arXiv:2303.01965v1 [math.NA] 1 Mar 2023

深度神经网络反演的提升式布雷格曼公式

Xiaoyu Wang [1], * **和 Martin Benning** [2,3]

[1]应用数学与理论物理系，剑桥大学，剑桥，英国[2]数学科学学院，理学与工程学院，伦敦玛丽女王大学，伦敦，英国[3]艾伦·图灵研究所，伦敦，英国

通讯作者*：Martin Benning
m.benning@qmul.ac.uk

摘要

我们提出了一种用于深度神经网络正则化反演的新框架。该框架基于作者近期关于无需对激活函数求导即可训练前馈神经网络的研究。通过引入辅助变量，该框架将参数空间提升到更高维的空间，并使用专门设计的布雷格曼距离对这些变量进行惩罚。我们提出了一系列基于这些布雷格曼距离的变分正则化，给出了理论结果，并通过数值示例验证其实际应用。具体而言，据我们所知，我们首次给出了单层感知机正则化反演的收敛性结果；该结果仅假设逆问题的解属于正则化算子的值域，并证明当测量误差趋于零时，正则化逆解必然收敛到真实逆解。

关键词：逆问题、正则化理论、提升式网络训练、Bregman 距离、感知机、多层感知机、变分正则化、全变差正则化

1 引言

在过去十年中，神经网络作为计算系统彻底改变了众多研究领域，并且表现优于许多传统机器学习方法（参见 LeCun 等人，2015年；Goodfellow 等人，2016年）。这种性能提升往往以牺牲神经网络针对给定输入所产生输出的可解释性（或者说，缺乏可解释性）为代价。因此，许多研究致力于理解神经网络的表示，并探索解释这些表示的策略，其中主要采用显著性图（Simonyan 等人，2013年；Fong 和 Vedaldi，2017年；Chang 等人，2018年；Fong 等人，2019年）。另一种方法则通过反演深度图像表示来加以理解（Mahendran 和 Vedaldi，2015年）。作者提出了一种基于全变差的变分优化方法，旨在通过正则化反演，根据网络输出推断网络输入。

尽管神经网络求逆这一概念当然并不新鲜（参见 Linden 和 Kindermann, 1989；Kindermann 和 Linden, 1990；Jensen 等人, 1999；Lu 等人, 1999），但近年来人们对此兴趣日益浓厚，这主要得益于非线性降维和生成

建模领域的发展，其中包括但不限于（变分）自编码器（Kingma 和 Welling, 2013年）、归一化流（Rezende 和 Mohamed, 2015年；Dinh 等人, 2015年）、循环一致生成对抗网络（Zhu 等人, 2017年）以及概率扩散模型（Sohl-Dickstein 等人, 2015年；Ho 等人, 2020年）。

尽管已有多种神经网络反演方法，尤其是在生成建模背景下提出了这些方法（例如，归一化流领域的 (Behrmann 等人，2019年，2021年)、生成对抗网络领域的 (Xia 等人，2022年)，以及概率扩散模型领域的 (Gal 等人，2022年)），但一个常被忽视的重要方面是：仅有可逆运算并不能保证其对数据中的微小变化具有稳定性。例如，在固定的终止时间后求解热方程，对于初始条件的变化是稳定的；但根据热方程的终端条件估计初始条件，则无法保证对终端条件中的扰动保持稳定。若不使用一族连续算子对逆运算进行近似，就无法解决这一问题；这种近似也称为正则化。反问题与病态问题 领域及其分支正则化理论 主要研究如何通过正则化 (Engl 等人，1996年)和所谓的变分正则化(Scherzer 等人，2009年；Benning 和 Burger，2018年)稳定地逼近病态和病态条件反演；后者是（非线性）正则化的一种特殊类别。(Mahendran 和 Vedaldi, 2015年)提出的优化模型可视为一种采用全变差正则化的变分正则化方法；然而，(Mahendran 和 Vedaldi，2015年)的研究纯属经验性工作，据我们所知，尚无研究严格证明所提出的方法属于变分正则化。

在这项工作中，我们提出了一种新颖的正则化框架，该框架基于提升并采用量身定制的 Bregman 距离；我们证明，对于从单层感知机估计输入的逆问题，或依次估计多层感知机中隐变量的逆问题，所提出的框架是一种收敛的变分正则化方法。过去几年中，已有大量研究致力于在变分正则化方法中将神经网络用作非线性算子（Lunz 等，2018年；Arridge 等，2019年；Schwab 等，2019年；Li 等，2020年；Mukherjee 等，2021年）；据我们所知，这是首项为神经网络的稳定、基于模型的反演提供理论保证的工作。

我们的贡献主要有三点。1）我们提出了一种多层感知机（即前馈神经网络）正则化反演的新框架，该框架基于作者近期在 Wang 和 Benning（2022）中提出的提升的布雷格曼框架。2）我们证明，对于单层感知机的情形，在非常温和的假设下，所提出的变分正则化方法可证明是收敛的。据我们所知，这是首次提出一种反演方法，它不仅能在经验上实现反演，而且还能证明所提方法是一种收敛的正则化方法，同时无需采用过于严格的假设，例如激活函数可微以及满足切锥条件。3）我们提出了一种近端一阶优化策略来求解所提出的变分正则化方法，并给出了若干数值示例，以验证所提出的基于模型的正则化方法的有效性。

本文结构如下。第2节介绍用于前馈神经网络基于模型的反演的提升的布雷格曼公式。第3节证明，对于单层感知机的情形，所提出的模型是一种收敛的变分正则化方法，并给出一般误差估计，以及针对一个采用 ReLU 激活函数的具体感知机实例的误差估计。第4节讨论如何以计算方式实现所提出的变分正则化

方法，并将原始-对偶混合梯度法和坐标下降推广到单层和多层感知机两种情形。随后，我们在第 5 节展示数值结果，以经验方式证明所提出的方法是一种基于模型的正则化；最后，在第 6 节简要总结并展望未来工作。

2 基于模型的前馈网络反演

假设给定一个具有 L 层的前馈神经网络 $\mathcal{N}\colon\ \mathbb{R}^n \times \mathcal{P} \to \mathbb{R}^m$，其形式为

$$\mathcal{N}(x, \boldsymbol{\Theta}) = \sigma_L(f(\sigma_{L-1}(f(\ldots \sigma_1(f(x, \Theta_1))\ldots)), \Theta_L)), \tag{1}$$

对于输入数据 $x \in \mathbb{R}^n$ 和**预训练**参数 $\boldsymbol{\Theta} \in \mathcal{P}$。这里，$\{\sigma_l\}_{l=1}^L$ 表示非线性激活函数的集合，f 表示由参数 $\{\Theta_l\}_{l=1}^L$ 参数化的通用函数。为简化记号，我们用 $\boldsymbol{\Theta}$ 表示所有参数 $\{\Theta_l\}_{l=1}^L$。对于给定的网络输出 $y \in \mathbb{R}^m$，我们的目标是求解逆问题

$$\mathcal{N}(x, \boldsymbol{\Theta}) = y$$

以求得未知输入 $x \in \mathbb{R}^n$。我们提出通过最小化如下形式的提升式 Bregman 公式，来近似求解这一非线性且可能适定性不良的逆问题的逆

$$\begin{pmatrix} x^\alpha \\ x_1^\alpha \\ \vdots \\ x_{L-1}^\alpha \end{pmatrix} \in \underset{x, x_1, \ldots, x_{L-1}}{\arg\min} \left\{ \sum_{l=1}^{L} B_{\Psi_l}(x_l, f(x_{l-1}, \Theta_l)) + \alpha R(x) \right\}, \tag{2}$$

其中，为简化记号，我们假设 $x_0 = x$ 和 $x_L = y^\delta$。数据 y^δ 是 y 的扰动版本，对此我们假设 $B_{\Psi_L}(y^\delta, f(x_{L-1}^\dagger,\ \Theta_L)) \leq \delta^2$，其中 $\delta \geq 0$ 为某个常数，且 $y = \sigma_L(f(x_{L-1}^\dagger,\ \Theta_L))$。对于 $l = 1,\ \ldots,\ L$，函数 B_{Ψ_l} 定义为

$$B_{\Psi_l}(x, z) = \frac{1}{2}\|x\|^2 + \Psi_l(x) + \left(\frac{1}{2}\|\cdot\|^2 + \Psi_l\right)^*(z) - \langle x, z\rangle, \tag{3}$$

其中 $\Psi_l\colon\ \mathbb{R}^{n_l} \to \mathbb{R} \cup \{\infty\}$ 是一个正常、凸且下半连续函数。记号 $\left(\frac{1}{2}\|\cdot\|^2 + \Psi_l\right)^*$ 表示 $\frac{1}{2}\|\cdot\|^2 + \Psi_l$ 的凸共轭或 Fenchel 共轭，即 $\left(\frac{1}{2}\|\cdot\|^2 + \Psi_l\right)^*(z) = \sup_y \langle z, y\rangle - \frac{1}{2}\|y\|^2 - \Psi_l(y)$。最后，函数 $R\colon\ \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ 是一个正常、凸且下半连续函数，使我们能够将先验信息纳入反演过程。其影响由参数 $\alpha > 0$控制。

请注意，函数 B_{Ψ_l} 具有一些有用的性质，并且与所选的激活函数 $\{\sigma_l\}_{l=1}^L$直接相关。根据（Wang 和 Benning，2022年），我们观察到

$$B_{\Psi_l}(x, z) \geq \frac{1}{2}\|\sigma_l(z) - x\|^2,$$

其中 $\sigma_l\colon\ \mathbb{R}^{n_l} \to \mathbb{R}^{n_l}$ 是关于 Ψ_l的近端映射，即

$$\sigma_l(z) = \operatorname*{arg\,min}_{y\in\mathbb{R}^{n_l}} \left\{ \frac{1}{2}\|y-z\|^2 + \Psi_l(y) \right\},$$

对于所有 $l \in \{1,\dots,L\}$。这意味着我们将仅关注具有**作为近端映射的非线性激活函数**的前馈神经网络。

另一个有用的性质是，函数 B_{Ψ_l} 关于其第二个参数连续可微。如果我们定义 $F_x^l(z) := B_{\Psi_l}(x,\ z)$，则可观察到

$$\nabla F_x^l(z) = \sigma_l(z) - x\,. \tag{4}$$

请注意，目标函数族 B_{Ψ_l} 还满足其他几个有趣的性质；我们建议感兴趣的读者参阅（Wang 和 Benning，2022年，定理 10）。

在本文余下部分，我们假设参数化函数 f 关于第一个参数是仿射线性的，其参数为 Θ_l。一个具体例子是仿射线性变换 $f(x,\ \Theta_l) = W_l x + b_l$，其中 $W_l \in \mathbb{R}^{n_l\times n_{l-1}}$为（权重）矩阵，$b_l \in \mathbb{R}^{n_l}$ 为（偏置）向量，$\Theta_l = (W_l,\ b_l)$为参数集合。

在下一节中，我们将证明，对于 $L = 1$，(2) 是一种变分正则化方法，并证明当 δ趋于零时，(2) 的解以某一收敛速率趋近于感知机的真实输入。

3 收敛性分析与误差估计

本节中，我们将证明，对于特定的选择 $L = 1$ 以及假设 $f(x,\ \Theta) = Wx + b$ for $\Theta = (W,\ b)$，所提出的模型 (2) 是一个收敛的变分正则化方法；在此条件下，(2) 可化为 Wang 和 Benning（2020）研究的感知机情形下的变分正则化模型。与 Wang 和 Benning（2020）不同，我们不打算估计感知机参数 W 和 b，而是假设它们固定，并研究正则化算子

$$\mathcal{R}_\alpha\colon \operatorname{dom}(\Psi) \rightrightarrows \mathbb{R}^n\,, \qquad \mathcal{R}_\alpha\colon y^\delta \rightrightarrows x_\alpha \in \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \left\{ B_\Psi\left(y^\delta, Wx+b\right) + \alpha R(x) \right\}, \tag{5}$$

其中，$\operatorname{dom}(\Psi)$ 定义为 $\operatorname{dom}(\Psi) := \{y \in \mathbb{R}^m \mid \Psi(y) < \infty\}$。我们首先要确定在什么假设下，(5) 对所有 y^δ都有定义。

3.1 良定性

为简便起见，我们聚焦于网络输入属于 $\mathbb{R}^n$ 且输出属于 $\operatorname{dom}(\Psi)$ 的有限维情形。然而，在对算子 W 施加额外假设的情况下，以下分析也可推广到更一般的巴拿赫空间情形，参见例如（Benning 和 Burger，2018年，第 5.1 节）。沿用（Benning 和 Burger，2018年）的设定，我们假设 R 是非负的，并且是真函数的极函数，即对于真函数 $H\colon\ \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$，有 $R = H^*$。注意，这自动意味着 R 的凸性。此外，我们假设 Ψ 是一个真函数、非负凸函数，并且在 $\operatorname{dom}(\Psi)$ 上连续，这意味着 B_Ψ 是真函数、非负函数、关于其第二个自变量凸，并且

在其首个 对于每个 $y^\delta \in \text{dom}(\Psi)$，其关于第一个自变量的函数。于是，对于每个 $g \in \text{dom}(\Psi)$，存在 x，使得

$$B_\Psi(g, Wx + b) + \alpha R(x) < \infty \,.$$

最后但同样重要的是，我们假设 R 和 Ψ 的选取满足：对于每个 $g \in$ 定义域(Ψ) 和 $\alpha > 0$，我们都有

$$\|x\| \leq c(a, b, \|g\|), \qquad \text{if} \quad B_\Psi(g, Wx + b) \leq a \quad \text{and} \quad \alpha R(x) \leq d \,,$$

对于常数 a，d 以及一个常数 c，该常数关于所有参数单调不减。基于这些假设，我们随后可以验证以下引理。

定理 1. 设上一段中概述的假设均得到满足
1. 那么，对于每个 $y \in \left\{ g \in \text{定义域}(\Psi) \;\middle|\; \arg\min_{x \in \mathbb{R}^n, R(x) < \infty} B_\Psi(g, Wx + b) \neq \emptyset \right\}$ 选择算子

$$\mathcal{S}(y) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ R(x) \;\middle|\; x \in \operatorname*{arg\,min}_{\tilde{x} \in \mathbb{R}^n} B_\Psi(y, W\tilde{x} + b) \right\}$$

是良定义的。
2. 按 (2) 定义的正则化算子 $\mathcal{R}_\alpha$ 是良定义的，其含义是：对于每个 $y \in$ 定义域(Ψ)，都存在 $x_\alpha \in \mathbb{R}^n$ 使得 $x_\alpha \in \mathcal{R}_\alpha(y)$。此外，集合 $\mathcal{R}_\alpha(y)$ 是一个凸集。
3. 对于每个序列 $y_n \to y \in$ 定义域(Ψ)，都存在一个子序列 $x_{n_k} \in \mathcal{R}_\alpha(y_{n_k})$，收敛到某个元素 $x^* \in \mathcal{R}_\alpha(y)$。

证明。结论可直接由 Benning 和 Burger，2018 的引理 5.5、定理 5.6 和定理 5.7 得出。后一个结论原本只能推出弱星拓扑下的收敛；然而，由于我们处于有限维希尔伯特空间中，这在此处自动意味着强收敛。

3.2 误差估计

在证明了 (2) 是一个正则化算子之后，我们现在要证明，根据如下估计，它也是一个收敛正则化算子

$$D_R(x^\dagger, x^\alpha) \leq C\delta \,, \tag{6}$$

使得

$$\limsup_{\delta \to 0} \left\{ D_R(x^\dagger, x^\alpha) \;\middle|\; x^\alpha \in \mathcal{R}_\alpha(y^\delta),\, y^\delta \in \text{dom}(\Psi),\, B_\Psi(y^\delta, y) \leq \delta^2 \right\} = 0 \,.$$

这里，术语 D_R 表示关于 R的（广义）Bregman 距离（或散度）（参见 (Bregman, 1967; Kiwiel, 1997)），即

$$D_R(x, \tilde{x}) = R(x) - R(\tilde{x}) - \langle \tilde{q}, x - \tilde{x} \rangle \,,$$

对于两个参数 x，$\tilde{x} \in$ 定义域(R) 和一个次梯度 $\tilde{q} \in \partial R(\tilde{x}) = \{q \in \mathbb{R}^n \,|\, R(x) \geq R(\tilde{x}) + \langle q, x - \tilde{x} \rangle, \quad \forall x \in$ 定义域$(R)\}$。向量 x^α 是带有数据 y^δ 的式 (2) 的一个解，其中我们假设 $B_\Psi(y^\delta, y) \leq \delta^2$，且 $C \geq 0$ 是一个常数。向量 $x^\dagger$ 是引理1.1中所规定的选择算子的一个元素，即对于 $y \in$ 定义域(Ψ)，$x^\dagger \in \mathcal{S}(y)$。注意，$x^\dagger \in \mathcal{S}(y)$ 等价于 $x^\dagger$ 是所有满足条件的向量中一个 R-最小化向量

满足 $0 = W^*\left(\sigma(Wx^\dagger + b) - y\right)$的所有向量，其中 σ 表示关于 Ψ的近端映射。这是因为 $x^\dagger \in \arg\min_{\tilde{x}\in\mathbb{R}^n} B_\Psi(y,\ W\tilde{x} + b)$ 等价于 $0 = \nabla B_\Psi(y,\ Wx^\dagger + b) = W^*\left(\sigma(Wx^\dagger + b) - y\right)$。假设 $\sigma(Wx^\dagger + b) - y$不属于 W^*的零空间，则进一步推出 $y = \sigma(Wx^\dagger + b)$。

为了能够推导出形如（6）的误差估计，我们将研究范围限制在满足 $x^\dagger$ 且属于 $\mathcal{R}_\alpha$值域的解上。这意味着存在 $y^\dagger$，使得 $x^\dagger \in \mathcal{R}_\alpha(y^\dagger)$。考虑（2）关于 $y^\dagger$的最优性条件，可得

$$W^*\left(\frac{y^\dagger - \sigma(Wx^\dagger + b)}{\alpha}\right) \in \partial R(x^\dagger)\,,$$

对于 $v^\dagger :\,= (y^\dagger - \sigma(Wx^\dagger + b))/\alpha = (y^\dagger - y)/\alpha$，这等价于存在满足源条件（参见 Engl 等（1996）；Benning 和 Burger（2018））的源条件元素 $v^\dagger$

$$W^*v^\dagger \in \partial R(x^\dagger)\,, \tag{SC}$$

下面，我们验证：如果数据误差趋于零，则正则化算子的解与逆问题的解之间关于 R 的对称 Bregman 距离也趋于零。两个向量 x和 $\tilde{x}$ 之间的对称 Bregman 距离或 Jeffreys 距离，就是交换参数后两个 Bregman 距离之和，即

$$D_R^{\text{symm}}(x,\tilde{x}) := D_R(x,\tilde{x}) + D_R(\tilde{x},x) = \langle x - \tilde{x}, q - \tilde{q}\rangle\,,$$

对于 $q \in \partial R(x)$ 和 $\tilde{q} \in \partial R(\tilde{x})$；因此，对对称 Bregman 距离的误差估计也意味着对经典 Bregman 距离的误差估计。

在开始分析之前，我们回顾 Jensen-Shannon 散度（Lin，1991）的概念。对于一般的正常、凸且下半连续函数 $F\colon \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$，它推广为所谓的 ***Burbea-Rao*** 散度（Burbea 和 Rao，1982a,b；Nielsen 和 Boltz，2011），定义如下。

定义 1（Burbea-Rao 散度）。假设 $F\colon \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ 是一个正常、凸且下半连续函数。相应的 ***Burbea-Rao*** 散度定义为

$$J_F(x,\tilde{x}) := \frac{1}{2}\left(F(x) + F(\tilde{x}) - 2F\left(\frac{x+\tilde{x}}{2}\right)\right)\,, \tag{7}$$

对于所有 $x, \tilde{x} \in$ 定义域(F)。

我们还需要借助另一个重要概念来建立误差估计，即 Fenchel 共轭（参见 Beck（2017））。

DEFINITION 2（Fenchel 共轭）。函数 $F^*\colon\ \mathbb{R}^n \to \mathbb{R}\cup\{-\infty,\infty\}$ $F\colon \mathbb{R}^n \to \mathbb{R}\cup\{-\infty,+\infty\}$ 的 ***Fenchel***（或凸）共轭定义为

$$F^*(w)\colon\ =\ \sup_{x\in\mathbb{R}^n} \langle x, w\rangle - F(x)\,.$$

我们尤其关注的是函数 $B_\Psi(y,\ z)$ 关于第二个自变量的共轭，下面的引理对其进行了刻画。

LEMMA 1. $F_y(z)$ 关于第二个参数 z 的 *Fenchel* 共轭为 $= B_\Psi(y, z)$

$$F_y^*(w) = \left(\frac{1}{2}\|\cdot\|^2 + \Psi\right)(y+w) - \left(\frac{1}{2}\|\cdot\|^2 + \Psi\right)(y)\,.$$

证明。根据 Fenchel 共轭的定义，我们观察到

$$\begin{aligned}
F_y^*(w) &= \sup_{z\in\mathbb{R}^m} \langle z, w\rangle - F_y(z) \\
&= \sup_{z\in\mathbb{R}^m} \langle z, w\rangle - \left(\frac{1}{2}\|\cdot\|^2 + \Psi\right)(y) - \left(\frac{1}{2}\|\cdot\|^2 + \Psi\right)^*(z) + \langle y, z\rangle \\
&= -\left(\frac{1}{2}\|\cdot\|^2 + \Psi\right)(y) + \sup_{z\in\mathbb{R}^m} \langle z, w+y\rangle - \left(\frac{1}{2}\|\cdot\|^2 + \Psi\right)^*(z) \\
&= -\left(\frac{1}{2}\|\cdot\|^2 + \Psi\right)(y) + \left(\frac{1}{2}\|\cdot\|^2 + \Psi\right)(w+y)\,,
\end{aligned}$$

由此完成证明。

定义了 Burbea-Rao 散度，并求得固定 y 时， $B_\Psi(y, z)$ 关于第二个参数 z 的 Fenchel 共轭后，我们现在可以给出并验证受（Benning 和 Burger，2011）启发的主要结果。

定理 2. 假设 R 和 Ψ 满足第 *3.1* 节所述的假设。那么，对于满足 $B_\Psi(y^\delta, Wx^\dagger + b) \le \delta^2$ 且 $\delta \ge 0$的数据 y^δ 和 $x^\dagger$，若 $x^\alpha \in \mathcal{R}_\alpha(y^\delta)$ 是变分正则化问题(2)的一个解， $x^\dagger$ 是满足 $x^\dagger \in \mathcal{S}(y)$ 和 (SC)的感知机问题 $y = \sigma(Wx^\dagger + b)$ 的一个解，则可得误差估计

$$\begin{aligned}
(1-c)B_\Psi(y^\delta, Wx_\alpha + b) + \alpha D_R^{symm}(x_\alpha, x^\dagger) \le (1+c)\delta^2 + \frac{\alpha^2}{c}\|v^\dagger\|^2 \\
+ 2cJ_\Psi\left(y^\delta + \frac{\alpha}{c}v^\dagger, y^\delta - \frac{\alpha}{c}v^\dagger\right),
\end{aligned} \tag{8}$$

其中常数为 $c \in (0, 1]$。

证明。每个满足 $x_\alpha \in \mathcal{R}_\alpha(y^\delta)$ 的解 x_α 都可等价地由最优性条件刻画：

$$W^*\left(\sigma(Wx_\alpha + b) - y^\delta\right) + \alpha p_\alpha = 0\,,$$

对于任意次梯度 $p_\alpha \in \partial R(x_\alpha)$。从等式两边减去 $p^\dagger \in \partial R(x^\dagger)$，并与 $x_\alpha - x^\dagger$ 取对偶积，可得

$$\langle \sigma(Wx_\alpha + b) - y^\delta, Wx_\alpha - Wx^\dagger\rangle + \alpha D_R^{\text{symm}}(x_\alpha, x^\dagger) = -\alpha\langle p^\dagger, x_\alpha - x^\dagger\rangle\,. \tag{9}$$

我们很容易验证

$$D_{B_\Psi(y^\delta,W\cdot+b)}(x^\dagger,x_\alpha)=B_\Psi(y^\delta,Wx^\dagger+b)-B_\Psi(y^\delta,Wx_\alpha+b)-\langle\sigma(Wx_\alpha+b)-y^\delta,Wx^\dagger-Wx_\alpha\rangle\,;$$

因此，我们可以在（9）中用 $D_{B_\Psi(y^\delta,W\cdot+b)}(x^\dagger,x_\alpha)+B_\Psi(y^\delta,Wx_\alpha+b)-B_\Psi(y^\delta,Wx^\dagger+b)$ 替换 $\langle\sigma(Wx_\alpha+b)-y^\delta,Wx_\alpha-Wx^\dagger\rangle$，从而得到

$$D_{B_\Psi(y^\delta,W\cdot+b)}(x^\dagger,x_\alpha)+B_\Psi(y^\delta,Wx_\alpha+b)+\alpha D_R^{\text{symm}}(x_\alpha,x^\dagger)=B_\Psi(y^\delta,Wx^\dagger+b)-\alpha\langle p^\dagger,x_\alpha-x^\dagger\rangle$$

由于 $B_\Psi(y^\delta,W\cdot+b)$ 的凸性，我们知道 $0\le D_{B_\Psi(y^\delta,W\cdot+b)}(x^\dagger,x_\alpha)$，并且我们还知道 (SC) 使我们能够选择 $p^\dagger=W^*v^\dagger$。因此，我们可以估计

$$B_\Psi(y^\delta,Wx_\alpha+b)+\alpha D_R^{\text{symm}}(x_\alpha,x^\dagger)\le B_\Psi(y^\delta,Wx^\dagger+b)-\alpha\langle v^\dagger,Wx_\alpha-Wx^\dagger\rangle\,.$$

接下来，我们引入常数 $c\in(0,1]$，将损失函数 $B_\Psi(y^\delta,Wx_\alpha+b)$ 和 $B_\Psi(y^\delta,Wx^\dagger+b)$ 分别拆分为 $(1-c)B_\Psi(y^\delta,Wx_\alpha+b)+cB_\Psi(y^\delta,Wx_\alpha+b)$ 和 $(1+c)B_\Psi(y^\delta,Wx^\dagger+b)-cB_\Psi(y^\delta,Wx^\dagger+b)$。这意味着我们估计

$$\begin{aligned}(1-c)B_\Psi(y^\delta,Wx_\alpha+b)+\alpha D_R^{\text{symm}}(x_\alpha,x^\dagger)\le{}&(1+c)B_\Psi(y^\delta,Wx^\dagger+b)\\&+\langle\alpha v^\dagger,Wx^\dagger+b\rangle-cB_\Psi(y^\delta,Wx^\dagger+b)\\&-\langle\alpha v^\dagger,Wx_\alpha+b\rangle-cB_\Psi(y^\delta,Wx_\alpha+b)\,.\end{aligned}$$

接下来，我们利用引理 1 进行估计

$$\langle\alpha v^\dagger,Wx^\dagger+b\rangle-cB_\Psi(y^\delta,Wx^\dagger+b)\le c\left(\left(\frac12\|\cdot\|^2+\Psi\right)\left(y^\delta+\frac{\alpha}{c}v^\dagger\right)-\left(\frac12\|\cdot\|^2+\Psi\right)\left(y^\delta\right)\right),$$

and

$$-\langle\alpha v^\dagger,Wx_\alpha+b\rangle-cB_\Psi(y^\delta,Wx_\alpha+b)\le c\left(\left(\frac12\|\cdot\|^2+\Psi\right)\left(y^\delta-\frac{\alpha}{c}v^\dagger\right)-\left(\frac12\|\cdot\|^2+\Psi\right)\left(y^\delta\right)\right).$$

将两个估计相加可得

$$\begin{aligned}&\langle\alpha v^\dagger,Wx^\dagger+b\rangle-cB_\Psi(y^\delta,Wx^\dagger+b)-\langle\alpha v^\dagger,Wx_\alpha+b\rangle-cB_\Psi(y^\delta,Wx_\alpha+b)\,,\\&\le\frac{\alpha^2}{c}\|v^\dagger\|^2+c\left(\Psi\left(y^\delta+\frac{\alpha}{c}v^\dagger\right)+\Psi\left(y^\delta-\frac{\alpha}{c}v^\dagger\right)-2\Psi(y^\delta)\right),\\&=\frac{\alpha^2}{c}\|v^\dagger\|^2+2cJ_\Psi\left(y^\delta+\frac{\alpha}{c}v^\dagger,y^\delta-\frac{\alpha}{c}v^\dagger\right),\end{aligned}$$

结合误差界 $B_\Psi(y^\delta,Wx^\dagger+b)\le\delta^2$ 即可完成证明。

备注 1. 我们想强调，对于连续的 Ψ 和 $c>0$，我们会自动观察到 e

$$\lim_{\alpha\to0}J_\Psi\left(y^\delta+\frac{\alpha}{c}v^\dagger,y^\delta-\frac{\alpha}{c}v^\dagger\right)=0\,,$$

在这种情况下，从误差估计的角度来看，重要的问题是该项是否比 α 更快地收敛到零，因为我们需要确保 $\lim_{\alpha\to 0} J_\Psi\left(y^\delta + \frac{\alpha}{c}v^\dagger, y^\delta - \frac{\alpha}{c}v^\dagger\right)/\alpha = 0$，才能保证(8) 中的对称 *Bregman* 距离对于 $\alpha \to 0$ 收敛到零。

例1 （ReLU 感知器）。让我们考虑一个具体例子，以说明相对于 R 的对称 *Bregman* 距离，(5) 是一种收敛正则化。我们知道，要使 $\sigma(z) = \mathit{prox}_\Psi(z) = \max(0, z)$ 成立，必须选择 $\Psi(z) = \begin{cases} 0 & z \in [0,\infty)^m \\ \infty & \mathit{else} \end{cases}$。这意味着，为了使 $B_\Psi(y^\delta, z)$ 对任意 z 都有良好定义，我们要求对所有 $i \in \{1,\dots,m\}$ 均有 $y_i^\delta \geq 0$。为了使 *Burbea-Rao* 散度有良好定义，我们还要求

$$-\frac{c}{\alpha}y_i^\delta \leq v_i^\dagger \leq \frac{c}{\alpha}y_i^\delta\,,$$

对所有 $i \in \{1,\dots,m\}$ 均成立，或者用更紧凑的记法表示为 $\|v^\dagger\|_\infty \leq (c\|y^\delta\|_\infty/\alpha)$。如果能够保证 $\|v^\dagger\|_\infty \leq (c\|y^\delta\|_\infty/\alpha)$，我们观察到 $J_\Psi\left(y^\delta + \frac{\alpha}{c}v^\dagger, y^\delta - \frac{\alpha}{c}v^\dagger\right) = 0$。因此，我们可以将估计式 (8) 简化为

$$\frac{1-c}{\alpha}B_\Psi(y^\delta, Wx_\alpha + b) + D_R^{\mathit{symm}}(x_\alpha, x^\dagger) \leq \frac{1+c}{\alpha}\delta^2 + \frac{\alpha}{c}\|v^\dagger\|^2\,,$$

其中，我们还将不等式两边都除以 α。如果选择 $\alpha(\delta) = \sqrt{c(1+c)}\delta/\|v^\dagger\|$，则可得到估计式

$$\frac{(1-c)\|v^\dagger\|}{\delta\sqrt{c(1+c)}}B_\Psi(y^\delta, Wx_{\alpha(\delta)} + b) + D_R^{\mathit{symm}}(x_{\alpha(\delta)}, x^\dagger) \leq 2\sqrt{\frac{1+c}{c}}\|v^\dagger\|\delta\,,$$

只要我们能够确保

$$\left\|\frac{v^\dagger}{\|v^\dagger\|}\right\|_\infty \leq \sqrt{\frac{c}{1+c}}\left\|\frac{y^\delta}{\delta}\right\|_\infty.$$

结合 $D_R^{\mathit{symm}}(x_{\alpha(\delta)}, x^\dagger) \geq D_R(x^\dagger, x_{\alpha(\delta)})$，我们已经建立了具有常数 $C = 2\sqrt{\frac{1+c}{c}}\|v^\dagger\|$ 的形式为 (6) 的估计。因此，我们已经验证，在这个具体例子中，变分正则化方法 (2) 不仅是一种正则化方法，甚至还是一种收敛正则化方法。

我们想简要评论一下如何将收敛分析推广到一般情形 $L > 1$，具体见以下备注。

R备注 2. 所提出的收敛分析很容易推广到逐层顺序反演方法。假设我们有 L 层，并从最后一层开始，那么可以构造变分问题

$$x_{L-1}^\alpha \in \operatorname*{arg\,min}_{x_{L-1}} \left\{ B_{\Psi_L}(y^\delta, W_L x_{L-1} + b_L) + \alpha_{L-1}\Psi_{L-1}(x_{L-1}) \right\},$$

这同样具有(5)的形式，但其中 R 已被 Ψ_{L-1}替换。或者，如果我们对辅助变量 x_{L-1} 有良好的先验知识，也可以用另一个函数 R_{L-1} 替换 Ψ_{L-1}。

一旦我们估计出 x_{L-1}^{α}，便可以递归地估计

$$x_{l-1}^{\alpha} \in \arg\min_{x_{l-1}} \left\{ B_{\Psi_l}(x_l^{\alpha}, W_l x_{l-1} + b_l) + \alpha_{l-1} \Psi_{l-1}(x_{l-1}) \right\} ,$$

对于 $l = L-1$、.....、2，随后将 x_{α} 作为 (2) 的解求出，但使用数据 x_1^{α} 而非 y^{δ}。

这种顺序方法的优点是，每个单独的正则化问题都是凸的，并且此前提出的定理和保证仍然适用。其缺点是，要使这种方法在理论上成立，我们需要对每个辅助变量都存在如下形式的界 $B_{\Psi_l}(x_l^{\alpha}, W_l x_{l-1}^{\alpha} + b_l) \leq \delta_l^2$，而这是一项相当不切实际的假设。此外，假设存在关于辅助变量的良好先验知识也不现实。

请注意，证明同时方法 (2) 是一种（收敛的）变分正则化，超出了本文的研究范围，因为这更为困难，而且可能需要额外的假设，原因如下。(2) 中的整体目标函数不再保证关于所有变量同时为凸，这意味着我们不能将单层情形的分析直接套用于多层感知机情形。

至此，感知机反演模型的理论分析告一段落。下一节将重点介绍如何实现（5）及其更一般的对应形式（2）。

4 实现

本节将介绍如何在单层和多层感知机的情形下，以计算方式实现所提出的变分正则化。具体而言，我们将说明，所提出的变分正则化可以通过广义原始-对偶混合梯度法和坐标下降方法高效求解。

4.1 感知机反演

首先，我们考虑反演一个（单层）感知机的例子。对于 $L = 1$，问题（2）可化简为（5）；对于复合正则化函数 $R \circ K$，其形式为

$$x^{\alpha} \in \arg\min_{x} \left\{ B_{\Psi} \left(y^{\delta}, f(x, \Theta) \right) + \alpha R(Kx) \right\} . \tag{10}$$

这里，K 是一个矩阵，而 $\alpha R(Kx)$表示作用于参数 x上的正则化函数。上述问题（10）可改写为以下鞍点问题

$$\min_{x} \max_{z} B_{\Psi} \left(y^{\delta}, f(x, \Theta) \right) + \langle z, Kx \rangle - \alpha R^*(z) , \tag{11}$$

其中，R^* 表示 R的凸共轭。在计算上，我们可以使用广义原始-对偶混合梯度（PDHG）方法求解该鞍点问题（Zhu 和 Chan，2008年；Pock 等人，2009年；Esser 等人，2010年；Chambolle 和 Pock，2011年、2016年；Benning 和 Riis，2021年）：

$$x^{k+1} = x^k - \tau_x \left(\left(\text{prox}_{\Psi} \left(f(x^k, \Theta) \right) - y^{\delta} \right) \mathcal{J}_f^x(x^k, \Theta) + \alpha K^{\top} z^k \right) \tag{12a}$$

$$z^{k+1} = \text{prox}_{\tau_z R^*} \left(z^k + \tau_z \alpha K \left(2x^{k+1} - x^k \right) \right) . \tag{12b}$$

其中，我们交替对 x 变量执行下降步，并对对偶变量 z执行上升步。由于（10）是一个凸最小化问题，对于任意初始点，只要选择 τ_x 和 τ_z 使得 $\tau_x\tau_z < 1/\|K\|$成立，（12）就能保证全局收敛。

本文将采用离散全变差 $\|\nabla x\|_{p,\,1}$（Rudin 等人，1992；Chambolle 和 Lions，1997）作为正则化函数 $R(Kx)$，当然也可以选择其他函数。若考虑二维标量图像 $x \in \mathbb{R}^{H\times W}$，则可以将梯度算子 ∇ 的有限前向差分离散化定义为：$\mathbb{R}^{H\times W} \to \mathbb{R}^{H\times W\times 2}$，即

$$(\nabla x)_{i,j,1} = \begin{cases} x_{i+1,j} - x_{i,j} & \text{if } 1 \le i < H, \\ 0 & \text{else,} \end{cases}, \qquad (\nabla x)_{i,j,2} = \begin{cases} x_{i,j+1} - x_{i,j} & \text{if } 1 \le j < W, \\ 0 & \text{else.} \end{cases}.$$

离散全变差定义为像素级图像梯度的 p-范数的 ℓ_1 范数，即

$$\|\nabla x\|_{p,1} = \sum_{i=1}^{H}\sum_{j=1}^{W} |\,(\nabla x)_{i,j}\,|_p = \sum_{i=1}^{H}\sum_{j=1}^{W} \left((\nabla x)_{i,j,1}^p + (\nabla x)_{i,j,2}^p\right)^{1/p}.$$

在数值结果中，我们采用各向同性全变差，因此选择 $p = 2$。因此，对于具有仿射线性变换 $f(x, \Theta) = Wx + b$的感知机，并以 $\sigma = \text{prox}_\Psi$表示激活函数，求解感知机逆问题（5）的原始-对偶混合梯度法（12）可概括如下：

$$x^{k+1} = x^k - \tau_x \left(W^\top \left(\sigma(Wx^k + b) - y\right) - \alpha \text{div} z^k\right), \tag{13a}$$

$$z^{k+1} = \text{prox}_{\tau_z \|\cdot\|_{2,1}^*}\left(z^k + \tau_z \left(2\alpha\nabla x^{k+1} - \alpha\nabla x^k\right)\right). \tag{13b}$$

请注意，我们定义了散度 div 的离散近似，使其满足 $\text{div} = -\nabla^\top$，从而与连续情形类似，成为梯度离散有限差分近似的负转置；这就是 (13a) 中的符号与 (12a) 相比发生反转的原因。关于 $\|\cdot\|_{2,\,1}^*$ 的凸共轭的近端映射很简单：如果每个像素的欧几里得向量范数最大值不超过一，则映射结果就是参数本身；否则，结果就是投影到这个单位球上。

4.2 多层感知机的反演

现在我们讨论如何实现对 (2) 中所述、具有 L 层的多层感知机进行反演。请注意，在这种情况下，为了针对 x 进行最小化，我们还需要针对辅助变量 $x_1, \ldots, x_{L-1}$ 进行优化。

对于式 (2) 的最小化，我们考虑一种交替最小化方法，也称为坐标下降 （Beck 和 Tetruashvili, 2013年；Wright，2015年；Wright 和 Recht，2022年）。在这种方法中，我们每次针对一个变量最小化目标函数。具体而言，我们关注一种半显式坐标下降算法，该算法针对整体目标函数中的光滑函数进行线性化。这将整体最小化问题分解为 L 个子问题，其中对于 x_0 和每个 x_l 变量

对于 $l \in \{1, \dots, L-1\}$，我们有以下形式的独立最小化问题：

$$x_0^{k+1} = \arg\min_{x_0} \left\{ \left(\frac{1}{2}\|\cdot\|^2 + \Psi_1 \right)^* (f(x_0, \Theta_1)) - \left\langle x_1^k, f(x_0, \Theta_1) \right\rangle + \alpha R(Kx_0) \right\}, \tag{14a}$$

$$\begin{aligned} x_l^{k+1} = \arg\min_{x_l} \Bigg\{ & \left(\frac{1}{2}\|\cdot\|^2 + \Psi_l \right)(x_l) - \left\langle x_l, f(x_{l-1}^{k+1}, \Theta_l) \right\rangle + \frac{1}{2\tau_{x_l}}\|x_l - x_l^k\|^2 \\ & + \left\langle x_l, \left(\text{prox}_{\Psi_{l+1}} \left(f(x_l^k, \Theta_{1+1}) \right) - x_{l+1}^k \right) \mathcal{J}_f^x(x_l^k, \Theta_{1+1}) \right\rangle \Bigg\}. \end{aligned} \tag{14b}$$

请注意，采用这种方法的一个优点是：当其他所有变量保持固定时，我们利用了整体目标函数关于每个单独变量都是凸的这一性质。接下来，我们将讨论以计算方式求解各个子问题的不同策略。

对输入变量 x_0进行优化时，子问题 (14a) 的结构与我们在第4.1节讨论过的感知机逆问题相同。因此，我们可以用式 (11) 近似 x_0^{k+1}，但现在是相对于x_1^k 而非 y^δ，从而得到如下迭代：

$$x_0^{k+1} = x_0^k - \tau_{x_0} \left(\left(\text{prox}_{\Psi} \left(f(x_0^k, \Theta_1) \right) - x_1^k \right) \mathcal{J}_f^x(x_0^k, \Theta_1) + \alpha K^\top z^k \right), \tag{15a}$$

$$z^{k+1} = \text{prox}_{\tau_z R^*} \left(z^k + \tau_z \alpha K \left(2x^{k+1} - x^k \right) \right). \tag{15b}$$

对于每个满足 $l \in \{1, \dots, L-1\}$的辅助变量 x_l，与 (14b) 相关的子问题相当于执行一个步长τ_{x_l}适当的近端梯度步，我们可以将其改写为

$$\begin{aligned} x_l^{k+1} = \text{prox}_{\frac{\tau_{x_l}}{1+\tau_{x_l}}\Psi_l} \Bigg(\frac{1}{1+\tau_{x_l}} \Big(x_l^k - \tau_{x_l} \Big(\Big(\text{prox}_{\Psi_l} \left(f(x_l^k, \Theta_{l+1}) \right) - x_{l+1}^k \Big) \mathcal{J}_f^x(x_l^k, \Theta_{l+1}) \\ - f(x_{l-1}^{k+1}, \Theta_l) \Big) \Big) \Bigg) \end{aligned}. \tag{16}$$

至此，我们结束了对正则化单层和多层感知机反演实现的讨论。下一节将展示一些数值结果，以实证说明所提出方法的有效性。

5 数值结果

本节将展示采用式（13）所述的原始-对偶混合梯度算法实现的感知机逆问题，以及采用式（15）和式（16）所述的坐标下降方法实现的多层感知机反演问题的数值结果。所有结果均使用 PyTorch 3.7 在 Intel Xeon CPU E5-2630 v4 上计算得出。

5.1 感知机

我们展示了两个实验的结果：第一个实验是根据感知机的含噪输出，对圆形图像进行感知器逆演，并比较 Landweber 正则化与基于全变差的变分正则化 (5)。第二个实验中，我们对 MNIST 数据集中的样本进行感知器逆演，并将其与线性和非线性解码器的性能进行比较。

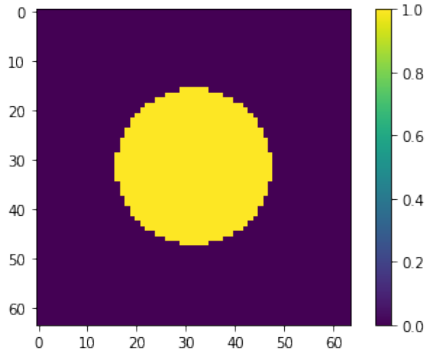


图 1. 圆的真实图像 $x^\dagger$。

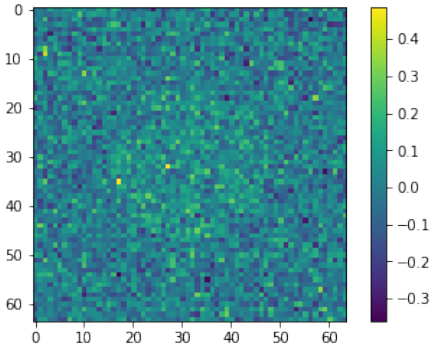
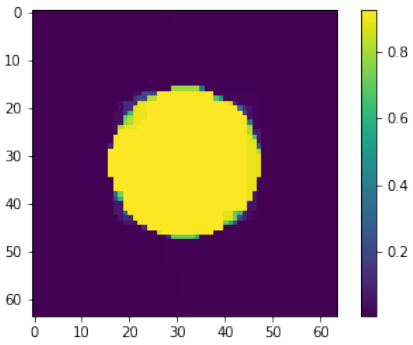


图 2.反演图像 Landweber 正则化。



via **图 3.** 通过全变差正则化得到的反演图像。

圆 我们先从一个简单示例开始：根据 ReLU 感知机的含噪测量恢复圆形图像。为准备实验，我们生成一幅圆形图像 $x^\dagger \in \mathbb{R}^{64\times64}$，如图 1 所示。我们使用随机权重和偏置构建一个具有 ReLU 激活函数的感知机，其中 $W \in \mathbb{R}^{512\times4096}$，$b \in \mathbb{R}^{512\times1}$。这些权重作用于 x 的列向量表示，其中$x \in \mathbb{R}^{4096\times1}$。通过模型的前向运算生成无噪数据，即 $y = \sigma(Wx^\dagger + b)$。我们通过添加均值为 0 、标准差为 0。005 的高斯噪声，生成含噪数据 y^δ。注意，我们将 y^δ 的所有负值截断为零，以确保 $y^\delta \in \text{dom}(\Psi)$。

解决这一适定性不良的感知机逆问题的首次尝试，是采用 Landweber 正则化（Landweber，1951）。在图 2 中，我们看到结合提前停止并依据莫罗佐夫偏差原理 Morozov（2012）；Engl 等（1996）得到的 Landweber 正则化重建图像。尽管 Landweber 正则化重建与数据相符，直到偏差值$\|\sigma(Wx^K + b) - y^\delta\|$，但恢复出的图像并不像图像 $x^\dagger$。我们将在下文简要讨论这种视觉效果较差的反演的原因。相比之下，在图 3 中，我们看到遵循（13）的全变差正则化方法所得到的正则化反演。该重建的正则化参数选择为 $\alpha = 1.5 \times 10^{-2}$。$x_0$ 和 z 均使用零向量进行初始化。步长参数选择为 $\tau_x = 1.99/\|W\|_2^2$ 和 $\tau_z = 1/(8\alpha)$，参见（Chambolle，2004）。当 x_0 和 z 的范数变化小于阈值 10^{-5}，或达到我们设定为 10000 的最大迭代次数时，我们停止迭代。如图 3 所示，全变差正则化方法能够找到一个在（视觉上）更有意义的解。

为了解释为什么在这个特定例子中，兰德韦伯迭代的表现不如全变差正则化，我们比较了每个解与真实图像 $x^\dagger$ 的 ℓ_2 范数。图 2 中兰德韦伯解的 ℓ_2 范数为 6.58，而图 3 中经全变差正则化的解和真实图像 $x^\dagger$ 的对应数值分别为 25.69 和 28.07。这并不令人意外，因为众所周知，当噪声水平趋于零时，兰德韦伯迭代会收敛到一个具有最小欧几里得范数的解。另一方面，比较各个解的全变差半范数时，真实图像的数值为 128.0，而图 2 中的兰德韦伯解和图 3 中经全变差正则化的解的数值分别为 707.02 和 114.93。这表明，对于反演卡通风格图像（例如 $x^\dagger$）而言，全变差半范数是更合适的正则化函数。

MNIST 在这个第二个示例中，我们在 MNIST 数据集上进行感知器逆演（LeCun 等人，1998年）。具体而言，我们考虑以下实验设置。我们首先训练一个自编码器 $\mathcal{A}(x) = \mathcal{D}(\mathcal{E}(x, \Theta_\mathcal{E}) \Theta_\mathcal{D})$，其中 $\mathcal{D}(\cdot, \Theta_\mathcal{D})$ 和 $\mathcal{E}(\cdot, \Theta_\mathcal{E})$ 分别表示解码器和编码器，其参数为

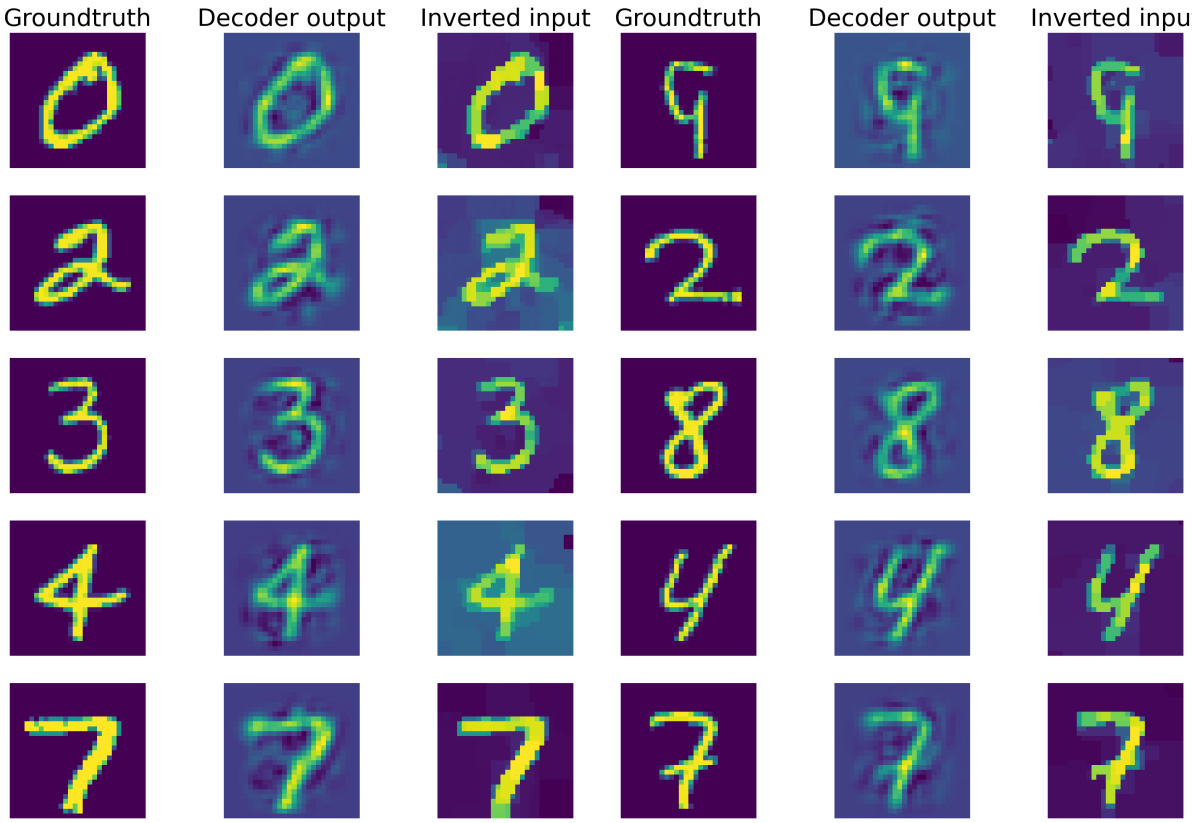


图 4. MNIST 训练数据集中的真实输入图像，以及**图 5.** MNIST 验证数据集中的真实输入图像，以及相应的自编码器输出图像和相应的自编码器输出图像，以及感知机的反演输入图像和感知机的反演输入图像

参数 $\Theta_{\mathcal{D}}$和 $\Theta_{\mathcal{E}}$。我们对自编码器 $\mathcal{A}$进行预训练，计算编码 $\mathcal{E}(x, \Theta_{\mathcal{E}})$ 并将其赋给无噪声数据变量 y，然后根据扰动编码 y^{δ}求解输入 x 对应的逆问题

$$\mathcal{E}(x, \Theta_{\mathcal{E}}) = y^{\delta} \ .$$

更准确地说，我们首先使用普通随机梯度法（SGM），通过最小化 MNIST 训练数据集上的均方误差（MSE），训练一个两层全连接自编码器 $y = W_2(\sigma(W_1 x + b_1)) + b_2$。我们将编码维度设为 100，并使用 ReLU 作为激活函数。因此，$\Theta_{\mathcal{E}} = (W_1, \ b_1)$，其中 $W_1 \in \mathbb{R}^{784\times 100}$ 和 $b_1 \in \mathbb{R}^{100\times 1}$。

所有 MNIST 图像都经过居中处理，作为预处理的一种方式。在算法上，我们遵循 (13) 来计算求解 (10)。步长参数选为 $\tau_x = 1.99/\|W_1\|_2^2$ 和 $\tau_z = 1/(8\alpha)$。我们将正则化参数 α 选在值域 $[10^{-4}, 10^{-2}]$ 内，并对训练集中的所有样本图像设为 5×10^{-3}，对验证集中的所有样本图像设为 $\alpha = 5 \times 10^{-2}$。这些选择在反演图像的视觉质量方面效果良好。

在图 4 和图 5 中，我们分别展示了训练集和验证集中五张样本图像的可视化结果。作为比较，我们也对解码器输出进行了可视化。可以看到，使用包含相同压缩信息的编码时，反演图像呈现出的

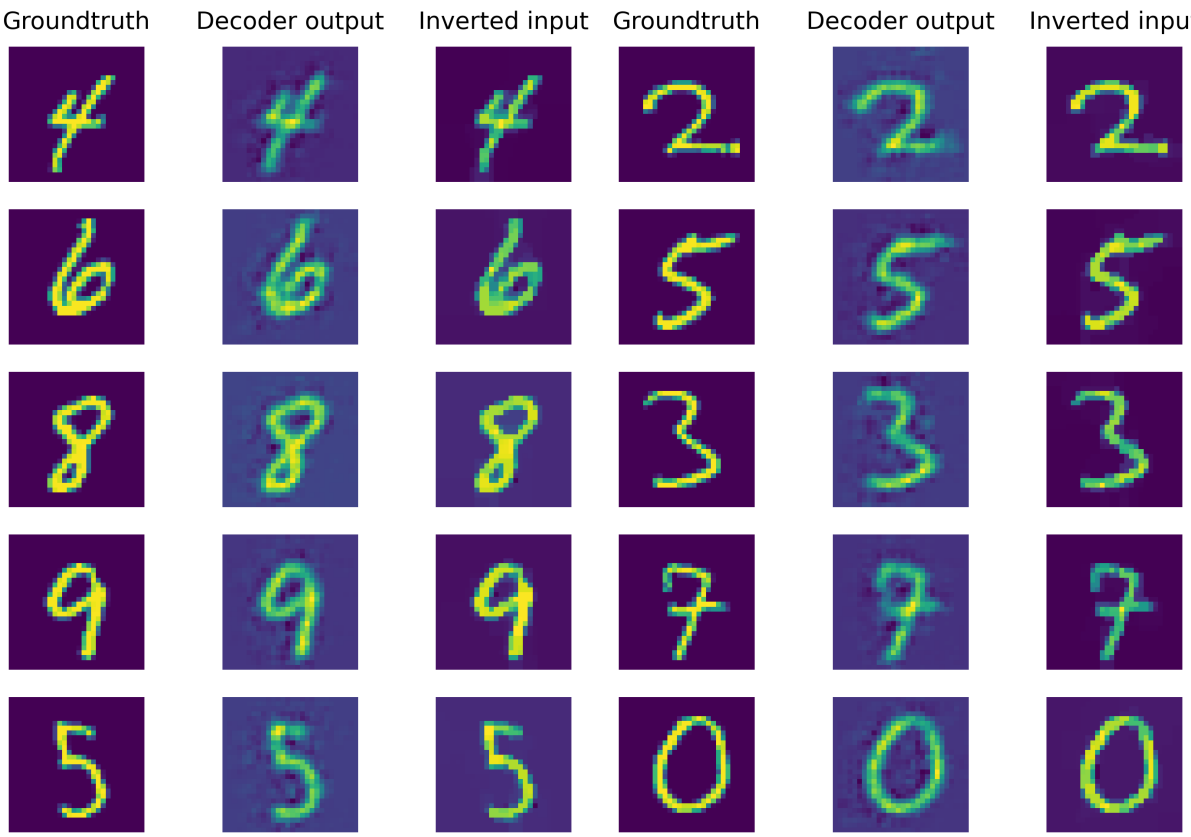


图 6。 卷积神经网络的反演输入图像**图 7。** 来自 MNIST 训练数据集的卷积神经网络反演输入图像，以及训练良好的自编码器输出图像和真实输入图像；来自 MNIST 验证数据集的卷积神经网络反演输入图像，以及训练良好的自编码器输出图像和真实输入图像

边缘更加清晰，视觉质量也优于解码输出。对此并不意外，因为我们比较的是非线性正则化反演方法与线性解码器。

5.2 多层感知机

本节展示多层感知机反演的数值结果。具体而言，我们考虑包含卷积层的前馈神经网络（卷积神经网络），其中网络架构使用二维卷积运算来表示仿射线性函数 $f(x,\ \Theta)$ 中的线性运算。与第 5.1 节所述的实验设计类似，我们考虑一个多层神经网络反演问题，即从噪声扰动编码 y^{δ} 推断输入图像 x 。

更具体地说，我们首先采用随机梯度法在 MNIST 训练数据集上训练六层卷积自编码器，以最小化均方误差。编码器 $\mathcal{E}(x,\ \Theta_{\mathcal{E}})$由两个卷积层组成，两层均采用步幅为 2 的 4×4 卷积，且每层之后都应用 ReLU 激活函数。随着图像空间维度减半，我们将特征通道数从 8 加倍至 16。我们使用一个全连接层，通过权重 $W_3 \in \mathbb{R}^{300\times 784}$ 和偏置 $b_3 \in \mathbb{R}^{300\times 1}$ 生成编码。解码器网络首先通过权重$W_4 \in \mathbb{R}^{784\times 300}$ 和偏置 $b_4 \in \mathbb{R}^{784\times 1}$的仿射线性变换扩展编码。随后使用两个转置卷积层，其卷积核大小为 4×4，并且每层之后都应用 ReLU 激活函数。随着空间维度每次加倍，特征通道数则减半。

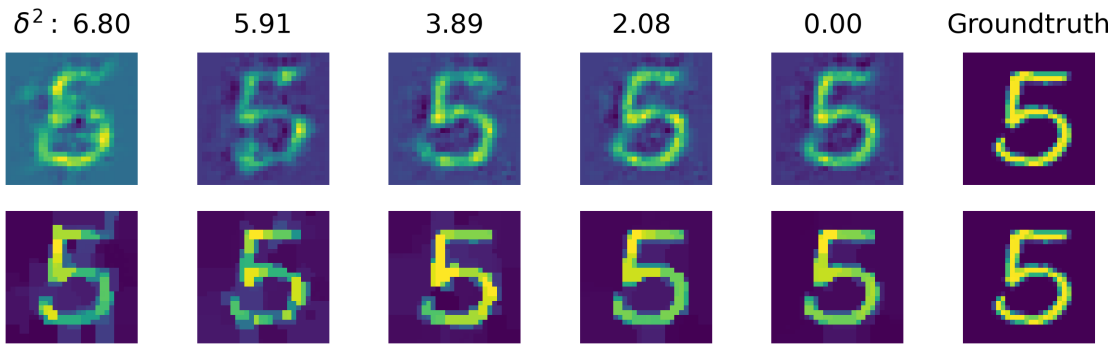


图8。 可视化比较不同噪声水平下的反演图像与解码图像。**上：** 由训练好的卷积自编码器生成的解码器输出图像。**下：** 通过带全变差正则化的卷积神经网络反演得到的输入图像。

按照第4.2节所述的实现细节，我们迭代计算更新步骤 (15) 和 (16)。对于PDHG方法，我们选择步长参数为 $\tau_x = 1.99/\|W_1\|_2^2$ 和 $\tau_z = 1/(8\alpha)$。初始值 x_0 和 z 均为零。更新步骤在达到 1500 次的最大迭代次数后停止；或者，当 x_0 和 z 的改进量的范数小于 10^{-5} 时停止。对于坐标下降算法，每一层的步长参数均设为 $\tau_{x_l} = 1.99/\|W_{l+1}\|_2^2$。

在图 6 和图 7 中，我们分别展示了训练数据集和验证数据集中的反演图像、解码器输出图像以及真实图像。对于每张图像，为了获得最佳的视觉反演效果，α 在范围 $[10^{-4}, 10^{-2}]$ 内选取，并针对训练样本图像和验证样本图像均设为 9×10^{-3}。

在图8中，我们进一步比较全变差正则化和解码器对不同数据噪声水平的响应。噪声数据通过向每张图像的编码添加高斯噪声生成。我们从均值为零、标准差为 0.33 的高斯噪声开始，逐渐降低噪声水平，这相当于将 δ^2 从 6.80 逐步降至 0.00。

请注意，对于每个噪声水平，正则化因子 α 都是在范围 $[10^{-4}, 10^{-2}]$ 内手动选取的，以获得最佳峰值信噪比。我们可以看到，对于标准差为 0.33 的噪声水平（此时 δ^2 为 6.80），解码器只能生成模糊失真的输出，而反演图像则更清晰地显示了数字的结构。当我们将噪声水平降低至 0.00 时，反演图像变得更加清晰利落，而解码图像的轮廓仍不够清晰。

图9绘制了随着噪声水平降低，解码图像和反演图像的峰值信噪比变化。我们想强调，按照式（8）中的误差估计界所建议的做法，计算并比较 $D_R^{\text{symm}}(x_{\alpha(\delta)},\ x^\dagger)$ 会更加严谨，但从经验上看，峰值信噪比也支持正则化收敛这一观点。

6 结论与展望

我们提出了一种新颖的变分正则化框架，该框架基于提升的布雷格曼公式，用于稳定地反演前馈神经网络（也称为多层感知机）。我们证明，在逆问题解属于正则化算子的值域这一较弱假设下，所提出的框架对于单层感知器情形是一种收敛正则化。我们推导出了一般误差估计，以及在激活函数为 ReLU 激活函数时的特定误差估计。我们还探讨了将该理论推广至

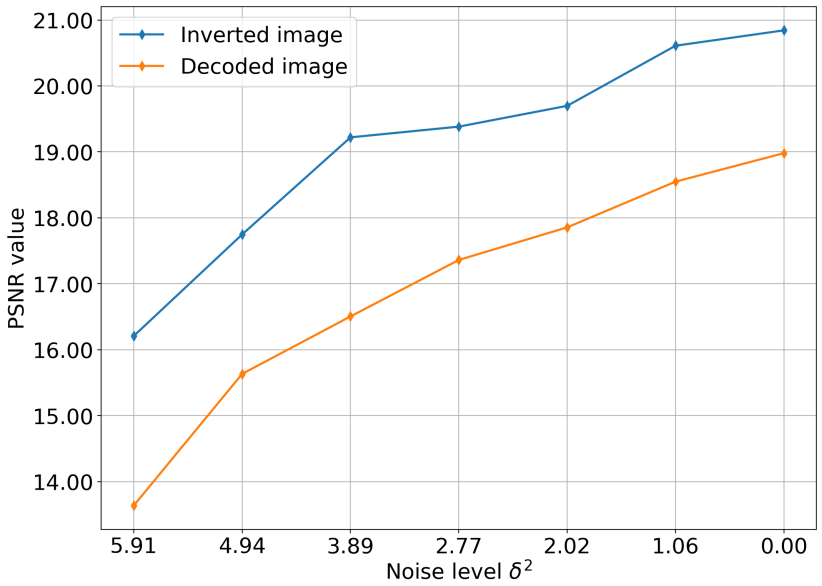


图9. 比较不同噪声水平下基于全变差的重建结果与解码器输出的峰值信噪比。每条曲线展示了随着高斯噪声水平逐渐降低，峰值信噪比的变化，其中 δ^2 的范围为 0.00 至 6.80。

多层感知机情形；尽管需要不切实际的假设，但这一推广可以按顺序进行。我们讨论了求解所提方案的计算实现策略，并给出了数值结果，展示了采用全变差正则化，如何根据单层和多层感知机的输出对圆形图像以及手写数字的分段常数图像进行正则化反演。

尽管本文取得了许多积极成果，所提出的框架仍存在一些局限。目前，该框架仅适用于采用仿射线性变换和近端激活函数的前馈架构。将该框架扩展到其他架构（如残差网络（He 等人，2016）或 U-Net（Ronneberger 等人，2015））并不困难，但要纳入无法表示为凸函数近端映射的非线性操作（如最大池化）则并非易事。不过，在许多情况下存在补救方法，例如在前面的示例中使用平均池化代替最大池化。

一个尚待解决的问题是，如何在不作限制性且不切实际假设的情况下，为多层情形建立收敛理论。一个问题是所提出公式的非凸性。可以通过采用不同的架构来补救，从而得到关于所有辅助变量联合凸的提升的 Bregman 形式。

最后但同样重要的是，人们还希望考虑其他形式的正则化，例如迭代正则化、数据驱动正则化（Kabri 等人，2022），甚至二者的组合（Aspriet 等人，2020）。然而，这类方法的收敛分析目前仍是一个开放问题。

利益冲突声明

作者声明，本研究是在不存在任何可能被视为潜在利益冲突的商业或财务关系的情况下开展的。

作者贡献

XW 编写了程序并完成了所有数值结果，以及第4节和第 5 节。MB 负责撰写引言（第1节）以及理论结果（第3节）。两位作者共同完成了第2节和第6节，贡献相同。

致谢

作者感谢 Cantab Capital 信息数学研究所、剑桥分析中心（CCA）和艾伦·图灵研究所（ATI）提供的支持。

数据可用性声明

本研究的编程编码可在剑桥大学数据存储库中找到，网址为 10.17863/CAM.94404。

参考文献

¨Arridge, S., Maass, P., Oktem, O. 和 Sch¨onlieb, C.-B.（2019）。使用数据驱动模型求解逆问题。*Acta Numerica*28，1–174¨Aspri, A., Banert, S., Oktem, O. 和 Scherzer, O.（2020）。数据驱动的迭代正则化兰德韦伯迭代。*Numerical Functional Analysis andOptimization* 41，1190–1227。doi:10.1080/01630563.2020. 1740734 Beck, A.（2017）。优化中的一阶方法（SIAM）Beck, A. 和 Tetruashvili, L.（2013）。关于块坐标下降类型方法的收敛性。*SIAM* 优化杂志 23，2037–2060 Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D. 和 Jacobsen, J.-H.（2019）。可逆残差网络。载于国际机器学习会议（PMLR），573–582 Behrmann, J., Vicol, P., Wang, K.-C., Grosse, R. 和 Jacobsen, J.-H.（2021）。理解并缓解可逆神经网络中的逆映射爆炸问题。载于国际人工智能与统计会议（PMLR），1792–1800 Benning, M. 和 Burger, M.（2011）。一般保真项的误差估计。数值分析电子汇刊38，77 Benning, M. 和 Burger, M.（2018）。逆问题的现代正则化方法。*Acta Numerica*27，1–111 Benning, M. 和 Riis, E. S.（2021）。用于大规模优化及其在成像中应用的 Bregman 方法。计算机视觉与成像中的数学模型和算法手册，1–42 Bregman, L. M.（1967）。寻找凸集公共点的松弛法及其在求解凸规划问题中的应用。苏联计算数学与数学物理7，200–217 Burbea, J. 和 Rao, C.（1982a）。基于熵函数的高阶 Jensen 差的凸性（通信）。*IEEE* 信息论汇刊28，961–963 Burbea, J. 和 Rao, C.（1982b）。基于熵函数的某些散度度量的凸性。*IEEE* 信息论汇刊28，489–495 Chambolle, A.（2004）。一种全变分最小化算法及其应用。*Journal ofMathematical imaging and vision*20，89–97

Chambolle, A. 和 Lions, P.-L.（1997）。通过全变差最小化进行图像恢复及相关问题。*NumerischeMathematik* 76，167–188 Chambolle, A. 和 Pock, T.（2011）。一种用于凸问题的一阶原始-对偶算法及其在成像中的应用。*Journal of mathematical imaging and vision*40，120–145 Chambolle, A. 和 Pock, T.（2016）。成像连续优化导论。*ActaNumerica*25，161–319 Chang, C.-H., Creager, E., Goldenberg, A. 和 Duvenaud, D.（2018）。通过反事实生成解释图像分类器。*arXiv* 预印本 *arXiv:1807.08024*Dinh, L., Krueger, D. 和 Bengio, Y.（2015）。NICE：非线性独立成分估计。载于国际学习表征会议Engl, H. W., Hanke, M. 和 Neubauer, A.（1996）。逆问题的正则化，第375卷（Springer Science & Business Media）Esser, E., Zhang, X. 和 Chan, T. F.（2010）。成像科学中一类凸优化一阶原始-对偶算法的通用框架。*SIAM Journal on Imaging Sciences* 3，1015–1046 Fong, R., Patrick, M. 和 Vedaldi, A.（2019）。通过极值扰动和平滑掩码理解深度网络。载于*IEEE/CVF* 国际计算机视觉会议。2950–2958 Fong, R. C. 和 Vedaldi, A.（2017）。通过有意义的扰动解释黑箱。载于*IEEE* 国际计算机视觉会议。3429–3437 Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G. 等（2022）。一张图像胜过千言万语：使用文本反演实现文生图个性化。*arXiv* 预印本 *arXiv:2208.01618*Goodfellow, I., Bengio, Y. 和 Courville, A.（2016）。深度学习（MIT 出版社）He, K., Zhang, X., Ren, S. 和 Sun, J.（2016）。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。770–778 Ho, J., Jain, A. 和 Abbeel, P.（2020）。去噪扩散概率模型。神经信息处理系统进展33，6840–6851 Jensen, C. A., Reed, R. D., Marks, R. J., El-Sharkawi, M. A., Jung, J.-B., Miyamoto, R. T. 等（1999）。前馈神经网络的反演：算法与应用。*IEEE* 会议录 87，1536–1549 Kabri, S., Auras, A., Riccio, D., Bauermeister, H., Benning, M., Moeller, M. 等（2022）。用于 CT 重建的收敛数据驱动正则化。*arXiv* 预印本 *arXiv:2212.07786*Kindermann, J. 和 Linden, A.（1990）。通过梯度下降反演神经网络。并行计算14，277–286 Kingma, D. P. 和 Welling, M.（2013）。自编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*Kiwiel, K. C.（1997）。带广义 Bregman 函数的近端最小化方法。*SIAM* 控制与优化杂志35，1142–1168 Landweber, L.（1951）。第一类 Fredholm 积分方程的迭代公式。*Americanjournal of mathematics*73，615–624 LeCun, Y., Bengio, Y. 和 Hinton, G.（2015）。深度学习。*Nature*521，436–444 LeCun, Y., Bottou, L., Bengio, Y. 和 Haffner, P.（1998）。应用于文档识别的基于梯度的学习。*IEEE* 会议录 86，2278–2324 Li, H., Schwab, J., Antholzer, S. 和 Haltmeier, M.（2020）。NETT：使用深度神经网络求解逆问题。*Inverse Problems*36，065005

Lin, J. (1991). 基于香农熵的散度度量。*IEEE* 信息论汇刊37, 145–151 Linden, A. and Kindermann, J. (1989). 多层网络反演。载于神经网络国际联合会议论文集。第 2 卷，425–430 Lu, B.-L., Kita, H., and Nishikawa, Y. (1999). 使用线性规划和非线性规划反演前馈神经网络。*IEEE* 神经网络汇刊10, 1271–1290¨Lunz, S., Oktem, O., and Sch¨onlieb, C.-B. (2018). 逆问题中的对抗正则化器。神经信息处理系统进展31 Mahendran, A. and Vedaldi, A. (2015). 通过反演深度图像表示来理解它们。载于*IEEE* 计算机视觉与模式识别会议论文集。5188–5196 Morozov, V. A. (2012). 求解不适定问题的方法 (Springer Science & Business Media)¨ Mukherjee, S., Dittmer, S., Shumaylov, Z., Lunz, S., Oktem, O., and Sch¨onlieb, C.-B. (2021). 用于逆问题的学习型凸正则化器。*arXiv* 预印本 *arXiv:2008.02839v2*Nielsen, F. and Boltz, S. (2011). Burbea-Rao 质心与 Bhattacharyya 质心。*IEEE* 信息论汇刊57, 5455–5466 Pock, T., Cremers, D., Bischof, H., and Chambolle, A. (2009). 一种最小化 Mumford-Shah 泛函的算法。载于*2009*年 *IEEE* 第 *12* 届国际计算机视觉会议 (IEEE)，1133–1140 Rezende, D. and Mohamed, S. (2015). 使用归一化流进行变分推断。载于国际机器学习会议(PMLR)，1530–1538 Ronneberger, O., Fischer, P., and Brox, T. (2015). 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干预——*MICCAI 2015*：第 *18* 届国际会议，德国慕尼黑，*2015*年*10*月*5*日至*9*日，论文集，第 *III* 部分 *18* (Springer)，234–241 Rudin, L. I., Osher, S., and Fatemi, E. (1992). 基于非线性全变分的去噪算法。《物理学 *D*：非线性现象》60, 259–268 Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., and Lenzen, F. (2009). 成像中的变分方法 Schwab, J., Antholzer, S., and Haltmeier, M. (2019). 用于逆问题的深度零空间学习：收敛分析与收敛速率。逆问题35, 025008 Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). 卷积网络深处：可视化图像分类模型与显著性图。*arXiv* 预印本 *arXiv:1312.6034*Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). 利用非平衡热力学进行深度无监督学习。载于国际机器学习会议(PMLR)，2256–2265 Wang, X. and Benning, M. (2020). 广义感知机学习。第 *12* 届机器学习优化年度研讨会 Wang, X. and Benning, M. (2022). 神经网络的提升的 Bregman 训练。*arXiv* 预印本 *arXiv:2208.08772*Wright, S. J. (2015). 坐标下降算法。《数学规划》151, 3–34 Wright, S. J. and Recht, B. (2022). 数据分析优化(剑桥大学出版社) Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., and Yang, M.-H. (2022). GAN 反演综述。*IEEE* 模式分析与机器智能汇刊

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). 使用循环一致性对抗网络进行无配对图像到图像转换。载于*IEEE* 国际计算机视觉会议论文集。2223–2232 Zhu, M. and Chan, T. (2008). 一种高效的全变分图像复原原始-对偶混合梯度算法。加州大学洛杉矶分校技术报告34, 8–34