢复** 是一个非线性逆问题，可以表示为 $\boldsymbol{A}(\boldsymbol{x}) = |\boldsymbol{B}\boldsymbol{x}|^2$，其中 $\boldsymbol{B} \in \mathbb{C}^{m\times n}$ 是一个线性算子，可根据应用场景 [7]采用随机或结构化形式。

- **反散射** 是一个与亥姆霍兹方程相关的复杂非线性逆问题，可写作 [8]

$$\boldsymbol{A}(\boldsymbol{x}) = \text{circ}(\boldsymbol{g})\text{diag}\,(\boldsymbol{x})\,(\boldsymbol{I} - \text{circ}(\boldsymbol{g})\text{diag}\,(\boldsymbol{x}))^{-1}\boldsymbol{v}$$

其中 $\text{circ}(\boldsymbol{g})$ 表示与格林函数核 $\boldsymbol{g} \in \mathbb{C}^n$ 的卷积，而 $\boldsymbol{v} \in \mathbb{C}^n$ 是入射场。

1.2 从解析重建到机器学习

求解逆问题需要设计一个重建函数 $f:\ \mathbb{R}^m \mapsto \mathbb{R}^n$，以去除噪声并反演前向算子的作用，使得 $f(\boldsymbol{y}) \approx \boldsymbol{x}$能够近似恢复底层信号。在许多成像与传感场景中，前向模型 (1.1) 是线性的，早期系统经过专门设计，确保能够获取充分的测量值，从而可以通过迭代或直接反演 $f(\boldsymbol{y}) = \boldsymbol{A}^{-1}\boldsymbol{y}$ 进行重建。然而，随着传感与成像问题变得更具挑战性，人们需要更复杂的重建技术。

1.2.1 为什么难以反演?

解决任何病态逆问题的关键挑战有两个方面。首先，获取的测量值通常并非没有噪声。例如，低通量成像意味着每次测量时只能观测到有限数量的光子——通常会用泊松噪声对这种情况进行统计建模。因此，反演过程的一部分作用是能够从含噪观测中推断出干净信号。

第二个主要挑战在于无法获取一组“完整”的测量值。有时这是因为明确进行了欠采样，例如上文提到的加速磁共振成像。另一些情况下，测量的不完整程度则更为隐蔽，例如反卷积问题中，前向算子可能是满秩的，但问题严重病态；又如超分辨率的例子，其中欠采样的概念在某种程度上由用户定义。

无论哪种情况，测量值不完整都意味着测量值不足以直接求解逆问题。例如，对于线性问题，前向算子可能秩亏且具有非平凡零空间，从而产生无穷多个测量一致解；或者前向算子可能是满秩的，但严重病态，这意味着不存在通用的稳定逆算子。

从几何角度看，要解决不完整测量问题，至少需要隐式地将信号模型限制在一个低维集合上，因为稳定（即 Lipschitz）映射 $f:\mathbb{R}^m \mapsto \mathbb{R}^n$ 的像的维数至多为 m。例如，这正是压缩感知革命背后的核心思想 [9], ，它使稀疏信号模型的概念广为人知。

1.2.2 基于模型的重建

从历史上看，处理带噪声和不完整测量的逆问题通常采用统计技术，将基于模型的一致性损失与某种统计约束相结合，以捕捉感兴趣信号的期望特性。例如，这通常通过求解一个正则化变分优化问题来实现，该问题由 ℓ_2 一致性损失[1]、 $\|\boldsymbol{y}-\boldsymbol{A}(\boldsymbol{x})\|^2$以及正则化项 $\rho(\boldsymbol{x})$组成，其中正则化项包含了关于感兴趣信号集合的先验知识：

$$f(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y}-\boldsymbol{A}(\boldsymbol{x})\|^2 + \rho(\boldsymbol{x}). \tag{1.2}$$

[1]在非高斯噪声的情况下， ℓ_2 一致性通常会被负对数似然或其他稳健替代方案取代。

根据具体应用的不同，人们采用了许多不同的正则化器，其范围从经典的蒂霍诺夫正则化，到鼓励稀疏解或低秩解的正则化器，例如全变分或核范数正则化。然而，这种手工设计的正则化很少能够捕捉问题内部所有复杂的统计依赖关系，因此研究人员开始探索开发更优数据驱动解的可能性。

1.3 监督学习

从数据中学习逆问题求解器的标准（监督）方法是使用神经网络 f 作为重建函数 $\hat{\boldsymbol{x}} = f_{\boldsymbol{\theta}}(\boldsymbol{y})$，其权重 $\boldsymbol{\theta} \in \mathbb{R}^p$ 直接通过训练数据学习得到。训练数据由真实信号及其对应测量值组成：$\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$。通常通过最小化某种监督损失来实现，

$$f^* = \arg\min_f \sum_{i=1}^N \mathcal{L}_{\text{SUP}}\left(\boldsymbol{x}_i, \boldsymbol{y}_i, f\right) \tag{1.3}$$

例如 ℓ_2 损失

$$\mathcal{L}_{\text{SUP}}\left(\boldsymbol{x}, \boldsymbol{y}, f\right) = \frac{1}{n}\|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 \tag{1.4}$$

从而得到学习得到的解 f^*。这里我们省略了对权重向量 $\boldsymbol{\theta}$ 的显式依赖，转而在可容许函数空间中进行优化。

原则上，如果可容许神经网络函数类具有足够的灵活性且我们有充足的训练数据，这种方法应该能够近似最优重建函数；对于 ℓ_2 损失而言，最优重建函数就是条件均值估计器：

$$f^*(\boldsymbol{y}) \approx \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\} \tag{1.5}$$

其中 t 这里的期望[2] 是相对于后验分布 $p(\boldsymbol{x}|\boldsymbol{y})$ 取的。

这种学习方法将问题转化为回归问题，并有望使我们充分利用训练数据中的结构。在实践中，正如我们将在第 1.5 节中简要讨论的那样，神经网络架构的选择也将在所学习的逆映射的性能中发挥重要作用。

1.3.1 常用网络模型

针对逆映射 f，人们已为逆问题求解器提出了多种不同的神经网络配置。这里，我们重点介绍已考虑过的两大类解决方案，同时指出，在这个快速发展的领域中，这种介绍不可避免地并不全面。

大多数成像解决方案都采用高效的低层视觉子网络结构，该结构提供图像到图像映射，$g: \mathbb{R}^n \mapsto \mathbb{R}^n$，通常通过 ResNet [10] 或 UNet [11, 12] 风格的架构实现；较新的版本还融入了注意力机制，例如 [13]。随后，该子网络会以多种方式使用，主要区别在于采集模型 $\boldsymbol{A}$ 如何被纳入整体解决方案。总体而言，有两种主要方法。

[2]在全文中，我们将使用记号 $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\phi(\boldsymbol{x})\}$ 表示 $\phi(\boldsymbol{x})$ 在 $p(\boldsymbol{x}|\boldsymbol{y})$ 下的期望。

反投影网络 一种广受欢迎、简单而有效的解决方案是：首先使用反投影算子将测量值映射回图像/信号域。在线性情形下，可以使用线性伪逆 $\boldsymbol{A}^\dagger$ 或某种易于计算的替代方法来实现，例如 $\boldsymbol{A}^\top$。随后，使用子网络 $\phi(\boldsymbol{u})$ 将反投影信号映射为清晰的重建结果。完整的重建函数随后可表示为 $f(\boldsymbol{y}) = \phi(\boldsymbol{A}^\dagger \boldsymbol{y})$，并以端到端方式进行训练。

展开架构 这类网络旨在模仿迭代优化算法的结构，即将算法展开为少量迭代，并由图像到图像映射发挥近端型算子的作用。这类算法中，最简单的可能是采用如下形式的近端梯度下降变体：

$$\boldsymbol{x}^{(k+1)} = \phi_k\Big(\boldsymbol{x}^{(k)} - \tau \frac{\partial \boldsymbol{A}}{\partial \boldsymbol{x}}^\top (\boldsymbol{A}(\boldsymbol{x}^{(k)}) - \boldsymbol{y})\Big) \tag{1.6}$$

其中，每次迭代中子网络的权重 k可以共享，也可以独立训练。多种不同的优化算法都以这种方式展开，包括原始-对偶方法 [14]以及用于变分损失函数 [15]的梯度求解器。这类网络的表现通常优于简单的反投影网络。

在所有这些情况下，为获得最佳效果，通常会使用（1.3）以端到端方式进行训练。

1.3.2 监督学习的局限性

尽管监督学习方法似乎能够学习统计最优估计器的近似值，但这需要大量真实数据来训练模型。这使其应用范围受限于那些实际上已被解决过的问题（以便生成真实数据）。在重要的科学、医学和工程场景中，这一点尤其成问题，例如天文成像或显微成像，以及在复杂环境中运行的系统；在这些场景中，真实数据十分稀缺，而预测准确率至关重要。

机器学习社区常采用的一种解决方案，是通过模拟生成数据，以应对真实标注训练数据不足的问题。尽管这种方法能够提供近乎无限量的数据，但这些数据仅适用于生成模拟的模型；而且，即使是先进的模拟，也无法完全捕捉现实环境中存在的细微复杂性和依赖关系。

一个相关问题是分布偏移，即训练数据的分布与测试时获取的测量值之间发生变化。例如，对于一组不同但相关的信号或图像，可能存在真实数据，但这些数据无法准确代表测试时所针对的信号或图像。即使表面上已有真实数据，这类数据也往往是通过延长或重复采集生成的，例如在磁共振成像中，或通过提高照明/辐射水平生成，例如在 X射线成像或显微成像中。这可能会显著影响成像过程的性质，并导致所获取的训练数据与测试时获取的测量值之间出现分布偏移。遗憾的是，监督学习在泛化到此类分布偏移方面表现一向很差 [16]。

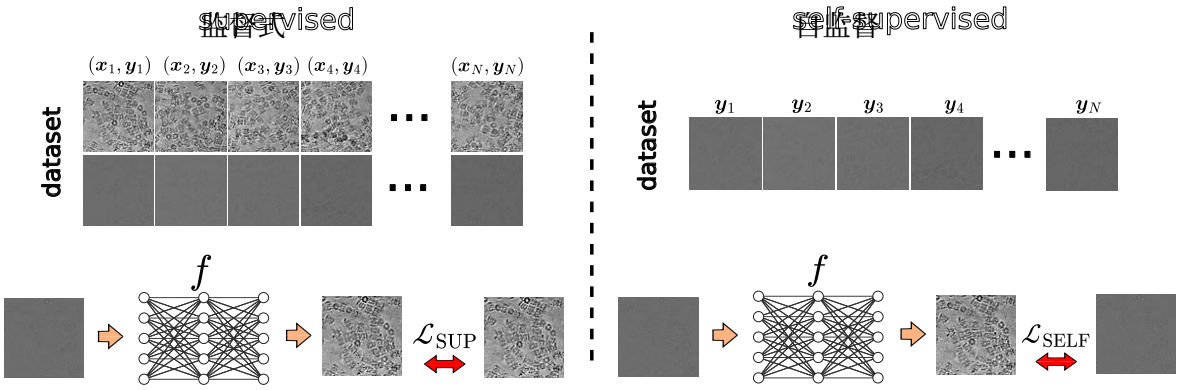

图 1.1：**监督学习与自监督学习**。监督学习需要一个由配对数据 $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}$构成的数据集，而本书的重点——自监督学习——仅依赖测量数据 $\{\boldsymbol{y}_i\}_{i=1}$，其做法是构建无需真实数据、且能近似监督损失的损失函数。

这些挑战促使研究人员寻求开发全新的自监督学习方法，这些方法仅依赖测量数据和对采集过程的了解。无论是仅使用测量值从头开始训练，还是用于微调基于模拟或相关数据训练的现有模型 [17], ，这类方法都有望在科学成像中学习不存在真实图像的结构和模式 [18]。

1.4 自监督学习

成像与传感逆问题中的自监督学习方法，其核心目标是用自监督损失 $\mathcal{L}(\boldsymbol{y}, f)$取代式 (1.4) 中所需的监督损失函数 $\mathcal{L}_{\text{SUP}}(\boldsymbol{x}, \boldsymbol{y}, f)$；如图 1.1 所示，该自监督损失仅是测量数据的函数。

总体策略是开发一种代理指标，以自监督的方式用于复现或近似监督损失。在这里，我们将看到，采集物理机制和噪声模型对于构建适当的自监督损失至关重要。相关研究日益增多，应用包括音频复原 [19], 、点云 [20] 、图像去噪 [21–24], 以及图像重建 [25–27]。

在许多应用中，我们可以预期有许多 $n \gg 1$ 样本/像素，而占主导地位的统计依赖关系往往是局部的。因此，大数定律告诉我们，即使考虑单个训练样本的损失 $\mathcal{L}(\boldsymbol{y}, f) \approx \mathbb{E}_{\boldsymbol{y}}\{\mathcal{L}(\boldsymbol{y}, f)\}$，我们也能用适量的样本获得相对稳定的期望损失估计值 N （有时甚至只用一个样本就足够）。因此，在本专著的大多数分析中，我们将假设可以获取足够大的测量值数据集 $\{\boldsymbol{y}_i\}_{i=1}^N$，从而能够用测量分布上的期望替代对数据集的求和$p_{\boldsymbol{y}}$。有限训练数据集所产生的影响将在第4章中讨论。

取决于　基于对噪声分布以及产生测量值的正向算子范围的了解，

我们可以获得不同程度的监督损失近似。

无偏损失函数我们所能期待的最佳结果，是构造一种自监督损失，使其成为监督损失的无偏估计值，即:

$$\mathbb{E}_{\boldsymbol{y}}\left\{\mathcal{L}(\boldsymbol{y}, f)\right\} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left\{\mathcal{L}_{\text{SUP}}\left(\boldsymbol{x}, \boldsymbol{y}, f\right)\right\} + \text{const.} \tag{1.7}$$

其中，常数通常是噪声方差的函数。在测量数据足够多的情况下，我们可以期望学到一个与通过监督学习所学到的重建函数同样出色的重建函数。

在某些特定情况下，我们甚至可以得到更强的结果：损失函数是单个实例 $\boldsymbol{x}$ 重建误差的无偏估计值，即:

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\mathcal{L}(\boldsymbol{y}, f)\right\} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\mathcal{L}_{\text{SUP}}\left(\boldsymbol{x}, \boldsymbol{y}, f\right)\right\} + \text{const.} \tag{1.8}$$

其中，常数是已知的。这意味着，我们还可以在测试时使用自监督损失来量化 $\boldsymbol{x}$ 的重建误差。例如，Stein无偏风险估计量及其变体就是这种情况，我们将在第2章中讨论它们。

约束损失在某些情况下，我们无法构造一种在所有可能的重建函数空间上都无偏的损失函数，但可以在一个受约束的重建函数集合 $\mathcal{F}$ 上获得无偏估计:

$$\mathbb{E}_{\boldsymbol{y}}\left\{\mathcal{L}(\boldsymbol{y}, f)\right\} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left\{\mathcal{L}_{\text{SUP}}\left(\boldsymbol{x}, \boldsymbol{y}, f\right)\right\} + \text{const. for } f \in \mathcal{F} \tag{1.9}$$

选择约束集合的依据可以是：采用旨在使问题可学习的受限函数类，例如 [21, 28]，；或者在模型中纳入额外的先验信息，例如施加等变约束 [26]。如果最优的监督重建函数不属于约束集合 $\mathcal{F}$，那么学习得到的重建函数必然不是最优的，其性能也无法达到监督学习所能实现的水平。不过，在某些情况下，我们能够量化该约束引入的偏差，从而控制性能差距。

共享全局最小值的损失函数 最后一种情况是，自监督损失并非监督损失的无偏估计值，但这两种损失函数具有相同的全局最小值，即，

$$\arg\min_{f} \mathbb{E}_{\boldsymbol{y}}\left\{\mathcal{L}(\boldsymbol{y}, f)\right\} = \arg\min_{f} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left\{\mathcal{L}_{\text{SUP}}\left(\boldsymbol{x}, \boldsymbol{y}, f\right)\right\} \tag{1.10}$$

因此，我们可以期望学习到的重建函数接近通过监督学习得到的函数，但可能无法量化相关的重建误差。

在本综述中，我们将主要关注监督损失 **(1.4)** 的代理量，因为相关理论在这一领域发展得最为完善。不过，我们也会指出理论适用范围的拓展之处，以及研究人员以启发式方式运用其他损失函数和类似技术的实例。

1.4.1 学习生成模型

除了监督损失的自监督代理量，我们还可以进一步探讨：能否仅凭测量数据学习完整信号分布 $p_{\boldsymbol{x}}(\boldsymbol{x})$ 或后验分布$p(\boldsymbol{x}|\boldsymbol{y})$ 的生成模型。如果能够学习这样的生成模型，我们就不仅可以计算条件均值 $f^*(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$，还可以计算任何其他后验统计量。我们将看到，在某些情形下这确实可行，而且通常也依赖于在一种近似 ℓ_2 监督损失的自监督损失上进行训练（例如，扩散模型需要学习条件均值估计器）。不过，值得注意的是，学习生成模型通常是一项更困难的任务；在某些情况下，即使仍可构造近似监督损失的自监督损失，也可能无法学习这样的模型 [29] （另请参见附录 B 中的一些示例）。仅凭测量数据能否识别信号分布这一问题，将在第 3.4 节讨论。

1.5 本文不涉及的内容

自监督表征学习 此处有必要区分本文所讨论的成像与传感领域中的自监督学习概念，与自监督表征学习（SSRL）技术 [30], ，例如 Sim- CLR [31], BYOL [32], DINO [33], 或掩码自编码器 [34], 。这些技术通过在一组前置任务上训练来学习强大的表征。

SSRL 方法通常需要一个由干净数据组成的数据集 $\{\boldsymbol{x}_i\}_{i=1}^N$，旨在学习强大的高级表征，以用于分类或分割等下游任务。相反，本文介绍的自监督方法仅依赖带噪声和/或不完整的数据 $\{\boldsymbol{y}_i\}_{i=1}^N$，旨在恢复与这些测量相关的底层干净图像。此外，本文中的自监督损失充当式（1.3）中黄金标准监督重建损失的代理量，而 SSRL 所用的前置任务并不旨在逼近监督分类或分割损失。

尽管存在这些差异，前置任务设计背后的一些基本原则，例如对变换或掩码的不变性，也是成像逆问题中所用自监督损失的支柱；而更深入地理解这两个领域之间的联系，仍是一个有待研究的问题。

深度图像先验与归纳偏置 虽然我们通常会重点关注期望损失的行为，但在实践中，训练数据量有限，因此学习系统的归纳偏置也会对实际观察到的性能发挥重要作用。

神经网络系统中的归纳偏置有多种来源，从模型架构和权重初始化的选择，到在损失函数中加入正则化项（例如权重衰减），乃至优化过程本身（例如使用 Adam 而非随机梯度下降，或采用提前停止）。

例如，Ulyanov 等人 [35]表明，各种卷积神经网络架构都可以利用单组测量值（即待恢复图像的测量值）进行训练，以解决逆问题。他们将此称为深度图像先验（DIP）。这促使

许多研究人员 [36–38] 尝试利用这一概念进行无监督图像重建。然而，人们对这种归纳偏置的性质了解甚少 [39], ；而且，正如原始 DIP 论文所示，其性能高度依赖于特定的网络架构，通常远不及本文所述自监督方法的性能，例如，参见 [40]中的实际比较。因此，尽管 DIP 确实是一种引人入胜的现象，我们在此不再进一步讨论。

预训练扩散模型与即插即用模型 去噪扩散模型 [41] 和即插即用（PnP）解决方案 [42]已成为解决逆问题的常用方法。此类方法通常依赖于预训练的 去噪神经网络，其中去噪器通过得分函数和 Tweedie 公式来定义隐式信号先验。虽然这些解决方案通常被称为无监督方法，但这并不完全准确，因为创建预训练去噪器需要访问真实数据。尽管如此，近期已有研究尝试以自监督方式学习去噪器 [43, 44], ，我们也将在本书中对此进行介绍。

1.6 概要

第 1.6.1 节介绍了本综述中通用的记号。下面将概述本综述其余部分的内容。本专著中讨论的大多数自监督方法都通过 DeepInverse 开源库实现 [45], ，该库包含各种 Jupyter 笔记本示例，涵盖了本书中的许多主题。

第2章聚焦于使用前向算子进行去噪的自监督学习问题， $\boldsymbol{A} = \boldsymbol{I}$。我们考察了各种自监督损失函数；在一系列不同的噪声模型下，这些损失函数可作为监督损失的代理量，涵盖从假定已知各种常见噪声分布（高斯分布、泊松分布等）到部分指定的噪声模型。全文始终强调新提出的自监督学习策略与经典统计学理论结果之间的联系。最后，我们讨论更一般但可逆的前向算子情形。

第3章 继续讨论在测量值不完整，即前向算子不可逆时可以采取的措施。这里，我们重点关注存在非平凡零空间的线性前向算子情形。我们介绍解决这一问题的两种方法。第一种方法依赖于可获取一组多个前向算子 $\{\boldsymbol{A}_g\}_{g=1}^{G}$，例如在加速磁共振成像中能够选择不同的采样模式 [27],，其中不同的算子通常具有不同的零空间。第二种方法应对的是更具挑战性的单个秩亏前向算子问题， $\boldsymbol{A}$，转而利用这样一个假设：感兴趣的信号分布对于某个变换群具有不变性，例如，图像的平移版本仍然是一个有效图像。

第4章 前两章重点探讨了期望损失函数在实现自监督学习方案中的作用，本章则研究在只能获得有限数量的训练样本时，如何准确地近似这些期望。我们将考察简单的 Noise2Noise 算法，以探究自监督学习的样本复杂度

与监督学习情形之间的关系。我们还将说明，如何把多数监督学习中用于避免过拟合的标准留出法扩展到自监督学习场景，以及如何利用预训练模型减少获得良好性能所需的测量样本数量。

第5章 阐述了该领域的一些开放问题以及未来可能的研究方向。

1.6.1 符号说明

遵循标准数学记法，向量用粗体小写字母表示，矩阵用粗体大写字母表示。向量$\boldsymbol{x}$ 的第 i个分量记作 x_i。单位矩阵记作 $\boldsymbol{I}$，矩阵 $\boldsymbol{A}$的转置记作 $\boldsymbol{A}^\top$ ，其伪逆记作 $\boldsymbol{A}^\dagger$。本文中经常使用的其他记法见下表。

符号	描述
$\mathbb{R}$	实数集。
$\mathbb{C}$	复数集。
$p_{\boldsymbol{x}}(\boldsymbol{x})$	信号分布。
$p_{\boldsymbol{y}}(\boldsymbol{y})$	测量分布。
$\mathcal{X}$	信号分布的支撑集。
$\mathcal{Y}$	测量分布的支撑集。
$\boldsymbol{x}$	表示真实信号或图像的向量。
$\boldsymbol{y}$	表示观测测量值的向量。
$\boldsymbol{b}$	表示掩码的二进制向量。
n	信号向量的维度， $\boldsymbol{x} \in \mathbb{R}^n$。
m	测量向量的维度， $\boldsymbol{y} \in \mathbb{R}^m$。
k	信号集的维度， $\mathcal{X}$。
N	训练样本的数量。
$\boldsymbol{A}$	前向算子， $\boldsymbol{A}: \mathbb{R}^n \to \mathbb{R}^m$ 其中 $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x})$。
f	将 $\boldsymbol{y}$ 映射为 $\boldsymbol{x}$的估计值的重建网络。
$\boldsymbol{\theta}$	神经网络的权重， f。
$\|\cdot\|$	ℓ_2 范数。
$\|\cdot\|_F$	矩阵的 Frobenius 范数。
$\mathbb{E}_{\boldsymbol{u}}\{g(\boldsymbol{u})\}$	$g(\boldsymbol{u})$在分布 $p(\boldsymbol{u})$下的期望。
$\mathbb{E}_{\boldsymbol{u}\|\boldsymbol{v}}\{g(\boldsymbol{u}, \boldsymbol{v})\}$	$g(\boldsymbol{u}, \boldsymbol{v})$在分布 $p(\boldsymbol{u}\|\boldsymbol{v})$下的期望。
$\mathbb{V}_{\boldsymbol{u}\|\boldsymbol{v}}\{\boldsymbol{u}\}$	$\boldsymbol{u}$ 在分布 $p(\boldsymbol{u}\|\boldsymbol{v})$下的方差。
$\mathcal{L}_{\text{SUP}}(\boldsymbol{x}, \boldsymbol{y}, f)$	监督损失。
$\mathcal{L}_{\text{X}}(\boldsymbol{y}, f)$	与技术 X 相关的自监督损失。
∇	标量场的梯度。
常数。	未进一步量化的常数项。
$\mathcal{N}(\boldsymbol{x}, \boldsymbol{\Sigma})$	均值为 $\boldsymbol{x}$ 、协方差为 $\boldsymbol{\Sigma}$ 的多元高斯分布。
$\mathcal{P}(\boldsymbol{x})$	率为 $\boldsymbol{x}$的泊松分布。
$\text{Ber}(\boldsymbol{x})$	概率为 $\boldsymbol{x} \in [0,1]^n$ 的伯努利分布。

第2章

从含噪测量中学习

我们首先关注去噪问题，其中前向算子只是恒等映射 $\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{x}$，因此图像和测量值都属于同一空间。我们介绍了多种自监督损失，这些损失只需要测量数据，旨在使其期望值近似于监督损失。我们表明，损失的设计取决于对噪声分布的了解程度：如果我们完全了解噪声分布，通常就能够构造监督损失的无偏估计器；而当噪声分布并非完全已知时，我们仍然可以构造自监督损失，但其性能不及监督学习。

本章分为三部分：第一部分假设，对于每幅图像 $\boldsymbol{x}$，我们都观测到两个相互独立的含噪实现（$\boldsymbol{y}_1$，$\boldsymbol{y}_2$）。第二部分将放宽这一假设，仅依赖每幅图像的单个含噪实现 $\boldsymbol{y}$，但假设我们完全了解噪声分布。第三部分则讨论这样的情况：每幅图像只有一个含噪实现 $\boldsymbol{y}$，且噪声分布部分未知。这里介绍的大多数结果都与重建网络 f的架构或参数化方式无关。

2.1 从独立噪声对中学习

在某些应用中，可以观测到同一底层信号 $\boldsymbol{x}$的两个（或更多）相互独立的含噪实现（$\boldsymbol{y}_1$，$\boldsymbol{y}_2$）。即使不了解噪声分布 [46]，也可以利用这些实现以自监督的方式学习一个估计器。Noise2Noise [40] 提出了这样一种方法[1]：将其中一个含噪测量值作为重建网络的输入，另一个作为目标，并构建如下损失函数：

$$\mathcal{L}_{\mathrm{N2N}}\left(\boldsymbol{y}_1, \boldsymbol{y}_2, f\right) = \frac{1}{n}\|f(\boldsymbol{y}_1) - \boldsymbol{y}_2\|^2. \tag{Noise2Noise}$$

由于输入噪声与输出噪声相互独立，我们可以证明，(Noise2Noise) 是监督损失的无偏估计量，二者仅相差一个常数：

[1]利用同一潜在参数的独立观测进行模型选择这一想法，最早可追溯到 Mallows 在 20 世纪 70 年代的研究 [46]。Noise2Noise [40] 在计算机视觉领域重新发现了这一想法。

命题 2.1。 设 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 是给定 $\boldsymbol{x}$时条件独立的两个随机变量，并假设 $\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\boldsymbol{y}_2\} = \boldsymbol{x}$，则

$$\mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{x}}\left\{\frac{1}{n}\|f(\boldsymbol{y}_1) - \boldsymbol{y}_2\|^2\right\} = \mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{\frac{1}{n}\|f(\boldsymbol{y}_1) - \boldsymbol{x}\|^2\right\} + const. \tag{2.1}$$

其中，该常数与 f 无关。

证明。

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{x}}\left\{\|f(\boldsymbol{y}_1) - \boldsymbol{y}_2\|^2\right\} \\
&= \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{x}}\left\{\|(f(\boldsymbol{y}_1) - \boldsymbol{x}) - (\boldsymbol{y}_2 - \boldsymbol{x})\|^2\right\} \\
&= \mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{\|f(\boldsymbol{y}_1) - \boldsymbol{x}\|^2\right\} - 2\,\mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{x}}\left\{(f(\boldsymbol{y}_1) - \boldsymbol{x})^\top(\boldsymbol{y}_2 - \boldsymbol{x})\right\} + \text{const.} \\
&= \mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{\|f(\boldsymbol{y}_1) - \boldsymbol{x}\|^2\right\} - 2\left(\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\{f(\boldsymbol{y}_1) - \boldsymbol{x}\}\right)^\top\left(\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\boldsymbol{y}_2 - \boldsymbol{x}\}\right) + \text{const.} \\
&= \mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{\|f(\boldsymbol{y}_1) - \boldsymbol{x}\|^2\right\} + \text{const.}
\end{aligned}$$

其中，第四行利用了 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 条件独立这一事实，最后一行则依赖于 $\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\boldsymbol{y}_2 - \boldsymbol{x}\} = \boldsymbol{0}$。 □

这一结果可以扩展到超出 ℓ_2 范数 [47]，的任意 Bregman 散度，但对某些其他常用的损失函数（例如 ℓ_1 范数）并不成立。直观地说，估计器 f 无法对 $\boldsymbol{y}_2$ 中的噪声进行过拟合，因为它观测到的是一个独立的噪声实现 $\boldsymbol{y}_1$。命题 2.1 对噪声仅要求极少的假设（即目标具有零均值噪声，即 $\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\boldsymbol{y}_2\} = \boldsymbol{x}$），因此对于现实世界中的问题非常有吸引力，因为在这些问题中，噪声分布通常未知，但有可能获得同一对象的两个独立观测。我们通过一些成像示例对此进行了说明：

- 在冷冻电子显微镜中，我们会观测到同一底层对象的一系列噪声非常大的图像（显微照片）。Bepler等人 [48] 表明，使用 Noise2Noise 方法可以大幅提升信噪比。

- 在合成孔径雷达（SAR）中，我们观测到的复数图像服从圆对称复正态分布，其实部和虚部包含相互独立的噪声。Dalsasso等人 [49]表明，可以将实部作为输入、虚部作为目标来训练去噪器。

- 在视频去噪中，连续帧之间的相似性几乎满足 Noise2Noise 准则。Ehret等人 [50] 表明，可以结合光流估计对一帧进行变形，使其与另一帧对齐，并使用 Noise2Noise 方法对预训练的视频去噪网络进行微调。 e

在许多应用中，观察到两个独立测量值的假设并不成立。不过，我们将在下一节中看到（或许令人惊讶！），如果已知噪声分布，我们通常可以从单个测量值 $\boldsymbol{y}$ 中获得两个独立的噪声实现（$\boldsymbol{y}_1$，$\boldsymbol{y}_2$），而无需知道底层图像 $\boldsymbol{x}$，并使用这些独立样本对应用相同的 Noise2Noise 损失。

2.2 已知噪声分布

在许多应用中，噪声分布是近似已知的，或者可以利用一些校准数据对其进行近似。有两种主要方法可以构建融合这些知识的自监督损失：第一种方法由 Noisier2Noise [51] 和 Recorrupted2Recorrupted [52]，首创，他们证明可以向观测值 $\boldsymbol{y}$ 添加合成噪声，从而生成两个独立的噪声实现（$\boldsymbol{y}_1$，$\boldsymbol{y}_2$）。第二种方法基于统计学中的一个经典结果，即 Stein无偏风险估计（SURE）[53]，；该方法通过惩罚网络 f 的散度来避免对噪声过拟合。在这两种情况下，我们都需要准确知道噪声分布，才能正确近似监督情形。我们还将看到，尽管乍看之下差异很大，Recorrupted2Recorrupted 与 SURE 实际上密切相关。

2.2.1 自举含噪测量

虽然 Noisier2Noise 框架 [51] 提出了自举含噪测量的原始方法，但我们将采用等价的[2] Recorrupted2Recorrupted [52] 方法，因为这为进一步的泛化奠定了基础。

假设采用高斯噪声模型，$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}$ 其中 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$，或者等价地，$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}, \boldsymbol{\Sigma})$，我们可以从原始测量值$\boldsymbol{y}$中重新采样出两个相互独立的含噪实现，表示为

$$\begin{cases} \boldsymbol{y}_1 = \boldsymbol{y} + \sqrt{\frac{\alpha}{1-\alpha}}\boldsymbol{\omega} \\ \boldsymbol{y}_2 = \boldsymbol{y} - \sqrt{\frac{1-\alpha}{\alpha}}\,\boldsymbol{\omega} \end{cases} \tag{2.2}$$

其中，$\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$服从与噪声 $\boldsymbol{\epsilon}$ 相同的分布，而 $\alpha \in (0,1)$是一个正标量参数。

命题 2.2 (Pang 等人 [52])。由(2.2) 定义的随机变量 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 在给定 $\boldsymbol{x}$ 的条件下，对于任意 $\alpha \in (0, 1)$都是相互独立的。

证明。设 $\tau = \sqrt{\frac{\alpha}{1-\alpha}}$。由于给定 $\boldsymbol{x}$ 时，$\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 服从高斯分布，因此只需证明它们不存在线性相关性，即可证明它们相互独立：

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{x}}\left\{(\boldsymbol{y}_1 - \boldsymbol{x})(\boldsymbol{y}_2 - \boldsymbol{x})^\top\right\} \\ &= \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\omega}}\left\{(\boldsymbol{\epsilon} + \tau\boldsymbol{\omega})(\boldsymbol{\epsilon} - \frac{1}{\tau}\boldsymbol{\omega})^\top\right\} \\ &= \mathbb{E}_{\boldsymbol{\epsilon}}\left\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right\} - \frac{1}{\tau}\mathbb{E}_{\boldsymbol{\epsilon}}\left\{\boldsymbol{\epsilon}\right\}\mathbb{E}_{\boldsymbol{\omega}}\left\{\boldsymbol{\omega}^\top\right\} + \tau\mathbb{E}_{\boldsymbol{\omega}}\left\{\boldsymbol{\omega}\right\}\mathbb{E}_{\boldsymbol{\epsilon}}\left\{\boldsymbol{\epsilon}^\top\right\} - \mathbb{E}_{\boldsymbol{\omega}}\left\{\boldsymbol{\omega}\boldsymbol{\omega}^\top\right\} \\ &= \mathbb{E}_{\boldsymbol{\epsilon}}\left\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right\} - \mathbb{E}_{\boldsymbol{\omega}}\left\{\boldsymbol{\omega}\boldsymbol{\omega}^\top\right\} \\ &= \mathbf{0} \end{aligned}$$

最后一行依赖于这样的假设：添加的噪声 $\boldsymbol{\omega}$ 与测量噪声具有相同的协方差，从而实现独立性。

□

[2] Noisier2Noise [51] 提出了在 Recor- rupted2Recorrupted [52], 之前向输入添加噪声的想法，但后者提出了一种简化的损失，证明模拟得到的成对样本 $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ 具有条件独立性。有关这两种方法之间密切联系的更多详情，请参阅附录 A。

遵循 与 Noise2Noise 方法类似，我们可以定义 Recorrupted2Recorrupted 损失 a s

$$\mathcal{L}_{\mathrm{R2R}}(\boldsymbol{y}, f) = \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{y}} \left\{ \frac{1}{n} \|f(\boldsymbol{y}_1) - \boldsymbol{y}_2\|^2 \right\} \tag{R2R}$$

由于 $(\boldsymbol{y}_1,\ \boldsymbol{y}_2)$ 与 $\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\boldsymbol{y}_2\} = \boldsymbol{x}$ 条件独立，因此它是监督式 ℓ_2 损失的无偏估计，其中网络的输入处含有 $\boldsymbol{y}_1$：

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \{\mathcal{L}_{\mathrm{R2R}}(\boldsymbol{y}, f)\} = \mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}} \left\{ \frac{1}{n} \|f(\boldsymbol{y}_1) - \boldsymbol{x}\|^2 \right\} + \mathrm{const.} \tag{2.3}$$

注意，(R2R) 是一种理想化的损失，因为它涉及对重新采样结果的期望。不过，实际应用中，我们可以在每个梯度步使用一组重新采样的样本对 $(\boldsymbol{y}_1,\ \boldsymbol{y}_2)$，由此得到的损失随机梯度估计仍然是无偏的。

$\boldsymbol{y}_1$ 与 $\boldsymbol{y}_2$ 的独立性也有代价：由于额外的合成噪声，网络的输入 $\boldsymbol{y}_1$的信噪比（SNR）低于原始测量值 $\boldsymbol{y}$。参数 $\alpha \in (0,\ 1)$用于权衡注入 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$中的额外噪声量。将随机变量 $\boldsymbol{z}$ 的信噪比定义为 $\mathrm{SNR}(\boldsymbol{z}) :=_{\mathbb{E}_{\boldsymbol{z}}\{\|\boldsymbol{z}-\mathbb{E}_{\boldsymbol{z}}\{\boldsymbol{z}\}\|^2\}}^{\|\mathbb{E}_{\boldsymbol{z}}\{\boldsymbol{z}\}\|^2}$，则新变量的信噪比为

$$\begin{aligned} \mathrm{SNR}(\boldsymbol{y}_1) &= (1-\alpha)\,\mathrm{SNR}(\boldsymbol{y}), \\ \mathrm{SNR}(\boldsymbol{y}_2) &= \alpha\,\mathrm{SNR}(\boldsymbol{y}). \end{aligned} \tag{2.4}$$

因此，一个好的策略是选择较小的 α，使输入 $\boldsymbol{y}_1$ 具有尽可能高的信噪比，但也不能太小，以免目标 $\boldsymbol{y}_2$ 的信噪比过低，从而导致损失估计值的噪声更大。在实践中，对于固定的 $\boldsymbol{y}$，我们只对成对样本（$\boldsymbol{y}_1$、$\boldsymbol{y}_2$）进行有限次重采样。实践中通常较好的选择落在区间 $\alpha \in (0.05,0.2)$内。有关 α最优选择的更多信息，请参见 Oliveira 等人 [54]的分析。

在测试时，我们可以通过对 $J > 1$ 次合成噪声添加结果取平均来改进估计，即

$$f^{\mathrm{test}}(\boldsymbol{y}) = \frac{1}{J} \sum_{j=1}^{J} f(\boldsymbol{y}_1^{(j)}) \tag{2.5}$$

其中 $\boldsymbol{y}_1^{(j)} \sim \mathcal{N}\left(\boldsymbol{y}, \frac{\alpha}{1-\alpha}\boldsymbol{\Sigma}\right)$ 适用于 $j = 1$、.....、J。

这种损失可以扩展到非高斯加性噪声：如果已知噪声分布的某些一阶矩，我们仍然可以使用（2.2），通过添加 $\boldsymbol{\omega}$ 与这些矩 [55]相匹配的合成噪声来生成成对样本（$\boldsymbol{y}_1$、$\boldsymbol{y}_2$）。

加性噪声之外的扩展 在许多应用中，影响测量值的噪声并非加性的。例如，光子计数设备（如单光子激光雷达）会产生泊松噪声 [56], ，而伽马分布常用于对合成孔径雷达图像中的散斑噪声进行建模 [49]。

高斯分布、泊松分布和伽马分布属于自然指数族（NEF） [57], ，可以写成

$$p(\boldsymbol{y}|\boldsymbol{x}) = h(\boldsymbol{y}) \exp(\boldsymbol{y}^\top \eta(\boldsymbol{x}) - \phi(\boldsymbol{x})), \tag{NEF}$$

对于某些 h、η 和 ϕ 函数（这些函数分别针对每种分布而定；请注意，这构成了关于自然参数 $\eta(\boldsymbol{x})$ 的 NEF，而不是关于图像 $\boldsymbol{x}$ 本身的 NEF ）。我们可以使用以下定理将 (R2R) 损失推广到 NEF:

定理 2.3 (Monroy 等人 [55])。令 $p(\boldsymbol{y}|\boldsymbol{x})$ 属于具有 $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\{\boldsymbol{y}\} = \boldsymbol{x}$ 和$\alpha \in (0,1)$ 的自然指数族（*NEF*）。我们可以对 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 进行采样，使得

$$\begin{cases} \boldsymbol{y}_1 \sim p(\boldsymbol{y}_1|\boldsymbol{y}, \alpha), \\ \boldsymbol{y}_2 = \frac{1}{\alpha}\boldsymbol{y} - \frac{(1-\alpha)}{\alpha}\boldsymbol{y}_1, \end{cases} \tag{2.6}$$

在给定 $\boldsymbol{x}$ 的条件下，$\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 是相互独立的随机变量，其均值分别为 $\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\{\boldsymbol{y}_1\} =\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\boldsymbol{y}_2\} = \boldsymbol{x}$，方差分别为 $\mathbb{V}_{\boldsymbol{y}_1|\boldsymbol{x}}\{\boldsymbol{y}_1\} = \frac{1}{1-\alpha}\mathbb{V}_{\boldsymbol{y}|\boldsymbol{x}}\{\boldsymbol{y}\}$和 $\mathbb{V}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\boldsymbol{y}_2\} = \frac{1}{\alpha}\mathbb{V}_{\boldsymbol{y}|\boldsymbol{x}}\{\boldsymbol{y}\}$，且它们的分布 $p_1(\boldsymbol{y}_1|\boldsymbol{x})$ 和 $p_2(\boldsymbol{y}_2|\boldsymbol{x})$ 也属于自然指数族（*NEF*）。

因此，广义 R2R 损失定义为

$$\mathcal{L}_{\text{GR2R}}(\boldsymbol{y}, f) = \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{y}}\left\{\frac{1}{n}\|f(\boldsymbol{y}_1) - \boldsymbol{y}_2\|^2\right\} \tag{GR2R}$$

其中 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 通过式 (2.6) 生成。由于给定 $\boldsymbol{x}$ 时合成对相互独立，我们可以再次使用命题 2.1 来证明

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\{\mathcal{L}_{\text{GR2R}}(\boldsymbol{y}, f)\} = \mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{\|f(\boldsymbol{y}_1) - \boldsymbol{x}\|^2\right\} + \text{const.}$$

[3] 高斯、泊松分布和伽马噪声分布的 $p(\boldsymbol{y}_1|\boldsymbol{y},\ \alpha)$定义收录于表 2.1 中。与高斯噪声的情况一样，$\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 的信噪比由公式 (2.4) 给出，并且应选择 α 使其大致落在 (0.05，0.2) 区间 [55]内。

近年来，在其他统计推断应用中，人们也探索了生成独立数据的合成对这一想法；该方法被称为数据分裂 [58] ，并且包含超出 NEF 的扩展。

超越 ℓ_2 损失我们还可以将关于分布 $p(\boldsymbol{y}_2|\boldsymbol{x})$的知识纳入损失度量的选择中。$\ell_2$ 度量非常适用于高斯损失，因为它与高斯模型下的负对数似然成正比。因此，如果噪声模型不是高斯模型，而是属于 NEF，我们可以使用相应噪声模型的负对数似然:

$$\begin{aligned} \mathcal{L}_{\text{GR2R-NLL}}(\boldsymbol{y}, f) &= \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{y}}\left\{\frac{1}{n}\phi\left(f(\boldsymbol{y}_1)\right) - \frac{1}{n}\boldsymbol{y}_2^\top\eta\left(f(\boldsymbol{y}_1)\right)\right\} \\ &= \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{y}}\left\{-\frac{1}{n}\log p_2\Big(\boldsymbol{y}_2|\hat{\boldsymbol{x}} = f(\boldsymbol{y}_1)\Big)\right\} + \text{const.} \end{aligned}$$

[3]一般而言，我们有 $p(\boldsymbol{y}_1|\boldsymbol{y},\ \alpha) = h_1(\boldsymbol{y}_1)h_2(\boldsymbol{y}-\boldsymbol{y}_1)/h(\boldsymbol{y})$，其中

$$h_1(\boldsymbol{y}_1) = \int e^{-\boldsymbol{s}^\top\boldsymbol{y}_1 + (1-\alpha)\phi(\eta^{-1}(\frac{\boldsymbol{s}}{1-\alpha}))}d\boldsymbol{s}$$
$$h_2(\boldsymbol{y}_2) = \int e^{-\boldsymbol{s}^\top\boldsymbol{y}_2 + \alpha\phi(\eta^{-1}(\frac{\boldsymbol{s}}{\alpha}))}d\boldsymbol{s}.$$

模型	高斯分布 $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x},\ \boldsymbol{\Sigma})$	泊松分布 $\boldsymbol{z} \sim \mathcal{P}(\boldsymbol{x}/\gamma)\ \boldsymbol{y} = \gamma \boldsymbol{z}$	伽马分布 $\boldsymbol{y} \sim \mathcal{G}(\ell,\ \ell/\boldsymbol{x})$
$\boldsymbol{y}_1$	$\boldsymbol{y}_1 = \boldsymbol{y} + \sqrt{\frac{\alpha}{1-\alpha}}\boldsymbol{\omega}$, $\boldsymbol{\omega} \sim \mathcal{N}(0,\ \boldsymbol{\Sigma})$	$\boldsymbol{y}_1 = \frac{\boldsymbol{y}-\gamma\boldsymbol{\omega}}{1-\alpha}$, $\boldsymbol{\omega} \sim$ 箱$(\boldsymbol{z},\ \alpha)$	$\boldsymbol{y}_1 = \boldsymbol{y} \circ (\mathbf{1} - \boldsymbol{\omega})/(1-\alpha)$ $\boldsymbol{\omega} \sim$ 贝塔$(\ell\alpha,\ \ell(1-\alpha))$
$\mathcal{L}_{\text{GR2R-NLL}}(\boldsymbol{y},\ f)$	$\|\sqrt{\boldsymbol{\Sigma}^{-1}}(f(\boldsymbol{y}_1) - \boldsymbol{y}_2)\|^2$	$-\gamma \boldsymbol{y}_2^\top \log f(\boldsymbol{y}_1)\ + \mathbf{1}^\top f(\boldsymbol{y}_1)$	对数 $f(\boldsymbol{y}_1)\ + \boldsymbol{y}_2 / f(\boldsymbol{y}_1)$
$\mathcal{L}_{\text{SURE}}(\boldsymbol{y},\ f)\ =$ $\lim_{\alpha\to 0} \mathcal{L}_{\text{GR2R}}(\boldsymbol{y},\ \alpha,\ f)$	$\|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 +$ 2 迹 $\left(\boldsymbol{\Sigma}\frac{\partial f}{\partial \boldsymbol{y}}(\boldsymbol{y})\right)$	$\|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 +$ $2\sum_{i=1}^n y_i(f_i(\boldsymbol{y})\ - f_i(\boldsymbol{y} - \gamma \boldsymbol{e}_i))$	$\|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 +$ $2\sum_{i=1}^n \sum_{k\geq 1} b(\ell,k)(-y_i)^{k+1} \frac{\partial^k f_i}{\partial y_i^k}(\boldsymbol{y})$

表 2.1：**广义 R2R 损失汇总**。 属于自然指数族的常见噪声分布及其对应的自举函数，其中 $\alpha \in (0,\ 1)$ 和负对数似然损失。

由于给定 $\boldsymbol{x}$ 时 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 相互独立，因此该损失是监督式负对数似然损失的无偏估计量：

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\mathcal{L}_{\text{GR2R-NLL}}(\boldsymbol{y}, f)\right\} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{-\frac{1}{n}\log p_2\Big(\boldsymbol{x}|\hat{\boldsymbol{x}} = f(\boldsymbol{y}_1)\Big)\right\} + \text{const.}$$

这些损失（包括 ℓ_2）对应于 Bregman散度，并且在对数据集取期望时具有相同的全局最小值，即后验均值 $f(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$ [55]。然而，在处理有限数据集时，使用 ℓ_2 或负对数似然会得到不同的网络 f。（各向异性）高斯分布、伽马分布和泊松分布噪声所对应的负对数似然损失列于表 2.1。

2.2.2 Stein无偏风险估计

现在我们来讨论一种不基于生成两个独立含噪对的自监督损失。让我们再次回到高斯噪声模型 $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x},\ \boldsymbol{\Sigma})$。Stein [53] 的以下开创性结果表明，无需知道真实值，我们就可以估计预测与噪声之间的相关性：

引理 2.4 (Stein [53])。 设 $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}, \boldsymbol{\Sigma})$ 和 f 均为弱可微。那么，我们有

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{f(\boldsymbol{y})^\top(\boldsymbol{y} - \boldsymbol{x})\right\} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{trace\left(\frac{\partial f}{\partial \boldsymbol{y}}\boldsymbol{\Sigma}\right)\right\} \tag{2.7}$$

其中，$\frac{\partial f}{\partial \boldsymbol{y}} \in \mathbb{R}^{n\times n}$ 是 f 在 $\boldsymbol{y}$处的雅可比矩阵。

证明依赖于分部积分，我们将其留作读者的练习。当噪声是各向同性的，即 $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ 的迹为 $\left(\frac{\partial f}{\partial \boldsymbol{y}}\right) = \sum_{i=1}^n \frac{\partial f_i}{\partial y_i}$时，这表明噪声与预测之间的相关性仅与估计器的散度 f成正比。

B基于这一引理，Stein无偏风险估计（SURE）可表示为 ：

$$\mathcal{L}_{\text{SURE}}(\boldsymbol{y}, f) = \frac{1}{n}\|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 + \frac{2}{n}\operatorname{trace}\left(\frac{\partial f}{\partial \boldsymbol{y}}\boldsymbol{\Sigma}\right) - \frac{1}{n}\operatorname{trace}(\boldsymbol{\Sigma}) \tag{SURE}$$

并且是监督损失的无偏估计值，表示为

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\frac{1}{n}\|f(\boldsymbol{y})-\boldsymbol{y}\|^2+\frac{2}{n}\operatorname{trace}\left(\frac{\partial f}{\partial \boldsymbol{y}}\boldsymbol{\Sigma}\right)-\frac{1}{n}\operatorname{trace}\left(\boldsymbol{\Sigma}\right)\right\}\\
&=\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\frac{1}{n}\|\left(f(\boldsymbol{y})-\boldsymbol{x}\right)-\left(\boldsymbol{y}-\boldsymbol{x}\right)\|^2+\frac{2}{n}\operatorname{trace}\left(\frac{\partial f}{\partial \boldsymbol{y}}\boldsymbol{\Sigma}\right)-\frac{1}{n}\operatorname{trace}\left(\boldsymbol{\Sigma}\right)\right\}\\
&=\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\frac{1}{n}\|f(\boldsymbol{y})-\boldsymbol{x}\|^2-\frac{2}{n}f(\boldsymbol{y})^\top(\boldsymbol{y}-\boldsymbol{x})+\frac{2}{n}\operatorname{trace}\left(\frac{\partial f}{\partial \boldsymbol{y}}\boldsymbol{\Sigma}\right)\right\}\\
&=\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\frac{1}{n}\|f(\boldsymbol{y})-\boldsymbol{x}\|^2\right\}
\end{aligned}$$

其中，第三行用到了斯坦因引理和 $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\|\boldsymbol{y}-\boldsymbol{x}\|^2\right\}=$ 迹 $(\boldsymbol{\Sigma})$。早在神经网络出现之前，SURE 就已在信号处理领域广受欢迎，参见例如 [59]。据我们所知，Metzler 等人 [60] 率先使用 SURE 训练深度网络，此后 SURE 一直被广泛用于学习，例如 [61, 62]。

高斯噪声之外的扩展 SURE损失已扩展到高斯噪声之外。在受泊松噪声污染的测量值情况下，增益为 $\gamma>0$ 的 $\boldsymbol{y}\sim\gamma\mathcal{P}(\frac{\boldsymbol{x}}{\gamma})$，我们可以使用以下由 Hudson [63] 提出的斯坦因引理扩展形式：

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{(\boldsymbol{y}-\boldsymbol{x})^\top f(\boldsymbol{y})\right\}=\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\sum_{i=1}^{n}y_i\left(f_i(\boldsymbol{y})-f_i(\boldsymbol{y}-\gamma\boldsymbol{e}_i)\right)\right\}\tag{2.8}$$

其中，$\boldsymbol{e}_i$ 是一个标准基向量，其第 i 个分量为 1，其余分量为 0。与（SURE）类似，该引理可用于推导以下自监督损失 [5]：

$$\mathcal{L}_{\mathrm{PURE}}(\boldsymbol{y},f)=\frac{1}{n}\|f(\boldsymbol{y})-\boldsymbol{y}\|^2+\frac{2}{n}\sum_{i=1}^{n}y_i\Big(f_i(\boldsymbol{y})-f_i(\boldsymbol{y}+\gamma\boldsymbol{e}_i)\Big)\tag{PURE}$$

当测量值受到泊松噪声污染时，该估计器是监督损失的无偏估计量。Le Montagner 等人 [64] 结合 Stein (2.7) 和 Hudson (2.8) 恒等式，以处理混合泊松-高斯噪声，即增益为 $\boldsymbol{\epsilon}\sim\mathcal{N}$ 的 $\boldsymbol{y}\sim\gamma\mathcal{P}(\frac{\boldsymbol{x}}{\gamma})+\boldsymbol{\epsilon}$ （$\mathbf{0},\boldsymbol{\Sigma}$），这种噪声出现在各种成像应用中，例如荧光显微镜 [65] 或低剂量计算机断层扫描 [66]。

还可以使用以下引理，将 SURE 的其他类似扩展应用于属于自然指数族的连续变量：

引理 2.5 (Hudson [63])。设 $\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x})$ 是一个分布属于（自然指数族）的连续随机变量，并且设 f 为弱可微函数。则

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\left(\nabla\log h(\boldsymbol{y})+\eta(\boldsymbol{x})\right)^\top f(\boldsymbol{y})\right\}=\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\frac{\partial f_i}{\partial y_i}(\boldsymbol{y})\right\}.\tag{2.9}$$

对于具有 $\eta(\boldsymbol{x})=-\frac{\boldsymbol{x}}{\sigma^2}$ 和 $h(\boldsymbol{y})=\exp(\frac{\|\boldsymbol{y}\|^2}{2\sigma^2})$ 的（各向同性）高斯噪声这一特殊情况，可以恢复斯坦因引理。

Eldar [67] 利用这一结果构造了广义 SURE（GSURE）损失。使用（NEF）的记法，该损失可写为：

$$\mathcal{L}_{\mathrm{GSURE}}(\boldsymbol{y}, f) = \frac{1}{n}\|f(\boldsymbol{y}) + \nabla \log h(\boldsymbol{y})\|^2 + \frac{2}{n}\sum_{i=1}^{n}\frac{\partial f_i}{\partial y_i}(\boldsymbol{y}) \tag{GSURE}$$

与高斯分布情形类似，我们可以使用引理 2.5 证明

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\mathcal{L}_{\mathrm{GSURE}}(\boldsymbol{y}, f)\right\} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\left\{\frac{1}{n}\|f(\boldsymbol{y}) - \eta(\boldsymbol{x})\|^2\right\} + \mathrm{const.} \tag{2.10}$$

不过请注意，这并不适合推导出与本章所介绍损失等价的损失，因为 GSURE 的目标是自然参数 $\eta(\boldsymbol{x})$，而不是 $\boldsymbol{x}$。因此，最优 f 是估计器 $f(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\eta(\boldsymbol{x})\}$；除高斯噪声情形外，它并不等于后验均值，此时 $\eta(\boldsymbol{x}) \propto \boldsymbol{x}$。

近似散度项 对于神经网络等复杂模型，SURE损失中的散度项通常没有解析解，因此通常采用蒙特卡罗估计。一种方法是使用自动微分 [68], 结合 Hutchinson 无偏迹估计量 [69]，计算 SURE损失中的散度项：

$$\mathrm{trace}\left(\boldsymbol{\Sigma}\frac{\partial f}{\partial \boldsymbol{y}}\right) \approx \boldsymbol{\omega}^\top \frac{\partial f}{\partial \boldsymbol{y}}\boldsymbol{\omega} \tag{2.11}$$

其中，$\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ 是一个随机标准高斯分布，$\frac{\partial f}{\partial \boldsymbol{y}}\boldsymbol{\omega}$ 通过自动微分计算为雅可比向量积。

或者，也可以通过简单的蒙特卡洛近似 [70]，对同一估计器使用有限差分近似。

$$\mathrm{trace}\left(\boldsymbol{\Sigma}\frac{\partial f}{\partial \boldsymbol{y}}\right) \approx (\boldsymbol{\Sigma}\frac{\boldsymbol{\omega}}{\tau})^\top \left(f(\boldsymbol{y} + \tau\boldsymbol{\omega}) - f(\boldsymbol{y})\right) \tag{2.12}$$

其中 $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 且 $\tau > 0$ 较小。若对 $\boldsymbol{\omega}$ 取期望并令 $\tau \to 0$，该近似便成为精确值。这种近似无需自动微分，但代价是每次计算都需要对 f 求值两次。实际上，τ 可以取一个较小的比例，例如噪声标准差的 1%。

尽管这两种估计器都是无偏的，但它们的方差可能较高；对 $\boldsymbol{\omega}$ 的多个样本取平均可以降低方差。不过，对于成像应用，正如 [70]，中所论述的，通常可以合理地假设单个蒙特卡洛样本就能提供方差足够低的估计值。直观地说，这是因为去噪函数往往只在局部起作用，因此我们实际上是在空间上对每个像素的许多近乎独立的估计值 $n \ll 1$ 进行平均。

对于泊松噪声，Luisier 等人 [5] 提出了泊松变体（PURE）的以下近似：

$$\begin{aligned}\sum_{i=1}^{n} y_i \left(f_i(\boldsymbol{y}) - f_i(\boldsymbol{y} - \gamma\boldsymbol{e}_i)\right) &\approx \mathrm{trace}\left(\gamma\mathrm{diag}\left(\boldsymbol{y}\right)\frac{\partial f}{\partial \boldsymbol{y}}\right) \\ &\approx (\gamma\mathrm{diag}\left(\boldsymbol{y}\right)\frac{\boldsymbol{\omega}}{\tau})^\top (f(\boldsymbol{y} + \tau\boldsymbol{\omega}) - f(\boldsymbol{y}))\end{aligned} \tag{2.13}$$

该近似基于这样的假设：泊松噪声近似服从具有信号相关协方差的高斯分布，即当 γ 较大时为 $\boldsymbol{\Sigma} \approx \text{diag}(\boldsymbol{y})$；而在 γ 较低的情况下，该近似并不适用。

与 Recorrupted2Recorrupted 的等价性 细心的读者可能已经注意到，R2R (2.2) 中的合成噪声与 SURE (2.12) 的蒙特卡罗近似相当相似。事实证明，(R2R) 可视为对解析形式 (SURE) 的另一种蒙特卡罗近似，如 $\alpha \to 0$ 所示。Oliveira 等人 [54] 最先在高斯噪声情形下提出了这一观察，随后将其推广到（离散）泊松情形 [71], ，最终推广到连续自然指数族，如下述命题所示：

命题 2.6 (Monroy 等人 [55])。 假设 f 是解析的， $p(\boldsymbol{y}|\boldsymbol{x})$ 是属于自然指数族的连续分布，并且 $a_k : \mathbb{R} \mapsto \mathbb{R}$ 满足

$$a_k(y_i) = \lim_{\alpha \to 0} \mathbb{E}_{y_{2,i}|y_i,\alpha} \left\{ (y_{2,i} - y_i)(\alpha y_{2,i})^k \right\}$$

对于所有 $i = 1, \ldots, n$，均满足 $|a_k(y_i)| < \infty$，其中所有正整数 $k \geq 1$ 均成立。于是，

$$\lim_{\alpha \to 0} \mathcal{L}_{GR2R}(\boldsymbol{y},\ f,\ \alpha) =$$

$$\frac{1}{n}\|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 + \frac{2}{n}\sum_{i=1}^{n}\sum_{k \geq 1}(-1)^{k+1} a_k(y_i)\frac{1}{k!}\frac{\partial^k f_i}{\partial y_i^k}(\boldsymbol{y}) + const.$$

其中，所得的类 *SURE* 损失是监督损失的无偏估计量，其输入为$\boldsymbol{y}$ ，而非 $\boldsymbol{y}_1$，即

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left\{ \lim_{\alpha \to 0} \mathcal{L}_{GR2R}\left(\boldsymbol{y}, f, \alpha\right) \right\} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left\{ \frac{1}{n}\|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 \right\}.$$

有趣的是，在各向同性高斯情形下，对于$k \geq 2$，我们有 $a_1(y_i) = \sigma^2$ 且 $a_k(y_i) = 0$ ，从而得到标准 SURE 公式。在泊松噪声情形下， [71], R2R 损失可恢复为 PURE，即 $\alpha \to 0$, 且无需依赖式（2.13）中的连续近似。与只需要一个散度项、但学习条件估计器$\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\eta(\boldsymbol{x})\}$ 的 GSURE 不同，命题 2.6 表明，存在用于学习条件均值 $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$的类 SURE 公式，尽管这些公式通常需要计算 f的高阶导数。

与 Tweedie 公式的联系 （SURE）中的第二项满足以下等式

$$\mathbb{E}_{\boldsymbol{y}} \left\{ \text{trace}\left(\frac{\partial f}{\partial \boldsymbol{y}} \boldsymbol{\Sigma} \right) \right\} = -\mathbb{E}_{\boldsymbol{y}} \left\{ f(\boldsymbol{y})^\top \boldsymbol{\Sigma} \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y}) \right\}$$

利用分部积分可以再次证明这一点。借助这一结果，我们可以通过求解以下问题，找到 SURE损失下的最优去噪器

$$\begin{aligned} f^* &= \arg\min_f \ \mathbb{E}_{\boldsymbol{y}} \left\{ \mathcal{L}_{\text{SURE}}(\boldsymbol{y}, f) \right\} \\ &= \arg\min_f \ \mathbb{E}_{\boldsymbol{y}} \left\{ \frac{1}{n}\|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 - \frac{2}{n} f(\boldsymbol{y})^\top \boldsymbol{\Sigma} \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y}) \right\} \\ &= \arg\min_f \ \mathbb{E}_{\boldsymbol{y}} \left\{ \frac{1}{n}\|f(\boldsymbol{y}) - \boldsymbol{y} - \boldsymbol{\Sigma} \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})\|^2 \right\} \end{aligned}$$

其中最后一步通过配方完成， $\nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})$ 被称为测量分布的得分。因此，最优解就是众所周知的 Tweedie 公式：

$$f^*(\boldsymbol{y}) = \boldsymbol{y} + \boldsymbol{\Sigma} \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y}) \tag{Tweedie}$$

由于 SURE 是有监督 ℓ_2 损失的无偏估计量，而该损失的全局极小值是条件均值估计器 $f^*(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$，因此在高斯噪声情况下，最优的最小二乘估计器为 $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\} = \boldsymbol{y} + \boldsymbol{\Sigma} \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})$。

我们还可以结合 (SURE) 和 (Tweedie) 来计算去噪问题的最小均方误差（MMSE），其为得分和噪声协方差的函数：

$$\begin{aligned} \mathrm{MMSE} &= \mathbb{E}_{\boldsymbol{y}}\left\{\mathcal{L}_{\mathrm{SURE}}(\boldsymbol{y}, f^*)\right\} && (2.14) \\ &= \frac{1}{n}\mathrm{trace}\left(\boldsymbol{\Sigma}\right) - \mathbb{E}_{\boldsymbol{y}}\left\{\frac{1}{n}\|\boldsymbol{\Sigma}\nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})\|^2\right\} && (2.15) \end{aligned}$$

受 SURE 与 Tweedie 公式之间紧密联系的启发，Noise2Score 方 [72, 73] 法提出首先训练一个网络 $s(\boldsymbol{y})$ 来近似得分，即 $s(\boldsymbol{y}) \approx \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})$，然后在测试时应用 (Tweedie) $f(\boldsymbol{y}) = \boldsymbol{y} + \boldsymbol{\Sigma} s(\boldsymbol{y})$。

Tweedie 公式在扩散模型中也发挥着重要作用，因为它使人们能够通过去噪器网络 $f(\boldsymbol{y})$ 间接评估得分函数。第 2.4 节将讨论利用这一联系的自监督扩散方法。

2.3 部分未知的噪声分布

在许多现实场景中，我们没有独立样本对 $(\boldsymbol{y}_1, \boldsymbol{y}_2)$，因此无法使用 (Noise2Noise)；我们也不知道确切的噪声分布，因此无法使用 (R2R) 或 (SURE) 损失。在某些情况下，我们会看到，仍然可以构建自监督损失，但代价是限制学习得到的去噪器的灵活性：我们只能期望最小化一个受约束的监督损失 [28], ，满足

$$\arg\min_{f} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left\{\|f(\boldsymbol{y}) - \boldsymbol{x}\|^2\right\} \quad \text{s.t. } f \in \mathcal{F} \tag{2.16}$$

其中 $\mathcal{F}$ 是一个受约束的函数集。因此，我们通常无法逼近预言估计器，即条件均值 $f^*(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$，如果它不属于$\mathcal{F}$。不过，正如我们将看到的，在某些情况下，约束可能相对宽松，从而使我们能够找到性能接近预言机的 f。如图 2.1 所示，我们对分布了解得越少，所需的约束就越强，得到的性能也就越偏离监督性能。例如，我们可能知道噪声是各向同性的高斯分布，但不知道噪声水平 σ^2；更极端的情况是，我们可能完全不知道分布，只知道像素间独立这一假设。

2.3.1 未知噪声水平的 Stein无偏风险估计

在某些应用中，噪声模型可以假设为高斯分布 $\boldsymbol{y} = \boldsymbol{x} + \sigma\boldsymbol{\epsilon}$ with$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$，但噪声水平 σ 未知。一种简单的方法是先从含噪数据中估计 σ，然后使用 (SURE) 或 (R2R) 进行训练。如图 2.2 所示，

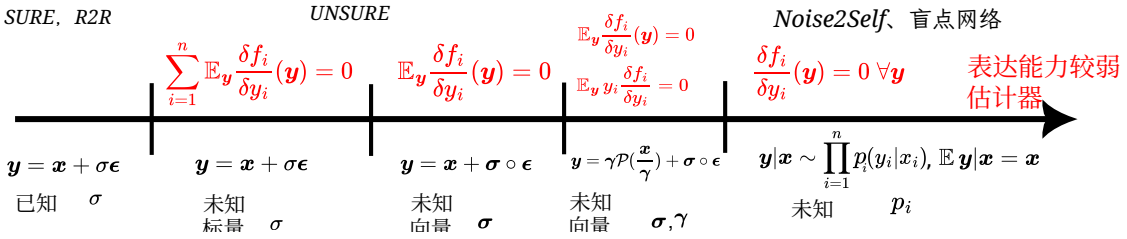

图 2.1：**自监督去噪中的表达能力-鲁棒性权衡** [28]。随着对噪声分布的假设逐渐放宽，学习到的估计器需要降低表达能力，以避免对噪声过拟合。

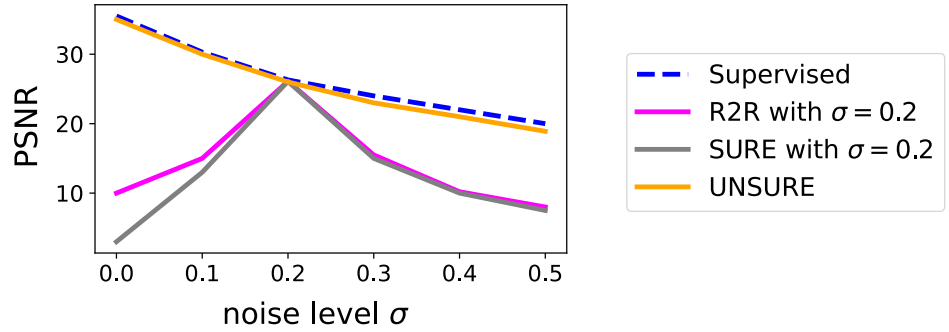

图 2.2：**不同噪声水平下的自监督去噪** [28]。在 MNIST 高斯去噪任务中，使用 U-Net 去噪器对监督损失、（SURE）、（R2R）和（UNSURE）损失函数进行比较。如果 SURE 和 R2R 中的噪声水平 σ 设定正确，性能与监督情形接近。然而，如果噪声级别设定错误，性能会显著下降。UNSURE 损失对噪声级别设定错误具有鲁棒性，且性能与监督情形接近。

这两种损失函数对设定错误的 σ 都非常敏感，误差超过 10% 就可能导致性能显著下降。

我们也可以构建一种自监督损失，使（SURE）中的散度项为零，从而消除损失对未知噪声水平 σ的依赖：

$$\arg\min_{f}\ \mathbb{E}_{\boldsymbol{y}}\left\{\frac{1}{n}\|f(\boldsymbol{y})-\boldsymbol{y}\|^2\right\}\ \text{ s.t. }\ \mathbb{E}_{\boldsymbol{y}}\left\{\sum_{i=1}^{n}\frac{\partial f_i}{\partial y_i}(\boldsymbol{y})\right\}=0 \tag{2.17}$$

应用引理 2.4，我们可以证明该最小化问题等价于带约束的监督学习场景：

$$\arg\min_{f}\ \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left\{\frac{1}{n}\|f(\boldsymbol{y})-\boldsymbol{x}\|^2\right\}\ \text{ s.t. }\ f\in\mathcal{F}$$

其中，我们只寻找具有零期望散度（ZED）的去噪器，即属于约束集合的去噪器

$$\mathcal{F}=\left\{f:\mathbb{E}_{\boldsymbol{y}}\left\{\sum_{i=1}^{n}\frac{\partial f_i}{\partial y_i}(\boldsymbol{y})\right\}=0\right\}.$$

使用拉格朗日乘子 $\eta \in \mathbb{R}$，我们可以将（UN-SURE）中的约束学习问题改写为

$$\min_{f} \max_{\eta} \mathbb{E}_{\boldsymbol{y}} \left\{ \frac{1}{n} \|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 + \frac{2\eta}{n} \sum_{i=1}^{n} \frac{\partial f_i}{\partial y_i}(\boldsymbol{y}) \right\}. \tag{UNSURE}$$

有趣的是，所得损失与 (SURE) 非常相似，但我们不使用固定的噪声水平 σ，而是将其替换为乘数 η 并对其进行最大化。

分析约束估计量 在学习得到的去噪器上添加零期望散度约束的代价是什么？容易证明，除图像分布仅由单幅图像组成这一平凡情况外，最优去噪器的散度 [74], 为正。然而，我们将看到，该约束相当温和，与监督学习之间的差距可能很小。

遵循与 (Tweedie) 中类似的推导，我们再次可以用得分函数表示约束学习问题的最优解，写作

$$f^{\mathrm{ZED}}(\boldsymbol{y}) = \boldsymbol{y} + \frac{n \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})}{\mathbb{E}_{\boldsymbol{y}} \|\nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})\|^2}.$$

我们还可以通过将零期望散度估计量的定义直接代入 (SURE) 来计算其期望误差:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}} \left\{ \frac{1}{n} \|f^{\mathrm{ZED}}(\boldsymbol{y}) - \boldsymbol{x}\|^2 \right\} &= \mathbb{E}_{\boldsymbol{y}} \left\{ \mathcal{L}_{\mathrm{SURE}}\Big(\boldsymbol{y}, f^{\mathrm{ZED}}(\boldsymbol{y})\Big) \right\} \\ &= \frac{n}{\mathbb{E}_{\boldsymbol{y}} \|\nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})\|^2} - \sigma^2 \\ &= \mathrm{MMSE} \left(1 - \frac{\mathrm{MMSE}}{\sigma^2}\right)^{-1} \\ &= \mathrm{MMSE} + \sigma^2 \sum_{j \geq 2} \left(\frac{\mathrm{MMSE}}{\sigma^2}\right)^j \end{aligned} \tag{2.18}$$

其中，第三行使用了 (2.14) 中的 MMSE 表达式（即最优估计器 $f(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$的误差），最后一行则利用了等比级数公式。由于$\mathrm{MMSE}_{\sigma^2}$ 是最优估计器在信噪比方面的提升，因此我们始终有$\frac{\mathrm{MMSE}}{\sigma^2} < 1$，并且通常$\frac{\mathrm{MMSE}}{\sigma^2} \ll 1$。因此，与预言机相比，$f^{\mathrm{ZED}}$ 的额外误差可能非常小。

超越各向同性高斯噪声的扩展 该（UNSURE）方法还可以进一步扩展到噪声协方差未知的情形：考虑由一些合理的 协方差矩阵构成的 s 维集合

$$\mathcal{R} = \left\{ \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \in \mathbb{R}^{n \times n} : \boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \sum_{j=1}^{s} \eta_j \boldsymbol{\Psi}_j, \boldsymbol{\eta} \in \mathbb{R}^s \right\}$$

对于某些固定的基矩阵 $\{\boldsymbol{\Psi}_j \in \mathbb{R}^{n\times n}\}_{j=1}^s$，并希望真实协方差属于该集合，即 $\boldsymbol{\Sigma} \in \mathcal{R}$。在这种情况下，我们考虑 $s \geq 1$ 个约束，并最小化以下目标函数

$$\arg\min_f \mathbb{E}_{\boldsymbol{y}} \left\{ \frac{1}{n} \|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 \right\} \tag{2.19}$$

$$\text{s.t. } \mathbb{E}_{\boldsymbol{y}} \left\{ \text{trace} \left(\boldsymbol{\Psi}_j \frac{\partial f}{\partial \boldsymbol{y}}(\boldsymbol{y}) \right) \right\} = 0,\ j = 1, \ldots, s \tag{2.20}$$

注意，(UNSURE)可作为一种特殊情形恢复，其中 $s = 1$ 和 $\boldsymbol{\Psi}_1 = \boldsymbol{I}$。我们对协方差了解得越少，集合 $\mathcal{R}$就越大，从而对学习得到的估计器施加更多约束。因此，维度 $s \geq 1$ 体现了所得估计器的最优性与其对错误设定协方差的鲁棒性之间的权衡。与各向同性情形一样，我们可以求得最优受约束去噪器的闭式表达式：

定理 2.7 (Tachella 等人 [28])。 设 $p_{\boldsymbol{y}} = p_{\boldsymbol{x}} * \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$，并假设 $\{\boldsymbol{\Psi}_j \in \mathbb{R}^{n\times n}\}_{j=1}^s$线性无关。问题(2.19) 的最优解为

$$f(\boldsymbol{y}) = \boldsymbol{y} + \boldsymbol{\Sigma}_{\hat{\boldsymbol{\eta}}} \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y}) \tag{2.21}$$

其中，最优乘子为$\hat{}\boldsymbol{\eta} = \boldsymbol{Q}^{-1}\boldsymbol{v}$，且

$$Q_{i,j} = trace\left(\boldsymbol{\Psi}_i \mathbb{E}_{\boldsymbol{y}} \left\{ \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y}) \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})^\top \right\} \boldsymbol{\Psi}_j^\top \right)$$

并且，对于 $i, j = 1, \ldots, s$，有 $v_i =$迹$(\boldsymbol{\Psi}_i)$。

我们可以将这一泛化应用于具有相关噪声且相关性未知的问题，这类问题通常建模为

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\sigma} * \boldsymbol{\epsilon}$$

其中 $*$ 表示卷积算子，$\boldsymbol{\sigma} \in \mathbb{R}^p$ 是向量值的，$\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$。如果我们不知道确切的噪声相关性，可以考虑相关范围不超过 $\pm r$ 个抽头/像素的协方差集合。我们可以选择 $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ 为由滤波器 $\boldsymbol{\eta}$ 参数化的正定循环矩阵。在这种情况下，解为

$$f(\boldsymbol{y}) = \boldsymbol{y} + \hat{\boldsymbol{\eta}} * \nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})$$

最优乘子为 $2r + 1$ 得分 [28] 的自相关系数。

我们还可以利用引理 2.5，将 UNSURE 方法扩展到非高斯噪声分布，例如参数 [28] 未知的泊松-高斯噪声。

2.3.2 交叉验证方法

在许多应用中，噪声在像素/测量值 $p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n p_i(y_i|x_i)$之间是可分离的，但除以下假设外，我们不知道每个像素处的噪声分布 p_i ：

[4]为简单起见，我们在此考虑一维信号，但该结果可直接推广到二维情形。

均值为 $\mathbb{E}_{y_i|x_i}\{x_i\} = x_i$。在这种情况下，我们迄今介绍的损失函数都不适用，但我们仍然可以找到一种损失，它是受约束的有监督损失的无偏估计量。由于我们不知道噪声分布，因此需要施加比 (UNSURE) 中零期望散度条件更严格的约束；该条件依赖于高斯噪声假设。

许多近期的自监督方法，包括 Noise2Void [21], Noise2Self [75], 盲点网络 [23], Neighbor2Neighbor [76], ，大体上可归类为交叉验证方法 [47], ，它们最小化以下目标函数 [28]:

$$\begin{aligned} \arg\min_{f} \quad & \mathbb{E}_{\boldsymbol{y}} \left\{ \frac{1}{n} \|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 \right\} \\ & \text{s.t. } \frac{\partial f_i}{\partial y_i}(\boldsymbol{y}) = 0, \ \forall \boldsymbol{y} \in \mathbb{R}^n, \ i = 1, \dots, n \end{aligned} \tag{CV}$$

其中，导数约束等价于要求第 i 个输出 f_i 不依赖于第 i 个输入 y_i。

命题 2.8（改编自 Batson 和 Royer [75]）。设 f： $\mathbb{R}^n \to \mathbb{R}^n$ 是一个函数，其第 i 个输出不依赖于第 i 个条目，或者等价地，对于所有$\boldsymbol{y} \in \mathbb{R}^n$ 以及 $i = 1$、.....、n，均有 $\frac{\partial f_i}{\partial y_i}(\boldsymbol{y}) = 0$；进一步假设，对于所有 $i = 1$、.....、n，还有 $p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n p_i(y_i|x_i)$ 以及 $\mathbb{E}_{y_i|x_i}\{y_i\} = x_i$。那么，

$$\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left\{ \|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 \right\} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left\{ \|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 \right\} + const. \tag{2.22}$$

证明。

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left\{ \|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 \right\} \\ & = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left\{ \|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 \right\} + 2 \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left\{ f_i(\boldsymbol{y})^\top (y_i - x_i) \right\} + \text{const.} \\ & = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left\{ \|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 \right\} + 2 \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{y}_{-i}|\boldsymbol{x}} \left\{ f_i(\boldsymbol{y}) \right\}^\top \mathbb{E}_{y_i|x_i} \left\{ y_i - x_i \right\} + \text{const.} \\ & = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} \left\{ \|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 \right\} + \text{const.} \end{aligned}$$

其中， $\boldsymbol{y}_{-i}$ 表示一个向量，其所有条目均与 $\boldsymbol{y}$ 相同，但第 i 个条目除外；第三行利用了以下事实： $f_i(\boldsymbol{y})$ 与 y_i 在给定 $\boldsymbol{x}$ 的条件下相互独立；最后一行则利用了 $\mathbb{E}_{y_i|x_i}\{y_i\} = x_i$。 □

根据命题 2.8，(变分自编码器) 中的最小化问题等价于以下受约束的监督问题

$$\arg\min_{f} \ \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \frac{1}{n} \|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 \right\} \ \text{ s.t. } f \in \mathcal{F}$$

其中，我们只寻找不会使用第 i 个输入值来预测第 i 个输出值的去噪器，也就是说

$$\mathcal{F} = \left\{ f : \frac{\partial f_i}{\partial y_i}(\boldsymbol{y}) = 0, \ \forall \boldsymbol{y} \in \mathbb{R}^n, \ i = 1, \dots, n \right\}.$$

与 (UNSURE) 一样，条件均值 $f^*(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$ 不属于约束集合 $\mathcal{F}$，因此我们无法期望达到与监督学习相同的性能。在这种情况下，(CV) 的最优解为 $f_i^{\mathrm{CV}}(\boldsymbol{y}) = \mathbb{E}_{x_i|\boldsymbol{y}_{-i}}\{x_i\}$，其中 $i = 1$、.....、n。这里，与在 (2.18) 中以闭式形式给出次优性差距的 UNSURE 情况不同，f^*与 f_{CV}之间的差距没有简单的闭式表达式，并且高度依赖于信号分布：我们预计，具有较强空间相关性的信号，其差距较小；而稀疏信号的差距较大 [75]。

目前已提出两种用于施加零导数约束的主要方法：（1）盲点网络，它采用能够施加该约束的特定架构；（2）分裂损失函数，它在训练过程中施加该约束。

盲点网络 Laine 等人 [23] 提出了一种图像到图像网络架构 f，该架构仅依赖像素的邻域（而非像素本身）来预测其去噪值，这等价于对所有像素施加 $\frac{\partial f_i}{\partial y_i}(\boldsymbol{y}) = 0 \quad i = 1, \ldots\ldots, \ n$。该盲点网络依赖于一种全卷积架构，将经过偏移的（向上、向下、向左和向右）感受野结合起来。该方法非常高效，因为它允许直接在测量一致性损失 $f \ \|f(\boldsymbol{y}) - \boldsymbol{y}\|^2$上训练去噪器，但对 f 施加了超出零梯度约束之外的严格架构限制。

分裂方法 训练不依赖中心像素进行去噪的网络的另一种方法，是在训练过程中随机将该像素掩码掉 [21, 75]。该方法可写成如下损失函数

$$\mathcal{L}_{\mathrm{SPLIT}}(\boldsymbol{y}, f) = \mathbb{E}_{\boldsymbol{b}}\left\{\frac{1}{n}\|(\mathbf{1} - \boldsymbol{b}) \circ (f(\boldsymbol{b} \circ \boldsymbol{y}) - \boldsymbol{y})\|^2\right\} \tag{SPLIT}$$

其中，$\circ$ 表示逐元素（Hadamard）乘积，$\boldsymbol{b} \in \{0, 1\}^n$ 为随机二值掩码。关于如何选择分割分布 $p(\boldsymbol{b})$，已有大量研究；选择通常取决于数据的结构（例如图像、视频等）。一些最常用的选择包括：

1. Noise2Void [21] 将中心输入像素替换为一个随机邻近像素，并且只针对被替换的像素计算损失。2. Neighbor2Neighbor [76] 通过随机选择每个 2×2 图块中的一个像素作为输入，再从同一图块中选择另一个像素作为目标，构造图像的两个互不重叠的子采样版本。这种损失隐含地假设图像具有尺度不变性，因此去噪器可以在欠采样图像上训练，并在全分辨率下测试。3. Noise2Self [75] 使用 J 不相交掩码对图像进行划分，$\boldsymbol{b}^{(j)}$，使得对于 $i = 1, \ldots\ldots, \ n$，$\frac{1}{J}\sum_{j=1}^{J} b_i^{(j)} = \mathbf{1}$；然后按照以下方式构造估计器并进行训练：

f

$$f(\boldsymbol{y}) = \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{b}^{(j)} \circ g\Big((1 - \boldsymbol{b}^{(j)}) \circ \boldsymbol{y} + \boldsymbol{b}^{(j)} \circ \boldsymbol{u}\Big)$$

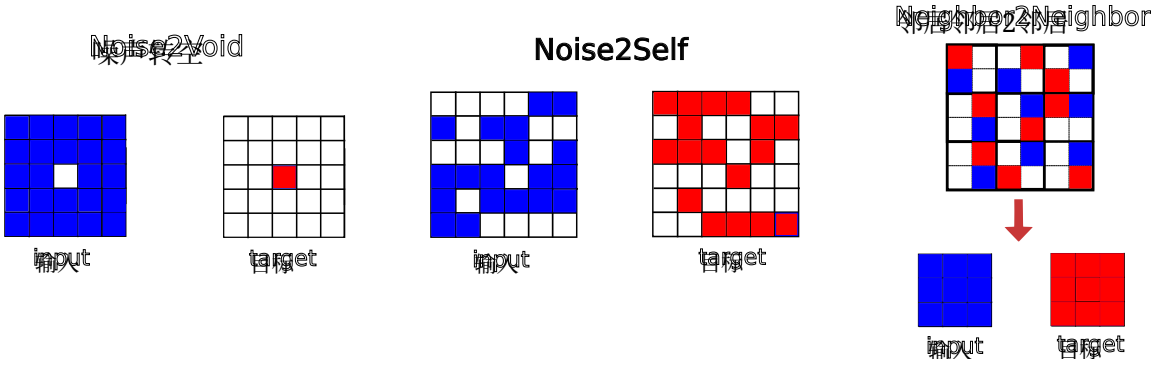

图 2.3：**像素拆分策略**。 Noise2Void 和 Noise2Self 会将输入图像中被移除的像素置零或填入其邻近值，而 Neighbor2Neighbor 则通过对每个 2×2 邻域进行两次随机子采样来拆分，其中一次作为输入，另一次作为目标。

其中， g 是一个基于神经网络的模型， $\boldsymbol{u}$ 是一个独立同分布均匀随机向量[5]。

图 2.3 展示了不同的掩码方法。也可以使用更通用的一组 J 随机掩码 $\boldsymbol{b}^{(j)}$，只要它们覆盖整个图像即可（即$\sum_{s=1}^{J}(1-b_i^{(s)}) > 0$对于所有像素 $i = 1,\ldots,n$）。然后，与 (R2R) 类似，在测试时，我们可以对多个拆分取平均值 [77]:

$$f^{\text{test}}(\boldsymbol{y}) = \sum_{j=1}^{J} \boldsymbol{w}^{(j)} \circ f(\boldsymbol{b}^{(j)} \circ \boldsymbol{y}) \tag{2.23}$$

权重

$$w_i^{(j)} = \frac{1-b_i^{(j)}}{\sum_{s=1}^{J}(1-b_i^{(s)})} \quad \text{for} \quad i = 1,\ldots,n$$

在下一章中，我们将介绍这一思想在一般逆问题中的扩展，此时前向算子并非平凡算子 $\boldsymbol{A}(\boldsymbol{x}) \neq \boldsymbol{x}$，并且分裂策略是以算子特定的方式制定的。

2.4 从噪声数据中学习生成模型

我们已经看到，只要噪声分布已知，就可以通过自监督损失来近似后验均值。我们还可以进一步探究，是否能近似均值之外的其他后验统计量；或者更一般地，是否能仅凭测量数据估计信号分布 $p_{\boldsymbol{x}}$。由于我们有

$$p_{\boldsymbol{y}}(\boldsymbol{y}) = \int_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x}) p_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x} \tag{2.24}$$

[5]虽然缺乏理论依据， [75] 也观察到，仅使用训练好的基础神经网络 g作为最终估计器，性能同样良好。

这个问题可以看作测度空间中的线性反问题，可写为 $p_{\boldsymbol{y}} = \mathcal{A}(p_{\boldsymbol{x}})$，其中 $\mathcal{A}$ 与核 $p(\boldsymbol{y}|\boldsymbol{x})$ 的积分相关。这种表述可追溯至 Robbins 在经验贝叶斯估计量方面的工作 [78]。

这里我们考虑最简单的设定：带有加性噪声的去噪问题，即 $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}$。由于噪声是加性的，因此测量分布是信号分布的卷积形式，即

$$p_{\boldsymbol{y}} = p_{\boldsymbol{x}} * p_{\boldsymbol{\epsilon}} \tag{2.25}$$

因此，模型识别可以视为一个反卷积问题。该问题的傅里叶对应形式为

$$\phi_{\boldsymbol{y}}(\boldsymbol{\omega}) = \phi_{\boldsymbol{x}}(\boldsymbol{\omega})\phi_{\boldsymbol{\epsilon}}(\boldsymbol{\omega}) \tag{2.26}$$

其中，$\phi_{\boldsymbol{y}}(\boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{y}}\left\{\exp(\mathrm{i}\boldsymbol{y}^\top\boldsymbol{\omega})\right\}$、$\phi_{\boldsymbol{x}}(\boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{x}}\left\{\exp(\mathrm{i}\boldsymbol{x}^\top\boldsymbol{\omega})\right\}$ 和 $\phi_{\boldsymbol{\epsilon}}(\boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left\{\exp(\mathrm{i}\boldsymbol{\epsilon}^\top\boldsymbol{\omega})\right\}$分别是 $p_{\boldsymbol{x}}$、$p_{\boldsymbol{y}}$ 和 $p_{\boldsymbol{\epsilon}}$ 及 $\boldsymbol{\omega} \in \mathbb{R}^n$的特征函数。通过简单检查式 (2.26)，我们可以推知，如果噪声模型已知，因此 $\phi_{\boldsymbol{\epsilon}}(\boldsymbol{\omega})$ 也已知，那么只要对所有 $\boldsymbol{\omega} \in \mathbb{R}^n$ 都有 $\phi_{\boldsymbol{\epsilon}}(\boldsymbol{\omega}) \neq 0$，就可以识别信号分布：

命题 2.9 (Tachella 等人 [79])。 如果噪声分布 $\phi_{\boldsymbol{\epsilon}}$ 的特征函数处处非零，那么干净测量分布与含噪测量分布的空间之间存在一一映射。

例如，高斯噪声分布的特征函数处处非零，因此我们可以从含噪测量[6]中唯一确定信号分布。现在，我们介绍一些近期仅利用含噪数据学习生成模型的方法。

变分自编码器 假设噪声模型 $p(\boldsymbol{y}|\boldsymbol{x})$已知（或已在校准步骤中估计），我们可以将测量数据的分布建模为

$$\begin{cases} \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}),\ \hat{\boldsymbol{x}} = f(\boldsymbol{z}) \\ \hat{\boldsymbol{y}} \sim p(\boldsymbol{y}|\boldsymbol{x} = \hat{\boldsymbol{x}}) \end{cases} \tag{2.27}$$

其中，f 是一个深度网络，$\boldsymbol{z}$ 是服从标准高斯分布的潜变量。Prakash 等人 [80, 81] 提出使用变分自编码器学习此类生成模型，这需要额外学习一个编码器网络，以近似分布 $p(\boldsymbol{z}|\boldsymbol{y})$。编码器 $p(\boldsymbol{z}|\boldsymbol{y})$和解码器 f均学习完成后，在推断时，我们既可以从 $p_{\boldsymbol{x}}$生成干净样本，形式为 $\hat{\boldsymbol{x}}^{(i)} = f(\boldsymbol{z}^{(i)})$，其中 $i = 1$、.....、N，且 $\boldsymbol{z}^{(i)} \sim \mathcal{N}(\boldsymbol{0},\ \boldsymbol{I})$；也可以先从编码器 $\boldsymbol{z}^{(i)} \sim p(\boldsymbol{z}|\boldsymbol{y})$中对潜变量进行采样，再将其输入解码器网络，从后验分布中生成干净样本。在下一章中，我们将介绍对抗方法 [82]，用于在测量值不完整的情况下学习类似的生成模型。

扩散方法（Tweedie）表明，最优高斯去噪器作为噪声水平的函数，可用于获取噪声信号分布的得分函数，$\nabla \log p_{\boldsymbol{y}}(\boldsymbol{y})$。扩散模型利用这一点，通过神经网络近似得分函数，构建

[6]请注意，这并未说明该问题的样本复杂度，也就是说，仅凭有限数据解决这一问题有多困难。

随机采样器，以便从干净信号分布 $p(\boldsymbol{x})$ 中进行近似采样，或在给定含噪样本 $\boldsymbol{y}$ [41] 的情况下，从条件后验分布 $p(\boldsymbol{x}|\boldsymbol{y})$ 密度中进行近似采样。

正如我们在本章中所见，仅使用含噪数据，我们就可以通过 (SURE) 或 (R2R) 以自监督的方式学习去噪器。然而，这只能让我们获得噪声水平大于或等于观测数据噪声水平 $\sigma \geq \sigma_n$时得分的近似。一些近期方法会在该噪声水平 [83], 处停止扩散，而另一些方法则尝试通过对学习得到的去噪器施加一致性约束 [44]，将噪声水平降得更低。

2.5 迈向一般逆问题

如果我们想将本章中的损失函数扩展到简单去噪问题之外，其中 $\boldsymbol{A}\colon \mathbb{R}^n \mapsto \mathbb{R}^m$ 是一个一般前向算子，会发生什么？将去噪函数定义为 $\boldsymbol{A} \circ f$后，我们可以在测量空间中使用上述任意损失函数。例如，如果我们观测到带有高斯噪声的测量值 $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{A}(\boldsymbol{x}),\ \sigma^2 \boldsymbol{I}\right)$，且已知噪声水平 σ，就可以将 (SURE) 改写为

$$\mathcal{L}_{\mathrm{SURE}}(\boldsymbol{y}, f) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{A} \circ f(\boldsymbol{y})\|^2 + \frac{2}{n}\operatorname{trace}\left(\boldsymbol{\Sigma}\,\frac{\partial \boldsymbol{A} \circ f}{\partial \boldsymbol{y}}(\boldsymbol{y})\right)$$

因此，我们得到了等价的测量监督损失：

$$\mathbb{E}_{\boldsymbol{y}}\left\{\mathcal{L}_{\mathrm{SURE}}(\boldsymbol{y}, f)\right\} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left\{\frac{1}{n}\|\boldsymbol{A}(\boldsymbol{x}) - \boldsymbol{A} \circ f(\boldsymbol{y})\|^2\right\} + \text{const.}$$

重要的是，如果 $\boldsymbol{A}$ 是一一映射[7]，那么极小化器为 $f^*(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$。然而，如果 $\boldsymbol{A}$ 不是一一映射，例如像素 n 多于测量值 m，那么$f^*(\boldsymbol{y}) \neq \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$，即使我们拥有包含无限多个测量向量的数据集，也不能指望学到与监督学习场景相同的解。下一章将介绍一些克服这一局限的方法。

2.6 总结

本章介绍了如何构建自监督损失，使其无需访问干净目标即可处理含噪数据。损失的选择取决于我们对噪声分布了解多少：了解得越多，就越有望达到接近全监督学习的性能；了解得越少，就越需要对学习得到的去噪器施加约束，且其性能与监督学习性能的差距也会越大。尽管如此，我们已经看到，在许多情况下，自监督学习与监督学习之间的差距可能很小，而且可以利用自监督损失训练出在实践中表现良好的去噪器。第 2.6 节总结了本章介绍的不同损失类别，并重点说明了各类损失所采用的不同噪声假设。

[7]在这种情况下，自监督损失与监督损失具有相同的极小化器，即满足（1.10）。

损失族	**噪声假设**	**学习最优** $f^*(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}\|\boldsymbol{y}}\{\boldsymbol{x}\}$**?**	**参考文献**
(Noise2Noise)	两个独立的 噪声实现 $(\boldsymbol{y}_1,\ \boldsymbol{y}_2)$	Yes	[40]
(R2R) (GR2R)	自然指数族 或加性噪声 已知参数	Yes	[51, 52] [54, 55]
（SURE）	自然指数族 和泊松-高斯 已知参数	Yes	[53, 63] [60, 67] [61, 62]
（UNSURE）	自然指数族 和泊松-高斯 未知参数	No, 但较小 间隔见（2.18）	[28]
(CV)	$p(\boldsymbol{y}\|\boldsymbol{x}) = \prod_{i=1}^{n} p_i(y_i\|x_i)$	不，差距取决于 空间相关性。	[75, 76] [21, 23] [84]

表 2.2：**本章所介绍的自监督损失函数总结**。自然指数族包含许多常见的噪声分布，例如高斯分布、泊松分布和伽马分布。所有损失函数都假设噪声满足 $\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\{\boldsymbol{y}\} = \boldsymbol{x}$（对于 Noise2Noise，则满足 $\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\{\boldsymbol{y}_1\} = \boldsymbol{x}$）。

第3章

从不完整测量中学习

上一章介绍的自监督损失可以处理各种类型的噪声模型，并适用于任意一对一前向算子。然而，当算子是多对一的，也就是不可逆时，会发生什么？很容易证明，即使在无噪声测量和线性算子 $\boldsymbol{A}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ 这一简单情形下，我们也无法获得 $\boldsymbol{A}$ 的零空间（线性子空间 $\{\boldsymbol{x} \in \mathbb{R}^n : \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$）中的任何信息，因此无法学习该空间部分上的重建函数：

命题 **3.1 (Chen 等人 [26]).** 任何形式为 $f\colon \mathbb{R}^m \mapsto \mathbb{R}^n$ 的重建函数

$$f(\boldsymbol{y}) = \boldsymbol{A}^\dagger \boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{A}^\dagger \boldsymbol{A})\, v(\boldsymbol{y}) \tag{3.1}$$

其中 $\boldsymbol{A}^\dagger$ 是 $\boldsymbol{A}$ 的伪逆，且 $v\colon \mathbb{R}^n \to \mathbb{R}^n$ 是任意函数，满足测量一致性 $\boldsymbol{A}f(\boldsymbol{y}) = \boldsymbol{y}$，因此是 $\mathbb{E}_{\boldsymbol{y}}\left\{\frac{1}{m}\|\boldsymbol{A}f(\boldsymbol{y}) - \boldsymbol{A}\boldsymbol{x}\|^2\right\}$ 的全局极小值。 —

因此，为了解决这个问题，我们需要更多信息。克服这一局限主要有两种方式：第一种方式将在第3.1节中介绍，即使用与不同前向算子 $\{\boldsymbol{A}_g\}_{g=1}^G$ 相关的测量值进行训练，每个算子都可能具有不同的零空间。第二种方式将在第3.2节中介绍，即假设图像集合具有某种变换群 $\{\boldsymbol{T}_g\}_{g=1}^G$ 不变性，我们将证明，这等价于观测由算子集合 $\{\boldsymbol{A} \circ \boldsymbol{T}_g\}_{g=1}^G$ 得到的测量值。

3.1 利用多个算子

我们假设测量值是通过以下模型获得的

$$\begin{cases} \boldsymbol{x} \sim p(\boldsymbol{x}),\ \boldsymbol{A} \sim p(\boldsymbol{A}) \\ \boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}) \end{cases} \tag{3.2}$$

其中我们仅考虑线性前向算子，并假设每次测量都会采样不同的算子 $\boldsymbol{A}$。在实践中，算子 $p(\boldsymbol{A})$ 的分布通常是离散且有限的。一些实际示例包括（见图 3.1）：

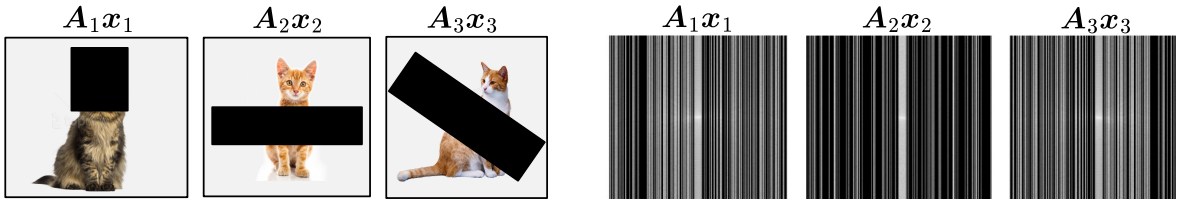

图 3.1：**利用多个前向算子进行学习**。在某些成像场景中，例如图像修复（左）和加速磁共振成像（右），掩膜算子会随数据集中的样本而变化，从而提供信号分布的不同视图。

- 在加速磁共振成像应用中，[27], 加速掩模可能会随扫描而变化，其中 $\boldsymbol{A} = \text{diag}\,(\boldsymbol{b})\,\boldsymbol{F}$，$\boldsymbol{F}$为离散傅里叶变换，掩模则按照 $\boldsymbol{b} \sim p(\boldsymbol{b})$ 进行随机采样。

- 在图像修复问题中，$\boldsymbol{A} = \text{diag}\,(\boldsymbol{b})$，缺失像素通常会因图像而异 $[51, 82]$, ，从而产生一组不同的掩码。

多个算子提供的信息可以帮助我们克服命题 3.1 的局限。设有一组最多包含 G 个可供选择的前向算子，$p(\boldsymbol{A}) = \frac{1}{G}\sum_{g=1}^{G}\delta_{\boldsymbol{A}_g}$，且信号分布仅由一个信号组成，$p(\boldsymbol{x}) = \delta_{\boldsymbol{x}_0}$。在这个简单的情形下，我们可以获得 G 个不同的测量向量，$\boldsymbol{y}_g = \boldsymbol{A}_g\boldsymbol{x}_0$ 分别对应于 $g = 1, \ldots, G$，它们测量的都是同一个底层信号 $\boldsymbol{x}_0$，但各自使用了 G 个不同前向算子中的一个 $\boldsymbol{A}_g \in \mathbb{R}^{m\times n}$。于是，我们可以尝试通过求解以下方程组来恢复 $\boldsymbol{x}_0$：

$$\begin{bmatrix}\boldsymbol{y}_1\\ \vdots\\ \boldsymbol{y}_G\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}_1\\ \vdots\\ \boldsymbol{A}_G\end{bmatrix}\boldsymbol{x}_0 \tag{3.3}$$

其包含 mG 个方程和 n 个未知数。由于这是我们能够获得的最大测量值数量，若要根据测量数据识别 $\boldsymbol{x}_0$，则系统 (3.3) 必须具有最大秩；等价地说，$\mathbb{E}_{\boldsymbol{A}}\left\{\boldsymbol{A}^\top\boldsymbol{A}\right\} = \frac{1}{G}\sum_{g=1}\boldsymbol{A}_g^\top\boldsymbol{A}_g$ 必须是可逆矩阵。

这一论证给出了在信号分布非平凡的一般情形下，利用不完整测量进行学习所需算子的数量和多样性的必要条件。它告诉我们，要利用不完整测量进行学习，至少需要 $G \geq n/m$ 个不同的前向算子。然而，这既不是一种实用算法，也无法在信号分布非平凡的一般情形下提供充分条件，因为通常再次观测到同一图像的概率为零。

假设算子已知，我们可以明确重建网络对前向算子的依赖关系，将其表示为 $f(\boldsymbol{y}, \boldsymbol{A})$。正如第一章所述，将前向算子的知识融入架构有多种方法，其中最常见的是展开优化算法，例如 $[14, 15]$。处理多个算子的一种朴素方法是最小化测量一致性：

$$\arg\min_{f}\mathbb{E}_{\boldsymbol{y},\boldsymbol{A}}\left\{\mathcal{L}_{\text{MC}}(\boldsymbol{y}, \boldsymbol{A}\circ f)\right\} \tag{3.4}$$

with

$$\mathcal{L}_{\mathrm{MC}}(\boldsymbol{y}, \boldsymbol{A} \circ f) = \frac{1}{m}\|\boldsymbol{A} f(\boldsymbol{y}, \boldsymbol{A}) - \boldsymbol{y}\|^2. \tag{MC}$$

遗憾的是，这种思路可能会失效，因为 $f(\boldsymbol{y},\ \boldsymbol{A}) = \boldsymbol{A}^{\dagger}\boldsymbol{y}$ 是损失的一个全局极小值。这种平凡解的出现，是因为 f 无需学习任何关于信号分布的知识，就能达到零训练误差。我们可以通过两种方式避免这种解：使用分割损失，或强制各算子之间保持一致性。

使用无噪声测量进行拆分 我们可以从输入中移除部分测量值，使 f必须预测未观测部分，从而避免测量一致性损失的平凡解。将测量值划分为两个不重叠的部分，即 $\boldsymbol{y} = [\boldsymbol{y}_1^{\top}, \boldsymbol{y}_2^{\top}]^{\top}$ 和 $\boldsymbol{A}^{\top} = [\boldsymbol{A}_1^{\top}, \boldsymbol{A}_2^{\top}]^{\top}$，我们可以构建一个自监督损失，要求网络根据 $\boldsymbol{y}_1$预测 $\boldsymbol{y}$：

$$\mathcal{L}_{\mathrm{MSPLIT}}(\boldsymbol{y}, \boldsymbol{A}, f) = \mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{A}_1 | \boldsymbol{y}, \boldsymbol{A}} \left\{ \frac{1}{m}\|\boldsymbol{A} f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{y}\|^2 \right\} \tag{MSPLIT}$$

其中，期望是对随机拆分的某个分布 $p(\boldsymbol{y}_1,\ \boldsymbol{A}_1|\boldsymbol{y},\ \boldsymbol{A})$ 求平均。实际训练网络时，每个梯度步都会采样一次拆分。关于如何选择拆分分布，已有大量研究；拆分分布通常取决于具体问题。例如，在加速磁共振成像中，一种常用策略 [27] 是将加速掩码拆分为互不重叠的子掩模，通常将大部分低频信息保留在 $\boldsymbol{A}_1$ 中，以避免损失过多信息。

不难验证，平凡解 $f(\boldsymbol{y}_1,\ \boldsymbol{A}_1) = \boldsymbol{A}_1^{\dagger}\boldsymbol{y}_1$ 并非期望（MSPLIT）损失的全局极小值。因此，一个重要的问题是：该损失是否能近似监督损失？假设观测模型满足式 (3.2)，且测量值无噪声$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$，则有如下结果：

命题 3.2 （改编自 Daras 等人 [85]）。假设观测模型由(3.2) 给出，且测量值无噪声，即 $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$。多算子分割损失是加权监督损失的无偏估计量，即

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}, \boldsymbol{A}} \left\{ \mathcal{L}_{MSPLIT}(\boldsymbol{y}, \boldsymbol{A}, f) \right\} \\
&\quad = \mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{A}_1, \boldsymbol{x}} \left\{ \left(f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{x} \right)^{\top} \boldsymbol{Q}_{\boldsymbol{A}_1} \left(f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{x} \right) \right\}
\end{aligned}$$

其中

$$\boldsymbol{Q}_{\boldsymbol{A}_1} = \mathbb{E}_{\boldsymbol{A} | \boldsymbol{A}_1} \left\{ \boldsymbol{A}^{\top} \boldsymbol{A} \right\} \tag{3.5}$$

且期望损失的全局极小值为

$$f^*(\boldsymbol{y}_1, \boldsymbol{A}_1) = \boldsymbol{Q}_{\boldsymbol{A}_1}^{\dagger} \boldsymbol{Q}_{\boldsymbol{A}_1} \mathbb{E}_{\boldsymbol{x} | \boldsymbol{y}_1, \boldsymbol{A}_1} \{\boldsymbol{x}\} + (\boldsymbol{I} - \boldsymbol{Q}_{\boldsymbol{A}_1}^{\dagger} \boldsymbol{Q}_{\boldsymbol{A}_1}) v(\boldsymbol{y}_1) \tag{3.6}$$

其中 $v:\ \mathbb{R}^n \to \mathbb{R}^n$ 是任意函数。

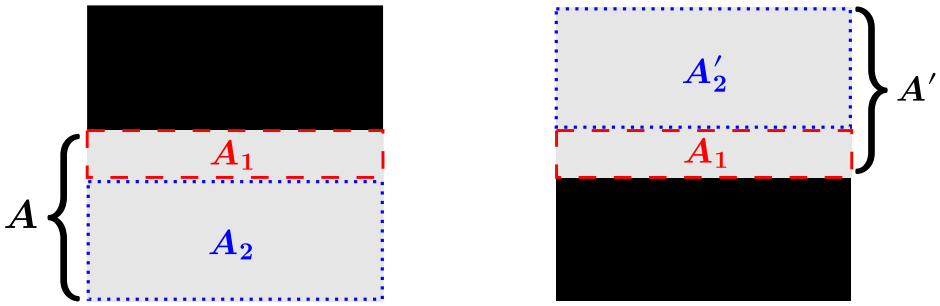

图 3.2：**命题 3.2 的示意图** 考虑一个简单的图像修复问题$\boldsymbol{A} = \operatorname{diag}(\boldsymbol{b})$，其中掩码为 $\boldsymbol{b}$。图像随机由两个灰色掩码之一进行测量，$\boldsymbol{b} \sim p(\boldsymbol{b})$。红色虚线表示的 $\boldsymbol{A}_1$ 拆分同时存在于两个掩码中，因此我们有 $\boldsymbol{Q}_{\boldsymbol{A}_1} = \boldsymbol{A}^\dagger \boldsymbol{A} + (\boldsymbol{A}')^\dagger \boldsymbol{A}'$ 是单位矩阵。因此，期望划分损失的全局极小值为条件均值 $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{A}_1 \boldsymbol{x}}\{\boldsymbol{x}\}$。

证明。

$$\mathbb{E}_{\boldsymbol{y},\boldsymbol{A}}\left\{\mathcal{L}_{\text{MSPLIT}}\left(\boldsymbol{y}, \boldsymbol{A}, f\right)\right\} \tag{3.7}$$

$$= \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{A}_1,\boldsymbol{y},\boldsymbol{A}}\left\{\|\boldsymbol{A} f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{y}\|^2\right\} \tag{3.8}$$

$$= \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{A}_1,\boldsymbol{x}}\left\{\mathbb{E}_{\boldsymbol{A}|\boldsymbol{A}_1}\left\{\frac{1}{m}\|\boldsymbol{A} f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{A}\boldsymbol{x}\|^2\right\}\right\} \tag{3.9}$$

$$= \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{A}_1,\boldsymbol{x}}\left\{\frac{1}{m}\left(f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{x}\right)^\top \boldsymbol{Q}_{\boldsymbol{A}_1}\left(f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{x}\right)\right\} \tag{3.10}$$

其中，第二行使用了观测模型的定义，第三行依赖于无噪声测量假设，最后一行则利用群使用了 $\boldsymbol{Q}_{\boldsymbol{A}_1}$ 的定义。由于这是加权 ℓ_2 损失，我们可以应用命题 C.2，得出该损失的任何极小化器均由 (3.6) 给出。

□

如果矩阵 $\boldsymbol{Q}_{\boldsymbol{A}_1}$ 对某个拆分 $\boldsymbol{A}_1$ 可逆，那么 $\boldsymbol{Q}_{\boldsymbol{A}_1}^\dagger \boldsymbol{Q}_{\boldsymbol{A}_1} = \boldsymbol{I}$，且期望下的极小化器（参见 (1.10) ）是唯一的，并由条件均值给出

$$f^*(\boldsymbol{y}_1, \boldsymbol{A}_1) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1,\boldsymbol{A}_1}\{\boldsymbol{x}\}. \tag{3.11}$$

因此，拥有可逆的 $\boldsymbol{Q}_{\boldsymbol{A}_1}$ 是获得类似于监督情形的条件均值估计器的充分条件，但并非必要条件，因为它依赖于对观测矩阵 $\boldsymbol{A}$进行一次拆分 $\boldsymbol{A}_1$ ，而我们可以在测试时对观测到的 $(\boldsymbol{y}, \boldsymbol{A})$ 的所有 J 可能拆分$\{(\boldsymbol{y}_1^{(j)}, \boldsymbol{A}_1^{(j)})\}_{j=1}^J$取平均。如果这个平均值

$$\bar{\boldsymbol{Q}}_{\boldsymbol{A}} = \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{Q}_{\boldsymbol{A}_1^{(j)}} = \mathbb{E}_{\boldsymbol{A}_1|\boldsymbol{A}}\left\{\boldsymbol{Q}_{\boldsymbol{A}_1}\right\} \tag{3.12}$$

是可逆的，因此我们可以在测试时计算平均预测：

$$f^{\text{test}}(\boldsymbol{y}, \boldsymbol{A}) = \frac{1}{J}\sum_{j=1}^{J} \bar{\boldsymbol{Q}}_{\boldsymbol{A}}^{-1} \boldsymbol{Q}_{\boldsymbol{A}_1^{(j)}}\, f(\boldsymbol{y}_1^{(j)}, \boldsymbol{A}_1^{(j)}) \tag{3.13}$$

其中，$\bar{\boldsymbol{Q}}_{\boldsymbol{A}}^{-1}\boldsymbol{Q}_{\boldsymbol{A}_1^{(j)}}$ 可视为加权项，其和为恒等算子，$\frac{1}{J}\sum_{j=1}^{J}\bar{\boldsymbol{Q}}_{\boldsymbol{A}}^{-1}\boldsymbol{Q}_{\boldsymbol{A}_1^{(j)}}=\boldsymbol{I}$。应用命题 3.2，我们可以证明，测试时的估计值近似于以下条件均值的平均值：

$$
\begin{aligned}
&f^{\text{test}}(\boldsymbol{y},\boldsymbol{A})\\
&\approx \frac{1}{J}\sum_{j=1}^{J}\bar{\boldsymbol{Q}}_{\boldsymbol{A}}^{-1}\boldsymbol{Q}_{\boldsymbol{A}_1^{(j)}}\left(\boldsymbol{Q}^{\dagger}_{\boldsymbol{A}_1^{(j)}}\boldsymbol{Q}_{\boldsymbol{A}_1^{(j)}}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1,\boldsymbol{A}_1^{(j)}}\{\boldsymbol{x}\}+(\boldsymbol{I}-\boldsymbol{Q}^{\dagger}_{\boldsymbol{A}_1^{(j)}}\boldsymbol{Q}_{\boldsymbol{A}_1^{(j)}})v(\boldsymbol{y}_1)\right)\\
&=\frac{1}{J}\sum_{j=1}^{J}\bar{\boldsymbol{Q}}_{\boldsymbol{A}}^{-1}\boldsymbol{Q}_{\boldsymbol{A}_1^{(j)}}\,\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1,\boldsymbol{A}_1^{(j)}}\{\boldsymbol{x}\}\\
&=\mathbb{E}_{\boldsymbol{A}_1|\boldsymbol{A}}\left\{\bar{\boldsymbol{Q}}_{\boldsymbol{A}}^{-1}\boldsymbol{Q}_{\boldsymbol{A}_1}\,\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1,\boldsymbol{A}_1}\{\boldsymbol{x}\}\right\}.
\end{aligned}
$$

一般而言，$\boldsymbol{Q}_{\boldsymbol{A}_1}$ 和 $\boldsymbol{Q}_{\boldsymbol{A}}$ 的闭式表达式难以处理，实践中通常会在测试时使用单个拆分，或使用多个等权重拆分 $f^{\text{test}}(\boldsymbol{y},\boldsymbol{A})=\frac{1}{J}\sum_{j=1}^{J}f(\boldsymbol{y}_1^{(j)},\boldsymbol{A}_1^{(j)})$，但这并无任何理论保证。然而，在某些特定情况下，例如对角算子 [86]，则可以对它们进行显式计算，如下面的例子所示。

例 *3.3*. 考虑一个带有随机掩码的图像修复问题 $\boldsymbol{A}=\mathit{diag}(\boldsymbol{b})$，其中 $b_i\sim\mathit{Ber}(p_i)$ 取值于 $\{0,1\}$，对于 $i=1,\ldots\ldots,n$。我们可以通过额外的掩码操作来拆分测量值，使得 $\boldsymbol{A}_1=\mathit{diag}(\boldsymbol{b}_1)$ 和 $\boldsymbol{A}_2=\mathit{diag}(\boldsymbol{b}_2)$ 也是掩膜算子，并且具有 $\boldsymbol{b}_1=\boldsymbol{b}\circ\boldsymbol{\omega}$ 和 $\boldsymbol{b}_2=\boldsymbol{b}\circ(\boldsymbol{1}-\boldsymbol{\omega})$，其中拆分掩码的采样方式为 $\omega_i\sim\mathit{Ber}(q_i)$。在这种情况下，$\boldsymbol{Q}_{\boldsymbol{A}_1}$ 和 $\boldsymbol{Q}_{\boldsymbol{A}}$ 都是对角矩阵，因为它们是对角矩阵的平均值。由于掩码采样分布在像素之间具有可分离性，我们可以将分析集中于单个条目 $i\in\{1,\ldots,n\}$。令$[\boldsymbol{A}_2]_{i,\,i}=b_i(1-\omega_i)$，其中$[\boldsymbol{A}_1]_{i,\,i}=b_{1,i}$ 固定，则 $\boldsymbol{Q}_{\boldsymbol{A}_1}$ 的对角元素为

$$[\boldsymbol{Q}_{\boldsymbol{A}_1}]_{i,i}=\mathbb{E}_{b_i,\omega_i|b_i\omega_i=b_{1,i}}\{b_i\}$$

由于 $\mathbb{E}_{b_i,\,\omega_i|b_i\omega_i=0}\{b_i\}=\frac{p_i}{(1-q_i)\,1-p_iq_i}$ 和 $\mathbb{E}_{b_i,\,\omega_i|b_i\omega_i=1}\{b_i\}=1$，我们有

$$[\boldsymbol{Q}_{\boldsymbol{A}_1}]_{i,i}=\begin{cases}1 & if\ b_{1,i}=1\\ \frac{p_i(1-q_i)}{1-p_iq_i} & otherwise\end{cases}.$$

因此，如果对于所有 $i=1,\ldots,n$，$p_i>0$ 且 $q_i<1$，那么对于所有拆分 $\boldsymbol{A}$，$\boldsymbol{Q}_{\boldsymbol{A}_1}$ 都是可逆的 1.

现在我们可以计算 $\boldsymbol{A}^{-}$，满足$[\boldsymbol{A}]_{i,\,i}=b_i$，$\boldsymbol{Q}_{\boldsymbol{A}}$ 的对角元素为

$$[\bar{\boldsymbol{Q}}_{\boldsymbol{A}}]_{i,i}=\begin{cases}\frac{p_i(1-q_i)}{1-p_iq_i} & if\ b_i=0\\ \frac{p_i(1-q_i)^2}{1-p_iq_i} & if\ b_i=1\end{cases}.$$

再次，如果对于所有 $i=1,\ p_i>0$和 $q_i<1$，……，n，都能得到一个可逆的 $\boldsymbol{Q}_{\boldsymbol{A}}$，则适用于所有可能的修复掩码[1] $\boldsymbol{A}$ *[51]*。

[1]虽然这里将其表述为一个多算子修复问题，但它最初在 [51] 中被提出，作为 Noisier2Noise 自监督去噪技术的乘法版本。

这个例子也适用于具有 $\boldsymbol{A} = \operatorname{diag}(\boldsymbol{b})\,\boldsymbol{F}$ 的问题，其中 $\boldsymbol{F}$ 是一个固定可逆矩阵，例如加速磁共振成像 [77]，中的离散傅里叶变换，且对角线值按 $\boldsymbol{b} \sim p(\boldsymbol{b})$ 采样。此外，在这些问题中，如果 $\boldsymbol{Q}_{\boldsymbol{A}_1}$ 对所有可能的拆分 $\boldsymbol{A}_1$ 都是可逆的，则可以考虑分割损失 [77] 的加权版本：

$$\mathcal{L}_{\text{MSPLIT}}\left(\boldsymbol{y}, \boldsymbol{A}, f\right) = \mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{A}_1 | \boldsymbol{y}, \boldsymbol{A}} \left\{ \frac{1}{m} \| \boldsymbol{Q}_{\boldsymbol{A}_1}^{-\frac{1}{2}} \left(\boldsymbol{A} f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{y} \right) \|^2 \right\}$$

以获得监督损失的无偏估计值，即

$$\mathbb{E}_{\boldsymbol{y}, \boldsymbol{A}} \left\{ \mathcal{L}_{\text{MSPLIT}}\left(\boldsymbol{y}, \boldsymbol{A}, f\right) \right\} = \mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{A}_1} \left\{ \frac{1}{n} \| f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{x} \|^2 \right\} + \text{const.}$$

这一额外加权不会改变期望损失的全局极小值，但在数据集有限的实际情况下，它可以提升所学习到的 f 的性能。

带噪测量的分裂 如果测量值含有噪声，会发生什么？我们可以通过将 (MSPLIT) 损失分解为以下形式来分析这种情况：

$$\mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{A}_1, \boldsymbol{A}_2 | \boldsymbol{y}, \boldsymbol{A}} \left\{ \mathcal{L}_{\text{MC}}(\boldsymbol{y}_1, \boldsymbol{A}_1 \circ f) + \frac{1}{m} \| \boldsymbol{A}_2 f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{y}_2 \|^2 \right\}$$

其中，第一项用于惩罚输入 $\boldsymbol{y}_1$ 的测量一致性，与 (MC) 中相同；第二项则与对未观测部分 $\boldsymbol{y}_2$ 的预测有关。如果我们的带噪测量数据在各测量值之间可分离，即 $p(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{a}_i^\top \boldsymbol{x})$，其中 $\boldsymbol{a}_i \in \mathbb{R}^n$ 表示 $\boldsymbol{A}$ 的第 i 行，那么第二项与无噪声情形等价，因为在给定底层信号 $\boldsymbol{x}$ 的条件下，$\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 相互独立，即

$$\mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{y}_2} \left\{ \| \boldsymbol{A}_2 f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{y}_2 \|^2 \right\} = \mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{x}} \left\{ \| \boldsymbol{A}_2 f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{A}_2 \boldsymbol{x} \|^2 \right\}$$

根据命题 2.1。

然而，测量一致性项并不等同于无噪声情形，因为输入和目标中的噪声相同：

$$\mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{A}_1} \left\{ \mathcal{L}_{\text{MC}}(\boldsymbol{y}_1, \boldsymbol{A}_1 \circ f) \right\} \neq \mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{A}_1, \boldsymbol{x}} \left\{ \frac{1}{m} \| \boldsymbol{A}_1 f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{A}_1 \boldsymbol{x} \|^2 \right\}.$$

为了考虑测量数据中的噪声，我们需要将一致性损失 $\mathcal{L}_{\text{MC}}$替换为第2章介绍的、适用于含噪数据的自监督损失函数之一，具体选择取决于我们对噪声分布的了解程度：

a) 如果测量值包含来自已知分布（例如泊松分布、高斯分布等）的噪声，我们可以使用 (GR2R) 或 (SURE) 损失，即

$$\begin{aligned} &\mathcal{L}_{\text{GR2R-MSPLIT}}\left(\boldsymbol{y}, \boldsymbol{A}, f\right) = \\ &\mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{A}_1, \boldsymbol{A}_2 | \boldsymbol{y}, \boldsymbol{A}} \left\{ \mathcal{L}_{\text{GR2R}}\left(\boldsymbol{y}_1, \boldsymbol{A}_1 \circ f\right) + \frac{1}{m} \| \boldsymbol{A}_2 f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{y}_2 \|^2 \right\} \end{aligned}$$

正如我们在第2.5节所见，第一项是无噪声测量一致性的无偏估计量，即，

$$\mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{A}_1 | \boldsymbol{x}} \left\{ \mathcal{L}_{\text{GR2R}}\left(\boldsymbol{y}_1, \boldsymbol{A}_1 \circ f\right) \right\} = \mathbb{E}_{\boldsymbol{y}_1, \boldsymbol{A}_1 | \boldsymbol{x}} \left\{ \frac{1}{m} \| \boldsymbol{A}_1 f(\boldsymbol{y}_1, \boldsymbol{A}_1) - \boldsymbol{A}_1 \boldsymbol{x} \|^2 \right\}$$

因此，我们采用与命题 3.2 类似的分析，得出该损失的极小化器近似于 $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1,\ \boldsymbol{A}_1}\{\boldsymbol{x}\}$ 若 $\boldsymbol{Q}_{\boldsymbol{A}_1}$ 可逆且在测试时使用单次拆分；或者近似于 $\mathbb{E}_{\boldsymbol{A}_1|\boldsymbol{A}}\left\{\bar{\boldsymbol{Q}}_{\boldsymbol{A}}^{-1}\boldsymbol{Q}_{\boldsymbol{A}_1}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1,\ \boldsymbol{A}_1}\{\boldsymbol{x}\}\right\}$ 若 $\bar{\boldsymbol{Q}}_{\boldsymbol{A}}$ 可逆且在测试时使用多次拆分。

b) 如果测量值具有未知噪声分布，且可以假设该分布在各测量值之间可分离，那么我们只需移除测量一致性损失。由此得到的损失称为 SSDU [27], ，可视为 (SPLIT) 的多算子扩展：

$$\mathcal{L}_{\text{SSDU}}\left(\boldsymbol{y},\boldsymbol{A},f\right)=\mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{A}_1,\boldsymbol{A}_2|\boldsymbol{y},\boldsymbol{A}}\left\{\frac{1}{m}\|\boldsymbol{A}_2 f(\boldsymbol{y}_1,\boldsymbol{A}_1)-\boldsymbol{y}_2\|^2\right\}$$

在这种情况下，我们可以推导出与命题 3.2 类似的结果，但其中 $\boldsymbol{Q}_{\boldsymbol{A}_1}=\mathbb{E}_{\boldsymbol{A}_2|\boldsymbol{A}_1}\left\{\boldsymbol{A}_2^\top\boldsymbol{A}_2\right\}$ 取代了 $\mathbb{E}_{\boldsymbol{A}|\boldsymbol{A}_1}\left\{\boldsymbol{A}^\top\boldsymbol{A}\right\}$。由于 $\boldsymbol{A}_2$ 和 $\boldsymbol{A}_1$ 互不重叠， $\boldsymbol{Q}_{\boldsymbol{A}_1}$ 不包含 $\boldsymbol{A}_1^\top$的值域，因此对于任何拆分， $\boldsymbol{Q}_{\boldsymbol{A}_1}$ 都不可逆。尽管如此，我们仍可以像 (3.13) 那样在测试时对多个拆分取平均，使得$\bar{\boldsymbol{Q}}_{\boldsymbol{A}}=\frac{1}{J}\sum_{j=1}^{J}\boldsymbol{Q}_{\boldsymbol{A}_1^{(j)}}$ 变为可逆；测试时估计器 f^{test}则近似于条件估计器 $\mathbb{E}_{\boldsymbol{A}_1|\boldsymbol{A}}\left\{\bar{\boldsymbol{Q}}_{\boldsymbol{A}}^{-1}\boldsymbol{Q}_{\boldsymbol{A}_1}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1,\ \boldsymbol{A}_1}\{\boldsymbol{x}\}\right\}$。

跨算子一致性 基于拆分的方法的主要缺点是，它们可能没有将所有可供网络使用的测量信息作为输入。多算子成像（MOI）损失 [87] 通过假设每个成像问题都是可逆的来解决这一问题，即对于每个 $\boldsymbol{A}$，存在一个 f^*，使得 $\boldsymbol{x}\approx f^*(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{A})$，然后强制不同算子之间保持估计器一致性。也就是说，对于属于 $p(\boldsymbol{A})$ 的任意一对算子 $\boldsymbol{A}\neq\boldsymbol{A}'$，都有 $f(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{A})\approx f(\boldsymbol{A}'\boldsymbol{x},\ \boldsymbol{A}')$：

$$\mathcal{L}_{\text{MOI}}\left(\boldsymbol{y},\boldsymbol{A},f\right)=\mathbb{E}_{\boldsymbol{A}'\sim p(\boldsymbol{A})}\left\{\frac{1}{n}\|f\left(\boldsymbol{A}'f(\boldsymbol{y},\boldsymbol{A}),\boldsymbol{A}'\right)-f(\boldsymbol{y},\boldsymbol{A})\|^2\right\}\tag{MOI}$$

将该损失与测量一致性共同进行最小化，可得到以下总损失：

$$\mathcal{L}\left(\boldsymbol{y},\boldsymbol{A},f\right)=\mathcal{L}_{\text{MC}}(\boldsymbol{y},\boldsymbol{A}\circ f)+\lambda\,\mathcal{L}_{\text{MOI}}\left(\boldsymbol{y},\boldsymbol{A},f\right)\tag{3.14}$$

其中， $\lambda>0$ 是一个权衡超参数。很容易验证，只要对于某个 $\boldsymbol{A}\neq\boldsymbol{A}'$ 有 $\boldsymbol{A}'\boldsymbol{A}^\dagger\neq\boldsymbol{A}\boldsymbol{A}'^\dagger$，平凡重建 $f(\boldsymbol{y},\ \boldsymbol{A})=\boldsymbol{A}^\dagger\boldsymbol{y}$就不是该损失的极小化器。在含噪测量的情况下，可以根据我们对噪声分布的了解程度，以第2章中介绍的任意损失替换 $\mathcal{L}_{\text{MC}}$，其方式与（MSPLIT）的鲁棒扩展类似。

虽然 MOI 损失不像我们在分裂损失中看到的那样，与监督损失具有良好的等价性，但它能够利用所有可用的测量值，并且可以由第 3.4 节讨论的模型可辨识性理论提供理论依据。

3.2 利用变换不变性

在只有一个不可逆前向算子的应用中，Chen 等人 [26] 表明，假设信号分布对某个变换群 $\boldsymbol{T}_g$保持不变，我们仍然可以在其零空间中学习： $\mathbb{R}^n\to\mathbb{R}^n$ 对于 $g=1,\ \ldots\ldots,\ G$，例如对 G进行旋转或

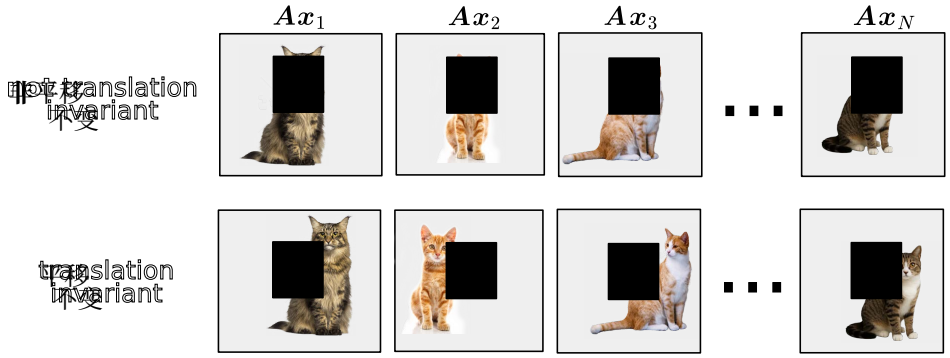

图3.3：**等变成像原理示例**。考虑这样一个玩具场景：我们通过一个固定的掩膜算子对猫进行成像。如果猫分布对平移不具有不变性（上排），猫在图像中就有一个规范位置，而我们永远无法观测到分布中的一部分（猫头）。如果分布具有不变性（下排），猫的位置就不总是居中；尽管从未观测到被掩蔽的像素，我们仍然可以学习猫分布（包括猫头）。

平移变换，作用于由 n 个像素构成的离散化图像。从数学角度来说，如果分布的支撑集具有不变性，那么对于每个变换 $\boldsymbol{T}_g$，若 $\boldsymbol{x} \in \text{supp}(p_{\boldsymbol{x}})$，则$\boldsymbol{T}_g(\boldsymbol{x}) \in \text{supp}(p_{\boldsymbol{x}})$。这一条件比要求分布具有不变性更宽松，即对于所有变换 $\boldsymbol{T}_g$和信号 $\boldsymbol{x}$，均有 $p(\boldsymbol{T}_g(\boldsymbol{x})) = p(\boldsymbol{x})$。基于这一不变性，我们得到以下关键观察：

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A} \circ \boldsymbol{T}_g \circ \boldsymbol{T}_g^{-1}(\boldsymbol{x}) = \boldsymbol{A} \circ \boldsymbol{T}_g(\boldsymbol{x}') \tag{3.15}$$

其中 $\boldsymbol{x}' = \boldsymbol{T}_g^{-1}(\boldsymbol{x})$ 也属于信号集。因此，不变性这一性质隐式地通过一族不同的算子 $\{\boldsymbol{A}_g = \boldsymbol{A} \circ \boldsymbol{T}_g\}_{g=1}^{G}$ 为我们提供了额外的虚拟观测，使我们的设置与第3.1节中的设置相似，但受到的约束略多（图3.3对此进行了说明）。我们将看到，这使我们能够利用一种称为等变性的强大性质。

等变性 等变性这一概念在机器学习中得到了广泛研究，尤其是在将对称性融入神经网络架构的背景下 [88]。通俗地说，如果对输入应用某种变换后，输出也会发生相应的变换，则称函数对某个变换群具有等变性。更形式化地说，定义如下：

定义 3.4 (等变性 [89])。 函数 $\phi\colon \mathbb{R}^n \to \mathbb{R}^m$ 对群作用 $\tilde{}$ $\tilde{}$具有等变性，当且仅当： $\boldsymbol{T}_g(\phi(\boldsymbol{x})) = \phi(\boldsymbol{T}_g(\boldsymbol{x}))$ 对所有 $g = 1, \ldots, G$均成立，其中 $\boldsymbol{T}_g\colon \mathbb{R}^n \to \mathbb{R}^n$和 $\boldsymbol{T}_g\colon \mathbb{R}^m \to \mathbb{R}^m$ 是满足群性质的变换（若 $n = m$，则二者可能相同）。

在逆问题的语境中，由于重建函数 f 与前向算子之间存在微妙的相互作用，我们无法直接应用这一定义。我们可以改用一种更具体的定义，将这种相互作用考虑在内：

定义 3.5 (等变重建器 [90])。 如果 $f(\boldsymbol{y}, \boldsymbol{A})$ 是等变重建器，则称其为

$$f(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{T}_g) = \boldsymbol{T}_g^{-1} f(\boldsymbol{y}, \boldsymbol{A}),\ \forall \boldsymbol{y} \in \mathbb{R}^m, \forall g, \forall \boldsymbol{A} \in \mathbb{R}^{m \times n}. \tag{3.16}$$

下面我们将说明，只要常用重建器的构建模块按照定义 3.4 的意义是等变的，或者信号分布 $p_{\boldsymbol{x}}$ 是不变的，这些重建器就是等变的。证明见 [90]。

1. **反投影网络**。 如果图像到图像映射 ϕ 是等变的，那么反投影网络 $f(\boldsymbol{y}, \boldsymbol{A}) = \phi(\boldsymbol{A}^\dagger \boldsymbol{y})$ 就是等变重建器。

2. **展开网络**。 对于任意步长 $\tau \in \mathbb{R}$，如果近端算子 ϕ_1、……、ϕ_k 是等变的，那么式 (1.6) 中定义的展开网络就是等变重建器。

3. **雷诺平均**。 设 $f(\boldsymbol{y}, \boldsymbol{A})$ 是任意一个（非等变）重建器，则

$$f(\boldsymbol{y}, \boldsymbol{A}) = \frac{1}{G} \sum_{g=1}^{G} \boldsymbol{T}_g \tilde{f}(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{T}_g) \tag{3.17}$$

是一个等变重建器。

4. **变分方法**。 变分估计值

$$f(\boldsymbol{y}, \boldsymbol{A}) = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x})\|^2 + \rho(\boldsymbol{x}). \tag{3.18}$$

若正则化项 ρ 是不变分布，则该重建器是等变重建器。特别地，对于具有各向同性高斯噪声的观测，且 $\rho(\boldsymbol{x}) = -\log p_{\boldsymbol{x}}(\boldsymbol{x})$，那么 f 对应最大后验估计，并且当 $p_{\boldsymbol{x}}$ 是不变分布时，该估计具有等变性。

5. **MMSE**。MMSE估计值

$$f(\boldsymbol{y}, \boldsymbol{A}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{A}} \{\boldsymbol{x}\} \tag{3.19}$$

若 $p_{\boldsymbol{x}}$ 是不变分布，则该重建器是等变重建器。

前三个示例展示了如何构建实用的等变重建器，而后两个示例则表明，当信号分布是不变分布时，最优贝叶斯重建器具有等变性。

此时，人们可能会问，定义 3.5 中的等变重建器性质，加上简单的测量一致性损失（MC），对于使用单个不可逆算子 $\boldsymbol{A}$ 进行学习是否充分。答案是否定的，因为简单的线性伪逆 $f(\boldsymbol{y}, \boldsymbol{A}) = \boldsymbol{A}^\dagger \boldsymbol{y}$ 是一种能够最小化测量一致性损失的等变重建器，但无法恢复 $\boldsymbol{A}$ 零空间中的任何信息。与多算子情形一样，我们可以使用分割损失，或强制执行跨变换一致性，从而进一步约束解空间。

等变拆分损失 我们可以采用与上一节中的（MSPLIT）损失类似的方法，但这次使用由变换关联的算子，从而得到等变拆分损失 [90]:

$$\begin{aligned}\mathcal{L}_{\text{ESPLIT}}(\boldsymbol{y}, f) &= \frac{1}{G}\sum_{g=1}^{G}\mathcal{L}_{\text{MSPLIT}}(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{T}_g, f) \\ &= \frac{1}{G}\sum_{g=1}^{G}\mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{A}_1|\boldsymbol{y},\boldsymbol{A}}\left\{\|\boldsymbol{A}\boldsymbol{T}_g f(\boldsymbol{y}_1, \boldsymbol{A}_1\boldsymbol{T}_g) - \boldsymbol{y}\|^2\right\}\end{aligned} \tag{ESPLIT}$$

其中， $\boldsymbol{A}_1 \sim p(\boldsymbol{A}_1|\boldsymbol{A})$ 是 $\boldsymbol{A}$ 的随机拆分。与多算子拆分情形下的命题 3.2 一样，在信号分布不变$p_{\boldsymbol{x}}$的假设下最小化该损失，即可恢复最小均方误差估计器，前提是 $\boldsymbol{Q}_{\boldsymbol{A}_1} = \sum_{g\in\mathcal{S}_{\boldsymbol{A}_1}}\boldsymbol{T}_g^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{T}_g$，其中 $\mathcal{S}_{\boldsymbol{A}_1}$是¯由所有满足以下条件的变换构成： $\boldsymbol{A}_1$ 可能作为 $\boldsymbol{A}\boldsymbol{T}_g$的随机拆分出现，或者 $\boldsymbol{Q}_{\boldsymbol{A}} = \mathbb{E}_{\boldsymbol{A}_1|\boldsymbol{A}}\left\{\boldsymbol{Q}_{\boldsymbol{A}_1}\right\}$ 满秩。由于对整个变换群求和的计算开销可能很大，因此可以在每次训练迭代中随机采样一个变换来计算损失。与多算子分裂损失一样，测试时估计器可以按照式 (3.13)，通过对多个拆分和变换取平均来计算；此外，也可以用 (GR2R) 或 (SURE) 替换用于确保与输入 $\boldsymbol{y}_1$ 一致的项，以类似方式推导出具有噪声鲁棒性的损失扩展。

如果我们有意将 f 设计为等变重建器，例如使用搭配等变去噪器架构的反投影网络，那么式 (ESPLIT) 中的等变拆分损失就会简化为使用单个算子的式 (MSPLIT) 中的分割损失，正如下述命题所述:

命题 3.6. 如果 f 是等变重建器，那么 (ESPLIT) 等价于分割损失

$$\mathcal{L}_{\text{ESPLIT}}(\boldsymbol{y}, f) = \mathcal{L}_{\text{MSPLIT}}(\boldsymbol{y}, \boldsymbol{A}, f). \tag{3.20}$$

证明直接来自定义 3.5 中对等变重建器的定义。该结果表明，即使在测量值不完整的情况下，只要网络在构造上是等变重建器，使用单个算子的简单分割损失进行训练也能奏效。

跨变换一致性 沿用式 (MOI) 背后的相同思路，假设成像问题近似可逆，即存在一个函数 f^*，使得对于所有 $\boldsymbol{x} \in \in\text{supp}(p_{\boldsymbol{x}})$， $\boldsymbol{x} \approx f^*(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{A})$ 成立。那么，如图 3.4 所示，一个良好的重建函数 f 应满足：完整的成像/感测系统 $f\circ\boldsymbol{A}$ 近似等变

$$f(\boldsymbol{A}\boldsymbol{T}_g(\boldsymbol{x}), \boldsymbol{A}) \approx \boldsymbol{T}_g f(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}). \tag{3.21}$$

等变性　　抗成像（EI）损失旨在通过训练 [26] 来强制系统具备等变性　　:

$$\mathcal{L}_{\text{EI}}(\boldsymbol{y}, f) = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{n}\|f(\boldsymbol{A}\boldsymbol{T}_g f(\boldsymbol{y}, \boldsymbol{A}), \boldsymbol{A}) - \boldsymbol{T}_g f(\boldsymbol{y}, \boldsymbol{A})\|^2. \tag{EI}$$

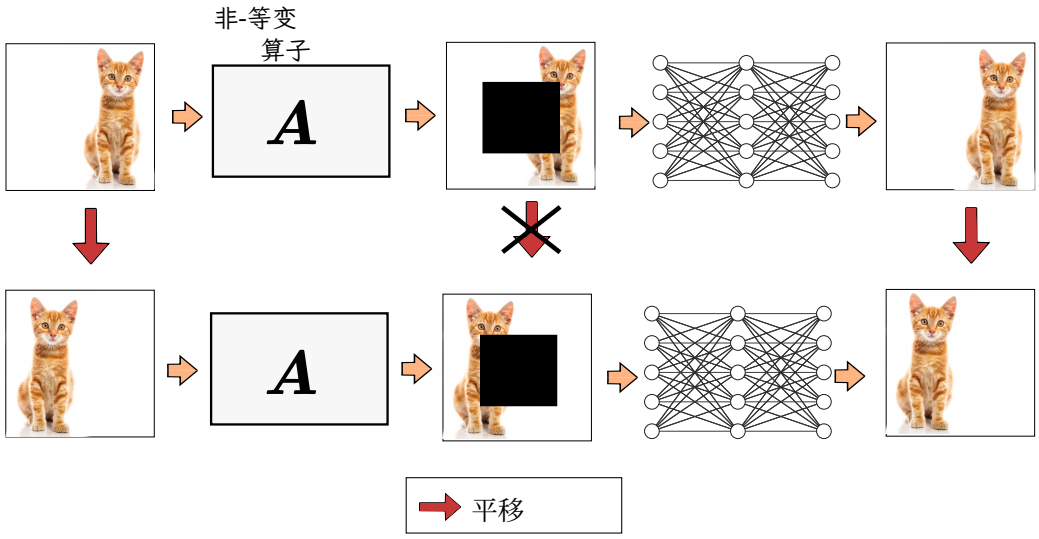

图 3.4：等变成像损失试图强制成像系统对变换群（此处为平移）满足等变性。这种损失可以使模型在算子的零空间（此处为缺失像素）中进行学习，因为这些变换会隐式地平移零空间。

等变系统条件比要求 f 成为定义 3.5 中的等变重建器更为严格，并且在某些情况下，只有当成像问题可逆时才会完全成立[2]。因此，即使使用等变重建器，EI损失仍然是必要的，因为它进一步约束了解空间。

与 MOI 损失一样，这种损失应与测量一致性损失一同最小化：

$$\mathcal{L}(\boldsymbol{y}, f) = \mathcal{L}_{\mathrm{MC}}(\boldsymbol{y}, f) + \lambda \mathcal{L}_{\mathrm{EI}}(\boldsymbol{y}, f) \tag{3.22}$$

其中，$\lambda > 0$ 是一个超参数，而在测量含噪的情况下，$\mathcal{L}_{\mathrm{MC}}$ 可以替换为第2章中介绍的任意损失函数，例如 (SURE) [62]。

关于 $\boldsymbol{A}$ 线性变换 旋转或平移等典型变换都是线性算子，我们可以将它们视为一组可逆矩阵 $\{\boldsymbol{T}_g\}_{g=1}^{G}$。在这种情况下，我们可以仿照上一节的多算子设置进行类似分析，其中算子定义为 $\boldsymbol{A}_g = \boldsymbol{A}\boldsymbol{T}_g$。正如我们在第3.1节中所见，我们需要 $\boldsymbol{A}\boldsymbol{T}_g$ 具有不同的零空间，换言之，$\sum_g \boldsymbol{T}_g^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{T}_g$ 是可逆矩阵。一个重要的推论是，前向算子不 应对变换群满足等变性：

命题3.7 (Tachella 等人 [79])。 如果 $\boldsymbol{A}^\top\boldsymbol{A}$ 对变换群$\{\boldsymbol{T}_g\}_{g=1}^{G}$满足等变性，那么所有算子 $\boldsymbol{A}\boldsymbol{T}_g$ 都具有相同的零空间。

等变性意味着，对于所有 $g = 1, \ldots\ldots, G$，$\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{T}_g = \boldsymbol{T}_g\boldsymbol{A}^\top\boldsymbol{A}$，因此，对于所有 g，$\boldsymbol{A}$与 $\boldsymbol{A}\boldsymbol{T}_g$ 具有相同的零空间。

[2]例如，在不变分布 $p_{\boldsymbol{x}}$ 下，MMSE估计始终是一个等变重建器，但只有当成像问题可逆时，它才满足这一更强条件，即 $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y},\ \boldsymbol{A}}\{\boldsymbol{x}\} = \boldsymbol{x}$。

如下例所示，不能想当然地认为 $\boldsymbol{A}$ 不具有等变性；在任何给定情境中，群作用的选择都取决于前向算子的具体性质：

- 压缩随机算子 $\boldsymbol{A} \sim \mathcal{N}\left(\mathbf{0}, \ \frac{1}{m}\boldsymbol{I}\right)$ 具有 $m < n$ 时，除幅度缩放 $\boldsymbol{T}_g = g\boldsymbol{I}$ 外，几乎必然不具有任何（非平凡）变换集合 $\{\boldsymbol{T}_g\}_{g=1}^{G}$ 下的等变性。

- 图像修复 $\boldsymbol{A} = \text{diag}(\boldsymbol{b})$ 通常不具有像素平移等变性，因为缺失像素在图像中的位置是固定的。

- 可对角化为 $\boldsymbol{A} = \boldsymbol{Q}\,\text{diag}\,(\boldsymbol{b})\,\boldsymbol{F}$ 的算子，其中 $\boldsymbol{F}$ 是傅里叶变换， $\boldsymbol{Q}$ 是任意酉基，例如任何模糊操作或磁共振成像，都具有像素平移或平移等变性。然而，如果模糊具有特定方向，或者磁共振成像掩码采用笛卡尔下采样模式进行加速，这些算子就不具有旋转等变性。

- 各向同性模糊和使用抗混叠滤波器的下采样对旋转和平移均具有等变性。然而，它们不具有缩放变换等变性，而缩放变换可用于学习重建缺失的高频 [91]。

尽管 EI损失与监督损失之间并没有很强的等价性，但它与 MOI 有着相同的动机，而 MOI 源自第 3.4 节讨论的模型可辨识性理论。

3.3 从不完整测量值中学习生成模型

到目前为止，我们已经介绍了多种能够近似该问题后验均值的损失函数。接下来，我们进一步探讨能否仅凭不完整的测量数据来识别信号分布 $p_{\boldsymbol{x}}$。本节将介绍迄今为止已有研究探索的主要方法，包括生成对抗网络（GANs）和扩散模型。

生成对抗网络在无条件生成的场景中，我们旨在训练一个生成器 $f : \mathbb{R}^k \mapsto \mathbb{R}^n$ ，将服从简单分布（如各向同性高斯分布）的潜变量 $\boldsymbol{z} \in \mathbb{R}^k$ 映射为图像分布 $p_{\boldsymbol{x}}$ 的样本 $\boldsymbol{x} \in \mathbb{R}^n$ 。我们可以尝试匹配与生成分布 $p_{\hat{\boldsymbol{y}}}$ 相关联的测量值分布，其定义为

$$\begin{cases} \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \ \boldsymbol{A} \sim p(\boldsymbol{A}) \\ \hat{\boldsymbol{x}} = f(\boldsymbol{z}) \\ \hat{\boldsymbol{y}} \sim p(\boldsymbol{y}|\boldsymbol{A}\hat{\boldsymbol{x}}) \end{cases} \tag{3.23}$$

使其符合观测测量值的分布 $p_{\boldsymbol{y}}$。AmbientGAN [82]提出训练 Wasserstein生成对抗网络 [92] 来实现这一目标：

$$\min_{f} \max_{d} \ \mathbb{E}_{\boldsymbol{z},\boldsymbol{A}} \left\{ d(\boldsymbol{A}f(\boldsymbol{z}), \boldsymbol{A}) \right\} - \mathbb{E}_{\boldsymbol{y},\boldsymbol{A}} \left\{ d(\boldsymbol{y}, \boldsymbol{A}) \right\} \tag{3.24}$$

其中 d： $\mathbb{R}^n \times \mathbb{R}^{m\times n} \mapsto \mathbb{R}$ 是一个与生成器联合训练的 1-Lipschitz 判别器网络，它可以整合有关前向算子的信息。(3.24) 中的第一项促使 f 通过欺骗判别器来生成逼真的测量值，而第二项则训练 d 将真实测量值与生成的测量值区分开来。

这一思路也可以扩展到条件生成对抗网络模型 [93,94], ，其中生成器 $f(\boldsymbol{z}、\ \boldsymbol{y}、\ \boldsymbol{A})$ 以测量值和算子为条件，旨在生成具有不同潜变量的后验样本 $\boldsymbol{z}$。

扩散模型 AmbientDiffusion [85] 将 AmbientGAN 的思路扩展到扩散模型，适用于来自多个前向算子的< b>无噪声但不完整的测量值。在此设置中，通过向输入测量值 $\boldsymbol{y}_1$添加标准差为 σ 的合成高斯噪声，在不同噪声级别下使用（MSPLIT）损失训练重建网络，从而近似条件估计器 $f(\boldsymbol{y}_1、\ \boldsymbol{A}_1、\ \sigma) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1+\sigma\epsilon、\ \boldsymbol{A}_1}\{\boldsymbol{x}\}$。网络训练完成后，固定一个随机拆分 $\boldsymbol{A}_1$，并将 f 用作运行扩散所需的高斯去噪器$\mathbb{E}_{\boldsymbol{x}|\boldsymbol{x}+\sigma\epsilon}\{\boldsymbol{x}\}$ 的代理指标，即可获得信号分布 [85] 或后验分布 [95] 的样本。

在类似的多算子设定中，Rozet 等人 [43] 提出了另一种基于期望最大化的方法。该方法在以下两步之间迭代：i）使用固定的去噪器网络，通过扩散方法生成后验样本；随后 ii）利用生成的样本更新去噪器。初始化时，去噪器被设定为从高斯分布中采样。

3.4 模型识别理论

尽管我们已经看到，在某些假设下，分割损失可以近似监督损失 ℓ_2，但仍有必要探究：学习生成模型这一更困难的任务是否有明确定义？换言之，如果不对 $p_{\boldsymbol{x}}$ 作任何额外假设，我们能否仅凭不完整测量值 [3], 唯一确定 $p_{\boldsymbol{x}}$？我们可以通过分析信号分布特征函数 $\phi_{\boldsymbol{x}}(\boldsymbol{\omega})$ 的可用信息来回答这个问题。对于每个算子 $\boldsymbol{A} \sim p(\boldsymbol{A})$，我们有

$$\phi_{\boldsymbol{y}|\boldsymbol{A}}(\tilde{\boldsymbol{\omega}}) = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{A}}\left\{e^{\mathrm{i}\tilde{\boldsymbol{\omega}}^\top \boldsymbol{y}}\right\} = \mathbb{E}_{\boldsymbol{x}}\left\{e^{\mathrm{i}\tilde{\boldsymbol{\omega}}^\top \boldsymbol{A}\boldsymbol{x}}\right\} = \mathbb{E}_{\boldsymbol{x}}\left\{e^{\mathrm{i}(\boldsymbol{A}^\top\tilde{\boldsymbol{\omega}})^\top \boldsymbol{x}}\right\} \tag{3.25}$$

$$= \phi_{\boldsymbol{x}}(\boldsymbol{\omega} = \boldsymbol{A}^\top\tilde{\boldsymbol{\omega}}) \text{ with } \tilde{\boldsymbol{\omega}} \in \mathbb{R}^m \tag{3.26}$$

因此，对于每个算子，我们都能在 $\boldsymbol{A}^\top$ 的值域内观测到 $\boldsymbol{x}$ 的特征函数；该值域是 $\mathbb{R}^n$ 中一个 m 维线性子空间，如 $m < n$ 所示。由于一个分布由其特征函数唯一确定，我们需要观测无穷多个算子，才能完全覆盖 $\boldsymbol{x}$ 的特征函数！这一观察可追溯至 Cramer 和 Wold 的研究，其重点是情形 $m = 1$:

定理 3.8 (克拉默和沃尔德 [96])。 概率分布 $p(\boldsymbol{x})$ 由其一维投影的全体 唯一确定。

[3]本部分不考虑噪声，因为我们可以通过分两步推理来将噪声纳入考虑：首先，针对每个前向算子，应用命题 2.9 中关于含噪数据的识别结果，从含噪测量分布中识别出无噪测量分布；其次，从无噪测量分布集合中识别出信号分布 [79]。

该定理指出，如果我们观测到（未配对的）标量测量值 $y = \boldsymbol{a}^\top \boldsymbol{x} \in \mathbb{R}$ with$\boldsymbol{x} \sim p_{\boldsymbol{x}}(\boldsymbol{x})$，以及覆盖所有可能投影空间（即在 $\mathbb{R}^n$ 中稠密）的测量分布 $\boldsymbol{a} \sim p(\boldsymbol{a})$，就能唯一确定 $p_{\boldsymbol{x}}$。遗憾的是，这一结果在实践中并不太实用，因为它只有在观测到$\mathbb{R}^n$中的所有可能的 投影时才成立；然而，我们通常只能通过有限个算子获得观测，因此该分布在 $\mathbb{R}^n$ 中并不稠密（例如，在图像修复或加速磁共振成像中，掩码的变化范围有限）。

3.4.1 低维分布的模型识别

为了在有限个算子的更现实情形下获得充分条件，我们需要对信号分布 $p_{\boldsymbol{x}}$考虑一些额外的假设。下面我们将看到，假设信号分布是低维的，或者换言之，假设 $p_{\boldsymbol{x}}$ 的支撑集是 $\mathbb{R}^n$的一个低维子集，对于获得模型识别保证是充分的。低维性是成像和数据科学中的一个常见假设，通常称为流形假设 [97]。

例 3.9。我们可以考虑一个简单的例子来说明低维性为何以及如何发挥作用： $p_{\boldsymbol{x}}$ 支撑在一个 k 维子空间上，该子空间的维度为 k，因此我们可以将任意信号写成 $\boldsymbol{x} = \boldsymbol{Q}\boldsymbol{z}$，其中 $\boldsymbol{z} \sim p_{\boldsymbol{z}}$ 是某个低维潜向量， $\boldsymbol{Q} \in \mathbb{R}^{n\times k}$是一个固定的线性解码器。假设线性解码器 $\boldsymbol{Q}$ 已知，我们可以观察潜变量的特征函数，即 $\phi_{\boldsymbol{y}|\boldsymbol{A}}(\tilde{\boldsymbol{\omega}}) = \phi_{\boldsymbol{z}}(\boldsymbol{\omega} = (\boldsymbol{A}\boldsymbol{Q})^\top \tilde{\boldsymbol{\omega}})$，对于所有 $\tilde{\boldsymbol{\omega}} \in \mathbb{R}^m$以及 $g = 1$、.....、G 均成立。如果我们进一步假设 $m \geq k$且秩$(\boldsymbol{A}\boldsymbol{Q}) = k$对某个 $\boldsymbol{A}$成立，那么我们可以沿用 (3.25) 中的类似论证，得出潜分布 $p_{\boldsymbol{z}}$可以被唯一识别；由于线性解码器已知，因此也可以唯一识别 $p_{\boldsymbol{x}}$。

线性子空间的直觉可以推广到更一般的 k维集合。在这个例子中，模型识别的两个关键步骤是：i) 识别分布的低维支撑集，我们记为 supp$(p_{\boldsymbol{x}}) := \mathcal{X}$（在线性情形中，支撑集由线性解码器 $\boldsymbol{Q}$ 的值域给出）；ii) 要求 $\boldsymbol{A}$ 限制在 $\mathcal{X}$ 上时为一一映射（在线性情形中，这等价于要求 rank$(\boldsymbol{A}\boldsymbol{Q}) = k$）。

在无噪声测量的情形下，如果测量集 supp$(p_{\boldsymbol{y}}) = \mathcal{Y}$ 与信号集 $\mathcal{X}$ 之间存在一一对应的重建映射 f，那么只需对 $\mathcal{Y}$中的所有测量值应用 f，即可识别信号分布 $p_{\boldsymbol{x}}$。因此，我们将重点放在如何从测量分布中识别支撑集 $\mathcal{X}$ 这一问题上。

集合的低维性有多种定义 [98], 。这里我们关注上盒维数[4], ，它便于建立理论结果，既适用于紧流形等性质良好的模型（在这类模型中，该定义与更直观的拓扑维数一致），也适用于更一般的集合。下面的定理给出了一个充分条件，可用于从与多个前向算子相关的测量值中唯一识别信号集（参见第3.1节）：

[4]盒计数维数 [98, 章 2] 对紧子集 $\mathcal{S} \subset \mathbb{R}^n$ 定义如下

$$\operatorname{boxdim}(\mathcal{S}) = \limsup_{\epsilon \to 0} \frac{\log \mathbb{N}(S, \epsilon)}{-\log \epsilon} \tag{3.27}$$

其中，N$(\mathcal{S}, \epsilon)$ 是覆盖 $\mathcal{S}$ 所需的闭球的最小数量，这些闭球的半径为 ϵ，且相对于范数 $\|\cdot\|$ 定义。

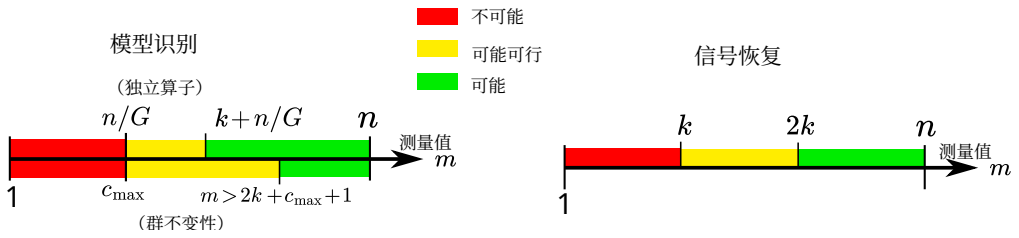

图3.5：模型识别与信号恢复的不同情形 [79] 随每个信号的部分观测数量 m 、模型维数 k、环境维数 n以及测量算子数量 G（当可以访问多个独立算子时）或不变子空间的最大重数 $c_{\max}$ （当信号具有群不变性或算子通过群作用相关联时）而变化。

定理 3.10 (Tachella 等人 [79])。 假设信号集 $\mathcal{X}$ 是一个盒计数维数为 k的有界集合。对于几乎所有的算子集合 G $\boldsymbol{A}_1$、……、 $\boldsymbol{A}_G \in \mathbb{R}^{m\times n}$，如果每个算子的测量值数量满足 $m > k + n/G$，则可以根据测量集 $\{\boldsymbol{A}_g\mathcal{X}\}_{g=1}^G$ 唯一识别信号模型 $\mathcal{X}$ 。

值得注意的是，该结果并不直接适用于任意 G 个算子集合（例如特定的磁共振成像或图像修复算子），而是要求使用 G 一般算子，从而排除退化情形。尽管如此，它仍为唯一识别信号分布所需的测量值和算子数量提供了基本界限。

对于算子由一个变换群关联这一约束更强的情形（参见第3.2节），即 $\{\boldsymbol{A}_g = \boldsymbol{A}\boldsymbol{T}_g\}_{g=1}^G$, 我们还可以进一步细化这一结果。以下定理依据 $\mathbb{R}^n$ 中对变换群[5], 不变的最大线性子空间的维数，该维数记为 $c_{\max}$。

定理 3.11 (Tachella 等人 [79])。 设 $\{\boldsymbol{T}_g\}_{g=1}^G$ 是与紧循环群相关联的变换群，并假设信号集 $\mathcal{X}$ 是一个盒计数维数为 k的有界集。对于几乎所有的 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，当测量值数量满足 $m > 2k + c_{\max} + 1$时，便可以从集合 $\{\boldsymbol{A}\boldsymbol{T}_g\mathcal{X}\}_{g=1}^G$ 中唯一确定该信号集。

利用以下事实：对于任意紧循环群 $c_{\max} \geq n/G$，当等式成立时，我们得到界 $m > 2k + n/G + 1$ ，它与定理 3.10 的界相似，但需要的测量值是$2k$ 个，而非 k个。事实证明，在某些情况下，这些额外的测量值是必要的，因为可以构造一个反例，其中无法使用任何具有 $m \leq 2k + c_{\max} - 2$ 个测量值的算子来确定 $p_{\boldsymbol{x}}$ 的支撑集 [79]。

与信号恢复理论的比较 压缩感知 [99] 和嵌入理论 [100], 中的一个著名结论是，对于一般的 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ ，若要使其在信号分布 $\mathcal{X}$的支撑集上一一对应，一个充分条件是测量值的数量大于集合维数的两倍，即 $m > 2k$ 其中 k 是 $\mathcal{X}$的盒计数维数。这些结果称为信号恢复定理，因为它们给出了能够保证存在重建函数的最少测量值

[5]群作用 $c_{\max}$ 的最大重数由满足以下条件的线性子空间 $\mathcal{S} \subseteq \mathbb{R}^n$ 的最大维数给出： $\boldsymbol{T}_g\mathcal{S} \subseteq \mathcal{S}$对所有 $g = 1$、……、G 均成立。更多详情请参见 [79] 。

以完美恢复所有合理信号。图3.5比较了本章提出的模型识别必要条件和充分条件与唯一信号恢复的相应条件。

与分裂的关系 基于命题 3.2，我们还可以探讨：为了通过分裂损失唯一识别信号分布，算子和测量值的最少数量分别是多少，以及它们与定理 3.10 中的结果有何比较。

G 满足命题 3.2 中条件的算子的最佳情形可以按如下方式构造：考虑由 $\boldsymbol{A} = [\boldsymbol{A}_1^\top, \boldsymbol{A}_2^\top]^\top \in \mathbb{R}^{m\times n}$给出的算子，其中 $\boldsymbol{A}_1 \in \mathbb{R}^{(2k+1)\times n}$ 固定，而 $\boldsymbol{A}_2 \in \mathbb{R}^{(m-2k+1)\times n}$ 从 G 个算子中随机采样一个，即 $\boldsymbol{A}_2 \sim \frac{1}{G}\sum_{g=1}^{G} \delta_{\boldsymbol{A}_{2,\,g}}$，且独立于 $\boldsymbol{A}_1$。为了学习条件估计器$\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1,\ \boldsymbol{A}_1}\overline{\{\boldsymbol{x}\}}$，命题 3.2 要求 $\boldsymbol{Q}_{\boldsymbol{A}_1} = \mathbb{E}_{\boldsymbol{A}|\boldsymbol{A}_1}\left\{\boldsymbol{A}^\top\boldsymbol{A}\right\} \in \mathbb{R}^{n\times n}$ 可逆（因此具有满秩）。假设我们处于无噪声设定，且信号分布的维数为 $k \le n$，我们可以在 $\boldsymbol{A}_1$中选择 $2k+1$ 个测量值，使得能够实现唯一信号恢复 [100]。因此，条件估计器可以得到完美重建$\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1,\ \boldsymbol{A}_2}\{\boldsymbol{x}\} = \boldsymbol{x}$。一旦学会条件均值估计器，我们就可以通过从 $p_{\boldsymbol{y}}$ 简单地重建测量值来唯一识别 $p_{\boldsymbol{x}}$。由于$\boldsymbol{A}$的形式得到了简化，我们可以将命题 3.2 中的 $\boldsymbol{Q}_{\boldsymbol{A}_1}$ 以闭式计算为

$$
\begin{aligned}
\boldsymbol{Q}_{\boldsymbol{A}_1} &= \mathbb{E}_{\boldsymbol{A}|\boldsymbol{A}_1}\left\{\boldsymbol{A}^\top\boldsymbol{A}\right\} \\
&= \mathbb{E}_{\boldsymbol{A}_2}\left\{\boldsymbol{A}_2^\top\boldsymbol{A}_2\right\} + \mathbb{E}_{\boldsymbol{A}|\boldsymbol{A}_1}\left\{\boldsymbol{A}_1^\top\boldsymbol{A}_1\right\} \\
&= \frac{1}{G}\sum_{g=1}^{G} \boldsymbol{A}_{2,g}^\top\boldsymbol{A}_{2,g} + \boldsymbol{A}_1^\top\boldsymbol{A}_1.
\end{aligned}
$$

由于 $\boldsymbol{Q}_{\boldsymbol{A}_1}$由秩至多为 $m-(2k+1)$ 的 G 个矩阵之和以及一个秩至多为 $2k+1$ 的矩阵组成，因此有秩 $\left(\boldsymbol{Q}_{\boldsymbol{A}_1}\right) \le \min\{G(m-2k-1)\ + 2k+1,\ \ n\}$ 当 $\boldsymbol{A}_{2,\ g}$为一般选择时取等号。

由于我们需要秩$\left(\boldsymbol{Q}_{\boldsymbol{A}_1}\right) = n$，因此必要条件是 $G(m-2k-1)\ + 2k+1 \ge n$。于是我们得到条件

$$
m \ge \frac{n}{G} + \left(1 - \frac{1}{G}\right)(2k+1)
$$

即　　这是定理 3.10 中充分条件的一个类似但不那么严格的版本。

3.5 总结

在前向算子是多对一的逆问题中，只要算子在不同样本之间发生变化，或者我们可以假设信号分布对于某个变换群保持不变，例如平移、旋转或缩放，自监督学习就是可行的。

表 3.1 总结了本章介绍的不同类别自监督损失背后的假设。这些假设侧重于前向算子所需满足的条件，而非噪声，因为所有损失函数都可以根据第2章介绍的原则进行调整，以处理噪声。

族	假设	参考文献
（MSPLIT）	**必要** 多个前向算子 $\mathbb{E}_{\boldsymbol{A}}\left\{\boldsymbol{A}^\top\boldsymbol{A}\right\}$ 可逆 **充分**（参见命题 3.2） $\mathbb{E}_{\boldsymbol{A}_1\|\boldsymbol{A}}\left\{\mathbb{E}_{\boldsymbol{A}\|\boldsymbol{A}_1}\left\{\boldsymbol{A}^\top\boldsymbol{A}\right\}\right\}$ 可逆	[27, 85]
（MOI）	**必要** 多个前向算子 $\mathbb{E}_{\boldsymbol{A}}\left\{\boldsymbol{A}^\top\boldsymbol{A}\right\}$ 可逆 $\exists f^*$ 使得 $\boldsymbol{x}\approx f^*(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{A})$ **充分**（参见定理 3.10） G 泛型 $\boldsymbol{A}$，且满足 $m > k + n/G$	[87]
(ESPLIT)	**必要** 单一前向算子 $\boldsymbol{A}$ $p(\boldsymbol{x})$ 对变换不变 $\{\boldsymbol{T}_g\}_{g=}^{G}$ $_1$ $\sum_{g=1}^{G}\boldsymbol{T}_g\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{T}_g$ 可逆 **充分** $\mathbb{E}_{\boldsymbol{A}_1\|\boldsymbol{A}}\left\{\sum_{g\in\mathcal{S}_{\boldsymbol{A}_1}}\boldsymbol{T}_g^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{T}_g\right\}$ 可逆 其中 $\mathcal{S}_{\boldsymbol{A}\ 1} = \{g:\ \boldsymbol{A}_1$ 是 $\boldsymbol{A}\boldsymbol{T}_g\}$的拆分	[90]
(EI)	**必要** 单一前向算子 $\boldsymbol{A}$ supp$(p_{\boldsymbol{x}})$ 对变换不变 $\{\boldsymbol{T}_g\}_{g=1}^{G}$ $\sum_{g=1}^{G}\boldsymbol{T}_g\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{T}_g$ 可逆 $\exists f^*$ 使得 $\boldsymbol{x}\approx f^*(\boldsymbol{A}\boldsymbol{x})$ **充分**（参见定理 3.11） 通用 $\boldsymbol{A}$，带有 $m > 2k + n/G + 1$	[26, 62] [91, 101]

表 3.1：**基于不完整测量进行学习时的损失函数总结**。前两个损失函数依赖于使用多个算子进行测量，而后两个则假设信号分布在一组变换下保持不变，以获得一组虚拟算子。

第4章

有限数据集效应

到目前为止，我们已经了解了如何构造自监督损失函数，使其在期望意义下成为（受约束或无约束）监督损失的无偏估计量。然而，一个重要的问题仍未解答：在样本数量有限时，这些近似的效果如何？本章将讨论现有的答案，同时指出，自监督方法样本复杂度的完整理论刻画尚未得到充分理解。我们通过一些实际场景说明自监督方法对数据集大小的依赖性，结果显示其通常与监督学习场景中的情况具有相似的缩放规律。我们还将介绍一些处理有限数据集的实用工具：应用留出法以避免欠拟合或过拟合，以及从预训练模型出发，以减少获得良好性能所需的测量值数量。

4.1 使用自监督损失函数的留出法

机器学习中的标准做法是将数据集划分为互不重叠的训练集、验证集和测试集 [102]。验证集在避免欠拟合或过拟合方面起着关键作用：如果验证损失始终与训练损失接近，模型可能不够灵活，无法拟合全部数据；反之，如果验证损失大于训练损失，模型很可能对训练集过拟合。

即使我们没有真实标签验证样本 [75, 103]，也可以在自监督学习场景中采用类似的做法。由于自监督损失可作为监督损失的代理指标，因此也可以在没有真实标签的验证集上使用，以验证模型是否对数据欠拟合或过拟合。

图4.1展示了训练集和验证集上的自监督损失，以及测试集上的监督损失。自监督验证损失能够很好地反映模型在测试集上的性能，可用于在模型开始过拟合时停止训练，也就是说，当验证损失与训练损失之间的差距开始增大时停止训练。

4.2 损失及其梯度的方差

理解自监督损失对监督对应方法的近似程度，首先要研究损失及其梯度的方差。方差越大，我们对监督损失的估计就越差，进而导致

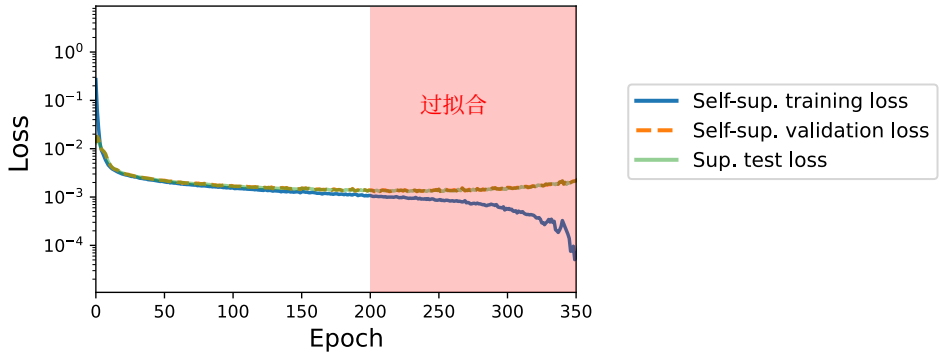

图4.1：在高斯去噪问题中，使用 UNet 网络和 $\sigma = 0.1$ 时的训练、验证和测试损失。训练和验证使用第2章介绍的 (SURE) 损失计算。测量值由 MNIST数据集生成，其中训练集包含 768 张带噪图像，验证集包含 256 张带噪图像，测试集包含 8192 对带噪和干净图像（即有监督样本对；在真实场景中无法获得）。自监督验证损失是有监督测试误差非常准确的代理指标；如果模型开始对训练数据过拟合，就可以用它来停止训练。

与监督情形相比，性能下降。图4.2展示了使用DRUNet去噪器 [104] 对Urban100数据集中的 512×512 图像块施加各向同性高斯噪声后，(Noise2Noise) 和 (SURE) 的损失及梯度估计的平均归一化均方误差。实验分别使用训练权重的网络和随机初始化权重的网络进行。在训练和未训练两种情形下，对于所有噪声级别，相对于监督损失的误差都低于30%。未训练模型的梯度估计比训练模型更准确。在未训练情形下，由于损失较大，额外方差可以忽略不计，因此所有损失的误差都约为10%。在训练情形下，损失较小，自监督损失的额外方差开始产生影响。Noise2Noise的估计优于SURE，因为它利用了更多信息，即每幅图像的两个独立噪声副本。

对 N 张带噪图像取平均时，如果图像样本相互独立，我们应预期所有损失的方差都按 $1/N$ 衰减。通过了解 N 需要增加多少才能达到有监督数据集的方差，我们可以估算自监督方法的有效样本复杂度。在图4.2的 $\sigma = .5$ 情形中，与监督情形相比，使用 SURE 获得相同方差大约需要多 $\sqrt{10} \approx 3$ 倍的带噪样本。

损失的方差 为简化起见，以下分析聚焦于（Noise2Noise）损失，但其中的直觉同样适用于前几章介绍的其他损失函数。关于（SURE）损失的方差的深入分析，见 [105]。由于在给定 $\boldsymbol{x}$ 的条件下， $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 相互独立，Noise2Noise 损失的方差可分解如下：

命题 **4.1.** 设 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 是两个满足条件的独立随机变量 al

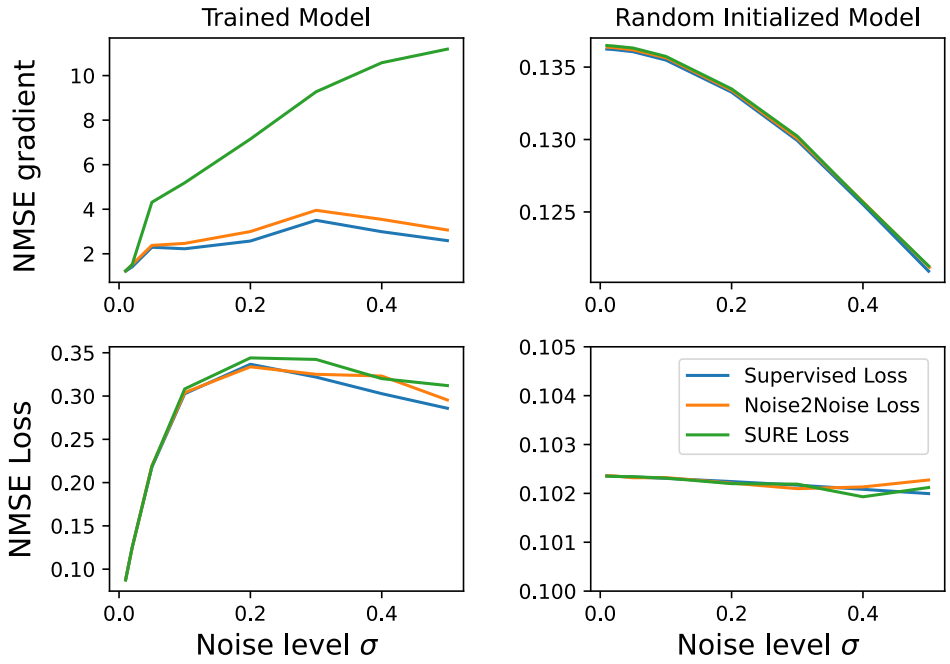

图4.2：**监督损失和自监督损失的梯度近似误差**。监督损失、Noise2Noise 和蒙特卡洛 SURE 损失，以及它们关于网络权重的梯度的归一化均方误差（NMSE）。实验采用 DRUNet 去噪器架构，在 Urban100 数据集的 512×512 图像块上进行评估。

给定 $\boldsymbol{x}$，使得 $\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\boldsymbol{y}_2\} = \boldsymbol{x}$。那么

$$\mathbb{V}_{\boldsymbol{y}_1,\boldsymbol{y}_2}\left\{\mathcal{L}_{N2N}(\boldsymbol{y}_1,\boldsymbol{y}_2,f)\right\} = \mathbb{V}_{\boldsymbol{x},\boldsymbol{y}_1}\left\{\frac{1}{n}\|f(\boldsymbol{y}_1)-\boldsymbol{x}\|^2\right\} + \Delta$$

其中，第一项是监督损失的方差，Δ 是相对于监督情形增加的方差，其表达式为

$$\begin{aligned}\Delta = {} & \mathbb{V}_{\boldsymbol{x},\boldsymbol{y}_2}\left\{\|\boldsymbol{y}_2-\boldsymbol{x}\|^4\right\} + \frac{4}{n^2}\mathbb{E}_{\boldsymbol{x}}\left\{trace\left(\boldsymbol{\Psi}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\right\} \\ & - \frac{2}{n^2}\mathbb{E}_{\boldsymbol{x}}\left\{\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{\|\boldsymbol{y}_2-\boldsymbol{x}\|^2(\boldsymbol{y}_2-\boldsymbol{x})^\top\right\}\left(\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\{f(\boldsymbol{y}_1)\}-\boldsymbol{x}\right)\right\}\end{aligned}$$

其中，$\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{(\boldsymbol{y}_2-\boldsymbol{x})(\boldsymbol{y}_2-\boldsymbol{x})^\top\}$ 是噪声目标 $\boldsymbol{y}_2$ 的协方差，而 $\boldsymbol{\Psi}_{\boldsymbol{x}} = \mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\{(f(\boldsymbol{y}_1)-\boldsymbol{x})(f(\boldsymbol{y}_1)-\boldsymbol{x})^\top\}$ 是图像 $\boldsymbol{x}$ 的误差协方差。

证明见附录 C。若噪声分布是对称的 $\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\|\boldsymbol{y}_2-\boldsymbol{x}\|^2(\boldsymbol{y}_2-\boldsymbol{x})\} = \boldsymbol{0}$，或者估计器 f 是无偏的 $\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\{f(\boldsymbol{y}_1)\} = \boldsymbol{x}$，对所有 $\boldsymbol{x}$ 均成立，则 Δ 中的最后一项为零。例如，对于带有各向同性高斯噪声的目标这一简单情形，$\boldsymbol{y}_2 = \boldsymbol{x} + \sigma\boldsymbol{\epsilon}$ 且 $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ 时，我们有

$$\Delta = \frac{3}{n}\sigma^4 + \frac{4\sigma^2}{n}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}_1}\left\{\frac{1}{n}\|f(\boldsymbol{y}_1)-\boldsymbol{x}\|^2\right\}$$

只要均方误差 $\mathbb{E}_{\boldsymbol{x},\,\boldsymbol{y}_1}\left\{\frac{1}{n}\|f(\boldsymbol{y}_1)-\boldsymbol{x}\|^2\right\}$ 近似独立于 n，它就会随着 n 的增大而趋于零。

梯度方差 为了计算梯度的方差，我们需要考虑去噪器的一种参数化方式 $f_{\boldsymbol{\theta}}$，其中 $\boldsymbol{\theta} \in \mathbb{R}^p$ 是去噪器的可训练参数（例如网络权重）。即使自监督损失的方差非常小，如果去噪器对参数变化高度敏感，梯度仍可能有显著变化。(Noise2Noise) 损失的梯度可以分解为两个相互独立的量：

$$\frac{\partial \mathcal{L}_{\mathrm{N2N}}}{\partial \boldsymbol{\theta}}(\boldsymbol{y}_1, \boldsymbol{y}_2, f_{\boldsymbol{\theta}}) \propto \frac{1}{n}\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\left(f_{\boldsymbol{\theta}}(\boldsymbol{y}_1) - \boldsymbol{y}_2\right) \tag{4.1}$$

$$\propto \underbrace{\frac{1}{n}\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\left(f_{\boldsymbol{\theta}}(\boldsymbol{y}_1) - \boldsymbol{x}\right)}_{\text{Supervised gradient}} - \underbrace{\frac{1}{n}\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\left(\boldsymbol{y}_2 - \boldsymbol{x}\right)}_{\text{Additional noise}} \tag{4.2}$$

其中，$\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{p \times n}$ 是去噪器在 $\boldsymbol{y}_1$处求得的雅可比矩阵。第一项对应监督损失的梯度，第二项则包含因使用噪声目标而产生的额外随机性。由于在给定 $\boldsymbol{x}$的条件下， $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 相互独立，因此损失梯度的方差为

$$\begin{aligned}
&\mathbb{V}_{\boldsymbol{y}_1,\boldsymbol{y}_2}\left\{\|\frac{\partial \mathcal{L}_{\mathrm{N2N}}}{\partial \boldsymbol{\theta}}(\boldsymbol{y}_1, \boldsymbol{y}_2, f_{\boldsymbol{\theta}})\|^2\right\} \\
&= \mathbb{V}_{\boldsymbol{y}_1,\boldsymbol{x}}\left\{\|\frac{1}{n}\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\left(f_{\boldsymbol{\theta}}(\boldsymbol{y}_1) - \boldsymbol{x}\right)\|^2\right\} + \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{x}}\left\{\|\frac{1}{n}\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\left(\boldsymbol{y}_2 - \boldsymbol{x}\right)\|^2\right\} \\
&= \underbrace{\mathbb{V}_{\boldsymbol{y}_1,\boldsymbol{x}}\left\{\|\frac{1}{n}\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\left(f_{\boldsymbol{\theta}}(\boldsymbol{y}_1) - \boldsymbol{x}\right)\|^2\right\}}_{\text{Variance of supervised loss}} + \underbrace{\frac{1}{n^2}\mathbb{E}_{\boldsymbol{x}}\left\{\operatorname{trace}\left(\boldsymbol{\Sigma}_{\boldsymbol{x}}\,\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}^{\top}\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\right\}\right)\right\}}_{\text{Additional variance}}
\end{aligned}$$

其中，第二行使用了式（4.2）中的分解，并利用了额外噪声的均值为零这一性质；第三行将 $\boldsymbol{\Sigma}_{\boldsymbol{x}} := \mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{(\boldsymbol{y}_2 - \boldsymbol{x})(\boldsymbol{y}_2 - \boldsymbol{x})^{\top}\right\}$ 定义为噪声目标 $\boldsymbol{y}_2$的协方差。例如，对于带有各向同性高斯噪声的目标， $\boldsymbol{y}_2 = \boldsymbol{x} + \sigma\boldsymbol{\epsilon}$且 $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0},\ \boldsymbol{I}\right)$，我们有

$$\text{Additional variance} = \sigma^2\,\mathbb{E}_{\boldsymbol{y}_1}\left\{\|\frac{1}{n}\frac{\partial f_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\|_F^2\right\}$$

其中，第二项是网络关于其参数的雅可比矩阵的 Frobenius 范数。噪声目标 $\boldsymbol{y}_2$ 引入的额外方差取决于噪声水平 σ^2，以及网络输出对输入权重变化的平均敏感度。

4.3 与监督学习的差距

比较监督方法和自监督方法时，一个关键问题是：自监督方法与监督对应方法的表现如何随训练数据量变化。这需要计算以下差距：

$$\operatorname{gap}(N) = \min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\|f_{\boldsymbol{\theta}}(\boldsymbol{y}_{1,i}) - \boldsymbol{y}_{2,i}\|^2 - \underbrace{\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}_1}\left\{\frac{1}{n}\|f_{\boldsymbol{\theta}}(\boldsymbol{y}_1) - \boldsymbol{x}\|^2\right\}}_{\approx \mathrm{MMSE}}$$

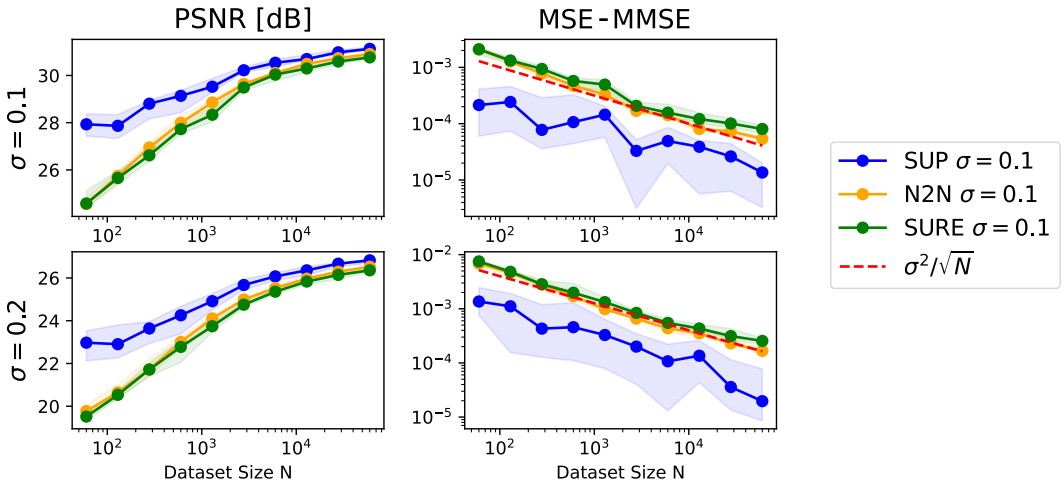

图 4.3：**使用 U-Net 去噪器对 MNIST 进行高斯去噪时，Noise2Noise 和 SURE 相对于监督学习的差距随数据集大小的变化**。左图展示了各学习方法获得的 PSNR，右图展示了与使用完整数据集的监督基线相比，测试均方误差的差距。对于每种数据集大小 N 以及两种不同的噪声水平 $\sigma_1 = .1$ 和 $\sigma_2 = .2$，实验均重复 15 次；阴影区域表示重复实验所得结果的 90% 区间。最优性差距近似遵循 $\sigma^2/\sqrt{N}$。

—

量化这一差距通常很困难，因为结果可能高度依赖于数据分布、估计器参数化以及用于估计参数的具体学习算法。理解学习方法的样本复杂度是一个活跃的研究领域 [102]，仍有许多未解问题 [106]。特别是，一个颇具挑战的问题是，如何获得有意义的界限，同时避免其高度依赖模型的参数数量，而深度网络的参数数量通常非常庞大。例如，一系列研究 [103, 107] 探讨了随机梯度下降方法的泛化误差。随机梯度下降是训练深度网络最常用的优化方法，这些研究得到了近似独立于参数数量的界限。

这里，我们使用第 4.1 节所述的（自监督）留出法，为每个数据集大小获取最佳模型，并对样本复杂度进行实证评估。图 4.3 展示了在不同数据集大小下，使用 Noise2Noise 或 SURE 在 MNIST 高斯去噪问题上训练的网络的实证差距。观察到的差距遵循渐近行为 $\text{gap}(N) \propto \sigma^2/\sqrt{N}$。这一率有多好？我们可以将其与一个由单个信号组成的平凡信号分布进行比较，从而获得一些直觉：在这种情况下，我们可以简单地对 N 个含噪观测取均值来估计该信号，并得到估计方差为 σ^2 的高斯分布均值时的标准率，即 σ^2/N。因此，处理非平凡信号分布的代价是 $\sqrt{N}$ 倍。

—

4.4 微调与测试时适应

尽管本手稿中介绍的大多数自监督损失都近似于监督损失，并且无需任何真实参考值即可从随机初始化开始训练重建网络，但它们也可用于在新的测量数据上对预训练网络进行微调，而这些测量数据可能与用于预训练的有监督数据集不同。这一过程通常称为测试时自适应 [108]。

与随机初始化模型相比，从预训练模型开始训练可以显著减少获得良好结果所需的样本数量和训练时间（例如，通常只需几个样本即可）。此外，微调可以显著提升预训练模型在域外测量数据上的性能 [17]。

第5章

扩展与开放问题

自监督学习的理念已经对成像与传感产生了相当大的影响，其应用正在 [77], 磁共振成像 [48], 显微成像 [49]和遥感等领域不断涌现。在本章最后，我们将回顾一些正在开展的工作，这些工作正探索将自监督学习框架扩展到逆问题，并介绍该领域的一些开放问题。

5.1 非线性逆问题

许多现实世界中的逆问题都是非线性的，例如量化传感 [109], 相位恢复 [110], 以及与偏微分方程相关的非线性问题，如逆散射问题 [8]。原则上，第3章介绍的大多数自监督损失都可以应用于非线性前向模型，但与这些损失相关的大多数理论分析仅限于线性情形，而针对非线性算子构建通用理论框架仍是一个开放问题。

非线性情形下的另一项重要挑战是，自监督损失需要对非线性算子 $\boldsymbol{A}$进行计算。与不涉及该算子的监督损失 $\boldsymbol{A}$相比，这既会带来较高的计算开销，也会导致损失景观更加复杂，其中包含许多局部极小值。我们认为，这些问题带来了新的挑战，例如探索自监督损失 [111, 112] 中的松弛方法 $\boldsymbol{A}$ ，以及针对具体问题设计优化算法。

Hu 等人 [113] 表明，多算子分割损失可用于线圈映射未知的加速磁共振成像场景，而该问题可视为一个双线性逆问题。

等变成像方法也已扩展到由$\boldsymbol{A}(\boldsymbol{x}) = \eta(\boldsymbol{x})$给出的去削波问题，其中 $\eta\colon \mathbb{R} \to [-1, 1]$ 是一个逐元素削波算子。在这种情况下，如果我们可以假设信号模型对幅度缩放具有不变性，即 $\boldsymbol{T}_g = g\boldsymbol{I}$ 且 $g > 0$，那么就有可能学习超出削波阈值的信息，因为前向算子对这些变换并不等变 [112]。

第 3.4 节中的模型识别理论已扩展至量化逆问题 [111], 。在极端情况下，每个测量值都被量化为单比特；这一问题可写为 $\boldsymbol{A}_g(\boldsymbol{x}) = \text{sign}(\boldsymbol{Q}\boldsymbol{x})$，其中 $\{\boldsymbol{Q}_g \in \mathbb{R}^{m\times n}\}_{g=1}^{G}$ 是线性算子。在这种情况下，即使集合的维度很低 $k \ll n$，也无法精确识别信号分布的支撑集，但仍然可以学习其近似

直至全局误差达到 $\mathcal{O}(\binom{k+n/G}{m}\log_{\frac{nm}{k+n/G}})$ 的量级。

5.2 迈向大规模自监督成像

深度学习成像解决方案依赖现代图形处理器提供的强大计算能力。尽管这项技术正在快速发展，但消费级图形处理器当前的内存容量限制了其能够处理的成像逆问题规模。例如，具有挑战性的高维医学成像就面临这一问题，例如超大规模三维或四维（3D +时间）CT 和磁共振成像成像 [114–116], ，以及电子显微镜中的叠层成像等应用 [117]。此类问题每次重建可能产生多达数十 GB 的测量值和/或图像数据（每次重建），图像和测量值的规模都很容易超过大多数图形处理器的内存容量。

因此，为这类问题训练深度学习解决方案还面临额外的计算难题。若要采用自监督学习技术，挑战就更大了，因为相关损失函数需要计算前向算子 $\boldsymbol{A}$（并通过该算子进行反向传播），导致训练时计算量和内存占用都很大，甚至评估时也可能如此。在监督式场景中训练展开式深度学习解决方案时，也会遇到同样的问题，例如 [114, 116], 。为缓解这些计算问题，人们已考虑采用各种方法，例如前向或反向重计算、在反向传播期间通过梯度检查点减少存储需求，以及传统上用于基于模型的图像重建的数据拆分方法 [118, 119]。

对于自监督学习中应对这些问题的最佳方法，目前关注较少，因此这是一个值得未来研究的方向。一个例外是 [120],，作者以自监督方式训练了一个用于低剂量三维螺旋计算机断层扫描的图像重建网络。他们的核心方法是采用一种测量拆分技术，其思路与 Noise2Inverse 方法类似 [25]，每次迭代仅需部分计算前向算子（他们还实现了其他多种计算技巧，例如梯度检查点、自定义 CUDA 模块等）。

Kosoma 等人 [120] 关注的是可逆成像算子；借助第3章介绍的相关拆分思路，将这些想法推广到多个不可逆算子应该并不困难。另一个有趣的方向是探索明确采用基于随机优化的重建网络的自监督学习方案 [121]。

5.3 鲁棒方案与部分定义的模型

在实际逆问题中，另一个需要解决的重要实际问题是，我们定义观测模型的准确程度。正如 George Box 所说：“ 所有模型都是错的，但有些是有用的。” 在任何统计学习或推断场景中，尽可能准确地捕捉变量之间的内在关系至关重要，同时还要确保解能够考虑到前向模型中的任何未知成分，并对各种近似或不完善之处保持稳健。这在自监督学习中尤为重要，因为监督损失的替代项严重依赖于逆问题提供的额外信息。

在第2章中，我们已经看到过一些考虑了部分定义的噪声模型的实例——这些模型会限制待学习的估计器类别，例如 [21, 28, 75]。这通常会导致非最优解，但即便如此，这些解仍可能优于基于错误设定假设得出的“最优”解。

将这些思路扩展到部分定义或设定错误的前向模型，无论是在理论上还是算法上，都更具挑战性，也是未来研究的一个有趣方向。例如，部分定义模型包括存在未知校准参数且必须对其进行估计的情况。一种方法是将未知参数视为成像问题中的额外未知量，在重建过程中对其进行估计或边缘化处理，例如，在磁共振成像中自监督估计线圈灵敏度图 [113]，或自监督盲去模糊解决方案 [122]。

前向模型中常见的一种近似来源，是通常将连续图像数字表示为有限数量的像素/体素。采集系统的设计通常会避免在此类数字表示中引入混叠，而这反过来可能导致图像噪声出现相关性。在自监督学习场景下，必须谨慎处理这种相关性，例如在 SAR 处理中 [49]。

5.4 超越 ℓ_2 损失

本文中介绍的大多数损失函数旨在近似监督 ℓ_2 损失。这种损失的优点是可以简单地分解为

$$\|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 = \|f(\boldsymbol{y}) - \boldsymbol{y}\|^2 - 2\, f(\boldsymbol{y})^\top (\boldsymbol{y} - \boldsymbol{x}) + \text{const.}$$

其中，第二项刻画了采用简单测量一致性时监督情形与非监督情形之间的差异。正如我们在第2章中所见，这一项可以通过使用独立噪声实现来处理，如 (Noise2Noise) 和 (R2R) 中所示；也可以使用斯坦因引理，如 (SURE) 中所示；或使用盲点网络，如 (CV) 中所示。虽然对于 Bregman 散度也有类似的表达式 [47, 71], 但这种分解并不适用于其他常见损失，例如 ℓ_1 或 ℓ_0 损失。

在 Noise2Noise 中， [40], 针对一般 ℓ_p 损失提出了处理非零均值噪声分布的方法，例如用于随机文本去除的 ℓ_1 损失，或用于椒盐噪声的 ℓ_0 损失。回想一下，Noise2Noise 依赖于独立含噪对（$\boldsymbol{y}_1, \boldsymbol{y}_2$），使用 ℓ_0 损失进行训练会（在期望意义上）得到众数估计量

$$\text{Mode}\{\boldsymbol{y}_2|\boldsymbol{y}_1\} = \arg\max_{\boldsymbol{y}_2} p(\boldsymbol{y}_2|\boldsymbol{y}_1).$$

一般而言，$\text{Mode}\{\boldsymbol{y}_2|\boldsymbol{y}_1\} \neq \text{Mode}\{\boldsymbol{y}_2|\boldsymbol{x}\}$，即使给定 $\boldsymbol{x}$时， $\boldsymbol{y}_2$ 和 $\boldsymbol{y}_1$ 相互独立。然而，在假设给定 $\boldsymbol{y}_1$ 时 $\boldsymbol{x}$ 的（后验）分布高度集中的情况下，我们有 $p(\boldsymbol{y}_2|\boldsymbol{y}_1) \approx p(\boldsymbol{y}_2|\boldsymbol{x})$，因此 $\text{Mode}\{\boldsymbol{y}_2|\boldsymbol{y}_1\} \approx \text{Mode}\{\boldsymbol{y}_2|\boldsymbol{x}\} = \boldsymbol{x}$。对于 ℓ_1 损失和其中位数由 $\boldsymbol{x}$给出的噪声分布，也可以采用类似的论证。尽管这一近似能够带来良好的经验结果 [40], ，但进一步理解这类近似在何种条件下是合理的，并确定如何将（SURE）及其他能够利用噪声分布信息的自监督损失扩展到一般的 ℓ_p 损失，都是值得未来研究的方向。

5.5 不确定性量化与生成式建模

本综述的大部分内容都集中于学习良好的重建映射或去噪器，通常以逆问题的条件均值 $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{x}\}$为目标。然而，在许多场景中，量化给定估计器的不确定性同样十分重要，这样就能有足够的信心将图像估计值用于下游分析。

本文介绍的大多数自监督损失都可作为监督损失的估计器，因此可用于在测试时估计重建误差。例如，(SURE) 或 (R2R) 可用于估计均方误差，而 SURE for SURE [105] 等高阶扩展则可用于量化该误差估计值的不确定性。也可以将 (Tweedie) 扩展应用于后验分布的高阶矩。例如，在高斯去噪场景中，Manor 和 Michaeli [123] 使用这些扩展，直接从最小均方误差估计器估计后验分布协方差的主成分（该估计器本身也可通过自监督方式估计），$p(\boldsymbol{x}|\boldsymbol{y})$。这能很好地为估计得到的去噪图像补充不确定性的主方向。处理不完整数据时，可使用 (EI) 量化前向算子 [124] 零空间中的重建误差。

这些思想在多大程度上可应用于本综述中的其他自监督学习方案，是一个值得探讨的开放问题。

生成模型 虽然点估计的预测误差协方差能够高效、简洁地表征不确定性，但在成像科学的某些领域，人们希望能够探索成像问题的完整后验分布，以便进行下游分析，例如，在存在歧义时刻画合理解，或检验统计假设。

在这种情况下，一种流行的机器学习方案是学习生成模型，例如变分自编码器 [125]，生成对抗网络 [126] 或扩散模型 [41]，这些模型可作为随机模拟器，并从后验分布中生成样本，$p(\boldsymbol{x}|\boldsymbol{y})$。因此，生成模型已成为解决成像逆问题的一种流行方法。然而，目前这类方案通常依赖于预训练 生成模型，这些模型是在现有真实数据上训练得到的。因此，一个值得探讨的研究方向是，生成模型在多大程度上能够以纯粹自监督的方式进行学习。目前在这方面已经取得了一些进展。例如，如第2章和第3章所述，Prakash 等人 [80, 81]开发了用于高斯去噪问题的变分自编码器，而生成对抗网络 [82] 和扩散模型 [43, 85] 则已被提出，可利用无噪但不完整的测量值进行训练。然而，这些方法通常依赖来自多个算子的低噪声测量值；一个悬而未决的问题是，能否以类似于 (EI) 的方式，利用来自单个病态测量算子的含噪测量值训练生成模型。

5.6 样本复杂度

关于自监督学习成像解决方案的大多数理论分析，要么是几何度量 [111]，要么侧重于考察学习问题中期望值行为的渐近性质。然而，这些分析并未说明从统计角度看，问题有多难，也没有说明需要多少测量训练样本才能获得良好的解——

这归根结底取决于我们能多准确地用经验风险近似期望风险。

我们在第4章讨论 Noise2Noise 时简要提到过这一点，并通过实验证明，监督学习与等价的（在期望意义下）自监督学习策略之间的差距，其尺度近似为 $\text{gap}(N) \propto \sigma^2/\sqrt{N}$。然而，即使在监督情形下，分析样本复杂度也是一个具有挑战性的问题 [102]。现有的大多数结果通常都会作出重要的简化假设，例如学习问题是凸的 [102], ，或者重建估计器是线性的 [103]。此外，还存在一个问题：样本复杂度如何随图像尺寸变化。在这里，我们有理由持乐观态度，认为或许能够受益于与图像通常具有的高维度相关的维数的祝福 [127] ，其方式类似于我们仅使用一个蒙特卡洛样本 [70]就能获得准确的 SURE损失估计值。

理解这种监督与自监督之间的差距的性质，将有助于成像从业者判断：是应当尝试收集少量监督训练数据，还是可以通过使用数量更多、通常更容易获取的测量数据来达到相同的结果。

5.7 选择合适的自监督方法

本文介绍了各种自监督技术，其能力范围从去噪到解决成像不适定逆问题。我们还看到，同一个问题有时可以通过审慎选择网络架构来解决（但应避免诉诸更加模糊的架构归纳偏置），也可以通过巧妙设计的损失函数来解决。不过，我们一直没有明确主张某一种解决方案优于另一种。这可能会让实践者因缺乏针对特定任务选择正确解决方案的指导而感到有些困惑。

根据具体情境，实践者可能获得各种形式的信息，例如测量噪声的性质及相关信息，或可用于创建训练数据的测量算子的数量和类型。在某些情况下，这些信息会自然地筛选出相关算法和技术的子集；第2章和第3章末尾的表格对此进行了重点说明。不过，还必须注意，并非所有相关算法都会以相同方式利用可用信息。例如，如果面对这样一个问题：训练和测试时都可以获得同一信号的成对含噪实现，人们可能会自然而然地选择 Noise2Noise 算法。然而，如果已知一些关于统计噪声模型的信息，即使这些信息并不完整，那么更好的做法可能是直接对多个实现进行平均（从而获得 3 dB 的信噪比增益），再应用一种能够纳入额外统计信息的不同技术[1]。

许多经过综述的技术会利用不同的信息，这也为探索混合方法提供了机会。例如，在上面的例子中，与其简单地对噪声图像对取平均，然后应用一种融入了更多噪声模型知识的图像处理技术，不如明确构建一个能够

[1]例如 ，Tachella 等人 [28] 表明，在
冷冻电子显微镜去噪中，使用 Noise2Noise 的性能更好，因为其噪声模型是近似已知的。

利用图像对[2]的版本。同样，只有在存在单个测量算子时，才将不变性属性纳入逆问题，这并非唯一选择。事实上，经验证据表明，在测量值中利用多个信息来源通常只会带来更好的效果 [128]。

在许多现实场景中，对于从业者而言，最具挑战性的是了解所利用的底层前向模型有多准确，该模型正如上文所述，用于实现自监督学习。这一点极为重要，从业者在选择合适的算法时，或许应当首先考虑这一点，因为定义得更精确的前向模型通常能带来更好的性能，但代价是对任何模型设定错误都更加敏感。

致谢

作者感谢与 Dongdong Chen、Laurent Jacques、Marcelo Pereyra、Lo¨ıc Denis、Andrew Wang、Victor Sechaud、J´er´emy Scanvic、C´esar Caiafa、Thomas Davies 和 Brett Levac 的讨论。J. Tachella 感谢法国国家研究署（ANR-23-CE23-0013）的支持。作者感谢里昂高等师范学院支持 Mike Davies 以客座教授身份开展研究访问。

[2]注：拥有成对的噪声图像，即可立即估计信号的信噪比。

附录 A

Noisier2Noise 与 R2R 的等价性

在本附录中，我们展示了高斯去噪情况下 Noisier2Noise [51] 与 Recorrupted2Recorrupted [52] 损失函数之间的渐近等价性，即 $\boldsymbol{y} = \boldsymbol{x} + \epsilon$ with$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{\Sigma})$。Noisier2Noise 提出训练一个网络

$$\mathcal{L}_{\text{Noisier2Noise}}(\boldsymbol{y}, f) = \mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{y}} \left\{ \| f(\boldsymbol{y}_1) - \boldsymbol{y} \|^2 \right\} \tag{A.1}$$

其中 $\boldsymbol{y}_1 = \boldsymbol{y} + \tau\boldsymbol{\omega}$ 满足 $\boldsymbol{\omega} \sim \mathcal{N}\left(\mathbf{0},\ \boldsymbol{\Sigma}\right)$，且 $\tau > 0$。该损失的期望极小化器为

$$\begin{aligned}
f^*(\boldsymbol{y}_1) &= \mathbb{E}_{\boldsymbol{y}|\boldsymbol{y}_1} \left\{ \boldsymbol{y} \right\} \\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}|\boldsymbol{y}_1} \left\{ \frac{\tau}{1+\tau}\boldsymbol{x} + \frac{1}{1+\tau}\boldsymbol{x} + \boldsymbol{\epsilon}) \right\} \\
&= \frac{\tau}{1+\tau}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1} \left\{ \boldsymbol{x} \right\} + \frac{1}{1+\tau}\Big(\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1} \left\{ \boldsymbol{x} \right\} + \mathbb{E}_{\boldsymbol{\epsilon}|\boldsymbol{y}_1} \left\{ \boldsymbol{\epsilon} \right\} + \frac{1}{1+\tau}\mathbb{E}_{\boldsymbol{\epsilon}|\boldsymbol{y}_1} \left\{ \boldsymbol{\epsilon} \right\} \Big) \\
&= \frac{\tau}{1+\tau}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1} \left\{ \boldsymbol{x} \right\} + \frac{1}{1+\tau}\Big(\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1} \left\{ \boldsymbol{x} \right\} + \mathbb{E}_{\boldsymbol{\epsilon}|\boldsymbol{y}_1} \left\{ \boldsymbol{\epsilon} \right\} + \tau\mathbb{E}_{\boldsymbol{\omega}|\boldsymbol{y}_1} \left\{ \boldsymbol{\omega} \right\} \Big) \\
&= \frac{\tau}{1+\tau}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1} \left\{ \boldsymbol{x} \right\} + \frac{1}{1+\tau}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1} \left\{ \boldsymbol{y}_1 \right\} \\
&= \frac{\tau}{1+\tau}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1} \left\{ \boldsymbol{x} \right\} + \frac{1}{1+\tau}\boldsymbol{y}_1
\end{aligned}$$

其中第三行利用了 $\mathbb{E}_{\boldsymbol{\epsilon}|\boldsymbol{y}_1} \left\{ \boldsymbol{x} \right\} = \mathbb{E}_{\boldsymbol{\omega}|\boldsymbol{y}_1} \left\{ \boldsymbol{\omega} \right\}$，因为 $\boldsymbol{\epsilon}$ 和 $\boldsymbol{\omega}$ 是独立同分布的。要使该结果成立，必须知道噪声的分布。

$$f^{\text{test}}(\boldsymbol{y}_1) = \frac{1+\tau}{\tau} f^*(\boldsymbol{y}_1) - \frac{1}{\tau}\boldsymbol{y}_1 \tag{A.2}$$

$$= \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1} \left\{ \boldsymbol{x} \right\} \tag{A.3}$$

(R2R) 损失定义为

$$\mathcal{L}_{\text{R2R}}(\boldsymbol{y}, f) = \mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{y}} \left\{ \| f(\boldsymbol{y}_1) - \boldsymbol{y}_2 \|^2 \right\} \tag{A.4}$$

其中 $\boldsymbol{y}_1 = \boldsymbol{y} + \tau\boldsymbol{\omega}$和$\boldsymbol{y}_2 = \boldsymbol{y} - \frac{1}{\tau}\boldsymbol{\omega}$,- 且 $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{\Sigma})$ 和 $\tau > 0$。正如我们在第2章中所示，这个损失是输入为 $\boldsymbol{y}_1$ 时监督式 ℓ_2 损失的无偏估计量，因此其极小化器为 $f^*(\boldsymbol{y}_1) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}_1} \left\{ \boldsymbol{x} \right\}$，这与 (A.3) 中的 Noisier2Noise 测试时函数相同。

附录 B

矩的识别

从测量分布 $p_{\boldsymbol{y}}$ 恢复信号分布 $p_{\boldsymbol{x}}$，可以看作一个由前向问题定义的无限维逆问题

$$p_{\boldsymbol{y}}(\boldsymbol{y}) = \int_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x}) p_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}. \tag{B.1}$$

当噪声分布的特征函数处处非零时，可以从 $p_{\boldsymbol{y}}$ 识别 $p_{\boldsymbol{x}}$（见第 2.4 节）；对于来自多个算子的不完整观测，如果 $p_{\boldsymbol{x}}$ 的支撑集是低维的，也可以进行识别（见第 3.4 节）。

然而，在某些情况下，我们可能无法识别完整的分布 $p_{\boldsymbol{x}}$，但仍可以求得它的一些高阶矩，或后验分布 $p(\boldsymbol{x}|\boldsymbol{y})$ 的矩。

例 B.1. 假设噪声模型为伯努利噪声模型 $\boldsymbol{y}|\boldsymbol{x} \sim Ber(\boldsymbol{x})$，其中 $p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} x_i^{y_i}(1-x_i)^{1-y_i}$。由于测量值是二元的，我们至多可以识别 $p_{\boldsymbol{y}}$的所有可能输入所对应的 $p_{\boldsymbol{x}}$的 $2^n - 1$ 个不同矩，它们由 $\mathbb{E}_{\boldsymbol{x}}\left\{\prod_{i\in\mathcal{I}} x_i\right\}$ 给出，其中 $\mathcal{I}$ 是从 $\{1,\ldots,n\}$中任意选取的指标。

由于我们 可以计算 $p_{\boldsymbol{y}}$ 的任意阶矩，根据全期望公式可得

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}}\{g(\boldsymbol{y})\} &= \mathbb{E}_{\boldsymbol{x}}\left\{\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\{g(\boldsymbol{y})\}\right\} &\text{(B.2)}\\ &= \mathbb{E}_{\boldsymbol{x}}\{r(\boldsymbol{x})\} &\text{(B.3)}\end{aligned}$$

其中我们定义 $r(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}\{g(\boldsymbol{y})\}$，其中 $g:\mathbb{R}^m \mapsto \mathbb{R}$ 是某个函数。

当 $p_{\boldsymbol{y}}$ 连续且可微时，我们可以通过对 (B.1) 求导来计算 $p_{\boldsymbol{x}}$ 的矩：

$$\begin{aligned}\frac{\partial^k p(\boldsymbol{y})}{\partial {y_i}^k} &= \mathbb{E}_{\boldsymbol{x}}\left\{\frac{\partial^k p(\boldsymbol{y}|\boldsymbol{x})}{\partial {y_i}^k}\right\} &\text{(B.4)}\\ &= \mathbb{E}_{\boldsymbol{x}}\{\tilde{r}_{i,k}(\boldsymbol{x},\boldsymbol{y})\} &\text{(B.5)}\end{aligned}$$

对于任意 $k \geq 0$和 $i = 1,..,n$，其中我们定义了 $\tilde{r}_{i,k}(\boldsymbol{x},\boldsymbol{y})$：$= \frac{\partial^k p(\boldsymbol{y}|\boldsymbol{x})}{\partial {y_i}^k}$。

示例 B.2。 考虑一个高斯噪声模型，其中我们有 $p(\boldsymbol{y}|\boldsymbol{x}) = {}^{1}{}_{(2\pi\sigma^2)_{n/2}} e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}}$，满足 $\tilde{r}_{i,1}(\boldsymbol{x},\boldsymbol{y}) = \frac{y_i - x_i}{\sigma^2} p(\boldsymbol{y}|\boldsymbol{x})$。利用这一结果，我们得到

$$\mathbb{E}_{\boldsymbol{x}}\left\{\frac{y_i - x_i}{\sigma^2} p(\boldsymbol{y}|\boldsymbol{x})\right\} = \frac{y_i}{\sigma^2} p(\boldsymbol{y}) - \frac{1}{\sigma^2}\mathbb{E}_{\boldsymbol{x}}\left\{x_i p(\boldsymbol{y}|\boldsymbol{x})\right\} \tag{B.6}$$

$$\frac{\partial p(\boldsymbol{y})}{\partial y_i} = \frac{y_i}{\sigma^2} p(\boldsymbol{y}) - \frac{1}{\sigma^2}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\left\{x_i\right\} p(\boldsymbol{y}) \tag{B.7}$$

$$\sigma^2 \frac{\partial \log p(\boldsymbol{y})}{\partial y_i} = y_i - \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}}\left\{x_i\right\} \tag{B.8}$$

这就是著名的（Tweedie）公式， $\mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}\} = \boldsymbol{y} + \sigma^2 \nabla$对数$p_{\boldsymbol{y}}$（$\boldsymbol{y}$），也是我们在第*2.2.2*节中推导出的、作为（SURE）损失的（期望意义下）极小化器的同一公式。高阶导数可用于估计更高阶的后验矩， *[123],* 例如后验方差，它等价于最小均方误差：

$$MMSE = \sigma^2 \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log p_{\boldsymbol{y}}}{\partial {y_i}^2}(\boldsymbol{y})\right).$$

附录 C

补充证明

命题 C.1. 设 $\boldsymbol{x} \in \mathbb{R}^n$ 和 $\boldsymbol{y} \in \mathbb{R}^n$ 是服从联合分布 $p(\boldsymbol{x}, \boldsymbol{y})$ 的两个随机变量。以下损失的极小化器为 ℓ_2

$$f^*(\boldsymbol{y}) = \arg\min_f \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 \right\} \tag{C.1}$$

由下式给出

$$f^*(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \{\boldsymbol{x}\}. \tag{C.2}$$

证明。令 $\mathcal{L}(f) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \|f(\boldsymbol{y}) - \boldsymbol{x}\|^2 \right\}$，则有

$$\begin{aligned}
\mathcal{L}(f) =& \ \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \| \left(f(\boldsymbol{y}) - f^*(\boldsymbol{y})\right) - \left(\boldsymbol{x} - f^*(\boldsymbol{y})\right) \|^2 \right\} \\
\propto& \ \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \|f(\boldsymbol{y}) - f^*(\boldsymbol{y})\|^2 \right\} - 2\,\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \left(f(\boldsymbol{y}) - f^*(\boldsymbol{y})\right)^\top \left(\boldsymbol{x} - f^*(\boldsymbol{y})\right) \right\} \\
=& \ \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \|f(\boldsymbol{y}) - f^*(\boldsymbol{y})\|^2 \right\} - 2\,\mathbb{E}_{\boldsymbol{y}} \left\{ \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \left\{ \left(f(\boldsymbol{y}) - f^*(\boldsymbol{y})\right)^\top \left(\boldsymbol{x} - f^*(\boldsymbol{y})\right) \right\} \right\} \\
=& \ \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \|f(\boldsymbol{y}) - f^*(\boldsymbol{y})\|^2 \right\} - 2\,\mathbb{E}_{\boldsymbol{y}} \left\{ \left(f(\boldsymbol{y}) - f^*(\boldsymbol{y})\right)^\top \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \left\{ \boldsymbol{x} - f^*(\boldsymbol{y}) \right\} \right\} \\
=& \ \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \|f(\boldsymbol{y}) - f^*(\boldsymbol{y})\|^2 \right\}
\end{aligned}$$

其中，第四行用到了 $\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \{\boldsymbol{x} - f^*(\boldsymbol{y})\} = 0$ 这一事实。因此，$\mathcal{L}(f)$ 的全局极小值为 $f(\boldsymbol{y}) = f^*(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \{\boldsymbol{x}\}$。 □

命题 C.2。 设 $\boldsymbol{x} \in \mathbb{R}^n$ 和 $\boldsymbol{y} \in \mathbb{R}^m$ 是服从联合分布 $p(\boldsymbol{x}, \boldsymbol{y})$的两个随机变量，并设 $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ 为一个线性算子。以下加权 ℓ_2 损失的极小化器

$$f^*(\boldsymbol{y}) \in \arg\min_f \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} \left\{ \|\boldsymbol{A}f(\boldsymbol{y}) - \boldsymbol{A}\boldsymbol{x}\|^2 \right\} \tag{C.3}$$

由下式给出

$$f^*(\boldsymbol{y}) = \boldsymbol{A}^\dagger \boldsymbol{A}\, \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \{\boldsymbol{x}\} + (\boldsymbol{I} - \boldsymbol{A}^\dagger \boldsymbol{A}) v(\boldsymbol{y}) \tag{C.4}$$

其中，$\boldsymbol{A}^\dagger$ 是 $\boldsymbol{A}$的线性伪逆，$\boldsymbol{A}^\dagger \boldsymbol{A}$ 是向 $\boldsymbol{A}^\top$的值域上的投影，且 $v: \ \mathbb{R}^n \mapsto \mathbb{R}^n$ 是任意函数。

证明。定义 $\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$ 和 $f = \boldsymbol{A} \circ f$，我们可以应用命题 C.1，得出 $f^*(\boldsymbol{y}) = \mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{y}} \{\tilde{\boldsymbol{x}}\}$，或等价地，$\boldsymbol{A}f^*(\boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} \{\boldsymbol{A}\boldsymbol{x}\}$。对两边应用 $\boldsymbol{A}$的线性伪逆，可得到式 (C.4) 中所需的等式。 □

命题 4.1。 设 $\boldsymbol{y}_1$ 和 $\boldsymbol{y}_2$ 是在给定 $\boldsymbol{x}$ 条件下的两个独立随机变量，且满足 $\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{\boldsymbol{y}_2\}=\boldsymbol{x}$。则

$$\mathbb{V}_{\boldsymbol{y}_1,\boldsymbol{y}_2}\{\mathcal{L}_{N2N}(\boldsymbol{y}_1,\boldsymbol{y}_2,f)\}=\mathbb{V}_{\boldsymbol{x},\boldsymbol{y}_1}\left\{\frac{1}{n}\|f(\boldsymbol{y}_1)-\boldsymbol{x}\|^2\right\}+\Delta$$

其中，第一项是监督损失的方差， Δ 是相对于监督情形的额外方差，其表示为

$$\begin{aligned}\Delta=&\mathbb{V}_{\boldsymbol{x},\boldsymbol{y}_2}\left\{\|\boldsymbol{y}_2-\boldsymbol{x}\|^4\right\}+\frac{4}{n^2}\mathbb{E}_{\boldsymbol{x}}\left\{trace\left(\boldsymbol{\Psi}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\right\}\\&-\frac{2}{n^2}\mathbb{E}_{\boldsymbol{x}}\left\{\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{\|\boldsymbol{y}_2-\boldsymbol{x}\|^2(\boldsymbol{y}_2-\boldsymbol{x})^\top\right\}\left(\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{f(\boldsymbol{y}_1)\right\}-\boldsymbol{x}\right)\right\}\end{aligned}$$

其中， $\boldsymbol{\Sigma}_{\boldsymbol{x}}=\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{(\boldsymbol{y}_2-\boldsymbol{x})\,\boldsymbol{y}_2-\boldsymbol{x}\,(^\top\}$ 是噪声目标 $\boldsymbol{y}_2$ 的协方差，而 $\boldsymbol{\Psi}_{\boldsymbol{x}}=\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\{(f(\boldsymbol{y}_1)-\boldsymbol{x})\,f(\boldsymbol{y}_1)-\boldsymbol{x})^\top\}$ 是图像 $\boldsymbol{x}$ 的误差协方差。

证明。 定义 $a_{\boldsymbol{x},\boldsymbol{y}_1}=\frac{1}{n}\|f(\boldsymbol{y}_1)-\boldsymbol{x}\|^2$, $b_{\boldsymbol{x},\boldsymbol{y}_2}=\frac{1}{n}\|\boldsymbol{y}_2-\boldsymbol{x}\|^2$ 以及 $c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}=-\frac{2}{n}(f(\boldsymbol{y}_1))\,\boldsymbol{y}_2-\boldsymbol{x}$, 我们有

$$\begin{aligned}\mathbb{V}_{\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{x}}\left\{\frac{1}{n}\|f(\boldsymbol{y}_1)-\boldsymbol{y}_2\|^2\right\}&=\mathbb{V}_{\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{x}}\left\{a_{\boldsymbol{x},\boldsymbol{y}_1}+b_{\boldsymbol{x},\boldsymbol{y}_2}+c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}\\&=\mathbb{V}\left\{a_{\boldsymbol{x},\boldsymbol{y}_1}\right\}+\Delta\end{aligned}$$

with

$$\begin{aligned}\Delta=&\mathbb{V}\left\{b_{\boldsymbol{x},\boldsymbol{y}_2}\right\}+\mathbb{V}\left\{c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}\\&+2\,\mathbb{E}_{\boldsymbol{x}}\left\{\mathrm{Cov}\left\{a_{\boldsymbol{x},\boldsymbol{y}_1},b_{\boldsymbol{x},\boldsymbol{y}_2}\right\}+\mathrm{Cov}\left\{a_{\boldsymbol{x},\boldsymbol{y}_1},c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}+\mathrm{Cov}\left\{b_{\boldsymbol{x},\boldsymbol{y}_2},c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}\right\}.\end{aligned}$$

其中，$\mathsf{Cov}\{\cdot,\cdot\}$ 表示关于 $p(\boldsymbol{y}_1,\ \boldsymbol{y}_2|\boldsymbol{x})$ 的两个一维随机变量之间的协方差。

第一个方差项就是 $\mathbb{V}\left\{b_{\boldsymbol{x},\boldsymbol{y}_2}\right\}=\mathbb{V}_{\boldsymbol{y}_2,\boldsymbol{x}}\left\{\frac{1}{n}\|\boldsymbol{y}_2-\boldsymbol{x}\|^2\right\}$, 第二个方差可以计算为

$$\begin{aligned}&\mathbb{V}_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\left\{c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}=\frac{4}{n^2}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\left\{\left((f(\boldsymbol{y}_1)-\boldsymbol{x})^\top(\boldsymbol{y}_2-\boldsymbol{x})\right)^2\right\}\\&=\frac{4}{n^2}\mathbb{E}_{\boldsymbol{x}}\left\{\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{(f(\boldsymbol{y}_1)-\boldsymbol{x})(f(\boldsymbol{y}_1)-\boldsymbol{x})^\top\right\}\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{(\boldsymbol{y}_2-\boldsymbol{x})(\boldsymbol{y}_2-\boldsymbol{x})^\top\right\}\right\}\\&=\frac{4}{n^2}\mathbb{E}_{\boldsymbol{x}}\left\{\mathrm{trace}\left(\boldsymbol{\Psi}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\right\}\end{aligned}$$

其中，第一行利用了以下事实: $\mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{x}}\left\{c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}=0$; 第三行则利用了定义 $\boldsymbol{\Sigma}_{\boldsymbol{x}}=\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\{(\boldsymbol{y}_2-\boldsymbol{x})(\boldsymbol{y}_2-\boldsymbol{x})^\top\}$ 和 $\boldsymbol{\Psi}_{\boldsymbol{x}}=\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\{(f(\boldsymbol{y}_1)-\boldsymbol{x})(f(\boldsymbol{y}_1)-\boldsymbol{x})^\top\}$。

我们有，给定$\boldsymbol{x}$的条件下，$\mathsf{Cov}\left\{a_{\boldsymbol{x},\boldsymbol{y}_1},b_{\boldsymbol{x},\boldsymbol{y}_2}\right\}=0$ 因为$a_{\boldsymbol{x},\ \boldsymbol{y}_1}$和$b_{\boldsymbol{x},\ \boldsymbol{y}_2}$相互独立。我们还知道

$$\begin{aligned}\mathrm{Cov}\left\{a_{\boldsymbol{x},\boldsymbol{y}_1},c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}&=\mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{x}}\left\{(a_{\boldsymbol{x},\boldsymbol{y}_1}-\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{a_{\boldsymbol{x},\boldsymbol{y}_1}\right\})c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}\\&=\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{(a_{\boldsymbol{x},\boldsymbol{y}_1}-\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{a_{\boldsymbol{x},\boldsymbol{y}_1}\right\})\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}\right\}\\&=0\end{aligned}$$

剩余的协方差项可计算为

$$
\begin{aligned}
&\operatorname{Cov}\left\{b_{\boldsymbol{x},\boldsymbol{y}_2}, c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\} \\
&= \mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{\left(b_{\boldsymbol{x},\boldsymbol{y}_2} - \mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{b_{\boldsymbol{x},\boldsymbol{y}_2}\right\}\right)\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}\right\} \\
&= \mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{b_{\boldsymbol{x},\boldsymbol{y}_2}\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}\right\} - \mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{b_{\boldsymbol{x},\boldsymbol{y}_2}\right\}\mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2|\boldsymbol{x}}\left\{c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\} \\
&= \mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{b_{\boldsymbol{x},\boldsymbol{y}_2}\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\}\right\} \\
&= -\frac{2}{n^2}\mathbb{E}_{\boldsymbol{y}_2|\boldsymbol{x}}\left\{\|\boldsymbol{y}_2 - \boldsymbol{x}\|^2(\boldsymbol{y}_2 - \boldsymbol{x})^\top\right\}\left(\mathbb{E}_{\boldsymbol{y}_1|\boldsymbol{x}}\left\{f(\boldsymbol{y}_1)\right\} - \boldsymbol{x}\right)
\end{aligned}
$$

其中第二行和第三行利用了这一事实：$\mathbb{E}_{\boldsymbol{y}_1,\ \boldsymbol{y}_2|\boldsymbol{x}}\left\{c_{\boldsymbol{x},\boldsymbol{y}_1,\boldsymbol{y}_2}\right\} = 0$。

□

参考文献

[1] J. A. Fessler，《基于模型的磁共振成像图像重建》，*IEEE*信号处理杂志，第27卷，第4期，第81–89页，2010年。[2] L. I. Rudin、S. Osher 和 E. Fatemi，《基于非线性全变分的去噪算法》，第60卷，第1期，第259–268页。[3] J. Zbontar、F. Knoll、A. Sriram、T. Murrell、Z. Huang、M. J. Muckley、A. Defazio、R. Stern、P. Johnson、M. Bruno、M. Parente、K. J. Geras、J. Katsnelson、H. Chandarana、Z. Zhang、M. Drozdzal、A. Romero、M. Rabbat、P. Vincent、N. Yakubova、J. Pinkerton、D. Wang、E. Owens、C. L. Zitnick、M. P. Recht、D. K. Sodickson 和 Y. W. Lui，《fastMRI：用于加速磁共振成像的开放数据集与基准测试》，2019年12月。[4] E. Shimron、J. I. Tamir、K. Wang 和 M. Lustig，《隐性数据犯罪：滥用公共数据导致的机器学习偏差》，《美国国家科学院院刊》，第119卷，第13期，文章编号 e2117203119，2022年。[5] F. Luisier、C. Vonesch、T. Blu 和 M. Unser，《泊松噪声污染图像的快速跨尺度小波去噪》，《信号处理》，第90卷，第2期，第415–427页，2010年。[在线]。可访问：https://www.sciencedirect.com/science/article/pii/S0165168409003016[6] S. A. Hussein、T. Tirer 和 R. Giryes，《单幅图像超分辨率的校正滤波器：增强现成深度超分辨率模型的鲁棒性》，2020年5月。[7] J. Dong、L. Valzania、A. Maillard、T.-a. Pham、S. Gigan 和 M. Unser，《相位恢复：从计算成像到机器学习》，*IEEE*信号处理杂志，第40卷，第1期，第45–57页，2023年1月。[8] E. Soubies、T.-A. Pham 和 M. Unser，《光学衍射层析成像中多重散射模型的高效求逆》，《*Optics Express*》，第25卷，第18期，第21 786–21 800页，2017年。[9] E. Candes 和 M. Wakin，《压缩采样导论》，*IEEE*信号处理杂志，第25卷，第2期，第21–30页，2008年3月。[10] K. He、X. Zhang、S. Ren 和 J. Sun，《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，2016年，第770–778页。

[11] K. H. Jin, M. T. McCann, E. Froustey 和 M. Unser, 《用于成像逆问题的深度卷积神经网络》, *IEEE*图像处理汇刊, 第26卷, 第9期, 第4509–4522页, 2017年9月。[12] O. Ronneberger、P. Fischer 和 T. Brox, 《U-net: 用于生物医学图像分割的卷积网络》, 载于医学图像计算与计算机辅助介入——*MICCAI 2015*, N. Navab、J. Hornegger、W. M. Wells 和 A. F. Frangi 编。Cham: 施普林格国际出版公司, 2015年, 第234–241页。[13] J. Liang、J. Cao、G. Sun、K. Zhang、L. Van Gool 和 R. Timofte, 《SwinIR: 使用Swin Transformer进行图像复原》, 载于*IEEE/CVF*国际计算机视觉大会论文集, 2021年, 第1833–1844页。.[14] J. Adler 和 O. Oktem, 《学习型原始-对偶重建》, *IEEE*医学成像汇刊, 第37卷, 第6期, 第1322–1332页, 2018年。[15]K. Hammernik、T. Klatzer、E. Kobler、M. P. Recht、D. K. Sodickson、T. Pock 和 F. Knoll, 《学习用于加速磁共振成像数据重建的变分网络》, 《医学磁共振》, 第79卷, 第6期, 第3055–3071页, 2018年。 [在线]。可访问: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.26977[16] B. Recht、R. Roelofs、L. Schmidt 和 V. Shankar, 《ImageNet分类器能否泛化到ImageNet? 》, 载于国际机器学习会议, 2019年。 [在线]。可访问: https://api.semanticscholar.org/CorpusID:67855879[17] M. Terris、S. Hurault、M. Song 和 J. Tachella。《Reconstruct Anything模型: 一种用于计算成像的轻量级基础模型》。[18] C. Belthangady 和 L. A. Royer, 《深度学习在荧光图像重建中的应用、前景与陷阱》, 《自然方法》, 第16卷, 第1215–1225页, 2019年2月。[19] M. M. Kashyap、A. Tambwekar、K. Manohara 和 S. Natarajan, 《无需干净训练数据的语音降噪: 一种Noise2Noise方法》, 载于*Interspeech 2021*论文集, 2021年, 第2716–2720页。[20] P. H. Casajus、T. Ritschel 和 T. Ropinski, 《全量降噪: 三维点云清理的无监督学习》, 载于*2019*年*IEEE/CVF*国际计算机视觉大会（*ICCV*）。首尔, 韩国: IEEE, 2019年10月, 第52–60页。[21] A. Krull、T.-O. Buchholz 和 F. Jug, 《Noise2Void——从单张含噪图像中学习去噪》, 载于*2019*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。美国加利福尼亚州长滩: IEEE, 2019年6月, 第2124–2132页。[22] A. Krull、T. Vič̆ar、M. Prakash、M. Lalit 和 F. Jug, 《概率Noise2Void: 无监督内容感知降噪》, 《计算机科学前沿》, 第2卷, 第5页, 2020年2月。[23] S. Laine、T. Karras、J. Lehtinen 和 T. Aila, 《高质量自监督深度图像降噪》, 载于神经信息处理系统进展, 2019年。

[24] E. Dalsasso、L. Denis 和 F. Tupin，《SAR2SAR：一种用于合成孔径雷达图像的半监督去斑算法》，*IEEE*《应用地球观测与遥感选题期刊》，第14卷，第4321–4329页，2021年。[25] A. A. Hendriksen、D. M. Pelt 和 K. J. Batenburg，《Noise2Inverse：用于断层成像的自监督深度卷积去噪》，*IEEE*计算成像汇刊，第6卷，第1320–1335页，2020年。[26] D. Chen、J. Tachella 和 M. E. Davies，《等变成像：超越值域的学习》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第4379–4388页。[27] B. Yaman、S. A. H. Hosseini、S. Moeller、J. Ellermann、K. Uˇgurbil 和 M. Akçakaya，《无需全采样数据的自监督基于物理的深度学习磁共振成像重建》，载于*2020*年*IEEE*第*17*届生物医学成像国际研讨会（*ISBI*），2020年4月，第921–925页。[28]J. Tachella、M. Davies 和 L. Jacques，《UNSURE：利用未知噪声水平和Stein无偏风险估计进行自监督学习》，载于第十三届国际学习表征会议，2025年。 [在线]。可访问：https://openreview.net/forum?id=ScVnYBaSEw[29] M. Raphan 和 E. P. Simoncelli，《无需先验或监督的最小二乘估计》，《神经计算》，第23卷，第2期，第374–420页，2011年2月。[30] L. Ericsson、H. Gouk、C. C. Loy 和 T. M. Hospedales，《自监督表征学习：简介、进展与挑战》，*IEEE*信号处理杂志，第39卷，第3期，第42–62页，2022年5月。[31] T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton，《视觉表征对比学习的简单框架》，载于第*37*届国际机器学习会议论文集。PMLR，2020年11月，第1597–1607页。[32] J.-B. Grill、F. Strub、F. Altch´e、C. Tallec、P. H. Richemond、E. Buchatskaya、C. Do-ersch、B. A. Pires、Z. D. Guo、M. G. Azar、B. Piot、K. Kavukcuoglu、R. Munos 和 M. Valko，《自举自己的潜在表示：一种新的自监督学习方法》，2020年9月。[33] M. Oquab、T. Darcet、T. Moutakanni、H. Vo、M. Szafraniec、V. Khalidov、P. Fernandez、D. Haziza、F. Massa、A. El-Nouby等，《DINOv2：无需监督学习鲁棒视觉特征》，*arXiv*预印本 *arXiv:2304.07193*，2023年。[34] K. He、X. Chen、S. Xie、Y. Li、P. Doll´ar 和 R. Girshick，《掩码自编码器是可扩展的视觉学习器》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第16 000–16 009页。[35] D. Ulyanov、A. Vedaldi 和 V. Lempitsky，《深度图像先验》，载于*IEEE*计算机视觉与模式识别会议论文集，2018年，第9446–9454页。

[36] R. Heckel 和 P. Hand，“深度解码器：来自未训练非卷积网络的简洁图像表示”，载于国际学习表征会议，2019年。 [在线]。网址：https://openreview.net/forum?id=rylV-2C9KQ[37] M. Z. Darestani 和 R. Heckel，“使用未训练神经网络加速磁共振成像”，*IEEE*计算成像汇刊，第7卷，第724–733页，2021年。[38] D. Ren、K. Zhang、Q. Wang、Q. Hu 和 W. Zuo，“使用深度先验的神经盲反卷积”，载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议（CVPR）。美国华盛顿州西雅图：IEEE，2020年6月，第3338–3347页。[39] J. Tachella、J. Tang 和 M. Davies，“卷积神经网络去噪器与非局部滤波器之间的神经切线联系”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第8618–8627页。[40] J. Lehtinen、J. Munkberg、J. Hasselgren、S. Laine、T. Karras、M. Aittala 和 T. Aila，“Noise2Noise：无需干净数据的图像复原学习”，载于第*35*届国际机器学习会议论文集。PMLR，2018年10月。[41] G. Daras、H. Chung、C.-H. Lai、Y. Mitsufuji、J. C. Ye、P. Milanfar、A. G. Dimakis 和 M. Delbracio，“逆问题扩散模型综述”，2024年9月。[42] U. S. Kamilov、C. A. Bouman、G. T. Buzzard 和 B. Wohlberg，“用于在计算成像中整合物理模型与学习模型的即插即用方法：理论、算法与应用”，第40卷，第1期，第85–97页。[43]F. Rozet、G. Andry、F. Lanusse 和 G. Louppe，“通过期望最大化从观测中学习扩散先验”，第37卷，第87 647–87 682页。 [在线]。网址：https://proceedings.neurips.cc/paper files/paper/2024/hash/9f94298bac4668db4dc77ddb0a244301-Abstract-Conference.html[44] G. Daras、A. Dimakis 和 C. C. Daskalakis，“一致扩散与Tweedie相遇：使用含噪数据训练精确的环境扩散模型”，载于第*41*届国际机器学习会议论文集，机器学习研究论文集系列，R. Salakhutdinov、Z. Kolter、K. Heller、A. Weller、N. Oliver、J. Scarlett 和 F. Berkenkamp 编，第235卷。PMLR，2024年7月，第10 091–10 108页。 [在线]。网址：https://proceedings.mlr.press/v235/daras24a.html[45]J. Tachella、M. Terris、S. Hurault、A. Wang、D. Chen、M.-H. Nguyen、M. Song、T. Davies、L. Davy、J. Dong、P. Escande、J. Hertrich、Z. Hu、T. I. Liaudat、N. Laurent、B. Levac、M. Massias、T. Moreau、T. Modrzyk、B. Monroy、S. Neumayer、J. Scanvic、F. Sarron、V. Sechaud、G. Schramm、R. Vo 和 P. Weiss，“DeepInverse：用于通过深度学习求解成像逆问题的Python软件包”，2025年。 [在线]。网址：https://arxiv.org/abs/2505.20160[46] C. L. Mallows，“关于CP的一些评论”，*Technometrics*，第15卷，第4期，第661–675页，1973年。

[47]B. Efron，“预测误差估计：协方差惩罚与交叉验证”，美国统计学会会刊，第99卷，第467期，第619–632页，2004年9月。[48] T. Bepler、K. Kelley、A. J. Noble 和 B. Berger，“Topaz-Denoise：适用于冷冻电镜和冷冻电子断层成像的通用深度去噪模型”，《自然·通讯》，第11卷，第1期，第5208页，2020年10月。[49] E. Dalsasso、L. Denis 和 F. Tupin，“ 宛如魔法：使用 MERLIN 对深度去斑网络进行自监督训练 ”，*IEEE* 地球科学与遥感汇刊，第60卷，第1–13页，2022年。[50] T. Ehret、A. Davy、J.-M. Morel、G. Facciolo 和 P. Arias，“通过帧间训练实现模型盲视频去噪”，载于*2019*年*IEEE/CVF*计算机视觉与模式识别会议（CVPR），2019年，第11 361–11 370页。[51] N. Moran、D. Schmidt、Y. Zhong 和 P. Coady，“Noisier2Noise：从未配对噪声数据中学习去噪”，载于*2020*年*IEEE/CVF*计算机视觉与模式识别会议（CVPR）。美国华盛顿州西雅图：IEEE，2020年6月，第12 061–12 069页。[52] T. Pang、H. Zheng、Y. Quan 和 H. Ji，“再污染到再污染：用于图像去噪的无监督深度学习”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第2043–2052页。[53] C. M. Stein，“多元正态分布均值的估计”，统计年刊，第9卷，第6期，第1135–1151页，1981年。[54] N. L. Oliveira、J. Lei 和 R. J. Tibshirani，“通过耦合自助法技术在正态均值问题中进行无偏风险估计”，2022年10月。[55] B. Monroy、J. Bacca 和 J. Tachella，“广义再污染到再污染：超越高斯噪声的自监督学习”，载于*2025*年*IEEE*计算机视觉与模式识别会议。IEEE，2024年12月5日。[56] J. Rapp、J. Tachella、Y. Altmann、S. McLaughlin 和 V. K. Goyal，“面向自动驾驶车辆的单光子激光雷达进展：工作原理、挑战与最新进展”，*IEEE*信号处理杂志，第37卷，第4期，第62–71页，2020年7月。[57] B. Efron，理论与实践中的指数族。剑桥大学出版社，2022年。[58] J. Leiner、B. Duan、L. Wasserman 和 A. Ramdas，“数据分裂：拆分单个数据点”，美国统计学会会刊，第120卷，第549期，第135–146页，2025年。[59] D. L. Donoho 和 I. M. Johnstone，“通过小波收缩适应未知光滑度”，美国统计学会会刊，第90卷，第432期，第1200–1224页，1995年。

[60] C. A. Metzler、A. Mousavi、R. Heckel 和 R. G. Baraniuk，《利用 Stein 无偏风险估计量进行无监督学习》，2016年7月。[61] M. Zhussip、S. Soltanayev 和 S. Y. Chun，《将 Stein 无偏风险估计量扩展至利用相关噪声图像对训练深度去噪器》，载于神经信息处理系统进展，第 32 卷。Curran Associates, Inc.，2019年。[62] D. Chen、J. Tachella 和 M. E. Davies，《鲁棒等变成像：一种从含噪和不完整测量中学习成像的完全无监督框架》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第 5647–5656 页。[63] H. M. Hudson，《指数族的一个自然恒等式及其在多参数估计中的应用》，统计年刊，第 6 卷，第 3 期，第 473–484 页，1978年。[64] Y. Le Montagner、E. D. Angelini 和 J.-C. Olivo-Marin，《混合泊松–高斯噪声下图像去噪的无偏风险估计器》，*IEEE*图像处理汇刊，第 23 卷，第 3 期，第 1255–1268 页，2014年3月。[65] Y. Zhang、Y. Zhu、E. Nichols、Q. Wang、S. Zhang、C. Smith 和 S. Howard，《包含真实荧光显微图像的泊松-高斯去噪数据集》，2019年4月。[66] Q. Ding、Y. Long、X. Zhang 和 J. A. Fessler，《利用混合泊松-高斯噪声模型进行X射线CT统计图像重建》，2018年1月。[67] Y. Eldar，《指数族的广义SURE：在正则化中的应用》，*IEEE*信号处理汇刊，第 57 卷，第 2 期，第 471–481 页，2009年2月。[68] S. Soltanayev、R. Giryes、S. Y. Chun 和 Y. C. Eldar，《关于基于 Stein 无偏风险估计量的深度去噪器无监督训练中的散度近似》，载于*ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2020年，第 3592–3596 页。[69] M. F. Hutchinson，《拉普拉斯平滑样条影响矩阵迹的随机估计器》，统计通讯：模拟与计算，第 18 卷，第 3 期，第 1059–1076 页，1989年。[70] S. Ramani、T. Blu 和 M. Unser，《蒙特卡罗SURE：通用去噪算法正则化参数的黑箱优化》，*IEEE*图像处理汇刊，第 17 卷，第 9 期，第 1540–1554 页，2008年9月。[71] N. L. Oliveira、J. Lei 和 R. J. Tibshirani，《通过耦合自助法技术对泊松均值问题进行无偏测试误差估计》，2023年8月。[72] K. Kim 和 J. C. Ye，《Noise2Score：在没有干净图像的情况下，利用 Tweedie 方法进行自监督图像去噪》，载于神经信息处理系统进展，2021年。

[73] K. Kim、T. Kwon 和 J. C. Ye，《利用 Tweedie 分布和得分匹配进行噪声分布自适应自监督图像去噪》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 2008–2016 页。[在线]。获取地址：http://openaccess.thecvf.com/content/CVPR2022/html/ Kim Noise Distribution Adaptive Self-Supervised Image Denoising Using Tweedie Distribution and CVPR 2022 paper.html[74] R. Gribonval，《惩罚最小二乘回归应被解释为最大后验估计吗？》*IEEE*信号处理汇刊，第 59 卷，第 5 期，第 2405–2410 页，2011年5月。[75] J. Batson 和 L. Royer，《Noise2Self：通过自监督实现盲去噪》，载于第*36*届国际机器学习会议论文集。PMLR，2019年6月。[76] T. Huang、S. Li、X. Jia、H. Lu 和 J. Liu，《Neighbor2Neighbor：利用单幅噪声图像进行自监督去噪》，载于*2021 IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*。美国田纳西州纳什维尔：IEEE，2021年6月，第 14 776–14 785 页。[77] C. Millard 和 M. Chiew，《利用可变密度子采样和 Noisier2Noise 进行自监督MR图像重建的理论框架》，*IEEE*计算成像汇刊，2023年。[78] H. Robbins，《统计决策问题的经验贝叶斯方法》，数学统计年刊，第 35 卷，第 1 期，第 1–20 页，1964年。[79] J. Tachella、D. Chen 和 M. Davies，《线性逆问题无监督学习的感知定理》，机器学习研究杂志 *(JMLR)*，2023年。[80]M. Prakash、A. Krull 和 F. Jug，《利用卷积变分自编码器实现完全无监督多样性去噪》，载于国际学习表征会议论文集。[在线]。获取地址：https://openreview.net/ forum?id=agHLCOBM5jP[81] M. Prakash、M. Delbracio、P. Milanfar 和 F. Jug，《可解释的无监督多样性去噪与伪影去除》，载于国际学习表征会议论文集，2021年10月。[82] A. Bora、E. Price 和 A. G. Dimakis，《AmbientGAN：基于有损测量的生成模型》，载于国际学习表征会议，2018年。[83] B. Kawar、N. Elata、T. Michaeli 和 M. Elad，《基于GSURE的受损数据扩散模型训练》，*arXiv*预印本 *arXiv:2305.13128*，2023年。[84] Z. Wang、J. Liu、G. Li 和 H. Han，《Blind2Unblind：利用可见盲点进行自监督图像去噪》，载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，2022年，第 2027–2036 页。

[85] G. Daras、K. Shah、Y. Dagan、A. Gollakota、A. Dimakis 和 A. Klivans，《环境扩散：从受损数据中学习干净分布》，神经信息处理系统进展，第36卷，2024年2月。[86] W. Gan、C. Ying、P. E. Boroojeni、T. Wang、C. Eldeniz、Y. Hu、J. Liu、Y. Chen、H. An 和 U. S. Kamilov，《具有理论保证并应用于磁共振成像重建的自监督深度均衡模型》，*IEEE*计算成像汇刊，第9卷，第796–807页，2023年。[87] J. Tachella、D. Chen 和 M. Davies，《用于逆问题的不完整测量值无监督学习》，神经信息处理系统进展，第35卷，第4983–4995页，2022年12月。[88] M. M. Bronstein、J. Bruna、T. Cohen 和 P. Veli˘ckovi´c，《几何深度学习：网格、群、图、测地线与规范》。arXiv，2021年5月。[89]T. S. Cohen 和 M. Welling，《可转向卷积神经网络》，载于国际学习表征会议，2017年。 [在线]。链接：https://openreview.net/forum? id=rJQKYt5ll[90] V. Sechaud、J. Scanvic、Q. Barth´elemy、P. Abry 和 J. Tachella，《等变分裂：从不完整数据中进行自监督学习》，*arXiv*预印本 *arXiv:2510.00929*，2025年。[91] J. Scanvic、M. Davies、P. Abry 和 J. Tachella，《用于图像超分辨率和去模糊的自监督学习》，2023年12月。[92] I. Gulrajani、F. Ahmed、M. Arjovsky、V. Dumoulin 和 A. Courville。《改进Wasserstein生成对抗网络的训练》。 [在线]。链接：http://arxiv.org/abs/1704.00028 [93] E. K. Cole、F. Ong、S. S. Vasanawala 和 J. M. Pauly，《利用生成对抗网络实现无需全采样真实数据的快速无监督磁共振成像重建》，载于*2021*年*IEEE/CVF*国际计算机视觉会议研讨会 *(ICCVW)*。加拿大不列颠哥伦比亚省蒙特利尔：IEEE，2021年10月，第3971–3980页。[94] M. Bendel、R. Ahmad 和 P. Schniter，《用于图像恢复问题后验采样的正则化条件生成对抗网络》，神经信息处理系统进展，第36卷，第68 673–68 684页，2023年。[95] A. Aali、G. Daras、B. Levac、S. Kumar、A. Dimakis 和 J. Tamir，《环境扩散后验采样：利用在受损数据上训练的扩散模型求解逆问题》，载于第十三届国际学习表征会议，2025年。[在线]。链接：https://openreview.net/forum?id=qeXcMutEZY[96] H. Cramer 和 H. Wold，《关于分布函数的一些定理》，伦敦数学学会杂志，第11卷，第4期，第290–294页，1936年。[97] C. Fefferman，《从输出重建神经网络》，载于神经信息处理系统进展，1993年。[98] K. Falconer，《分形几何：数学基础与应用》。John Wiley & Sons，2013 年。

[99] G. Puy、M. E. Davies 和 R. Gribonval，《从希尔伯特空间到 $\mathbb{R}^m$的稳定线性嵌入方法》，*IEEE* 信息论汇刊，第63卷，第4期，第2171–2187页，2017年4月。[100] T. Sauer、J. A. Yorke 和 M. Casdagli，《嵌入学》，统计物理学杂志，第65卷，第3期，第579–616页，1991年11月。[101] A. Wang 和 M. Davies，《透视等变成像：用于多光谱全色锐化的无监督框架》，2024年3月。[102] M. Hardt 和 B. Recht，《模式、预测与行动：机器学习基础》。普林斯顿大学出版社，2022年。[103] T. Klug、D. Atik 和 R. Heckel，《分析自监督图像重建方法的样本复杂度》，神经信息处理系统进展，第36卷，第65 869–65 893页，2023年12月。[104] K. Zhang、Y. Li、W. Zuo、L. Zhang、L. Van Gool 和 R. Timofte，《采用深度去噪器先验的即插即用图像恢复》，*IEEE* 模式分析与机器智能汇刊，第44卷，第10期，第6360–6376页，2021年。[105] P. C. Bellec 和 C.-H. Zhang，《二阶斯坦：用于SURE的SURE及其在高维推断中的其他应用》，统计年刊，第49卷，第4期，第1864–1903页，2021年。[106] C. Zhang、S. Bengio、M. Hardt、B. Recht 和 O. Vinyals，《理解深度学习（仍然）需要重新思考泛化》，美国计算机协会通讯，第64卷，第3期，第107–115页，2021年。[107] M. Hardt、B. Recht 和 Y. Singer，《训练更快，泛化更好：随机梯度下降的稳定性》，载于国际机器学习会议。PMLR，2016年，第1225–1234页。[108] S. Mohan、J. L. Vincent、R. Manzorro、P. Crozier、C. Fernandez-Granda 和 E. Simoncelli，《通过增益调优实现自适应去噪》，神经信息处理系统进展，第34卷，第23 727–23 740页，2021年。[109] L. Jacques、J. N. Laska、P. T. Boufounos 和 R. G. Baraniuk，《通过稀疏向量的二值稳定嵌入实现稳健的1比特压缩传感》，2015年11月。[110] Y. Shechtman、Y. C. Eldar、O. Cohen、H. N. Chapman、J. Miao 和 M. Segev，《用于光学成像的相位恢复：当代综述》，*IEEE* 信号处理杂志，第32卷，第3期，第87–109页，2015年5月。[111] J. Tachella 和 L. Jacques，《仅凭二值测量学习重建信号》，机器学习研究汇刊，2023年，精选认证。[在线]。链接：https://openreview.net/forum?id=ioFIAQOBOS[112] V. Sechaud、L. Jacques、P. Abry 和 J. Tachella，《基于等变性的自监督学习：从截幅测量中恢复音频信号》，载于*EUSIPCO 2024*，2024年8月。

[113] Y. Hu、W. Gan、C. Ying、T. Wang、C. Eldeniz、J. Liu、Y. Chen、H. An 和 U. S. Kamilov，“SPICE：采用自动线圈灵敏度估计的 MRI 自监督学习”，2022年10月。[114] M. Kellman、K. Zhang、E. Markley、J. Tamir、E. Bostan、M. Lustig 和 L. Waller，“面向大规模计算成像的内存高效学习”，*IEEE*计算成像汇刊，第 6 卷，第 1403–1414 页，2020年。[115] F. Ong、X. Zhu、J. Y. Cheng、K. M. Johnson、P. E. Z. Larson、S. S. Vasanawala 和 M. Lustig，“极限磁共振成像：基于连续非门控采集的大规模体积动态成像”，《医学磁共振》，第 84 卷，第 4 期，第 1763–1780 页，2020年10月。.[116] J. Rudzusika、B. Baji´c、T. Koehler 和 O. Oktem，“采用内存高效学习型原始-对偶架构的三维螺旋CT重建”，*IEEE*计算成像汇刊，第 10 卷，第 1414–1424 页，2024年。[117]C. Ophus，“四维扫描透射电子显微镜（4d- stem）：从扫描纳米衍射到叠层成像及更多”，《显微镜与微量分析》，第 25 卷，第 3 期，第 563–582 页，2019年6月。 [在线]。链接：https://doi.org/10.1017/S1431927619000497[118]H. Erdogan 和 J. A. Fessler，“透射断层成像的有序子集算法”，《医学与生物学物理学》，第 44 卷，第 11 期，第 2835 页，1999年11月。[在线]。链接：https://dx.doi.org/10.1088/0031-9155/44/11/311[119] J. Tang、K. Egiazarian、M. Golbabaee 和 M. Davies，“随机优化在成像逆问题中的实用性”，*IEEE*计算成像汇刊，第 6 卷，第 1471–1485 页，2020年。[120] O. Kosomaa、S. Laine、T. Karras、M. Aittala 和 J. Lehtinen，“基于模拟器的自监督学习：用于学习型三维断层成像重建”，2023年5月。[121] J. Tang、S. Mukherjee 和 C.-B. Sch¨onlieb，“随机原始-对偶深度展开”，2022年。 [在线]。链接：https://arxiv.org/abs/2110.10093[122] Z. Xia 和 A. Chakrabarti，“无需图像真实值训练图像估计器”，神经信息处理系统进展，第 32 卷，2019年。[123] H. Manor 和 T. Michaeli，“论去噪中的后验分布：在不确定性量化中的应用”，载于第十二届国际学习表征会议，2024年。[124] J. Tachella 和 M. Pereyra，“成像逆问题不确定性量化的等变自助法”，载于第二十七届国际人工智能与统计会议论文集。PMLR，2024年4月。[125] D. P. Kingma 和 M. Welling，“自编码变分贝叶斯”，载于第二届国际学习表征会议，*ICLR 2014*，加拿大艾伯塔省班夫，*2014*年*4*月*14*日至*16*日，会议论文集，2014年。

[126]I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio，“生成对抗网络”，载于神经信息处理系统进展，2014年，第 2672–2680 页。 [在线]。链接：http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf[127] D. Donoho，“高维数据分析：维数的祸福”，*AMS* 数学挑战讲座，第 1–32 页，2000年1月。[128]A. Wang、S. McDonagh 和 M. Davies，“加速磁共振成像重建自监督学习方法基准测试”，2025年。 [在线]。链接：https: //arxiv.org/abs/2502.14009