Gradient flow for deep equilibrium single-index models

Sanjit Dandapanthula* sanjitd@cmu.edu
Aaditya Ramdas† aramdas@cmu.edu

November 24, 2025

Abstract

Deep equilibrium models (DEQs) have recently emerged as a powerful paradigm for training infinitely deep weight-tied neural networks that achieve state of the art performance across many modern machine learning tasks. Despite their practical success, theoretically understanding the gradient descent dynamics for training DEQs remains an area of active research. In this work, we rigorously study the gradient descent dynamics for DEQs in the simple setting of linear models and single-index models, filling several gaps in the literature. We prove a conservation law for linear DEQs which implies that the parameters remain trapped on spheres during training and use this property to show that gradient flow remains well-conditioned for all time. We then prove linear convergence of gradient descent to a global minimizer for linear DEQs and deep equilibrium single-index models under appropriate initialization and with a sufficiently small step size. Finally, we validate our theoretical findings through experiments.

*Carnegie Mellon University, Department of Statistics
†Carnegie Mellon University, Department of Statistics and Machine Learning Department

1 Introduction

The deep equilibrium model (DEQ) framework introduced in Bai et al. (2019) has recently emerged as a powerful paradigm for training neural networks that has enjoyed practical success across many modern machine learning tasks, including image generation (Pokle et al., 2022; Geng et al., 2023), inverse problems (Gilton et al., 2021), graph neural networks (Gu et al., 2020; Liu et al., 2022), classification and semantic segmentation (Bai et al., 2020), representation learning (Huang et al., 2021), and large language modeling (Bai et al., 2019; Geiping et al., 2025). DEQs have been shown to have several advantages over traditional deep neural networks, including lower memory consumption during training (Geng et al., 2021; Fung et al., 2022), the ability to flexibly adapt to different computational budgets at inference time (Marwah et al., 2023; Wang et al., 2024; Geiping et al., 2025), and better representational power (Wu et al., 2024; Liu et al., 2025).

Instead of learning an explicit relation $y_i = f(x_i)$ from samples $\{(x_i, y_i)\}_{i=1}^n$, DEQs attempt to learn an implicit relation $y_i = g(x_i, y_i)$ over a parameterized class of functions $\{g_\theta\}_{\theta\in\Theta}$. Given an input x, the fixed point $y_\theta(x)$ of $g_\theta(x, \cdot)$ can often be interpreted as the limiting point of the sequence $y^{(t+1)} = g_\theta(x, y^{(t)})$ for any initial injection $y^{(0)}$; hence, we may view DEQs as a single layer which implicitly defines infinitely deep neural networks with input injection and weight tying across layers. In fact, Bai et al. (2019) showed that there is no advantage to stacking single-layer DEQs. The training of DEQs is typically performed using the implicit function theorem, and inference is

performed using a root-finding algorithm to solve for the fixed point $y_\theta(x)$.

Despite their practical success, theoretically understanding the gradient descent dynamics for training DEQs remains an area of active research (Kawaguchi, 2021; Ling et al., 2023; Wu et al., 2024). In this work, we initiate a rigorous study of the gradient descent dynamics for DEQs in the simple setting of linear models (where the target is $f(x) = \xi^\top x$) and single-index models (where the target is $f(x) = \sigma(\xi^\top x)$ for some activation function $\sigma : \mathbb{R} \to \mathbb{R}$). We concern ourselves with the following questions:

Q1: Due to the implicit nature of DEQs, the quantity $1 - \frac{\partial}{\partial y} g_\theta(x, y_\theta(x))$ must remain nonzero during training to ensure the existence and uniqueness of the fixed point $y_\theta(x)$. Why does the Jacobian remain well-conditioned during implicit training of linear DEQs? As an example, why do linear DEQs not converge to the trivial model $y_\theta(x) = y_\theta(x)$?

Q2: Under what conditions does gradient descent converge to the true parameter when training deep equilibrium single-index models?

Using techniques from dynamical systems theory and previously developed tools for analyzing gradient flow for explicit single-neuron neural networks (e.g., Yehudai and Shamir, 2020), we provide a full answer to Q1 and a partial answer to Q2 in the settings of well-specified linear and single-index models. Similarly to Yehudai and Shamir (2020), we work directly with the population risk and focus on the gradient flow and gradient descent dynamics for minimizing this risk, leaving the analysis of empirical risk minimization to future work. Our work makes similar assumptions to prior work on gradient flow for single-index models, but notably, our final rates depend on the *amount of nonlinearity* present in the activation function. By analyzing both linear and nonlinear single-index models, we are therefore able to highlight the differences in the training dynamics of DEQs that arise due to the presence of nonlinearity.

First, we summarize our main contributions in Section 1.1. Then in Section 1.2, we discuss related work on DEQs and gradient flow for neural networks and in Section 1.3, we introduce notation used throughout the paper. Next, we review the DEQ framework and the gradient flow framework for analyzing gradient descent dynamics in Section 2. In Section 3, we present our main results on the behavior of gradient flow and gradient descent for linear DEQs and deep equilibrium single-index models. Finally, we provide empirical validation of our theory in Section 4.

1.1 Contributions

We summarize our main contributions as follows:

- In the case of the linear model $f(x) = \xi^\top x$, we show that if we parameterize $y_\theta(x) = g_\theta(x, y) = \theta_1^\top x + \theta_2\, y_\theta(x)$ then gradient flow satisfies a conservation law that keeps the parameters trapped on spheres centered at $(0, \ldots, 0, 1)$ in $\mathbb{R}^{d+1}$.
- We prove that gradient flow for linear DEQs remains bounded away from the singularity $\theta_2 = 1$ and hence the DEQ remains well-defined during training. Further, we explicitly characterize the limiting parameter for any initialization.
- Using the above results, we show that gradient flow converges exponentially fast to a global minimizer of the population risk for linear DEQs as long as the initialization is not on the singular hyperplane $\theta_2 = 1$ and the data distribution has full-rank covariance.
- We extend our analysis to single-index models $f(x) = \sigma(\xi^\top x)$ parameterized as $g_\theta(x, y) = \sigma(\theta_1^\top x + \theta_2\, y)$. Under regularity conditions on the activation function (which are satisfied by the common hyperbolic tangent, sigmoid, and softplus activations) and on the data, we prove that gradient flow converges exponentially fast to the true parameter $\theta = (\xi, 0)$ as long as the initialization is sufficiently close to the true parameter.
- As long as the step size is chosen sufficiently small, we demonstrate that gradient descent converges linearly for linear DEQs, and for single-index models with appropriate initialization.

1.2 Related work

In this section, we briefly review related work on DEQs and gradient flow for neural networks.

Deep equilibrium models. DEQs were introduced in Bai et al. (2019) as a framework for training infinitely deep neural networks with weight tying across layers and input injection. Since then, DEQs have been successfully applied to image generation (Pokle et al., 2022; Geng et al., 2023), inverse problems (Gilton et al., 2021), graph neural networks (Gu et al., 2020; Liu et al., 2022), classification and semantic segmentation (Bai et al., 2020), representation learning (Huang et al., 2021), and large language modeling (Bai et al., 2019; Geiping et al., 2025). McCallum et al. (2025) recently suggest that it can be computationally beneficial to create an algebraically reversible fixed-point iteration, so that we can directly backpropagate through the fixed point iteration instead of approximating the gradient by the inverse problem in (1); however, their approach turns the implicit model into an explicit one, so it is not covered by our theory.

Gradient flow for neural networks. There is a large body of work establishing that gradient descent converges for non-convex problems which satisfy a *strict saddle property*, meaning that all suboptimal saddle points admit a direction of easy escape (Ge et al., 2015; Sun et al., 2015; Jin et al., 2017). For example, this phenomenon has been heavily studied for the phase retrieval problem (with $\sigma(z) = z^2$), where gradient descent is known to converge exponentially fast with appropriate spectral initialization (Candès et al., 2015; Sun et al., 2018). Gradient flow for *explicit* single-index models has been extensively studied in the literature (Mei et al., 2018; Oymak and Soltanolkotabi, 2019; Yehudai and Shamir, 2020), where typical assumptions include monotonicity of the activation function (which is assumed to have a strong gradient near zero) and that the data is sufficiently spread out near zero.

Gradient flow for DEQs. Kawaguchi (2021) studies the optimization landscape of implicit models, showing that gradient descent converges at a linear rate to the global optimum for over-parameterized linear implicit models. However, their results consider a stylized setting including a softmax nonlinearity on the weights in order to ensure well-posedness of the implicit model during training; we make no such modifications. In a similar vein, Ling et al. (2023) and Gao and Gao (2022) show that the loss landscape for DEQs is benign in the heavily overparameterized regime where the width of the implicit layer is much larger than the number of samples. Gao et al. (2023) shows that infinitely wide implicit models can be modeled by Gaussian processes, and Feng and Kolter (2023) shows that there is a deterministic neural tangent kernel (NTK) for implicit models under mild conditions and show how to find it explicitly by root-finding.

Closest to our work, Wu et al. (2024) analyzes the gradient descent dynamics for training *diagonal linear* DEQs and shows global convergence of gradient descent in the overparameterized regime. However, they analyze a different parameterization $y_\theta(x) = g_\theta(x, y_\theta(x)) = \theta_1^\top x + \theta_2^\top y_\theta(x) \in \mathbb{R}^d$ and only assume access to samples—they do not assume that the ground truth is linear. As a result, they do not obtain the conservation law that we derive in this work or our explicit characterization of the limiting parameters for linear DEQs.

Stability of training DEQs. There is a long line of work studying well-posedness of the Jacobian inverse in (1). For example, Winston and Kolter (2020), Revay et al. (2020), and El Ghaoui et al. (2021) study sufficient conditions on the form of the neural network g_θ to ensure that the fixed-point problem is well-posed. Bai et al. (2021) point out that regularizing the training of implicit models

can prevent overfitting and training instabilities arising from ill-conditioned Jacobians. Agarwala and Schoenholz (2022) show that initializing implicit models using orthogonal or symmetric matrices can help ensure that the Jacobian inverse is well-defined.

1.3 Notation

For an $\mathbb{R}^d$-valued random variable X, we write $X \in L^p(\mathbb{P})$ if $\mathbb{E}[\|X\|_2^p] < \infty$. We write

$$\|f\|_{\mathrm{Lip}} \coloneqq \sup_{x \neq y} \frac{\|f(x) - f(y)\|_2}{\|x - y\|_2}$$

for the Lipschitz norm of a function $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$. We denote by ∇_θ the gradient operator with respect to the parameter θ. We write $A \succeq 0$ if A is positive semidefinite and $A \succ 0$ if A is positive definite. We write $\lambda_{\min}(A)$ for the minimum eigenvalue of a symmetric matrix $A \in \mathbb{R}^{d \times d}$ and $\lambda_{\max}(A)$ for its maximum eigenvalue. We write I_d for the identity matrix over $\mathbb{R}^d$. We use $\mathbb{N}_0$ to denote the set of nonnegative integers and use $\mathrm{sign}(x) \coloneqq \mathbf{1}_{x \geq 0} - \mathbf{1}_{x < 0}$ for the sign function. We use $S^{d-1} \coloneqq \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ to denote the unit sphere in $\mathbb{R}^d$.

2 Background

In this section, we briefly review deep equilibrium models (DEQs) and the gradient flow framework for analyzing gradient descent dynamics. Suppose that the data is generated according to a function $f : \mathbb{R}^d \to \mathbb{R}$; our goal is to learn this function from samples. In the DEQ framework, we parameterize a class of functions $\{g_\theta\}_{\theta \in \Theta}$ and define the model output $y_\theta(x)$ for an input $x \in \mathbb{R}^d$ as the fixed point of the equation $y = g_\theta(x, y)$.

For the forward pass, we need to solve the fixed-point equation $y_\theta(x) = g_\theta(x, y(x))$ for $y_\theta(x)$, so we can use any black-box root-finding algorithm or fixed-point iteration. In particular, as long as g_θ is contractive in its second argument (i.e., $\|g_\theta(x, \cdot)\|_{\mathrm{Lip}} < 1$), the Banach fixed-point theorem guarantees that the fixed point exists and is unique. Furthermore, we may use simple fixed-point iteration $y^{(t+1)} = g_\theta(x, y^{(t)})$ to solve for $y_\theta(x)$ starting from any initial choice $y^{(0)}$. The fact that there are no assumptions about the root-finding algorithm allows for *test-time scaling*, since solving the equation by a more complex iterative scheme may enable convergence to a better solution.

For the backward pass, we would like to minimize $L(y_\theta(x))$ for some differentiable risk function L, by gradient descent. Given a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we may define the population risk as

$$R(\theta) \coloneqq \mathbb{E}[\ell(y_\theta(X), f(X))],$$

where X is an $\mathbb{R}^d$-valued random variable drawn according to the input data distribution. Although in practice we may only have access to finite samples, we focus on the problem of directly minimizing the population risk over $\theta \in \Theta$ using gradient descent. For gradient descent with step size $\eta > 0$ and initialization $\theta \in \Theta$, the parameter updates are given by

$$\theta(t+1) = \theta(t) - \eta \, \nabla_\theta \, R(\theta(t)).$$

To analyze the gradient descent dynamics, we may also consider the continuous-time limit of gradient descent, often referred to as gradient flow. In this framework, we consider the ordinary differential equation (ODE)

$$\theta'(t) = -\nabla_\theta \, R(\theta(t))$$

with initialization $\theta(0) = \theta$. Assuming that we can differentiate under the integral sign, it suffices to compute $\nabla_\theta \, \ell(y_\theta(x), f(x))$ for a fixed x to obtain the gradient of the population risk. If we assume $g_\theta \in C^1(U)$ for some open neighborhood U of $(x, y_\theta(x))$, then the implicit function theorem states that as long as $1 - \frac{\partial}{\partial y} \, g_\theta(x, y_\theta(x)) \neq 0$, there exists a neighborhood around x where $y_\theta(x)$ is uniquely defined and differentiable with

$$\nabla_\theta \, y_\theta(x) = \left(1 - \frac{\partial}{\partial y} \, g_\theta(x, y_\theta(x))\right)^{-1} \nabla_\theta \, g_\theta(x, y_\theta(x)). \tag{1}$$

In particular, we can then differentiate both sides of $y_\theta(x) = g_\theta(x, y_\theta(x))$ using the chain rule to get

$$\begin{aligned} \nabla_\theta \, \ell(y_\theta(x)) &= \left(\frac{\partial}{\partial y} \, \ell(y_\theta(x), f(x))\right) \nabla_\theta \, y_\theta(x) \\ &= \left(\frac{\partial}{\partial y} \, \ell(y_\theta(x), f(x))\right) \left(1 - \frac{\partial}{\partial y} \, g_\theta(x, y_\theta(x))\right)^{-1} \nabla_\theta \, g_\theta(x, y_\theta(x)). \end{aligned}$$

Hence, the Jacobian term $1 - \frac{\partial}{\partial y} \, g_\theta(x, y_\theta(x))$ plays a crucial role in determining the gradient of the loss with respect to the parameters. If this term becomes close to zero during training, the gradient may become unbounded, leading to instability in the training dynamics. Therefore, it is important to understand whether this Jacobian term remains well-conditioned during implicit training of DEQs.

3 Gradient flow for deep equilibrium models

In this section, we discuss our main results about the behavior of gradient flow and gradient descent for DEQs.

3.1 Linear models

First, we consider the linear function $f(x) = \xi^\top x$ for some fixed nonzero $\xi \in \mathbb{R}^d$.

3.1.1 Gradient flow

In this section, we analyze the gradient flow dynamics for fitting this function using a linear deep equilibrium model of the form

$$y_\theta(x) = g_\theta(x, y) \coloneqq \theta_1^\top x + \theta_2\, y_\theta(x). \tag{2}$$

In this case, we may explicitly solve for $y_\theta(x)$ as

$$y_\theta(x) = \frac{\theta_1^\top x}{1 - \theta_2},$$

as long as $\theta_2 \neq 1$, but our model still retains the implicit nature of the DEQ. We assume that the input data X is an $\mathbb{R}^d$-valued random variable and use the squared loss $\ell(y, \hat{y}) = (\hat{y} - y)^2$; our goal is now to minimize the population risk

$$R(\theta, \xi) \coloneqq \mathbb{E}[(y_\theta(X) - f(X))^2] = \mathbb{E}\left[\left(\frac{\theta_1^\top X}{1 - \theta_2} - \xi^\top X\right)^2\right],$$

assuming that $\theta_2 \neq 1$. Notice that any parameter along the line $(\alpha\xi,\, 1-\alpha)$ for $\alpha \in \mathbb{R}\setminus\{0\}$ is a global minimizer of the population risk for this model and represents the true relation $f(x) = \xi^\top x$. Now, we have the following theorem characterizing the gradient flow dynamics for minimizing $R(\theta, \xi)$.

Theorem 3.1 (Gradient flow is trapped on spheres)**.** *If $X \in L^2(\mathbb{P})$, then gradient flow for (2) satisfies the conservation law*

$$\|\theta_1(t)\|_2^2 + (\theta_2(t) - 1)^2 = \|\theta_1(0)\|_2^2 + (\theta_2(0) - 1)^2$$

for all $t \geq 0$ (as long as the gradient flow is well-defined).

Note that the conclusion of this theorem did not depend on the distribution of X beyond the existence of a second moment; hence, this conservation law holds for DEQs with a bias term as well. For the proof, we explicitly compute the gradient of the population risk, leading to a relation between the dynamics of $\theta_1(t)$ and $\theta_2(t)$.

Proof. We first differentiate the population risk with respect to θ, assuming $\theta_2 \neq 1$ so that the model is well-defined:

$$\nabla_\theta\, R(\theta, \xi) = \nabla_\theta\, \mathbb{E}\left[\left(\frac{\theta_1^\top X}{1 - \theta_2} - \xi^\top X\right)^2\right].$$

Since $X \in L^2(\mathbb{P})$ and $\theta_2 \neq 1$, the dominated convergence theorem justifies interchanging the gradient and expectation:

$$\begin{aligned}\nabla_\theta R(\theta,\xi) &= \mathbb{E}\left[\nabla_\theta \left(\frac{\theta_1^\top X}{1-\theta_2} - \xi^\top X\right)^2\right] \\ &= 2\,\mathbb{E}\left[\left(\frac{\theta_1^\top X}{1-\theta_2} - \xi^\top X\right) \nabla_\theta \left(\frac{\theta_1^\top X}{1-\theta_2} - \xi^\top X\right)\right] \\ &= 2\begin{pmatrix} \mathbb{E}[XX^\top]\left(\frac{\theta_1(t)}{(1-\theta_2(t))^2} - \frac{\xi}{1-\theta_2(t)}\right) \\ \frac{\theta_1(t)^\top}{(1-\theta_2(t))^2}\,\mathbb{E}[XX^\top]\left(\frac{\theta_1(t)}{1-\theta_2(t)} - \xi\right)\end{pmatrix}. \end{aligned} \tag{3}$$

Hence, the gradient flow dynamics are given by the ODE

$$\theta'(t) = -\nabla_\theta R(\theta(t),\xi) = -2\begin{pmatrix} \mathbb{E}[XX^\top]\left(\frac{\theta_1(t)}{(1-\theta_2(t))^2} - \frac{\xi}{1-\theta_2(t)}\right) \\ \frac{\theta_1(t)^\top}{(1-\theta_2(t))^2}\,\mathbb{E}[XX^\top]\left(\frac{\theta_1(t)}{1-\theta_2(t)} - \xi\right)\end{pmatrix}. \tag{4}$$

At this point, we may notice that

$$\frac{\theta_1(t)^\top \theta_1'(t)}{1-\theta_2(t)} = \theta_2'(t) \implies \theta_1(t)^\top \theta_1'(t) = (1-\theta_2(t))\,\theta_2'(t).$$

We integrate both sides with respect to t to obtain

$$\begin{aligned} &\int_0^t \theta_1(s)^\top \theta_1'(s)\,ds = \int_0^t (1-\theta_2(s))\,\theta_2'(s)\,ds \\ &\implies \sum_{i=1}^d \int_{(\theta_1)_i(0)}^{(\theta_1)_i(t)} (\theta_1)_i\,d(\theta_1)_i = \int_{\theta_2(0)}^{\theta_2(t)} (1-\theta_2)\,d\theta_2 \\ &\implies \sum_{i=1}^d \frac{(\theta_1)_i(t)^2 - (\theta_1)_i(0)^2}{2} = \theta_2(t) - \frac{1}{2}\theta_2(t)^2 - \left(\theta_2(0) - \frac{1}{2}\theta_2(0)^2\right) \\ &\implies \|\theta_1(t)\|_2^2 + (\theta_2(t)-1)^2 = \|\theta_1(0)\|_2^2 + (\theta_2(0)-1)^2. \end{aligned}$$

□

In particular, this result implies that $\theta(t)$ remains trapped on spheres centered at $(0,\ldots,0,1)$ in $\mathbb{R}^{d+1}$. In addition, we can guarantee that gradient flow converges to a global minimizer as long as the initialization is not on the hyperplane $\theta_2 = 1$. We begin with a lemma, which states that the gradient flow is well-defined for all time as long as the initialization is not on this hyperplane.

Lemma 3.2 (Gradient flow is well-defined)**.** *If $X \in L^2(\mathbb{P})$ has $\mathbb{E}[XX^\top] \succ 0$ and the initialization satisfies $\theta_2(0) \neq 1$, then the gradient flow for (2) is well-defined for all $t \geq 0$; in particular, $\theta_2(t) \neq 1$ for all $t \geq 0$.*

One may wonder whether the assumption that $\mathbb{E}[XX^\top] \succ 0$ is necessary for this result; indeed, simulations show that if $\mathbb{E}[XX^\top]$ is rank-deficient, then it is still possible for the gradient flow to

be well-defined for all time and converge. However, relaxing this assumption is nontrivial even for learning a single neuron with gradient flow (Vardi et al., 2021), so we leave this to future work.

The proof of this lemma relies on the simple observation that if $\theta_2(t)$ were to converge to 1 in finite time, then $\theta_1(t)$ would have to converge to a nonzero vector by Theorem 3.1, which would cause the population risk to diverge to infinity. However, since the population risk is non-increasing along the gradient flow, this leads to a contradiction.

Proof. We can rewrite the population risk as

$$R(\theta(t), \xi) = \mathbb{E}\left[\left(\frac{\theta_1(t)^\top X}{1-\theta_2(t)} - \xi^\top X\right)^2\right] = \left(\frac{\theta_1(t)}{1-\theta_2(t)} - \xi\right)^\top \mathbb{E}[XX^\top]\left(\frac{\theta_1}{1-\theta_2} - \xi\right). \tag{5}$$

If we were to have $\theta_2(t) \to 1$ as $t \uparrow T$ for some $T \in (0, \infty]$, we would have $\|\theta_1(t)\|_2^2 \to \|\theta_1(0)\|_2^2 + (\theta_2(0) - 1)^2 > 0$ by Theorem 3.1. However, this would imply that

$$\left\|\frac{\theta_1(t)}{1-\theta_2(t)} - \xi\right\|_2 \geq \frac{\|\theta_1(t)\|_2}{|1-\theta_2(t)|} - \|\xi\|_2 \to \infty$$

by the triangle inequality, and hence

$$R(\theta(t), \xi) = \left(\frac{\theta_1(t)}{1-\theta_2(t)} - \xi\right)^\top \mathbb{E}[XX^\top]\left(\frac{\theta_1(t)}{1-\theta_2(t)} - \xi\right) \geq \lambda_{\min}(\mathbb{E}[XX^\top])\left\|\frac{\theta_1(t)}{1-\theta_2(t)} - \xi\right\|_2^2 \to \infty.$$

But we know that

$$\frac{d}{dt}R(\theta(t), \xi) = \nabla_\theta R(\theta(t), \xi)^\top \theta'(t) = -\|\nabla_\theta R(\theta(t), \xi)\|_2^2 \leq 0,$$

so $R(\theta(t), \xi) \leq R(\theta(0), \xi)$, yielding a contradiction. □

Combining this lemma with Theorem 3.1, we obtain the following convergence result.

Theorem 3.3 (Gradient flow converges for linear DEQs)**.** *If $X \in L^2(\mathbb{P})$ has $\mathbb{E}[XX^\top] \succ 0$ and the initialization satisfies $\theta_2(0) \neq 1$, then the gradient flow for (2) is well-defined for all $t \geq 0$ and*

$$\lim_{t\to\infty} \theta(t) = \left(\frac{\operatorname{sign}(\theta_2(0))\,\|\theta(0)\|_2}{\sqrt{\|\xi\|_2^2 + 1}}\,\xi,\; 1 - \frac{\operatorname{sign}(\theta_2(0))\,\|\theta(0)\|_2}{\sqrt{\|\xi\|_2^2 + 1}}\right),$$

which is a global minimizer of the population risk and corresponds to the true relation $f(x) = \xi^\top x$.

This result follows from the observation that (by LaSalle's invariance principle and the fact that the flow is trapped on a hemisphere) the gradient flow must converge to a stationary point of the population risk.

Proof. Assume w.l.o.g. that $\theta_2(0) > 1$ and define the compact set $\Omega \coloneqq \{z \in \mathbb{R}^{d+1} : \|z\|_2^2 = \|\theta(0)\|_2^2,\, z_{d+1} > 0\}$, which we know by Theorem 3.1 and Lemma 3.2 is an invariant set under the dynamics. Define the function $V : \Omega \to \mathbb{R}$ by $V(\theta) \coloneqq R(\theta, \xi)$. From (3), we immediately see that $V \in C^1(\Omega)$:

$$\nabla_\theta V(\theta) = \nabla_\theta R(\theta, \xi) = 2 \begin{pmatrix} \mathbb{E}[XX^\top] \left(\frac{\theta_1(t)}{(1-\theta_2(t))^2} - \frac{\xi}{1-\theta_2(t)} \right) \\ \frac{\theta_1(t)^\top}{(1-\theta_2(t))^2} \mathbb{E}[XX^\top] \left(\frac{\theta_1(t)}{1-\theta_2(t)} - \xi \right) \end{pmatrix}.$$

Furthermore, V is nonincreasing along the flow because

$$\frac{d}{dt} V(\theta(t)) = \nabla_\theta R(\theta(t), \xi)^\top \theta'(t) = -\|\nabla_\theta R(\theta(t), \xi)\|_2^2 \le 0,$$

so V is a Lyapunov function for the gradient flow. Finally, the set of points where $\frac{d}{dt} V(\theta(t)) = 0$ is precisely the set of points $S = \{(\theta_1, \theta_2) \in \Omega : \theta_1 = \xi(1 - \theta_2)\}$ by (3). By LaSalle's invariance principle (Khalil and Grizzle, 2002, Theorem 4.4), we conclude that the gradient flow converges to a point in S as $t \to \infty$, which must be a global minimizer of the population risk since $\theta_1 = \xi(1-\theta_2)$. We can solve for this set by parameterizing $\theta = (\alpha\xi,\, 1 - \alpha)$ for $\alpha > 0$:

$$\|\theta\|_2^2 = \alpha^2 \|\xi\|_2^2 + \alpha^2 = \|\theta(0)\|_2^2 \implies \alpha = \frac{\|\theta(0)\|_2}{\sqrt{\|\xi\|_2^2 + 1}}.$$

Since the set S is a singleton, we deduce that the gradient flow must converge to this point as $t \to \infty$. The case $\theta_2(0) < 1$ follows by symmetry. □

Finally, we can use this result to show exponentially fast convergence by Grönwall's inequality, because convergence of the parameters implies that $|1 - \theta_2(t)|$ remains bounded away from zero.

Theorem 3.4 (Exponentially fast convergence for linear models)**.** *Define* $\varphi(t) \coloneqq \frac{\theta_1(t)}{1-\theta_2(t)}$. *Under the same assumptions as Theorem 3.3, we have*

$$\|\varphi(t) - \xi\|_2^2 \le \|\varphi(0) - \xi\|_2^2 \exp\left(-\frac{4}{\beta^2} \lambda_{\min}(\mathbb{E}[XX^\top])\, t \right)$$

for all $t \ge 0$, *where* $\beta > 0$ *is a constant (which exists under our previous assumptions) such that* $|1 - \theta_2(t)| \ge \beta$ *for all* $t \ge 0$.

Proof. Defining $r(t) \coloneqq \|\varphi(t) - \xi\|_2^2$, we compute

$$r'(t) = 2\,(\varphi(t) - \xi)^\top \varphi'(t). \tag{6}$$

We compute the derivative of $\varphi(t)$ by the chain rule:

$$\begin{aligned}\varphi'(t) &= \frac{\theta_1'(t)\,(1-\theta_2(t)) + \theta_1(t)\,\theta_2'(t)}{(1-\theta_2(t))^2} \\ &= \frac{-2\,\mathbb{E}[XX^\top]\left(\frac{\theta_1(t)}{(1-\theta_2(t))^2} - \frac{\xi}{1-\theta_2(t)}\right)(1-\theta_2(t)) - 2\,\theta_1(t)\,\frac{\theta_1(t)^\top}{(1-\theta_2(t))^2}\,\mathbb{E}[XX^\top]\left(\frac{\theta_1(t)}{1-\theta_2(t)} - \xi\right)}{(1-\theta_2(t))^2} \\ &= -\frac{2}{(1-\theta_2(t))^2}\,(I_d + \varphi(t)\varphi(t)^\top)\,\mathbb{E}[XX^\top]\,(\varphi(t)-\xi).\end{aligned}$$

Plugging this into (6), we obtain

$$\begin{aligned}r'(t) &= (\varphi(t)-\xi)^\top\left(-\frac{4}{(1-\theta_2(t))^2}\,(I_d + \varphi(t)\varphi(t)^\top)\,\mathbb{E}[XX^\top]\right)(\varphi(t)-\xi) \\ &\le -\lambda_{\min}\left(\frac{4}{(1-\theta_2(t))^2}\,(I_d + \varphi(t)\varphi(t)^\top)\,\mathbb{E}[XX^\top]\right)\|\varphi(t)-\xi\|_2^2 \\ &= -\lambda_{\min}\left(\frac{4}{(1-\theta_2(t))^2}\,(I_d + \varphi(t)\varphi(t)^\top)\,\mathbb{E}[XX^\top]\right) r(t).\end{aligned}$$

We know from Theorem 3.3 that $\theta(t) \to (\alpha\xi,\, 1-\alpha)$ as $t \to \infty$ for some $\alpha > 0$, so there exists a constant $\beta > 0$ such that $|1-\theta_2(t)| \ge \beta$ for all $t \ge 0$. In addition, since $\mathbb{E}[XX^\top] \succ 0$ and $I_d + \varphi(t)\varphi(t)^\top \succeq I_d$, we have

$$\lambda \coloneqq \lambda_{\min}\left(\frac{4}{(1-\theta_2(t))^2}\,(I_d + \varphi(t)\varphi(t)^\top)\,\mathbb{E}[XX^\top]\right) \ge \frac{4}{\beta^2}\,\lambda_{\min}(\mathbb{E}[XX^\top]) > 0.$$

At last, the result follows by Grönwall's inequality because

$$r'(t) \le -\lambda\, r(t) \implies r(t) \le r(0)\, e^{-\lambda t}. \qquad \square$$

3.1.2 Gradient descent

We can also analyze the gradient descent dynamics for fitting the linear deep equilibrium model (2). First, we recall the well-known descent lemma, which is a consequence of Theorem 2.1.5 of Nesterov (2018).

Lemma 3.5 (Descent lemma)**.** *If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable with $\|\nabla f\|_{\mathrm{Lip}} = L < \infty$, then each step of gradient descent $x(t+1) = x(t) - \eta\,\nabla f(x(t))$ with step size $0 < \eta \le 1/L$ satisfies*

$$f(x(t+1)) \le f(x(t)) - \frac{\eta}{2}\,\|\nabla f(x(t))\|_2^2.$$

In our setting, we now have the following theorem.

Theorem 3.6 (Gradient descent converges for linear DEQs)**.** *Under the assumptions of Theorem 3.3, let*

$$\kappa \coloneqq \frac{\lambda_{\max}(\mathbb{E}[XX^\top])}{\lambda_{\min}(\mathbb{E}[XX^\top])}$$

denote the condition number of the uncentered covariance matrix of X. Letting $\varphi(t) \coloneqq \frac{\theta_1(t)}{1-\theta_2(t)}$, there exists $\eta_0 > 0$ such that for all step sizes $0 < \eta \leq \eta_0$, gradient descent converges (almost) linearly:

$$\|\varphi(t) - \xi\|_2^2 \leq \kappa \left(1 - \frac{\eta}{4\kappa \|\theta(0) - (0,1)\|_2^2}\right)^t \|\varphi(0) - \xi\|_2^2$$

for all $t \in \mathbb{N}_0$.

The proof proceeds by first proving that the gradient descent iterates remain trapped in a compact set Θ for all time. Then, we prove that the risk satisfies a Polyak-Łojasiewicz (PŁ) inequality within Θ and thereby conclude the convergence of gradient descent. Hence, we start with the following lemma.

Lemma 3.7 (Invariant set for gradient descent in the linear model)**.** *Let $\theta(t)$ denote the gradient descent iterates initialized at $\theta(0)$ for (2) and define*

$$c_0 \coloneqq \sqrt{\frac{R(\theta(0), \xi)}{\lambda_{\min}(\mathbb{E}[XX^\top])}}, \qquad \beta \coloneqq \min\left\{|1 - \theta_2(0)|, \frac{\sqrt{w(0)}}{\sqrt{2}\,(c_0 + \|\xi\|_2 + 1)}\right\},$$

where

$$w(t) \coloneqq \left\|\theta(t) - \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right\|_2^2.$$

Then, there exists $\eta_0 > 0$ such that $\eta \leq \eta_0$ implies that

$$\Theta \coloneqq \{\theta \in \mathbb{R}^{d+1} : \beta \leq |1 - \theta_2| \leq 2\sqrt{w(0)},\ \|\theta_1\|_2 \leq 2\sqrt{w(0)}\}$$

contains $\theta(t)$ for all $t \in \mathbb{N}_0$.

To prove this lemma, we proceed by induction. First, we choose the step size small enough so that the gradient descent iterates $(\theta(0), \ldots, \theta(t+1))$ remain in $\Theta + \overline{B(0, \beta/2)}$, and then note that the gradient of the risk is Lipschitz over this set. Using the descent lemma, we obtain a bound on the sum of squared gradients over the first t iterations, as the risk cannot decrease below zero. Using this bound, we then show that $w(t+1)$ remains close to $w(0)$ up to an error of order η^2, which is an approximate version of the conservation law from Theorem 3.1. This then implies that $\theta(t+1) \in \Theta$, completing the induction.

Proof. We will prove the claim by induction on t. The base case $t = 0$ is trivial, so assume that $\theta(s) \in \Theta$ for all $0 \le s \le t$. It is clear that the map $(\theta \mapsto \nabla_\theta R(\theta, \xi)) \in C^1(\Theta + \overline{B(0, \beta/2)})$ has a finite Lipschitz norm $\|(\nabla_\theta R(\theta, \xi))|_{\Theta + \overline{B(0,\beta/2)}}\|_{\text{Lip}} = L_\Theta < \infty$ by compactness of $\Theta + \overline{B(0, \beta/2)}$. We will prove the claim by induction on t. The base case $t = 0$ follows easily from definitions, so assume that $\theta(s) \in \Theta$ for all $0 \le s \le t$. By the descent lemma (Lemma 3.5), we have

$$R(\theta(s+1), \xi) \le R(\theta(s), \xi) - \frac{\eta}{2} \|\nabla_\theta R(\theta(s), \xi)\|_2^2$$

whenever $0 < \eta \le \min\{1/L_\Theta, \beta/2\}$ (because $\theta(0), \dots, \theta(t+1) \in \overline{B(0, \beta/2)}$). Adding these inequalities yields a telescoping sum which implies the bound

$$0 \le R(\theta(t+1), \xi) \le R(\theta(0), \xi) - \frac{\eta}{2} \sum_{s=0}^{t} \|\nabla_\theta R(\theta(s), \xi)\|_2^2,$$

and rearranging shows that

$$\sum_{s=0}^{t} \|\nabla_\theta R(\theta(s), \xi)\|_2^2 \le \frac{2R(\theta(0), \xi)}{\eta}. \tag{7}$$

However, we also know that

$$\begin{aligned} w(s+1) = \left\| \theta(s+1) - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\|_2^2 &= \left\| \theta(s) - \eta \nabla_\theta R(\theta(s), \xi) - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\|_2^2 \\ &= w(s) - 2\eta \left(\theta(s) - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right)^\top \nabla_\theta R(\theta(s), \xi) + \eta^2 \|\nabla_\theta R(\theta(s), \xi)\|_2^2. \end{aligned}$$

The middle term vanishes because

$$\left(\theta(s) - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right)^\top \nabla_\theta R(\theta(s), \xi) = 2 \begin{pmatrix} \theta_1(s) \\ \theta_2(s) - 1 \end{pmatrix}^\top \begin{pmatrix} \mathbb{E}[XX^\top] \left(\frac{\theta_1(s)}{(1-\theta_2(s))^2} - \frac{\xi}{1-\theta_2(s)} \right) \\ \frac{\theta_1(s)^\top}{(1-\theta_2(s))^2} \mathbb{E}[XX^\top] \left(\frac{\theta_1(s)}{1-\theta_2(s)} - \xi \right) \end{pmatrix} = 0,$$

and we are left with

$$w(s+1) = w(s) + \eta^2 \|\nabla_\theta R(\theta(s), \xi)\|_2^2. \tag{8}$$

Roughly, this means that $w(t)$ is approximately conserved up to an error of order η^2. Unrolling this recursion and using (7), we obtain

$$w(t+1) = w(0) + \eta^2 \sum_{s=0}^{t} \|\nabla_\theta R(\theta(s), \xi)\|_2^2 \le w(0) + 2\eta R(\theta(0), \xi).$$

Now, suppose $\eta \le \beta^2/(4R(\theta(0), \xi))$; then, we have

$$w(t+1) \le w(0) + \frac{\beta^2}{2} \le 2w(0)$$

because $\beta \le \sqrt{w(0)}$. This shows that $\max\{\|\theta_1(t+1)\|_2, |1-\theta_2(t+1)|\} \le \sqrt{w(t+1)} \le 2\sqrt{w(0)}$. Finally, suppose for a contradiction that $|1-\theta_2(t+1)| < \beta$. In this case, we would have

$$\|\theta_1(t+1)\|_2^2 = w(t+1) - (\theta_2(t+1)-1)^2 > w(0) + \frac{\beta^2}{2} - \beta^2 \ge \frac{w(0)}{2}$$

and therefore

$$\|\varphi(t+1)\|_2 = \frac{\|\theta_1(t+1)\|_2}{|1-\theta_2(t+1)|} > \frac{\sqrt{w(0)}}{\sqrt{2}\,\beta} \ge c_0 + \|\xi\|_2 + 1.$$

By the triangle inequality,

$$\|\varphi(t+1) - \xi\|_2 \ge \|\varphi(t+1)\|_2 - \|\xi\|_2 > c_0 + 1.$$

But now this means that

$$\begin{aligned} R(\theta(t+1),\xi) &= (\varphi(t+1)-\xi)^\top \mathbb{E}[XX^\top]\,(\varphi(t+1)-\xi) \\ &\ge \lambda_{\min}(\mathbb{E}[XX^\top])\,\|\varphi(t+1)-\xi\|_2^2 \\ &> \lambda_{\min}(\mathbb{E}[XX^\top])\,(c_0+1)^2 \\ &> R(\theta(0),\xi), \end{aligned}$$

contradicting the descent lemma (Lemma 3.5). □

Using this lemma, we are now ready to prove Theorem 3.6 by proving a PŁ inequality for the risk within the invariant set Θ.

Proof of Theorem 3.6. First, choose $\eta \le \eta_0$ as in Lemma 3.7 so that the gradient descent iterates remain in the compact set Θ for all time. As a consequence, $R(\theta,\xi)$ and $\nabla_\theta R(\theta,\xi)$ are Lipschitz on Θ. Because there is an upper bound $|1-\theta_2(t)| \le \alpha \coloneqq 2\,\|\theta(0)-(0,1)\|_2$, notice that

$$\begin{aligned} \|\nabla_\theta R(\theta(t),\xi)\|_2^2 &= 2\left\| \frac{1}{1-\theta_2(t)} \begin{pmatrix} \mathbb{E}[XX^\top]\,(\varphi(t)-\xi) \\ \varphi(t)^\top \mathbb{E}[XX^\top]\,(\varphi(t)-\xi) \end{pmatrix} \right\|_2^2 \\ &= 2\left(\left\| \frac{1}{1-\theta_2(t)} \mathbb{E}[XX^\top]\,(\varphi(t)-\xi) \right\|_2^2 + \left\| \frac{\varphi(t)^\top}{1-\theta_2(t)} \mathbb{E}[XX^\top]\,(\varphi(t)-\xi) \right\|_2^2 \right) \\ &\ge \frac{2}{\alpha^2}\,\lambda_{\min}(\mathbb{E}[XX^\top])\,\|\varphi(t)-\xi\|_2^2. \end{aligned}$$

From (5), we also have the bounds

$$\lambda_{\min}(\mathbb{E}[XX^\top])\,\|\varphi(t)-\xi\|_2^2 \le R(\theta(t),\xi) \le \lambda_{\max}(\mathbb{E}[XX^\top])\,\|\varphi(t)-\xi\|_2^2, \tag{9}$$

which taken with our first inequality implies a PŁ inequality for $R(\theta(t), \xi)$ because

$$\|\nabla_\theta R(\theta(t), \xi)\|_2^2 \geq \frac{2}{\alpha^2} \lambda_{\min}(\mathbb{E}[XX^\top]) \|\varphi(t) - \xi\|_2^2 \geq \frac{2}{\kappa\alpha^2} R(\theta(t), \xi).$$

Finally, applying the descent lemma (Lemma 3.5) and applying the PŁ inequality yields

$$\begin{aligned} R(\theta(t+1), \xi) &\leq R(\theta(t), \xi) - \frac{\eta}{2} \|\nabla_\theta R(\theta(t), \xi)\|_2^2 \\ &= \left(1 - \frac{\eta}{\kappa\alpha^2}\right) R(\theta(t), \xi). \end{aligned}$$

Induction on t then yields

$$R(\theta(t), \xi) \leq \left(1 - \frac{\eta}{\kappa\alpha^2}\right)^t R(\theta(0), \xi),$$

and (9) gives the final result. □

Next, we extend our convergence results to nonlinear single-index models.

3.2 Nonlinear single-index models

Next, we consider the function $f(x) = \sigma(\xi^\top x)$ for some fixed nonzero $\xi \in \mathbb{R}^d$ and nonlinear activation function $\sigma : \mathbb{R} \to \mathbb{R}$.

3.2.1 Gradient flow

In this section, we analyze the gradient flow dynamics for fitting this function using a single implicit neuron

$$y_\theta(x) = g_\theta(x, y) \coloneqq \sigma(\theta_1^\top x + \theta_2\, y_\theta(x)). \tag{10}$$

Note that for $y_\theta(x)$ to be well-defined, it suffices to have σ differentiable with $\|\sigma\|_{\text{Lip}} = L$ and $|\theta_2| \leq \delta_2 < 1/L$. This way, we have the guarantee

$$\left|\frac{\partial g_\theta(x, y)}{\partial y}\right| = |\theta_2|\, |\sigma'(\theta_1^\top x + \theta_2\, y)| \leq L\delta_2 < 1,$$

meaning that $g_\theta(x, \cdot)$ is a global contraction mapping for each $x \in \mathbb{R}^d$ and the Banach fixed-point theorem implies that there exists a unique fixed-point $y_\theta(x)$ satisfying the implicit equation above. We assume that the input data X is an $\mathbb{R}^d$-valued random variable and use the squared loss $\ell(y, \hat{y}) = (\hat{y} - y)^2$; our goal is now to minimize the population risk

$$R(\theta, \xi) \coloneqq \mathbb{E}[(y_\theta(X) - f(X))^2] = \mathbb{E}[(y_\theta(X) - \sigma(\xi^\top X))^2].$$

Notice that $(\xi, 0)$ is a global minimizer of the population risk for this model because it is well-specified. Now, we have the following theorem characterizing the gradient flow dynamics for minimizing $R(\theta, \xi)$. In particular, for many activation functions and data distributions, we show that gradient flow initialized close enough to the target parameter $(\xi, 0)$ converges exponentially fast to this target.

Theorem 3.8 (Gradient flow converges for single-index DEQs)**.** *Fix constants $\delta_1 > 0$ and $\delta_2 \in (0,\, 1/L)$ and assume that*

1. *The data satisfies:*
   - *(i) $X \in L^2(\mathbb{P})$.*
   - *(ii) $\mathbb{E}[\|X\|_2^2\, \mathbf{1}_{\|X\|_2 \le \alpha}] \ge \beta$ for some constants $\alpha, \beta \in (0, \infty)$.*
2. *The activation function satisfies:*
   - *(i) σ is differentiable and monotonically increasing with $\|\sigma\|_{\mathrm{Lip}} = L < \infty$.*
   - *(ii) $\inf\{\sigma'(z) : |z| \le \alpha(1+\delta_1)\|\xi\|_2 + M\delta_2\} \ge \gamma > 0$ for some constant $\gamma > 0$, where*

$$M \coloneqq \frac{|\sigma(0)| + L\alpha(1+\delta_1)\|\xi\|_2}{1 - L\delta_2}.$$

3. *The activation function is nonlinear with respect to the data, in the sense that there do not exist $\kappa_1, \kappa_2 \in \mathbb{R}^d$ such that $\sigma(\kappa_1^\top X)\, \mathbf{1}_{\|X\|_2 \le \alpha} = \kappa_2^\top X\, \mathbf{1}_{\|X\|_2 \le \alpha}$ almost surely.*

Define the set

$$\Theta \coloneqq \{\theta \in \mathbb{R}^{d+1} : \|\theta_1\|_2 \le (1+\delta_1)\|\xi\|_2,\, |\theta_2| \le \delta_2 < 1/L\}.$$

Then, as long as the initialization satisfies $\|\theta(0) - (\xi, 0)\|_2 \le \min\{(1+\delta_1)\|\xi\|_2,\, \delta_2\}$, the gradient flow for (10) is well-defined and converges exponentially fast to the target parameter $(\xi, 0)$:

$$\|\theta(t) - (\xi,0)\|_2^2 \le \|\theta(0) - (\xi,0)\|_2^2 \exp\left(-\frac{2\rho\gamma^2}{1 + L\delta_2}\, t\right)$$

for all $t \ge 0$, where

$$\rho \coloneqq \inf_{\theta \in \Theta} \inf_{u \in S^d} \mathbb{E}\left[\left(u^\top \begin{pmatrix} X \\ y_\theta(X) \end{pmatrix}\right)^2 \mathbf{1}_{\|X\|_2 \le \alpha}\right] > 0$$

is a positive constant under the previous assumptions (quantifying the nonlinearity of σ).

Remark 3.1 (Discussion of assumptions). The assumptions on the data distribution are quite mild; for example, if $X \in L^2(\mathbb{P})$ is not almost surely equal to zero, then we can always find constants α, β satisfying the assumptions. The assumptions on the activation function rule out functions such as

the linear activation $\sigma(z) = z$ (due to the nonlinearity condition), the step function $\sigma(z) = \mathbf{1}_{z \geq 0}$ (due to differentiability), and the ReLU activation $\sigma(z) = \max\{0, z\}$ (due to differentiability and the lower bound on the derivative near zero).

However, note that Theorem 3.4 already characterizes the gradient flow dynamics for the linear activation, so this case is already covered in our analysis, and a modification of our proof technique can also handle the ReLU activation if we assume that the initialization satisfies $\theta_1(0)^\top \xi > 0$. Even so, other common activations such as the sigmoid $\sigma(z) = 1/(1+e^{-z})$, the hyperbolic tangent $\sigma(z) = \tanh(z)$, and the softplus $\sigma(z) = \log(1 + e^z)$ satisfy all of our assumptions for *any* initialization in $B((\xi, 0), 1/L)$ and random variable $X \in L^2(\mathbb{P})$ which witnesses the nonlinearity of σ (e.g., X can be any continuous random variable).

Proof. Our proof technique is inspired by Yehudai and Shamir (2020), and the goal is to apply Grönwall's inequality to $r(t) \coloneqq \|\theta(t) - (\xi, 0)\|_2^2$. Throughout the proof, we will use the notation $\omega_t(z) \coloneqq \theta_1(t)^\top z + \theta_2(t)\, y_{\theta(t)}(z)$. First, we compute

$$r'(t) = 2 \left(\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right)^\top \theta'(t) = -2 \left(\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right)^\top \nabla_\theta R(\theta(t), \xi). \tag{11}$$

To compute the gradient, we note that

$$\nabla_\theta R(\theta(t), \xi) = \nabla_\theta \mathbb{E}[(y_{\theta(t)}(X) - \sigma(\xi^\top X))^2]. \tag{12}$$

We will assume throughout the proof that $\theta(t) \in \Theta$ for all $t \geq 0$. Of course, this holds at $t = 0$ because

$$\max\{\|\theta_1(0) - \xi\|_2,\, |\theta_2(0)|\} \leq \sqrt{r(0)} = \left\|\theta(0) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right\|_2 \leq \min\{(1 + \delta_1)\|\xi\|_2,\, \delta_2\},$$

and it will hold for all $t \geq 0$ as long as we can show that $r(t)$ is nonincreasing; we will verify this at the end of the proof. Next, we will justify the interchange of the gradient with the expectation in (12). From Lipschitz continuity of σ, we have

$$\begin{aligned} |y_{\theta(t)}(X) - \sigma(\xi^\top X)| &= |\sigma(\theta_1(t)^\top X + \theta_2(t)\, y_{\theta(t)}(X)) - \sigma(\xi^\top X)| \\ &\leq L\, |(\theta_1(t) - \xi)^\top X + \theta_2(t)\, y_{\theta(t)}(X)|, \end{aligned}$$

and we obtain the bound

$$\begin{aligned} &\mathbb{E}[(y_{\theta(t)}(X) - \sigma(\xi^\top X))^2] \\ &\quad \leq L^2\, \mathbb{E}\left[\Big((\theta_1(t) - \xi)^\top X + \theta_2(t)\, y_{\theta(t)}(X)\Big)^2\right] \\ &\quad = L^2\, \mathbb{E}\left[\|\theta_1(t) - \xi\|_2^2\, \|X\|_2^2 + \theta_2(t)^2\, y_{\theta(t)}(X)^2 + 2\, \theta_2(t)\, (\theta_1(t) - \xi)^\top X\, y_{\theta(t)}(X)\right]. \end{aligned}$$

For this to be bounded, since we assumed $X \in L^2(\mathbb{P})$ it suffices by the Cauchy-Schwarz inequality to show that $y_\theta(X) \in L^2(\mathbb{P})$ in some neighborhood of $\theta(t)$. By the triangle inequality and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|y_{\theta(t)}(X)| - |\sigma(0)| &\leq |y_{\theta(t)}(X) - \sigma(0)| \\
&= |\sigma(\theta_1(t)^\top X + \theta_2(t)\, y_{\theta(t)}(X)) - \sigma(0)| \\
&\leq L\,|\theta_1(t)^\top X + \theta_2(t)\, y_{\theta(t)}(X)| \\
&\leq L(|\theta_1(t)^\top X| + |\theta_2(t)|\,|y_{\theta(t)}(X)|) \\
&\leq L(\|\theta_1(t)\|_2\,\|X\|_2 + |\theta_2(t)|\,|y_{\theta(t)}(X)|).
\end{aligned}$$

and rearranging yields

$$|y_{\theta(t)}(X)| \leq \frac{|\sigma(0)| + L\,\|\theta_1(t)\|_2\,\|X\|_2}{1 - L\,|\theta_2(t)|}. \tag{13}$$

Since the right-hand side is continuous in θ in some neighborhood of $\theta(t)$ and $X \in L^2(\mathbb{P})$, the right-hand side is bounded in some neighborhood of $\theta(t)$ and hence $y_\theta(X) \in L^2(\mathbb{P})$ for all θ in this neighborhood. This justifies the interchange of the gradient and expectation in (12), so we get

$$\nabla_\theta\, R(\theta(t), \xi) = \mathbb{E}[\nabla_\theta\,(y_{\theta(t)}(X) - \sigma(\xi^\top X))^2] = 2\,\mathbb{E}[(y_{\theta(t)}(X) - \sigma(\xi^\top X))\,\nabla_\theta\, y_{\theta(t)}(X)]. \tag{14}$$

We may compute the gradient of $y_{\theta(t)}(x)$ by the implicit function theorem, since $\theta(t) \in \Theta$ implies that

$$1 - \frac{\partial g_{\theta(t)}(x, y)}{\partial y} = 1 - \theta_2(t)\,\sigma'(\omega_t(x)) \geq 1 - L\delta_2 > 0,$$

giving the formula

$$\begin{aligned}
&\nabla_\theta\, y_\theta(x) = \nabla_\theta\,\sigma(\theta_1^\top x + \theta_2\, y_\theta(x)) = \sigma'(\theta_1^\top x + \theta_2\, y_\theta(x))\,\left(\begin{pmatrix} x \\ y_\theta(x) \end{pmatrix} + \theta_2\,\nabla_\theta\, y_\theta(x)\right) \\
&\implies (1 - \theta_2\,\sigma'(\theta_1^\top x + \theta_2\, y_\theta(x)))\,\nabla_\theta\, y_\theta(x) = \sigma'(\theta_1^\top x + \theta_2\, y_\theta(x))\,\begin{pmatrix} x \\ y_\theta(x) \end{pmatrix} \\
&\implies \nabla_\theta\, y_\theta(x) = \frac{\sigma'(\theta_1^\top x + \theta_2\, y_\theta(x))}{1 - \theta_2\,\sigma'(\theta_1^\top x + \theta_2\, y_\theta(x))}\,\begin{pmatrix} x \\ y_\theta(x) \end{pmatrix}.
\end{aligned}$$

Plugging this into (14) and (11), we obtain

$$\begin{aligned}
r'(t) &= -2\left(\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}^\top (\nabla_\theta\, R(\theta(t), \xi))\right) \\
&= -2\,\mathbb{E}\left[(y_{\theta(t)}(X) - \sigma(\xi^\top X))\,\frac{\sigma'(\omega_t(X))}{1 - \theta_2(t)\,\sigma'(\omega_t(X))}\,(\omega_t(X) - \xi^\top X)\right].
\end{aligned} \tag{15}$$

At this point, we notice that

$$\frac{\sigma'(\omega_t(X))}{1-\theta_2(t)\,\sigma'(\omega_t(X))} \geq \frac{\sigma'(\omega_t(X))}{1+L\delta_2} \geq 0$$

and

$$\begin{aligned}(y_{\theta(t)}(X) - \sigma(\xi^\top X)) \begin{pmatrix} X \\ y_{\theta(t)}(X) \end{pmatrix}^\top \left(\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right) \\ = (\sigma(\omega_t(X)) - \sigma(\xi^\top X)) \left(\omega_t(X) - \xi^\top X\right) \geq 0\end{aligned}$$

by monotonicity of σ. Hence, we may restrict the expectation in (15) with an inequality as

$$\begin{aligned} r'(t) &\leq -2\,\mathbb{E}\left[(y_{\theta(t)}(X) - \sigma(\xi^\top X))\,\frac{\sigma'(\omega_t(X))}{1-\theta_2(t)\,\sigma'(\omega_t(X))}\,(\omega_t(X) - \xi^\top X)\right] \\ &\leq -2\,\mathbb{E}\left[(y_{\theta(t)}(X) - \sigma(\xi^\top X))\,\frac{\sigma'(\omega_t(X))}{1-\theta_2(t)\,\sigma'(\omega_t(X))}\,(\omega_t(X) - \xi^\top X)\,\mathbf{1}_{\|X\|_2\leq\alpha}\right]. \end{aligned} \tag{16}$$

By the mean value theorem, we know that there exists some ζ between $\omega_t(X)$ and $\xi^\top X$ such that

$$\sigma(\omega_t(X)) - \sigma(\xi^\top X) = \sigma'(\zeta)\,(\omega_t(X) - \xi^\top X).$$

When $\|X\|_2 \leq \alpha$, we have from (13) that

$$|y_{\theta(t)}(X)| \leq \frac{|\sigma(0)| + L\,\|\theta_1(t)\|_2\,\|X\|_2}{1-L\,|\theta_2(t)|} \leq \frac{|\sigma(0)| + L\alpha(1+\delta_1)\|\xi\|_2}{1-L\delta_2} = M.$$

As a result, the Cauchy-Schwarz inequality implies that

$$|\omega_t(X)| \leq \|\theta_1(t)\|_2\,\|X\|_2 + |\theta_2(t)|\,|y_{\theta(t)}(X)| \leq \alpha(1+\delta_1)\|\xi\|_2 + M\delta_2$$

and

$$|\xi^\top X| \leq \|\xi\|_2\,\|X\|_2 \leq \alpha\|\xi\|_2 \leq \alpha(1+\delta_1)\|\xi\|_2 + M\delta_2.$$

Thus, by our assumption on the activation function and the fact that $|\zeta| \leq \alpha(1+\delta_1)\|\xi\|_2 + M\delta_2$, we have $\sigma'(\zeta) \geq \gamma > 0$, meaning that

$$|\sigma(\omega_t(X)) - \sigma(\xi^\top X)| \geq \gamma\,|\omega_t(X) - \xi^\top X|$$

and similarly

$$\frac{\sigma'(\omega_t(X))}{1-\theta_2(t)\,\sigma'(\omega_t(X))} \geq \frac{\gamma}{1+L\delta_2}.$$

Putting the pieces together, we get the following bound on (16):

$$\begin{aligned} r'(t) &\leq -2\,\mathbb{E}\left[(y_{\theta(t)}(X) - \sigma(\xi^\top X))\,\frac{\sigma'(\omega_t(X))}{1-\theta_2(t)\,\sigma'(\omega_t(X))}\,(\omega_t(X) - \xi^\top X)\,\mathbf{1}_{\|X\|_2\leq\alpha}\right] \\ &\leq -\frac{2\gamma^2}{1+L\delta_2}\,\mathbb{E}\left[(\omega_t(X) - \xi^\top X)^2\,\mathbf{1}_{\|X\|_2\leq\alpha}\right] \\ &= -\frac{2\gamma^2}{1+L\delta_2}\,\mathbb{E}\left[\Big((\theta_1(t)-\xi)^\top X + \theta_2(t)\,y_{\theta(t)}(X)\Big)^2\,\mathbf{1}_{\|X\|_2\leq\alpha}\right]. \end{aligned}$$

We rewrite this as a quadratic form:

$$\begin{aligned} r'(t) &\leq -\frac{2\gamma^2}{1+L\delta_2}\,\mathbb{E}\left[\Big((\theta_1(t)-\xi)^\top X + \theta_2(t)\,y_{\theta(t)}(X)\Big)^2\,\mathbf{1}_{\|X\|_2\leq\alpha}\right] \\ &\leq -\frac{2\gamma^2}{1+L\delta_2}\,\mathbb{E}\left[\begin{pmatrix}\theta_1(t)-\xi\\ \theta_2(t)\end{pmatrix}^\top \begin{pmatrix} XX^\top & y_{\theta(t)}(X)\,X \\ y_{\theta(t)}(X)\,X^\top & y_{\theta(t)}(X)^2\end{pmatrix}\begin{pmatrix}\theta_1(t)-\xi\\ \theta_2(t)\end{pmatrix}\mathbf{1}_{\|X\|_2\leq\alpha}\right] \\ &\leq -\frac{2\gamma^2}{1+L\delta_2}\,\lambda_{\min}\left(\mathbb{E}\left[\begin{pmatrix} XX^\top & y_{\theta(t)}(X)\,X \\ y_{\theta(t)}(X)\,X^\top & y_{\theta(t)}(X)^2\end{pmatrix}\mathbf{1}_{\|X\|_2\leq\alpha}\right]\right)\left\|\theta(t) - \begin{pmatrix}\xi\\ 0\end{pmatrix}\right\|_2^2. \end{aligned} \tag{17}$$

At this point, we may notice that

$$\lambda_{\min}\left(\mathbb{E}\left[\begin{pmatrix} XX^\top & y_{\theta(t)}(X)\,X \\ y_{\theta(t)}(X)\,X^\top & y_{\theta(t)}(X)^2\end{pmatrix}\mathbf{1}_{\|X\|_2\leq\alpha}\right]\right) = \inf_{u\in S^d}\,\mathbb{E}\left[\left(u^\top\begin{pmatrix}X\\ y_{\theta(t)}(X)\end{pmatrix}\right)^2\mathbf{1}_{\|X\|_2\leq\alpha}\right].$$

The integrand is uniformly dominated for $u \in S^d$ and $\theta(t) \in \Theta$ by Cauchy-Schwarz since

$$\left\|\begin{pmatrix}X\\ y_{\theta(t)}(X)\end{pmatrix}\right\|_2^2\mathbf{1}_{\|X\|_2\leq\alpha} \leq \alpha^2 + M^2.$$

Hence, the dominated convergence theorem implies that the expectation is continuous in u and the extreme value theorem implies that the infimum is attained at some $u_{\theta(t)} \in S^d$:

$$\inf_{u\in S^d}\,\mathbb{E}\left[\left(u^\top\begin{pmatrix}X\\ y_{\theta(t)}(X)\end{pmatrix}\right)^2\mathbf{1}_{\|X\|_2\leq\alpha}\right] = \mathbb{E}\left[\left(u_{\theta(t)}^\top\begin{pmatrix}X\\ y_{\theta(t)}(X)\end{pmatrix}\right)^2\mathbf{1}_{\|X\|_2\leq\alpha}\right].$$

Since $X\,\mathbf{1}_{\|X\|_2\leq\alpha}$ is not almost surely zero by our assumption on the data distribution, this expectation is strictly positive unless $y_{\theta(t)}(X)\,\mathbf{1}_{\|X\|_2\leq\alpha}$ is almost surely equal to some linear function of $X\,\mathbf{1}_{\|X\|_2\leq\alpha}$. Supposing for a contradiction that $y_{\theta(t)}(X)\,\mathbf{1}_{\|X\|_2\leq\alpha} = \kappa_2^\top X\,\mathbf{1}_{\|X\|_2\leq\alpha}$ almost surely for some $\kappa_2 \in \mathbb{R}^d$, we would have

$$\begin{aligned} \kappa_2^\top X\,\mathbf{1}_{\|X\|_2\leq\alpha} &= y_{\theta(t)}(X)\,\mathbf{1}_{\|X\|_2\leq\alpha} \\ &= \sigma(\theta_1(t)^\top X + \theta_2(t)\,y_{\theta(t)}(X))\,\mathbf{1}_{\|X\|_2\leq\alpha} \\ &= \sigma((\theta_1(t) + \theta_2(t)\,\kappa_2)^\top X)\,\mathbf{1}_{\|X\|_2\leq\alpha}, \end{aligned}$$

contradicting our nonlinearity assumption on σ if we pick $\kappa_1 = \theta_1(t) + \theta_2(t)\,\kappa_2$. Hence, the expectation is strictly positive for any fixed $\theta(t) \in \Theta$. The dominated convergence theorem again implies

that the expectation is continuous in θ over the compact set Θ, so the extreme value theorem implies that there exists some constant $\rho > 0$ such that

$$\inf_{\theta \in \Theta} \lambda_{\min} \left(\mathbb{E} \left[\begin{pmatrix} XX^\top & y_\theta(X)\, X \\ y_\theta(X)\, X^\top & y_\theta(X)^2 \end{pmatrix} \mathbf{1}_{\|X\|_2 \le \alpha} \right] \right) \ge \rho.$$

Thus, combining this with (17), we obtain

$$r'(t) \le -\frac{2\rho\gamma^2}{1 + L\delta_2}\, r(t).$$

Finally, by Grönwall's inequality, we have

$$r(t) \le r(0)\, \exp\left(-\frac{2\rho\gamma^2}{1 + L\delta_2}\, t \right),$$

which shows that $r(t)$ is nonincreasing and hence $\theta(t) \in \Theta$ for all $t \ge 0$, completing the proof. □

An immediate corollary of the above theorem is that gradient flow converges to zero risk exponentially fast as long as we initialize $\theta(0) \in B((\xi, 0), 1/L)$, for many common activation functions.

Corollary 3.1. *Suppose that the activation function $\sigma \in C^1(\mathbb{R})$ is strictly increasing with $\|\sigma\|_{\mathrm{Lip}} = L < \infty$. Furthermore, suppose that σ is not linear in any neighborhood of zero. Then, as long as X is a continuous random variable which is not almost surely equal to zero and we initialize $\theta(0) \in B((\xi, 0), 1/L)$, the gradient flow for (10) converges exponentially fast to the target parameter $(\xi, 0)$:*

$$\|\theta(t) - (\xi, 0)\|_2^2 \le \|\theta(0) - (\xi, 0)\|_2^2\, e^{-\lambda t}$$

for some constant $\lambda > 0$ depending on the data distribution and activation function.

Proof. Because X is not almost surely equal to zero, we can find constants $\alpha > 0$ and $\beta > 0$ such that $\mathbb{E}[\|X\|_2^2\, \mathbf{1}_{\|X\|_2 \le \alpha}] \ge \beta$. Then, we may choose $\delta_1 > \max\left\{0,\, \frac{\|\theta(0)\|_2}{\|\xi\|_2} - 1\right\}$ and $\delta_2 = \|\theta(0) - (\xi, 0)\|_2 < 1/L$. Since σ is continuously differentiable, we can find some $\gamma > 0$ such that

$$\inf\{\sigma'(z) : |z| \le \alpha(1 + \delta_1)\|\xi\|_2 + M\delta_2\} \ge \gamma > 0$$

by the extreme value theorem. The result then immediately follows from Theorem 3.8, because σ is not linear in any neighborhood of zero. □

For instance, the above corollary holds for the sigmoid, hyperbolic tangent, and softplus activation functions.

3.2.2 Gradient descent

By a similar analysis, we conclude that gradient descent with sufficiently small step size also converges to the target parameter exponentially fast when initialized close to the target. However, this approach requires a stronger moment assumption on the data distribution in order to ensure that the squared norm of the gradient is not too large. Instead of explicitly proving a PŁ inequality in this setting, we directly control the $\|\nabla_\theta R(\theta(t), \xi)\|_2^2$ by the squared distance to the target parameter.

Theorem 3.9 (Gradient descent converges for single-index DEQs)**.** *Suppose that the assumptions of Theorem 3.8 hold with $X \in L^4(\mathbb{P})$. Define*

$$\lambda_1 \coloneqq \frac{2\rho\gamma^2}{1 + L\delta_2}, \qquad \lambda_2 \coloneqq 4L^2 \left(\frac{L}{1 - L\delta_2}\right)^2 \sup_{\theta \in \Theta} \mathbb{E}\left[\left\|\begin{pmatrix} X \\ y_\theta(X) \end{pmatrix}\right\|_2^4\right],$$

both of which are finite positive constants under our assumptions. Then, as long as $\eta \leq \lambda_1/(2\lambda_2)$ and the initialization satisfies $\|\theta(0) - (\xi, 0)\|_2 \leq \min\{(1 + \delta_1)\|\xi\|_2, \delta_2\}$, gradient descent for (10) converges linearly:

$$\left\|\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right\|_2^2 \leq \left(1 - \frac{\eta\lambda_1}{2}\right)^t \left\|\theta(0) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right\|_2^2.$$

Proof. We begin by expanding the squared distance to the true parameter after $t+1$ steps of gradient descent:

$$\begin{aligned}\left\|\theta(t+1) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right\|_2^2 &= \left\|\theta(t) - \eta\,\nabla_\theta R(\theta(t), \xi) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right\|_2^2 \\ &= \left\|\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right\|_2^2 - 2\eta \left(\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right)^\top \nabla_\theta R(\theta(t), \xi) + \eta^2 \|\nabla_\theta R(\theta(t), \xi)\|_2^2.\end{aligned}$$

Now, notice that

$$-2\eta \left(\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right)^\top \nabla_\theta R(\theta(t), \xi) \leq -\frac{2\eta\rho\gamma^2}{1 + L\delta_2} \left\|\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right\|_2^2$$

from the proof of Theorem 3.8. Finally, we may bound the gradient norm as

$$\begin{aligned}\|\nabla_\theta R(\theta(t), \xi)\|_2^2 &= 4 \left\|\mathbb{E}\left[(y_{\theta(t)}(X) - \sigma(\xi^\top X)) \frac{\sigma'(\omega_t(X))}{1 - \theta_2(t)\,\sigma'(\omega_t(X))} \begin{pmatrix} X \\ y_{\theta(t)}(X) \end{pmatrix}\right]\right\|_2^2 \\ &\leq 4\,\mathbb{E}\left[(y_{\theta(t)}(X) - \sigma(\xi^\top X))^2 \left(\frac{\sigma'(\omega_t(X))}{1 - \theta_2(t)\,\sigma'(\omega_t(X))}\right)^2 \left\|\begin{pmatrix} X \\ y_{\theta(t)}(X) \end{pmatrix}\right\|_2^2\right]. \qquad (18)\end{aligned}$$

Using Lipschitz continuity of σ and the Cauchy-Schwarz inequality, we have

$$(y_{\theta(t)}(X) - \sigma(\xi^\top X))^2 \leq L^2\,((\theta_1(t) - \xi)^\top X + \theta_2(t)\,y_{\theta(t)}(X))^2 \leq L^2 \left\|\theta(t) - \begin{pmatrix} \xi \\ 0 \end{pmatrix}\right\|_2^2 \left\|\begin{pmatrix} X \\ y_{\theta(t)}(X) \end{pmatrix}\right\|_2^2.$$

By our assumptions on σ and $\theta(t) \in \Theta$, we also have

$$\left(\frac{\sigma'(\omega_t(X))}{1-\theta_2(t)\,\sigma'(\omega_t(X))}\right)^2 \leq \left(\frac{L}{1-L\delta_2}\right)^2.$$

Combining these bounds with (18), we obtain

$$\|\nabla_\theta R(\theta(t),\xi)\|_2^2 \leq 4L^2 \left(\frac{L}{1-L\delta_2}\right)^2 \left\|\theta(t) - \begin{pmatrix}\xi\\0\end{pmatrix}\right\|_2^2 \mathbb{E}\left[\left\|\begin{pmatrix}X\\ y_{\theta(t)}(X)\end{pmatrix}\right\|_2^4\right].$$

Because $X \in L^4(\mathbb{P})$, we have from (13) that $y_{\theta(t)}(X) \in L^4(\mathbb{P})$ for all $\theta(t) \in \Theta$, so the expectation is finite by Cauchy-Schwarz. Putting everything together, we get that

$$\left\|\theta(t+1) - \begin{pmatrix}\xi\\0\end{pmatrix}\right\|_2^2 \leq \left(1 - \frac{2\eta\rho\gamma^2}{1+L\delta_2} + 4\eta^2L^2\left(\frac{L}{1-L\delta_2}\right)^2 \mathbb{E}\left[\left\|\begin{pmatrix}X\\ y_{\theta(t)}(X)\end{pmatrix}\right\|_2^4\right]\right)\left\|\theta(t) - \begin{pmatrix}\xi\\0\end{pmatrix}\right\|_2^2.$$

To see that λ_2 is finite, we note that by (13) and the fact that $X \in L^4(\mathbb{P})$, the integrand is uniformly dominated for $\theta \in \Theta$. Hence, the dominated convergence theorem implies that the expectation is continuous in θ over the compact set Θ, so the extreme value theorem implies that the supremum defining λ_2 is finite. It is also clear that λ_1 and λ_2 are positive because X is not almost surely constant (by the assumptions on the data in Theorem 3.8). Then, we have

$$1 - \eta\lambda_1 + \eta^2\lambda_2 \leq 1 - \frac{\eta\lambda_1}{2} < 1 \iff \eta\left(\eta\lambda_2 - \frac{\lambda_1}{2}\right) \leq 0 \iff \eta \leq \frac{\lambda_1}{2\lambda_2}.$$

Hence, it suffices to choose any step size $\eta \leq \lambda_1/(2\lambda_2)$ to have the bound

$$\left\|\theta(t+1) - \begin{pmatrix}\xi\\0\end{pmatrix}\right\|_2^2 \leq \left(1 - \frac{\eta\lambda_1}{2}\right)\left\|\theta(t) - \begin{pmatrix}\xi\\0\end{pmatrix}\right\|_2^2.$$

Finally, since we initialize $\theta(0)$ such that $\|\theta(0) - (\xi, 0)\|_2 \leq \min\{(1+\delta_1)\|\xi\|_2,\, \delta_2\}$, then by induction we have $\theta(t) \in \Theta$ for all $t \in \mathbb{N}_0$, completing the proof. □

Next, we empirically validate our results.

4 Experiments

In this section, we provide empirical validation of our theory. All code for our experiments can be found at the following GitHub repository:

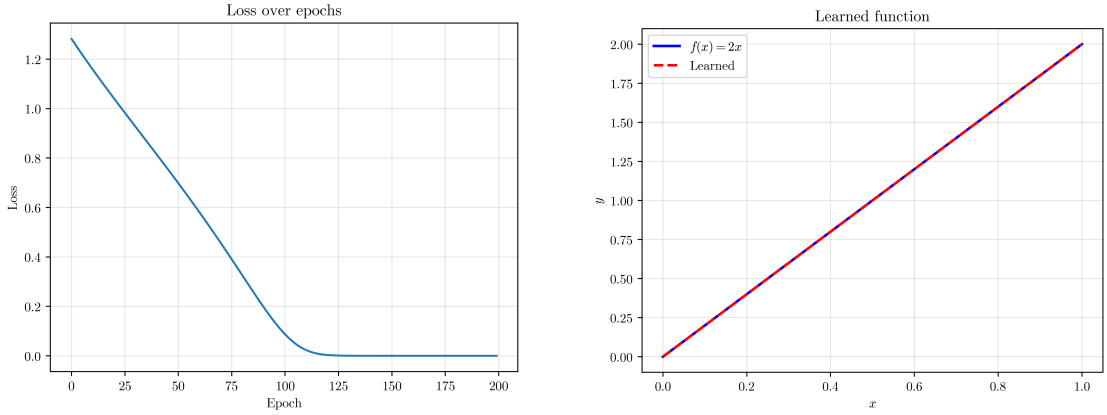
https://github.com/sanjitdp/single-index-deq

4.1 Linear models

We begin by considering the function $f(x) = \xi x$ with $\xi = 2$ and fitting an implicit linear model of the form

$$y_\theta(x) = g_\theta(x, y_\theta(x)) = \theta_1 \, x + \theta_2 \, y_\theta(x),$$

running gradient descent on this implicit model starting from $\theta(0) \sim \mathcal{N}(0, 0.1^2 \, I_2)$ using 1000 samples from $\mathrm{Unif}([0, 1])$. We run gradient descent for 200 epochs with a constant step size of $\eta = 0.01$. We depict the resulting loss over epochs and the learned function after training in Figure 1, showing that gradient descent is able to successfully learn the target function.

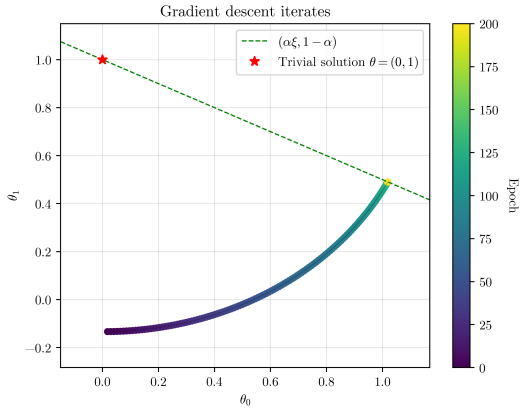


(a) Loss over epochs

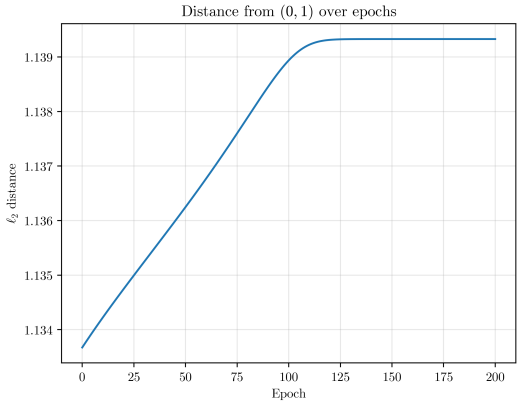
(b) Learned function

Figure 1: *(a) The loss of the linear DEQ converges to zero over epochs and (b) the learned function is close to the ground truth, so the implicit model is able to learn the target function.*

Then, we can plot the associated trajectory of the parameters in the (θ_1, θ_2)-plane, which we see spirals outward away from the trivial solution at $(0, 1)$ in Figure 2; this is what we expect from (8). We also plot the distance from the trivial solution over epochs, showing that the iterates move away from the trivial solution as training progresses.



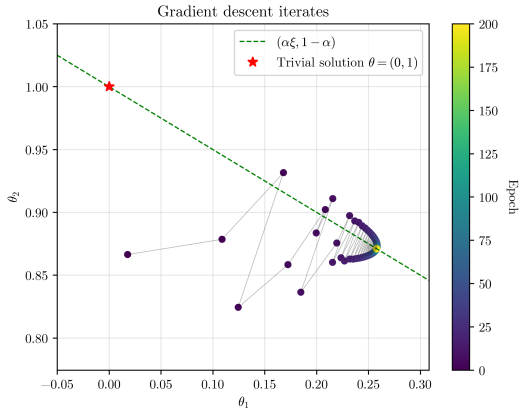
(a) Gradient descent trajectory



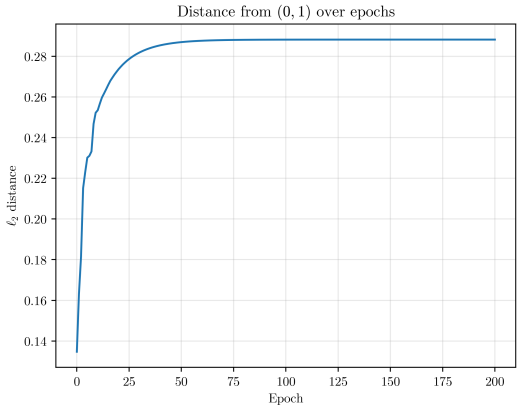
(b) Distance from $(0,1)$ over epochs

Figure 2: *(a) The gradient descent trajectory of the parameters roughly orbits around the ground truth, as predicted by our analysis of the gradient flow (Theorem 3.1). The line* $(\alpha\xi, 1-\alpha)$ *for* $\alpha \in \mathbb{R}$ *represents the set of valid solutions and* $\theta = (0,1)$ *corresponds to the trivial model* $y_\theta(x) = y_\theta(x)$*. (b) In addition, the parameters move further from the trivial solution* $(0,1)$*, as we demonstrated in (8).*

We observe the same phenomenon even when we initialize very close to the trivial solution by $\theta(0) \sim \mathcal{N}((0,1),\, 0.1^2 I_2)$, as shown in Figure 3.



(a) Gradient descent trajectory



(b) Distance from $(0,1)$ over epochs

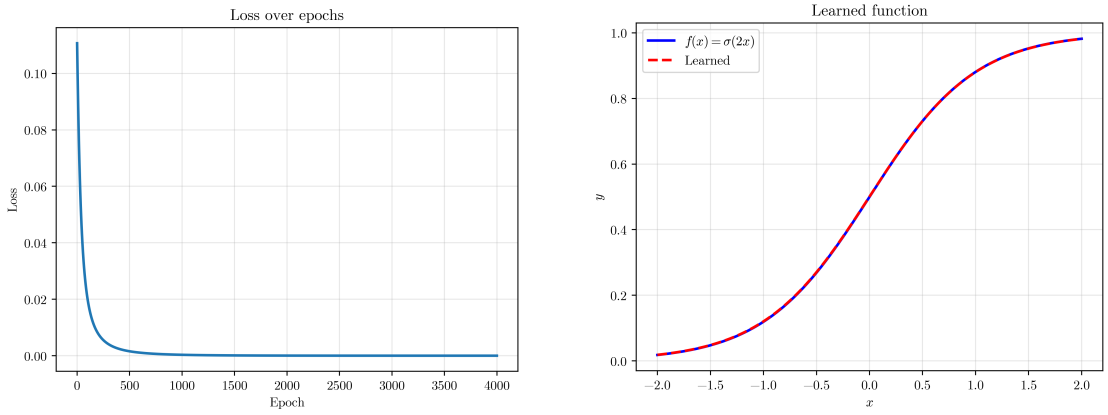
Figure 3: *Even when we initialize close to the trivial solution* $(0,1)$*, we observe that the gradient descent iterates (a) converge to the ground truth and (b) move further from* $(0,1)$ *over epochs. The line* $(\alpha\xi, 1-\alpha)$ *for* $\alpha \in \mathbb{R}$ *represents the set of valid solutions and* $\theta = (0,1)$ *corresponds to the trivial model* $y_\theta(x) = y_\theta(x)$*.*

4.2 Nonlinear single-index models

Next, we consider the function $f(x) = \sigma(2x)$ with the sigmoid activation function $\sigma(z) = 1/(1+e^{-z})$ and fit an implicit single-index model of the form

$$y_\theta(x) = g_\theta(x, y_\theta(x)) = \sigma(\theta_1\, x + \theta_2\, y_\theta(x)),$$

running gradient descent on this implicit model starting from $\theta(0) \sim \mathcal{N}(0,\, 0.1^2\, I_2)$ using 1000 training samples from Unif$([0, 1])$. For the forward pass, we use Brent's method as the root-finding algorithm (Brent, 2013). We run gradient descent for 4000 epochs with a constant step size of $\eta = 0.1$. We depict the resulting loss over epochs and the learned function after training in Figure 4, showing that gradient descent is able to successfully learn the target function.

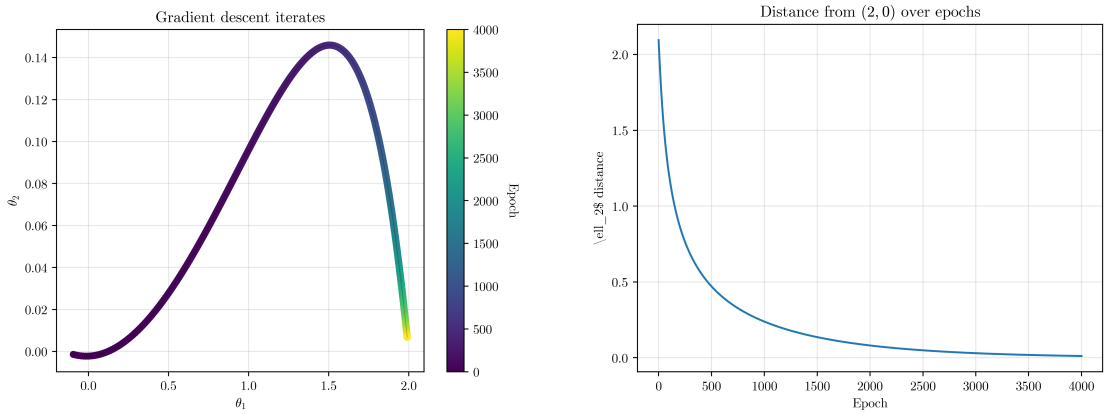


(a) Loss over epochs

(b) Learned function

Figure 4: *(a) The loss of the single-index DEQ converges to zero over epochs and (b) the learned function is close to the ground truth, so the implicit model is able to learn the target function.*

Finally, we plot the associated trajectory of the parameters in the (θ_1, θ_2)-plane in Figure 5, showing that the parameters converge to the target parameter $(2, 0)$ as training progresses. We also plot the distance from the target parameter over epochs, showing that the iterates move closer to the target parameter as training progresses.



(a) Gradient descent trajectory

(b) Distance from $(2, 0)$ over epochs

Figure 5: *(a) The gradient descent trajectory of the parameters converges to the target parameter* $(2, 0)$ *and (b) the distance to the target parameter decreases over epochs, as demonstrated in Theorem 3.9.*

Interestingly, the parameters do not move in a straight line towards the target parameter, but rather follow a curved path. One heuristic explanation for this is that $\sigma(\theta_1 x + \theta_2 y) \approx \sigma(0) + (\theta_1 x + \theta_2 y)\,\sigma'(0) = 1/2 + (\theta_1 x + \theta_2 y)/4$ by Taylor's theorem for small values of θ. If we further approximate $\sigma(2x) \approx 1/2 + x$ for $x \in [0, 1]$, then the training problem approximately reduces to fitting a linear model of the form $y_\theta(x) \approx (\theta_1 x + \theta_2 y_\theta(x))/4$ to the function $f(x) \approx x$. Hence, the training dynamics should approximately follow those of the linear model studied in Theorem 3.1, which explains why the parameters begin by moving away from the θ_1-axis to orbit around the approximate trivial solution $(0, 4)$.

As training progresses and the parameters grow larger in magnitude, the linear approximation becomes less accurate, and the parameters begin to curve back towards the target parameter $(2, 0)$ as guaranteed by Theorem 3.8. We know from the proof of Theorem 3.8 that the rate of convergence depends on the amount of nonlinearity in σ, which further corroborates this explanation. Even so, we know from Theorem 3.9 that the distance to the target parameter must decrease monotonically throughout the trajectory, which we observe in Figure 5b. We leave a more rigorous analysis of this phenomenon to future work. We also observe similar behavior when using the hyperbolic tangent and softplus activation functions, but we omit these experiments for brevity.

5 Conclusion

This work studies the gradient descent dynamics of training linear and single-index deep equilibrium models. We proved a conservation law for linear DEQs, which allowed us to explicitly characterize the trajectory of gradient flow and gradient descent. For single-index DEQs with nonlinear activation functions, we established local convergence guarantees for both gradient flow and gradient descent under suitable assumptions on the data distribution and activation function. Future work could extend these results to multi-neuron DEQs and explore the implications for practical training of deep equilibrium models.

References

Khalil, H. K., & Grizzle, J. W. (2002). *Nonlinear Systems* (3rd ed.). Prentice Hall.

Brent, R. P. (2013). *Algorithms for minimization without derivatives*. Courier Corporation.

Candès, E. J., Li, X., & Soltanolkotabi, M. (2015). Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, *61*(4), 1985–2007.

Ge, R., Huang, F., Jin, C., & Yuan, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. *Conference on Learning Theory*, 797–842.

Sun, J., Qu, Q., & Wright, J. (2015). When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., & Jordan, M. I. (2017). How to escape saddle points efficiently. *International Conference on Machine Learning*, 1724–1732.

Mei, S., Bai, Y., & Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, *46*(6A), 2747–2774.

Nesterov, Y. (2018). *Lectures on convex optimization* (Vol. 137). Springer.

Sun, J., Qu, Q., & Wright, J. (2018). A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, *18*(5), 1131–1198.

Bai, S., Kolter, J. Z., & Koltun, V. (2019). Deep equilibrium models. *Advances in Neural Information Processing Systems*, *32*.

Oymak, S., & Soltanolkotabi, M. (2019). Overparameterized nonlinear learning: Gradient descent takes the shortest path? *International Conference on Machine Learning*, 4951–4960.

Bai, S., Koltun, V., & Kolter, J. Z. (2020). Multiscale deep equilibrium models. *Advances in Neural Information Processing Systems*, *33*, 5238–5250.

Gu, F., Chang, H., Zhu, W., Sojoudi, S., & El Ghaoui, L. (2020). Implicit graph neural networks. *Advances in Neural Information Processing Systems*, *33*, 11984–11995.

Revay, M., Wang, R., & Manchester, I. R. (2020). Lipschitz bounded equilibrium networks. *arXiv preprint arXiv:2010.01732*.

Winston, E., & Kolter, J. Z. (2020). Monotone operator equilibrium networks. *Advances in neural information processing systems*, *33*, 10718–10728.

Yehudai, G., & Shamir, O. (2020). Learning a single neuron with gradient methods. *Conference on Learning Theory*, 3756–3786.

Bai, S., Koltun, V., & Kolter, J. Z. (2021). Stabilizing equilibrium models by Jacobian regularization. *arXiv preprint arXiv:2106.14342*.

El Ghaoui, L., Gu, F., Travacca, B., Askari, A., & Tsai, A. (2021). Implicit deep learning. *SIAM Journal on Mathematics of Data Science*, *3*(3), 930–958.

Geng, Z., Zhang, X.-Y., Bai, S., Wang, Y., & Lin, Z. (2021). On training implicit models. *Advances in Neural Information Processing Systems*, *34*, 24247–24260.

Gilton, D., Ongie, G., & Willett, R. (2021). Deep equilibrium architectures for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, *7*, 1123–1133.

Huang, Z., Bai, S., & Kolter, J. Z. (2021). $(\text{Implicit})^2$: Implicit layers for implicit representations. *Advances in Neural Information Processing Systems*, *34*, 9639–9650.

Kawaguchi, K. (2021). On the theory of implicit deep learning: Global convergence with implicit layers. *arXiv preprint arXiv:2102.07346*.

Vardi, G., Yehudai, G., & Shamir, O. (2021). Learning a single neuron with bias using gradient descent. *Advances in Neural Information Processing Systems*, *34*, 28690–28700.

Agarwala, A., & Schoenholz, S. S. (2022). Deep equilibrium networks are sensitive to initialization statistics. *International Conference on Machine Learning*, 136–160.

Fung, S. W., Heaton, H., Li, Q., McKenzie, D., Osher, S., & Yin, W. (2022). JFB: Jacobian-free backpropagation for implicit networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(6), 6648–6656.

Gao, T., & Gao, H. (2022). On the optimization and generalization of overparameterized implicit neural networks. *arXiv preprint arXiv:2209.15562*.

Liu, J., Hooi, B., Kawaguchi, K., & Xiao, X. (2022). MGNNI: Multiscale graph neural networks with implicit layers. *Advances in Neural Information Processing Systems*, *35*, 21358–21370.

Pokle, A., Geng, Z., & Kolter, J. Z. (2022). Deep equilibrium approaches to diffusion models. *Advances in Neural Information Processing Systems*, *35*, 37975–37990.

Feng, Z., & Kolter, J. Z. (2023). On the neural tangent kernel of equilibrium models. *arXiv preprint arXiv:2310.14062*.

Gao, T., Huo, X., Liu, H., & Gao, H. (2023). Wide neural networks as Gaussian processes: Lessons from deep equilibrium models. *Advances in Neural Information Processing Systems*, *36*, 54918–54951.

Geng, Z., Pokle, A., & Kolter, J. Z. (2023). One-step diffusion distillation via deep equilibrium models. *Advances in Neural Information Processing Systems*, *36*, 41914–41931.

Ling, Z., Xie, X., Wang, Q., Zhang, Z., & Lin, Z. (2023). Global convergence of over-parameterized deep equilibrium models. *International Conference on Artificial Intelligence and Statistics*, 767–787.

Marwah, T., Pokle, A., Kolter, J. Z., Lipton, Z., Lu, J., & Risteski, A. (2023). Deep equilibrium based neural operators for steady-state PDEs. *Advances in Neural Information Processing Systems*, *36*, 15716–15737.

Wang, H., Chang, J., Zhai, Y., Luo, X., Sun, J., Lin, Z., & Tian, Q. (2024). Lion: Implicit vision prompt tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(6), 5372–5380.

Wu, Z., Zhang, Y., Fang, C., & Lin, Z. (2024). Separation and bias of deep equilibrium models on expressivity and learning dynamics. *Advances in Neural Information Processing Systems*, *37*, 32476–32511.

Geiping, J., McLeish, S., Jain, N., Kirchenbauer, J., Singh, S., Bartoldson, B. R., Kailkhura, B., Bhatele, A., & Goldstein, T. (2025). Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*.

Liu, J., Ding, L., Osher, S., & Yin, W. (2025). Implicit models: Expressive power scales with test-time compute. *arXiv preprint arXiv:2510.03638*.

McCallum, S., Arora, K., & Foster, J. (2025). Reversible deep equilibrium models. *arXiv preprint arXiv:2509.12917*.