一些定理

定理 0.1 (平衡模型的梯度) 设 $\mathbf{z}^*_{1:T} \in \mathbb{R}^{T\times d}$ 是一个长度为 T，维度为 d 的一个平衡隐藏序列，$\mathbf{y}_{1:T} \in \mathbb{R}^{T\times q}$ 是对应真实目标序列，令 $g_\theta(\mathbf{z};\mathbf{x}) = f_\theta(\mathbf{z};\mathbf{x}) - \mathbf{z}$，其中 θ 是模型参数，$h : \mathbb{R}^d \to \mathbb{R}^q$ 是任意可微函数，$\mathcal{L} : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ 是损失函数，损失定义为 $\ell = \mathcal{L}(h(\mathbf{z}^*_{1:T}), \mathbf{y}_{1:T}) = \mathcal{L}(h(\text{RootFind}(g_\theta; \mathbf{x}_{1:T})), \mathbf{y}_{1:T})$，则关于任意待求梯度的变量 $(\cdot)$（如 θ 或 $\mathbf{x}_{1:T}$），有

$$\frac{\partial \ell}{\partial (\cdot)} = -\frac{\partial \ell}{\partial \mathbf{z}^*_{1:T}} \left(J^{-1}_{g_\theta}\big|_{\mathbf{z}^*_{1:T}} \right) \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial (\cdot)} = -\frac{\partial \ell}{\partial h}\frac{\partial h}{\partial \mathbf{z}^*_{1:T}} \left(J^{-1}_{g_\theta}\big|_{\mathbf{z}^*_{1:T}} \right) \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial (\cdot)},$$

其中 $\left(J^{-1}_{g_\theta}\big|_{\mathbf{z}^*_{1:T}} \right)$ 是 $\mathbf{x}$ 处 g_θ 的 Jacobi 矩阵的逆。

证明. 对平衡条件 $f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T}) = \mathbf{z}^*_{1:T}$ 两边同时对 $(\cdot)$ 求导，得到

$$\frac{\mathrm{d}\mathbf{z}^*_{1:T}}{\mathrm{d}(\cdot)} = \frac{\mathrm{d} f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\mathrm{d}(\cdot)} = \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial (\cdot)} + \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^*_{1:T}} \frac{\mathrm{d}\mathbf{z}^*_{1:T}}{\mathrm{d}(\cdot)},$$

记 I 是单位矩阵，$\dfrac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^*_{1:T}}$ 是 f_θ 隐藏状态的 Jacobi 矩阵，则有

$$\left(I - \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^*_{1:T}} \right) \frac{\mathrm{d}\mathbf{z}^*_{1:T}}{\mathrm{d}(\cdot)} = \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial (\cdot)} \qquad (*),$$

由于 $g_\theta(\mathbf{z};\mathbf{x}) = f_\theta(\mathbf{z};\mathbf{x}) - \mathbf{z}$，有 $J_{g_\theta}\big|_{\mathbf{z}_{1:T}} = \dfrac{\partial g_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^*_{1:T}} = \dfrac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^*_{1:T}} - I$，从而

$$\left(I - \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^*_{1:T}} \right)^{-1} = - J^{-1}_{g_\theta}\big|_{\mathbf{z}_{1:T}}$$

在 $(*)$ 左右两侧左乘 $\left(I - \dfrac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^*_{1:T}} \right)^{-1}$，得到

$$\frac{\mathrm{d}\mathbf{z}^*_{1:T}}{\mathrm{d}(\cdot)} = - \left(J^{-1}_{g_\theta}\big|_{\mathbf{z}_{1:T}} \right) \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial (\cdot)},$$

把结果带入链式法则 $\dfrac{\partial \ell}{\partial (\cdot)} = \dfrac{\partial \ell}{\partial \mathbf{z}^*_{1:T}} \dfrac{\mathrm{d}\mathbf{z}^*_{1:T}}{\mathrm{d}(\cdot)} = \dfrac{\partial \ell}{\partial h}\dfrac{\partial h}{\partial \mathbf{z}^*_{1:T}} \dfrac{\mathrm{d}\mathbf{z}^*_{1:T}}{\mathrm{d}(\cdot)}$，即得到

$$\frac{\partial \ell}{\partial (\cdot)} = -\frac{\partial \ell}{\partial \mathbf{z}^*_{1:T}} \left(J^{-1}_{g_\theta}\big|_{\mathbf{z}^*_{1:T}} \right) \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial (\cdot)} = -\frac{\partial \ell}{\partial h}\frac{\partial h}{\partial \mathbf{z}^*_{1:T}} \left(J^{-1}_{g_\theta}\big|_{\mathbf{z}^*_{1:T}} \right) \frac{\partial f_\theta(\mathbf{z}^*_{1:T}; \mathbf{x}_{1:T})}{\partial (\cdot)}.$$

- 意义：该定理对 DEQ 模型的训练至关重要。它揭示了即使网络被看作有“无限”层，反向传播的梯度仍然能够以一种高效且内存友好的方式计算。传统深度网络需要存储每一层的中间激活以进行反向传播，但有了此定理给出的公式，DEQ 在计算梯度时只需储存平衡态 z^* 和少量必要信息即可。这意味着 DEQ 的内存开销是恒定的，与网络层数无关，从而能够训练非常深（实际上无限深）的模型而不耗费比常规模型更多的内存。此外，该梯度公式与找到平衡点所采用的具体算法无关——无论前向求解平衡态用的是何种求根方法（如牛顿迭代、Broyden 方法等），梯度计算公式依然成立。这保证了我们可以把求解平衡点的过程看作一个黑盒，在不反复回溯计算路径的情况下直接获得梯度。这一特性是 DEQ 方法及其各项优势的核心基础。
- 正文联系：正文中的方法介绍部分直接依赖了本定理的结论来进行反向传播。利用这一梯度公式，DEQ 的反向过程可以简化为在平衡点处进行一次矩阵-向量乘法或求解一次线性方程组，代表了通过平衡态直接反传梯度。正文 3.1 节描述了这种常数内存的反传方案，并在 3.1.3 节讨论了如何使用 Broyden 等方法高效近似求解所需的 Jacobi 逆矩阵。在 3.2 节的“DEQ 的一些性质”中，作者强调由于有了定理 1 给出的解析梯度，DEQ 能够在不展开计算图的情况下完成训练，并将这一特性列为 DEQ 模型的主要优点之一（例如显著降低内存/显存占用，使得以前需要多块 GPU 的模型可以在单 GPU 上训练）。总之，定理 1 的结论奠定了 DEQ 训练算法的理论基础，其影响在正文中体现在算法设计、内存性能以及对比传统深度网络训练的讨论中。

定理 0.2 (单层 DEQ 的通用性) 设 $\mathbf{x}_{1:T} \in \mathbb{R}^{T\times p}$ 是输入序列。令 $f_{\theta^{[1]}} : \mathbb{R}^r \times \mathbb{R}^p \to \mathbb{R}^r, v_{\theta^{[2]}} : \mathbb{R}^d \times \mathbb{R}^r \to \mathbb{R}^d$ 是带参数 $\theta^{[1]}$ 和 $\theta^{[2]}$ 的稳定变换，则存在 $\Gamma_\Theta : \mathbb{R}^{d+r} \times \mathbb{R}^p \to \mathbb{R}^{d+r}, \Theta = \theta^{[1]} \cup \theta^{[2]}$ 满足 $\mathbf{z}^*_{1:T} = \text{RootFind}(g^v_{\theta^{[2]}}; \text{RootFind}(g^f_{\theta^{[1]}}; \mathbf{x}_{1:T})) = \text{RootFind}(g^\Gamma_\Theta; \mathbf{x}_{1:T})_{[:,-d:]}$，其中 $[\cdot]_{[:,-d:]}$ 表示取张量最后 d 个特征维度。

证明. 设 $\mathbf{z}^{[1]*}_{1:T} = \text{RootFind}(g^f_{\theta^{[1]}}; \mathbf{x}_{1:T}) \in \mathbb{R}^r$ 是第一层 DEQ 的平衡解，如下定义 Γ_Θ：

$$\Gamma_\Theta(\mathbf{w}_{1:T}; \mathbf{x}_{1:T}) = \Gamma_\Theta\left(\begin{bmatrix}\mathbf{w}^{(1)}_{1:T}\\ \mathbf{w}^{(2)}_{1:T}\end{bmatrix}; \mathbf{x}_{1:T}\right) \coloneqq \begin{bmatrix} f_{\theta^{[1]}}(\mathbf{w}^{(1)}_{1:T}; \mathbf{x}_{1:T})\\ v_{\theta^{[2]}}(\mathbf{w}^{(2)}_{1:T}; \mathbf{w}^{(1)}_{1:T})\end{bmatrix},$$

则 $\mathbf{w}^*_{1:T} \coloneqq \begin{bmatrix}\mathbf{z}^{[1]*}_{1:T}\\ \mathbf{z}^*_{1:T}\end{bmatrix}$ 是 Γ_Θ 的不动点，因此 Γ_Θ 即为所求。

- 思路：证明中，将两个顺序的 DEQ 模块合并为一个新的 DEQ：新 DEQ 的隐藏向量由原先第一和第二个 DEQ 的隐藏状态拼接而成（维度为 $r+d$，分别对应两个模块的隐藏维度）。同时，定义一个新的变换函数 Γ_Θ 使其对拼接隐藏向量的作用等价于同时执行原先的 $f_{\theta^{[1]}}$ 和 $v_{\theta^{[2]}}$ 两个变换。具体地，令联合隐藏向量 $w = \begin{bmatrix}w^{(1)}\\ w^{(2)}\end{bmatrix}$，其中 $w^{(1)} \in \mathbb{R}^r$ 对应第一个 DEQ 模块的部分，$w^{(2)} \in \mathbb{R}^d$ 对应第二个模块。可以看出，Γ_Θ 的上半部分输出相当于第一层 DEQ 变换 $f_{\theta^{[1]}}$，下半部分输出相当于第二层变换 $v_{\theta^{[2]}}$ 接受了第一层的结果作为输入。在这种构造下，如果 $\begin{bmatrix}z^{(1)}\\ z^{(2)}\end{bmatrix}$ 是 Γ_Θ 关于输入 x 的一个不动点，那么它满足：
  - 上半部分 $z^{(1)} = f_{\theta^{[1]}}(z^{(1)}, x)$，因此 $z^{(1)}$ 就是第一层 DEQ 的平衡解；

- 下半部分 $z^{(2)} = v_{\theta[2]}(z^{(2)}, z^{(1)})$，因此 $z^{(2)}$ 就是在 $z^{(1)}$ 作为输入时第二层 DEQ 的平衡解。

从而 $z^{(2)}$ 正是原两层 DEQ 堆叠后的输出。由此可知，$\begin{bmatrix} z^{(1)} \\ z^{(2)} \end{bmatrix}$ 是 Γ_Θ 的不动点当且仅当 $z^{(1)}$ 和 $z^{(2)}$ 分别是两层 DEQ 的不动点，并且 $z^{(2)}$ 以 $z^{(1)}$ 为输入达到平衡。因此 Γ_Θ 的平衡解下半部分（即最后 d 个维度）直接给出了两层 DEQ 堆叠的结果。这个证明过程的技巧在于引入一个增广的状态和联合变换，将串联的动态融合成单一步动态，形式上类似于把两个方程组并成一个方程组求解。

- 意义：这一定理在理论上回答了一个很自然的问题："如果一个 DEQ 模块效果很好，那么堆叠多个 DEQ 模块是否会更好?" 根据定理 2，答案是否定的。也就是说，增加 DEQ 的层数并不能提供额外的表示能力。这对于 DEQ 的设计具有重要意义：它表明采用单层平衡模型已经足够，一层 DEQ 可以取代传统深度网络中的多层堆叠。因此，在实践中，无需叠加多个 DEQ 模块来提升模型表达力，只需专注于设计一个稳定且强大的变换 f_θ 并求解其平衡点即可。这一发现使 DEQ 的架构更加简洁，并强化了"用不动点求解取代深度堆叠"这一思想在理论上的可行性和完备性。此外，该定理也暗示了深度并非提升表达力的唯一手段：通过非线性不动点方程，一个层（但内部相当于无限迭代）同样能够达到多层网络的效果。这为深度学习的网络设计提供了新的视角与可能性。
- 正文联系：该定理实质上说明了"堆叠多个 DEQ 并不会带来比单个 DEQ 更多的表示能力"。因此，在正文的实现和实验中，作者都只采用单层的 DEQ 结构，而没有尝试多层堆叠。定理 2 与正文的联系在于，它为作者选择极简的"单层 + 不动点求解"架构提供了保证，说明了将深度学习的"深"彻底移交给不动点求解并不会损失模型的表达能力。这个结论在正文第 3.2 节得到强调，并在后续实验中隐含地体现：无论是在 TrellisNet 还是 Transformer 的实例中，作者都使用一个 DEQ 模块获得了有竞争性的结果，因为理论上再增加同类模块也无益处。

定理 0.3 (权重共享深层网络的通用性) *假设一个 L 层前馈神经网络满足递归关系*

$$\mathbf{z}^{[i+1]} = \sigma^{[i]}(W^{[i]}\mathbf{z}^{[i]} + \mathbf{b}^{[i]}), i = 0, \dots, L-1, \mathbf{z}^{[0]} = \mathbf{x},$$

其中 $\mathbf{z}^{[i]}$ 表示第 i 层的隐藏特征；$W^{[i]}$、$\mathbf{b}^{[i]}$ 为该层的权重矩阵与偏置向量；$\sigma^{[i]}$ 为第 i 层的非线性激活函数；$\mathbf{x}$ 为原始输入，则存在一组参数 $\sigma, W_z, W_x, \tilde{\mathbf{b}}$，使得该网络可等价表示为同样深度的权重共享且输入注入网络

$$\tilde{\mathbf{z}}^{[i+1]} = \sigma(W_z\tilde{\mathbf{z}}^{[i]} + W_x\mathbf{x} + \tilde{\mathbf{b}}), i = 0, \dots, L-1,$$

其中 $\sigma, W_z, W_x, \tilde{\mathbf{b}}$ 在全部 L 层中保持恒定。

证明.

$$W_z = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ W^{[1]} & 0 & \cdots & 0 & 0 \\ 0 & W^{[2]} & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & W^{[L-1]} & 0 \end{bmatrix}, W_x = \begin{bmatrix} W^{[0]} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tilde{\mathbf{b}} = \begin{bmatrix} \mathbf{b}^{[0]} \\ \mathbf{b}^{[1]} \\ \mathbf{b}^{[2]} \\ \vdots \\ \mathbf{b}^{[L-1]} \end{bmatrix}, \sigma = \begin{bmatrix} \sigma^{[0]} \\ \sigma^{[1]} \\ \sigma^{[2]} \\ \vdots \\ \sigma^{[L-1]} \end{bmatrix}$$

显然经过 L 次迭代后隐藏向量 z 的取值将是 $\tilde{\mathbf{z}}^{[L]} = \begin{bmatrix} \mathbf{z}^{[1]} \\ \mathbf{z}^{[2]} \\ \vdots \\ \mathbf{z}^{[L]} \end{bmatrix}$，即权重共享网络计算的所有项都与原始网络相同，使用的深度也与原始网络相同，而隐藏单元的大小只是原始网络中各个隐藏单元大小的总和。这就得到了定理的证明。

- 思路：构造的关键在于引入更高维的隐藏状态，将原网络各层的隐藏单元串联起来，并借助特殊设计的权重矩阵在每个共享层中按层传递信息。具体来说，假设原始网络第 i 层的隐藏向量维度为 n_i。我们构建新网络的隐藏向量 $\tilde{z}$ 为原网络各层隐藏向量的级联，即 $\tilde{\mathbf{z}} = \begin{bmatrix} \mathbf{z}^{[0]} \\ \mathbf{z}^{[1]} \\ \vdots \\ \mathbf{z}^{[L-1]} \end{bmatrix}$，总维度为 $n_0 + n_1 + \ldots + n_{L-1}$（其中 $\mathbf{z}^{[0]} = x$ 是输入向量）。接着，构造一个分块形式的权重矩阵 W_z，使其在每次迭代（每层）作用在 $\tilde{\mathbf{z}}$ 上时，实现将 $\tilde{\mathbf{z}}$ 的下一段计算出来、实现前一段“移位”的效果。具体构造如下：W_z 被设计为一个块矩阵，尺寸为 $(\sum_i n_i) \times (\sum_i n_i)$。它的大部分块是零矩阵，唯有在一些特定位置放入原网络各层的权重矩阵 $W^{[i]}$。例如，可以令 W_z 在对应 $\mathbf{z}^{[i]} \to \mathbf{z}^{[i+1]}$ 的位置放置 $W^{[i]}$，其余位置为 0。这样，当 $\tilde{\mathbf{z}}$ 经过 $\sigma(W_z\tilde{\mathbf{z}} + W_x\mathbf{x} + \tilde{\mathbf{b}})$ 这一层的计算时：$\tilde{z}$ 的第 i 段（对应原网络第 i 层的隐藏状态) 将通过 $W^{[i]}$ 映射并加上偏置，形成 $\tilde{\mathbf{z}}$ 的第 $(i+1)$ 段；而 $\tilde{\mathbf{z}}$ 原来的第 i 段自身在 W_z 作用下并不会直接传递给下一层（因为对应的位置是零），只能通过被“移位”到 $\tilde{\mathbf{z}}$ 的下一段来间接体现。与此同时，引入一个固定的矩阵 W_x 来在每层都加入对输入 $\mathbf{x}$ 的线性作用（所谓“输入注入”），以及一个固定的偏置向量 $\tilde{\mathbf{b}}$，它实际上是将原网络各层的偏置 $[\mathbf{b}^{[0]}; \mathbf{b}^{[1]}; \ldots; \mathbf{b}^{[L-1]}]$ 纵向堆叠而成；激活函数 σ 则通过扩维为 $\begin{bmatrix} \sigma^{[0]} \\ \sigma^{[1]} \\ \vdots \\ \sigma^{[L-1]} \end{bmatrix}$ 的形式，作用于 $\tilde{\mathbf{z}}$ 的各对应区块。在这种设计下，每经过新网络的一层计算，就等价于依次在原网络中推进了一层：新网络隐藏状态的第 1 段经过 $\sigma^{[0]}(W^{[0]}\mathbf{x} + \mathbf{b}^{[0]})$ 得到原第一层输出，第 2 段经过 $\sigma^{[1]}(W^{[1]}\mathbf{z}^{[1]} + \mathbf{b}^{[1]})$ 得到原第二层输出，依此类推；同时原第 i 层的输出被“写入” $\tilde{\mathbf{z}}$ 的第 $i+1$ 段，准备在下一层计算中使用。经过 L 次这样的迭代，新网络的 $\tilde{\mathbf{z}}$ 最后一段正好是原网络第 L 层的输出 $\mathbf{z}^{[L]}$，而这也是新网络的输出 $\tilde{\mathbf{z}}^{[L]}$。由此可见，新网络成功地在 L 层共享权重的结构下重现了原网络每一层的效果，证明了二者的等价性。
- 意义：该定理为 DEQ 这类权重共享的无限深网络提供了重要的理论支持。直观来看，DEQ 相当于一个权重在各层都相同且层数趋于无限的极限情况，这种架构与传统每层独立的深度网络相比看似受到了很大限制。然而定理 3 证明了这种限制在表示能力上并不存在本质影响：权重共享的网络（即使层数受限，更何况无限深）同样具有万能逼近的能力，可以拟合任何原本通过逐层独立权重实现的功能。这个结论消除了人们对于“参数共享是否会降低网络表达力”的疑虑，表明通过增加隐藏层宽度，权重共享网络的表示能力可以与普通网络等价。因此，在理论层面，DEQ 选择采用层与层之间共享变换的架构并不会牺牲模

型的表达能力。这一点非常关键，它使作者能够将所有层权重设为相同（并通过不动点求解达至无限层），而不必担心模型因减少参数而变得“不万能”。另外，定理 3 也体现了深度学习中一个有趣的性质：模型容量可以在“深度”和“宽度”之间权衡转化——通过增大每层状态的维度，完全可以用权重共享的 L 层网络模拟出非共享权重的 L 层网络的行为（只需线性增大的宽度）。这一洞见对设计压缩参数或结构受限的网络（如循环神经网络中的参数共享模式）具有参考价值。

- 正文联系：作者提到 DEQ 对应的是一种权重共享且每层都有输入的无限深网络，乍看之下这似乎比一般深度网络更受限，但实际上在表示能力上没有显著损失。定理为 DEQ 模型选用“各层同构且参数共享”的架构提供了理论保证。正文在介绍 DEQ 时，强调这种无限深、参数层层共享的模型能够看作对现有各种序列模型的一种统一和延展，并在理论上不弱于它们。这背后隐含的就是定理的结论：任何传统深度网络都能等价地转换成权重共享形式。简而言之，定理解除了对于 DEQ 架构表达能力的疑虑，支持了正文关于 DEQ 普适性的论述，使得作者能够专注于 DEQ 的稳定求解和应用，而不必在意“共享权重会不会限制模型”。这加强了全文的逻辑闭环：通过定理 1 保证可训练性，定理 2 保证不需要多层，定理 3 保证共享权重不损失表达力，从而共同支撑起 DEQ 模型设计的合理性和优越性。