Deep limits of residual neural networks

Thorpe, Matthew; van Gennip, Yves

DOI
10.1007/s40687-022-00370-y

Publication date
2023

Document Version
Final published version

Published in
Research in Mathematical Sciences

Thorpe, van Gennip *Res Math Sci(2023)10:6*
https://doi.org/10.1007/s40687-022-00370-y

Research in the Mathematical Sciences

RESEARCH

Deep limits of residual neural networks

Matthew Thorpe[1,2*] and Yves van Gennip[3]

*Correspondence: matthew.thorpe-2@manchester.ac.uk
[2] The Alan Turing Institute, London NW1 2DB, UK
Full list of author information is available at the end of the article

Abstract

Neural networks have been very successful in many applications; we often, however, lack a theoretical understanding of what the neural networks are actually learning. This problem emerges when trying to generalise to new data sets. The contribution of this paper is to show that, for the residual neural network model, the deep layer limit coincides with a parameter estimation problem for a nonlinear ordinary differential equation. In particular, whilst it is known that the residual neural network model is a discretisation of an ordinary differential equation, we show convergence in a variational sense. This implies that optimal parameters converge in the deep layer limit. This is a stronger statement than saying for a fixed parameter the residual neural network model converges (the latter does not in general imply the former). Our variational analysis provides a discrete-to-continuum Γ-convergence result for the objective function of the residual neural network training step to a variational problem constrained by a system of ordinary differential equations; this rigorously connects the discrete setting to a continuum problem.

Keywords: Deep neural networks, Ordinary differential equations, Deep layer limits, Variational convergence, Gamma-convergence, Regularity

Mathematics Subject Classification: 34E05, 39A30, 39A60, 49J45, 49J15

1 Introduction

Recent advances in neural networks have proven immensely successful for classification and imaging tasks [81]. These practical successes have inspired many theoretical studies that try to understand why certain network architectures work better than others and what role the various parameters of the networks play. Over the years, these studies have come from such diverse areas as computational science [16,77,82], discrete mathematics [2], control theory and dynamical systems [23,36,53,72], approximation theory [15,50,51], frame theory [92], and statistical consistency. To the best of our knowledge, this and are the only papers to study variational limits of neural networks.

Stemming from the work of Haber and Ruthotto [36] and E [23], there has been recent interest in interpreting neural networks as dynamical systems. The connection with dynamical systems follows from an idealised infinitely deep interpretation of a neural network where one treats the depth as a time variable. There is a well-developed theory, such as Hamilton–Jacobi–Bellman equations and Pontryagin's maximum principle, which can be applied to analyse the dynamical system and therefore clarify the behaviour of the

discrete neural network [24]. Many more results have recently appeared in the literature, e.g. [4,11,33,66,67,80,87], and we refer to [8] for a more detailed overview. The aim of this paper is to connect discrete neural networks to a dynamical system using (a small modification) of the model presented in [36].

Classification of data is the task of assigning each element of a data set a label which indicates membership of one of several classes. Each of those classes has some a priori data assigned to it. A neural network approaches this task in two steps. First the a priori classified data is used to *train* the network. Then the trained network is used to classify other data. In this paper we will consider input data $x \in \mathbb{R}^d$ leading to network output $F(x) \in \mathbb{R}^m$ for some function $F : \mathbb{R}^d \to \mathbb{R}^m$. A neural network assigns a classification to some given input datum by performing a series of sequential operations to it, which are known as *layers*. Each layer is said to consist of *neurons*, by which it is meant that the output of each of the operations can be represented as a vector in $\mathbb{R}^d$ (encoding the state of d neurons). In our paper we assume there are n hidden layers[1] and each layer has the same number, d, of neurons. (Note that by making this assumption, the networks we consider cannot be used for dimensionality reduction; the network makes the classification decision based on the final layer, which contains a number of neurons equal to the dimension of the input datum.) We also assume that each input datum can be represented by a vector of the same dimension d. Hence, an input datum $x \in \mathbb{R}^d$ leads to a response in the first layer, $f_0(x) \in \mathbb{R}^d$, which in turn leads to a response in the second layer $f_1(f_0(x)) \in \mathbb{R}^d$, etc. After the response of the final layer $f_{n-1}(f_{n-2}(\dots f_0(x)\dots)) \in \mathbb{R}^d$ is obtained, a final function $\hat{f} : \mathbb{R}^d \to \mathbb{R}^m$ can be applied to map that response to the labels of the various pre-defined classes. The final output of the network then becomes $F(x) := \hat{f}(f_{n-1}(f_{n-2}(\dots f_0(x)\dots)))$.

In the training step training data $\{(x_s, y_s)\}_{s=1}^S$ is available, where $\{x_s\}_{s=1}^S \subset \mathbb{R}^d$ are inputs with class labels $\{y_s\}_{s=1}^S \subset \mathbb{R}^m$. The goal is to learn the form of the functions f_i such that the network's classifications $F(x_s)$ are close to the corresponding labels y_s. In this paper we restrict ourselves to functions f_i from a parametrised family of functions, as described in (4) below. The choice of cost function which is used to measure this "closeness" is one of many choices whose consequences are being studied, for example for classification [54] and image restoration tasks [98]. In this paper we consider a cost function with mild conditions, which allow for, for example, a quadratic error term (or loss function) $\sum_{s=1}^S \|F(x_s) - y_s\|^2$, together with regularisation terms which we will discuss later.

Implied in the architecture is the choice of parameterisation for f_i. A typical choice is to let f_i be of the form

$$f_i(x) = \sigma_i(K_i x + b_i), \tag{1}$$

where $K_i \in \mathbb{R}^{d\times d}$ is a matrix which determines the *weights* with which neurons in layer i activate neurons in layer $i+1$ and $b_i \in \mathbb{R}^d$ is a *bias* vector. The functions σ_i are called the *activation functions*. Many, although not all, activation functions used in practice are continuous approximations of a step function that effectively turn neurons "on" or "off" depending on the value of the input $K_i x + b_i$. In this paper, we assume every layer uses the same (Lipschitz continuous) activation function, $\sigma_i = \sigma$. Results from recent years have shown that the *rectified linear unit* (ReLU) activation function (or "positive part" [55])

[1] For simplicity we usually speak of n layers, even though it can be argued there are $n+2$ layers if one includes the 0^{th} input layer and the final classification layer from (7) in the count.

performs well in many situations [17,57,71]. It is given by

$$\sigma(x) = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } x \geq 0, \end{cases} \tag{2}$$

where its action on a vector should be interpreted componentwise (see Subsect. 1.1 for details). This, however, is not the only choice that can be made. The impact of the activation function on the performance of a given network is studied in many papers. For example, if ReLU is used the network trains faster than when some of the classical saturating nonlinear activation functions such as $x \mapsto \tanh x$ and $x \mapsto \frac{1}{1+e^{-x}}$ are used instead [57]. Moreover, ReLU has been observed to lead to sparsity in the resulting weights, with many of them being zero. These are sometimes referred to as "dead neurons" [68,89,96].

The activation function(s) are often[2] specified beforehand for a given network and are not a part of what should be "learned" by the network. That still leaves, however, a large number of parameters for the learning problem. Each layer contains $d \times d + d$ parameters in the form of K_i and b_i. Different types of networks restrict the admissible sets for the K_i and b_i. For example, some networks impose that the biases b_i are completely absent, such as the *Finite Impulse Response* (FIR) networks in [49,79,90,93] or that each layer has the same *shared bias* [75], while the traditional *convolutional neural networks* (CNN) restrict the choice of K_i to convolution matrices, i.e. matrices in which each row is a shifted version of a *filter* vector $(0, \ldots, 0, v_1, \ldots, v_k, 0, \ldots, 0)$, such that the product $K_i x$ becomes a discrete convolution of the vector $v = (v_1, \ldots, v_k)$ with x [47,59]. In this paper we will not restrict the choice of K_i and b_i by such hard constraints. Instead, we include regularisation terms in the cost function, which penalise K_i and b_i which vary too much between layers or whose entries in the first layer are too large (see Sect. 1.2 for details).

Oberman and Calder study, in a variational sense, the data rich limit $S \to \infty$. In particular, they consider, a sequence of variational problems of the form

$$\text{minimise: } L(F, \mu_S) + R(F), \tag{3}$$

where L is a loss term, μ_S is an empirical measure induced by the training data set $\{x_s, y_s\}_{s=1}^S$, and R a regularisation term; for example,

$$L(F, \mu_S) = \int_{\mathbb{R}^d \times \mathbb{R}^m} |F(x) - y|^2 \, \mathrm{d}\mu_S(x, y) = \frac{1}{S} \sum_{s=1}^{S} |F(x_s) - y_s|^2.$$

The set of admissible F is determined by a neural network. The main result of is to show that minimisers F_S of (3) converge as $S \to \infty$ to a solution of the variational problem

$$\text{minimise: } L(F, \mu) + R(F)$$

for an appropriate measure μ obtained as limit of the empirical measures μ_S.

In this paper we study the deep layer limit (i.e. the limit $n \to \infty$) of a *residual neural network* (ResNet) [45], which are related in spirit to the highway networks of [86]. A crucial way in which ResNet type neural networks differ from other networks such as CNNs, is the form of the functions f_i. Instead of assuming a form as in (1), in ResNet the assumption

$$f_i(x) = x + \sigma_i(K_i x + b_i) \tag{4}$$

[2] But not always; see for example [44] for parametric ReLU, which contains a learnable parameter.

is made. This can be interpreted as the network having *shortcut connections*: The additional term x on the right-hand side represents information from the previous layer "skipping a layer" (or, more accurately, skipping the processing associated with the layer) and being transmitted to the next layer without being transformed. The reason for introducing these shortcut connections is to tackle the *degradation* problem [43,45]: It has been observed that increasing the depth of a network (i.e. its number of layers) can lead to an increase in the error term instead of the expected decrease. Crucially, this behaviour appears while training the network, which indicates that it is not due to overfitting (as that would be an error which would be only present during the testing phase of an already trained network). In [45] it is argued that, if $\hat{f}_i(x)$ is the actual desired output in layer $i+1$, the *residual* $\hat{f}_i(x) - x$ is easier to learn in practice than $\hat{f}_i(x)$ itself. Deep networks using the architecture (1) can suffer from vanishing or exploding gradients during backpropagation [3,31,49,75], resulting in weights which either do not change much at all during the training phase or which change wildly in each step. In general, learning the residual does not suffer from vanishing/exploding gradients by approximately preserving the norm of the gradient between layers [95]. In [31] it is shown that these problems might be avoided by choosing a careful initialisation; [68] argues that using the ReLU activation function also helps in avoiding vanishing gradients.

Crucially for our purposes, the additional term x in (4) compared to (1) allows us to write

$$X_{i+1}^{(n)} - X_i^{(n)} = f_i(X_i^{(n)}) - X_i^{(n)} = \frac{1}{n}\sigma_i(K_i^{(n)}X_i^{(n)} + b_i^{(n)}), \tag{5}$$

where $X_{i+1}^{(n)} = f_i(X_i^{(n)}) \in \mathbb{R}^d$ is the output in layer $i+1$ and where we have introduced a factor $\frac{1}{n}$ with σ for scaling purposes. We have also added superscripts (n) to $X_i^{(n)}$, $K_i^{(n)}$ and $b_i^{(n)}$ to indicate that these weights and biases belong to the network with n layers. Remember that, in this paper, we will use the same activation function in each layer: $\sigma_i = \sigma$. As observed in [23,37,67], this setup describes an explicit Euler characterisation of the ordinary differential equation (ODE)

$$\dot{X}(t) = \sigma(K(t)X(t) + b(t)),$$

with time step $1/n$. Here X, K, and b denote real-valued functions on $[0, 1]$. This observation has been used to motivate new neural network architectures based on discretisations of partial/ordinary differential equations, e.g. [11,33,36,67,80,87].

Since the *forward pass* through ResNet is given by a discretised ODE in (5), a natural question is whether the deep limit ($n \to \infty$) of ResNet indeed gives us back the ODE. We need to be a bit more careful, however, when formulating this question, and distinguish between the training step and the use of a trained network. The latter consists of applying (5) through all layers (with known $K_i^{(n)}$ and $b_i^{(n)}$ obtained by training the network) with a single given input datum x as initial condition, $X_0 = x$. The deep limit question in this case then becomes whether solutions of this discretised process converge to the solution (Lipschitz continuity of $x \mapsto \sigma(Kx + b)$ guarantees a unique solution, by standard ODE theory) of the ODE. Our Corollary 2.3 shows that they do, in a pointwise sense. In order to derive this corollary we require the trained weights and biases, $K_i^{(n)}$ and $b_i^{(n)}$, to converge (up to a subsequence) to sufficiently regular weights and biases, K and b, which can be used in the ODE. This requires us to carefully analyse the training step. The main result of this paper, Theorem 2.1, does exactly that.

Theorem 2.1 uses techniques from variational methods to show that the trained weights and biases have (up to a subsequence) deep layer limits. In particular, it uses Γ-convergence, which is explained in further detail in Sect. 3.3. Variational calculus deals with problems which can be formulated in terms of minimisation problems. In this paper we formulate the training step (or *learning problem*) of an n-layer ResNet as a minimisation problem for the function $\mathcal{E}_n$ in (8), which consists of a quadratic cost function with regularisers for all the coefficients that are to be learned. We then identify the Γ-limit of the sequence $\{\mathcal{E}_n\}_{n=1}^{\infty}$, which is given by $\mathcal{E}_\infty$ in (12). Γ-convergence is a type of convergence which (in combination with a compactness result) guarantees that minimisers of $\mathcal{E}_n$ converge (up to a subsequence) to a minimiser of the Γ-limit $\mathcal{E}_\infty$. It has been successfully applied for discrete-to-continuum limits in a machine learning setting, for example in [88] and the references in the following sentence. The specific tools we use in this paper to obtain the discrete-to-continuum Γ-limit were developed in [30] and have been successfully applied in a series of papers since [22,27–29,83].

The impact of this Γ-convergence result is twofold. On the one hand it is an important ingredient in showing that the output of an already trained network for given input data is, in the sense made precise by Corollary 2.3, approximately the output of a dynamical system which has the input data as initial condition. On the other hand, it shows that the training step itself is a discrete approximation of a continuum variational problem. This opens up the possibility of using techniques from partial differential equations (PDEs) to solve the minimisation problem for $\mathcal{E}_\infty$ in order to obtain (approximate) solutions to the n-layer training step; such as Pontryagin's maximum principle [25,64]. It also opens up the possibility to construct different networks by using different discretisations of the ODE, as in the midpoint network in [9].

We note that connecting discrete difference equations to a continuum differential equation in the setting of recursive algorithms (i.e. $X_{i+1} = f_i(X_i, \theta)$ where θ are given parameters) is well studied, for example [13,65]. However, these results are in the pointwise convergence setting, i.e. the parameter θ is fixed. Pointwise convergence is not strong enough to imply convergence of minimisers, i.e. what we want is that the θ_n^* that minimises a variational problem converges as $n \to \infty$ to some θ that minimises a variational problem with the constraint $\dot{X} = f_\infty(X, \theta)$. This is the novelty of our result.

In the remainder of the introduction we introduce our framework; namely the neural network architecture, the choice and motivation of regularisation of the neural network parameters, and the continuum deep layer limit. In Sect. 2 we state our assumptions and main results connecting the discrete neural network with its continuum limit. In Sect. 3 we give some preliminary material which includes (1) defining the topology we use for convergence of the parameters $\mathbf{K}^{(n)}$, $\mathbf{b}^{(n)}$, i.e. we make precise $\mathbf{K}^{(n)} \to K$ and $\mathbf{b}^{(n)} \to b$, and (2) giving a brief background on variational methods and in particular Γ-convergence. Section 4 is devoted to the proofs of the main results. We conclude the paper in Sect. 5 with a brief discussion of open questions.

1.1 The finite layer neural network

We recap a simplified version of ResNet as presented in [36]. In this model there are n layers and the number of neurons in each layer is d. In particular, we let $X_i^{(n)} \in \mathbb{R}^d$ be the state of each neuron in the *ith* layer. For clarity we will denote with a superscript the

number of layers, this is to avoid confusion when talking about two versions of the neural network with different numbers of layers. The relationship between layers is given by

$$X_{i+1}^{(n)} = X_i^{(n)} + \frac{1}{n}\sigma(K_i^{(n)}X_i^{(n)} + b_i^{(n)}), \qquad i = 0, 1, \ldots, n-1, \tag{6}$$

where $\mathbf{K}^{(n)} = \{K_i^{(n)}\}_{i=0}^{n-1} \subset \mathbb{R}^{d\times d}$, $\mathbf{b}^{(n)} = \{b_i^{(n)}\}_{i=0}^{n-1} \subset \mathbb{R}^d$ determine an affine transformation at each layer and $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ is an activation function which characterises the difference between layers. We will assume that σ acts componentwise, i.e. $\sigma(x) = (\tilde{\sigma}(x_1), \tilde{\sigma}(x_2), \ldots, \tilde{\sigma}(x_d))^T$, for some $\tilde{\sigma} : \mathbb{R} \to \mathbb{R}$. For example, a valid, but not necessary choice for $\tilde{\sigma}$ is the ReLU function from (2). With a slight abuse of notation, $\tilde{\sigma}$ is sometimes also denoted by σ, as for example in (2). The layers $\{X_i^{(n)}\}_{i=1}^{n-1}$ are called hidden, $X_0^{(n)}$ is the input to the network, and $X_n^{(n)}$ is the output.

In order to apply the neural network (6) to labelling problems an additional, classification, layer is appended to the network. For example, one can add a linear regression model, that is we let $Y = WX_n^{(n)} + c$ where $W \in \mathbb{R}^{m\times d}$ and $c \in \mathbb{R}^m$. More generally, we assume the classification layer takes the form

$$Y = h(WX_n^{(n)} + c) \tag{7}$$

for a given function $h : \mathbb{R}^m \to \mathbb{R}^m$. Given all parameters, the forward model/classifier for input $X_0^{(n)} = x$ is $Y = h(WX_n^{(n)}[x; \mathbf{K}^{(n)}, \mathbf{b}^{(n)}] + c)$ where $X_n^{(n)}[x; \mathbf{K}^{(n)}, \mathbf{b}^{(n)}]$ is given by the recursive formula (6) with input $X_0^{(n)} = x$.

Given a set of training data $\{(x_s, y_s)\}_{s=1}^S$, where $\{x_s\}_{s=1}^S \subset \mathbb{R}^d$ are inputs with labels $\{y_s\}_{s=1}^S \subset \mathbb{R}^m$, one wishes to find parameters $\mathbf{K}^{(n)}$, $\mathbf{b}^{(n)}$, W, c that minimise the error of the neural network on the training data. There are clearly multiple ways to measure the error. To maximise generality, we define

$$E_n(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c; x, y) = \mathcal{L}\left(h(WX_n^{(n)}[x; \mathbf{K}^{(n)}, \mathbf{b}^{(n)}] + c), y\right),$$

where the function $\mathcal{L}$ is nonnegative and has to satisfy a continuity condition in its first argument, as detailed in Theorem 2.1 and Proposition 2.2. A typical allowed choice is $\mathcal{L}(z, y) = \|z - y\|^2$. The error $E_n(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c; x, y)$ should be interpreted as the error of the parameters $\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c$ when predicting x given that the true value is y. Naively, one may wish to minimise the sum of $E_n(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c; x_s, y_s)$ over $s \in \{1, \ldots, S\}$. However this problem is ill-posed once the number of layers, n, is large. In particular, the number of parameters being greater than the number of training data points leading to overfitting. The solution, as is common in the calculus of variations, is to include regularisation terms, e.g. (applicable to neural networks) [32,36,74,78], on each of $\mathbf{K}^{(n)}$, $\mathbf{b}^{(n)}$, W and c, this is discussed in the next section.

The finite layer objective functional, with regularisation weights $\alpha_1, \ldots, \alpha_4$, is given by

$$\begin{aligned}\mathcal{E}_n(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c) = \sum_{s=1}^S E_n(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c; x_s, y_s) + \alpha_1 R_n^{(1)}(\mathbf{K}^{(n)}) + \alpha_2 R_n^{(2)}(\mathbf{b}^{(n)}) \\ + \alpha_3 R^{(3)}(W) + \alpha_4 R^{(4)}(c).\end{aligned} \tag{8}$$

Here the $R^{(i)}$ are regularisation terms, which will be introduced in detail in Subsect. 1.2. The learning problem is to find $(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W^{(n)}, c^{(n)})$ which minimises $\mathcal{E}_n$.

The problem we concern ourselves with is the behaviour in the deep layer limit, i.e. what happens to $\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W^{(n)}, c^{(n)}$ as $n \to \infty$. The results of this paper are theoretical and in particular ignore the considerable challenge of finding such minimisers. However, we

do hope that a better understanding of the deep layer limit can aid the development of numerical methods by, for example, allowing PDE approaches to the minimisation of $\mathcal{E}_n$. Indeed, the authors of [56] view neural networks as inverse problems and apply filtering methods such as the ensemble Kalman filter which are gradient free. We note that theory is often developed for continuum models as it reveals what behaviour will be expected for large discrete problems. For example, the authors of [97] analyse stability properties of continuum analogues of neural networks.

In this paper we are not concerned with the actual numerical method used to compute the learning or training step, i.e. the method to compute minimisers of (8). However, for completeness we briefly point to some optimisation methods and potential pitfalls. Currently, a variety of different methods are being used to compute the training step; [89] gives an overview of various methods. One of the most popular ones is backpropagation [41,42,46,58,99] using stochastic gradient descent [47]. Since the minimisation problem is not convex, any gradient descent method risks running into critical points which are not minima. In [19] it is argued that in certain setups critical points are more likely to be saddle points than local minima and [61] proves that (under some assumptions on the objective function and the step size) gradient descent does not converge to a saddle point for almost all initial conditions. Moreover, [12] empirically verifies that in deep networks most local minima are close in value to the global minimum and the corresponding minimisers give good results. In some cases it can even be proven that all local minima are equal to the global minimum [60]. These results suggest that the critical points of the non-convex optimisation problem are not necessarily a major problem for gradient descent methods.

Variants of gradient descent, such as blended coarse gradient descent, which is not strictly speaking a gradient descent algorithm—rather it chooses an artificial ascent direction—have been explored in [94]. The authors of [10] show that the (local entropy) loss function satisfies a Hamilton-Jacobi equation and use this to analyse and develop stochastic gradient descent methods (in continuous time) which converge to gradient descent in the limit of fast dynamics. Outside of gradient-based methods the authors of [35] apply an Ensemble Kalman Filter method to the training of parameters.

Overfitting is also an issue to take into account during training. Techniques such as max pooling [81] (for a PDE-based interpretation of max pooling and ReLU as morphological convolutions in a CNN, see [84]) or Dropout [48] work well in practice to avoid overfitting. The former consists of downsampling a layer by pooling the neurons into groups and assigning to each group the maximum value of all its neurons. The latter consists of randomly omitting neurons on each presentation of each training case. The ReLU activation function works well with Dropout [17]. Recently [70] made the case that improvements can be obtained by using sparsely connected layers. Adding regularisation terms which encourage some level of smoothness to the cost functional can also help to avoid overfitting [47].

1.2 Regularisation

Explicit regularisation in neural networks dates back to at least [20], where the authors added a penalty on the rate of change of E_n with respect to the input x_s. Here we approximately follow [36]. We refer to for a more in-depth discussion on regularisation in machine learning.

We define regularisation terms $R_n^{(1)}(\mathbf{K}^{(n)})$, $R_n^{(2)}(\mathbf{b}^{(n)})$, $R^{(3)}(W)$, $R^{(4)}(c)$ by

$$\begin{aligned} R_n^{(1)}(\mathbf{K}^{(n)}) &= n\sum_{i=1}^{n-1} \|K_i^{(n)} - K_{i-1}^{(n)}\|^2 + \tau_1 \|K_0^{(n)}\|^2, \\ R_n^{(2)}(\mathbf{b}^{(n)}) &= n\sum_{i=1}^{n-1} \|b_i^{(n)} - b_{i-1}^{(n)}\|^2 + \tau_2 \|b_0^{(n)}\|^2, \\ R^{(3)}(W) &= \|W\|^2, \\ R^{(4)}(c) &= \|c\|^2 \end{aligned}$$

where $\tau_i > 0$. Since all norms on finite-dimensional vector spaces are topologically equivalent, many of the results in this paper do not depend on the specific choices for the norms $\|\cdot\|$ that are used in the definitions above. In some places, such as in Sect. 4.4 however, an inner product structure is assumed on certain norms, while in others, such as Lemma 4.16 and the derived Corollary 2.3, the constants in the estimates will depend on the specific choice of norm.

Since we eventually wish to interpret the n layers of the network as a discretisation of (one-dimensional) time with time step $\frac{1}{n}$, the scaling by n of the difference terms in $R_n^{(1)}$ and $R_n^{(2)}$ is the correct one to view these terms as discretised integrals of finite-difference approximations to squared gradients. This will be further clarified in Sect. 1.2.1 and follows as a consequence of the approximation

$$\|\dot{K}\|_{L^2}^2 \approx \frac{1}{n}\sum_{i=0}^{n-1} \|\dot{K}(i/n)\|^2 \approx \frac{1}{n}\sum_{i=1}^{n-1} \left\| \frac{K(i/n) - K((i-1)/n)}{1/n} \right\|^2. \tag{9}$$

We emphasise that $R^{(3)}$ and $R^{(4)}$ do not depend on n. We refer to $R_n^{(1)}$, $R_n^{(2)}$ as the non-parametric regularisers, and $R^{(3)}$, $R^{(4)}$ as the parametric regularisers (we consider $\mathbf{K}^{(n)}$ and $\mathbf{b}^{(n)}$ to be non-parametric as their complexity grows with the number of layers n, whilst W and c are parametric as their complexity is independent of n). The point of including regularisation is to enforce compactness in the minimisers; without compactness we cannot find converging sequences of minimisers which, in particular, can lead to objective functionals that become ill-posed in the deep layer limit. We justify the regularisation below; however, we note that the regularisation is quite strong. In particular, we are imposing H^1 bounds on $\mathbf{K}^{(n)}$ and $\mathbf{b}^{(n)}$—which are also suggested in [36]— as well as norm bounds on W and c. The cost of treating a wide range of activation functions σ and classification functions h is to include strong regularisation functions. In specific cases it may be possible to reduce the regularisation, for example by setting $\tau_i = 0$ and/or removing the terms $R^{(3)}$, $R^{(4)}$. In the next two subsections, we give a discussion on why these terms, in general, are necessary.

It should also be noted that it is sometimes observed that techniques such as stochastic node or layer dropout [52] can act as regularisers, without the need for explicitly added regularisation terms. A good mathematical understanding of this phenomenon is still missing from the literature, to the current knowledge of the authors, and in this paper we have restricted ourselves to adding explicit regularisation terms, as is common in the calculus of variations.

1.2.1 The nonparametric regularisation

By construction the regularisation terms on $\mathbf{K}^{(n)}$ and $\mathbf{b}^{(n)}$ resemble H^1 norms. These terms are used for compactness in order to apply the direct method of the calculus of variations. By standard Sobolev embeddings sequences bounded in H^1 are (pre-)compact in L^2. There is a little work to be done in order to match discrete sequences $\mathbf{K}^{(n)} = \{K_j^{(n)}\}_{j=0}^{n-1}$, $\mathbf{b}^{(n)} = \{b_j^{(n)}\}_{j=0}^{n-1}$ with continuum sequences $K^{(n)} : [0,1] \to \mathbb{R}^{d\times d}$, $b^{(n)} : [0,1] \to \mathbb{R}^d$, but with an appropriate identification we can show that $R_n^{(1)}(\mathbf{K}^{(n)}) \approx \|\dot{K}^{(n)}\|_{L^2}^2 + \tau_1\|K^{(n)}(0)\|^2$ and similarly for $\mathbf{b}^{(n)}$. In that sense they are very natural choices from a calculus of variations point of view as they allow us to conclude strong L^2 convergence of the parameters.

Of course, given $K : [0,1] \to \mathbb{R}^{d\times d}$ we can define $\tilde{K}_i^{(n)} = K(i/n)$ and then $R_n^{(1)}(\tilde{\mathbf{K}}^{(n)}) \to \|\dot{K}\|_{L^2}^2 + \|K(0)\|^2$. This we would call pointwise convergence. The main result of this paper is stronger, in particular we show variational convergence. Without the H^1 semi-norm part of our regularisation terms (i.e. without the L^2 norm of the gradient) we would a priori only get weak L^2 convergence. The question whether this suffices to still derive our results is a very interesting one, but goes beyond the scope of this paper, as it introduces a lot of extra technical difficulties. We note that $R_n^{(i)}$, $i = 1, 2$, are very similar to the choice of regularisation in [36], but we add the terms $\|K_0^{(n)}\|^2$, $\|b_0^{(n)}\|^2$.

The penalty on finite differences is natural; in order to achieve a limit it is necessary to bound oscillations in the parameters between layers. Physically this relates to imposing the condition that close layers discriminate similar features. For our analysis this is needed to establish compactness. It is interesting to note that one can obtain limits without including explicit regularisation terms. The limiting behaviour of the deep network, however, may no longer be given by a deterministic ODE system of the type we will describe in (10) and, in particular, could be stochastic [14]. The coefficients appearing in the limiting equations described in [14] are obtained through a different limiting procedure than the one we use (and describe in Sect. 3.2).

The additional terms, $\|K_0^{(n)}\|^2$, $\|b_0^{(n)}\|^2$, are perhaps less physically reasonable and introduce a bias into the methodology (meaning that preference is given to smaller values of $K_0^{(n)}$ and $b_0^{(n)}$). As examples of why it is necessary to have these additional terms, i.e. to have $\tau_1 > 0$ and $\tau_2 > 0$, consider the following. First assume $\tau_1 = 0$, let $d = m = 1$, $h = \mathrm{Id}$, $\sigma = \mathrm{Id}$, $\mathcal{L}(z,y) = |z-y|^2$, and fix $n \in \mathbb{N}$. Consider the set $\{(x_s, y_s)\}_{s=1}^S \subset \mathbb{R}\times\mathbb{R}$, where $y_s = x_s$, and the sequence $\{(\mathbf{K}_l^{(n)}, \mathbf{b}_l^{(n)}, W_l, c_l)\}_{l\in\mathbb{N}}$, with, for all i and l, $(K_i^{(n)})_l = l$, $(b_i^{(n)})_l = 0$, $W_l = (1+l)^{-n}$, $c_l = 0$. Then,

$$E_n(\mathbf{K}_l^{(n)}, \mathbf{b}_l^{(n)}, W_l, c_l; x_s, y_s) = |W_l X_n^{(n)} - y_s|^2 = |(1+l)^{-n}(1+l)^n x_s - y_s|^2 = 0.$$

Moreover, for all l, $R_n^{(1)}(\mathbf{K}_l^{(n)}) = R_n^{(2)}(\mathbf{b}_l^{(n)}) = R^{(4)}(c_l) = 0$ and $R^{(3)}(W_l) \to 0$ as $l \to \infty$. Therefore, $\{(\mathbf{K}_l^{(n)}, \mathbf{b}_l^{(n)}, W_l, c_l)\}_{l\in\mathbb{N}}$ is a minimising sequence for $\mathcal{E}_n$ (as $l \to \infty$ with n fixed) with no converging subsequence. As the elementary example shows, if one were to set $\tau_1 = 0$ then an additional assumption would be needed to guarantee relative compactness of minimising sequences. A second example showing a similar necessity to have $\tau_2 > 0$ is constructed by setting $\tau_1 \geq 0$, $\tau_2 = 0$, $(K_i^{(n)})_l = 0$, and $(b_i^{(n)})_l = l$ in the previous example.

1.2.2 The parametric regularisation

An example showing why $\alpha_3 > 0$ and $\alpha_4 > 0$ are necessary, can be constructed in a similar fashion. Let $d = m = 1$, $h = \mathrm{Id}$, $\sigma = \mathrm{Id}$, $\mathcal{L}(z,y) = |z-y|^2$, and fix $n \in \mathbb{N}$. Let $S = 1$, so that we have only one training pair $(x_1, y_1) = (x, y)$. Define the sequence $\{(\mathbf{K}_l^{(n)}, \mathbf{b}_l^{(n)}, W_l, c_l)\}_{l\in\mathbb{N}}$,

with, for all i and l, $(K_i^{(n)})_l = 0$, $(b_i^{(n)})_l = 0$, $W_l = l$, $c_l = y - lx$. Then,

$$E_n(\mathbf{K}_l^{(n)}, \mathbf{b}_l^{(n)}, W_l, c_l; x, y) = |W_l X_n^{(n)} + c_l - y_s|^2 = |lx + y - lx - y|^2 = 0.$$

Also, for all l, $R_n^{(1)}(\mathbf{K}_l^{(n)}) = R_n^{(2)}(\mathbf{b}_l^{(n)}) = 0$. We conclude, as before, that $\{(\mathbf{K}_l^{(n)}, \mathbf{b}_l^{(n)}, W_l, c_l)\}_{l\in\mathbb{N}}$ is a minimising sequence for $\mathcal{E}_n$ (as $l \to \infty$ with n fixed) with no converging subsequence.

1.3 The deep layer differential equation limit

By considering pointwise limits it is not difficult to derive our candidate limiting variational problem. Although pointwise convergence is not enough to imply convergence of minimisers, it is informative. Let $X : [0, 1] \to \mathbb{R}^d$ solve the differential equation

$$\dot{X}(t) = \sigma(K(t)X(t) + b(t)), \qquad t \in [0, 1] \tag{10}$$

for some given parameters $K : [0, 1] \to \mathbb{R}^{d\times d}$ and $b : [0, 1] \to \mathbb{R}^d$ (as is usual we understand $\dot{X}(0)$ to be the right-derivative of X at $t = 0$ and $\dot{X}(1)$ to be the left-derivative of X at $t = 1$). For shorthand we write $X(t; x, K, b)$ for the solutions of (10) with initial condition $X(0) = x$ and parameters K, b. One can see that (6) is the discrete analogue of (10) with $K_i^{(n)} = K(i/n)$ and $b_i^{(n)} = b(i/n)$. In fact one can show (under sufficient conditions) that $X_{\lfloor nt\rfloor}^{(n)}[x, \mathbf{K}^{(n)}, \mathbf{b}^{(n)}] \to X(t; x, K, b)$ as $n \to \infty$ (see Lemma 4.6).

Similarly, the regularisation terms $R_n^{(i)}$, $i = 1, 2$, are discretisations of the functionals

$$\begin{aligned} R_\infty^{(1)}(K) &= \|\dot{K}\|_{L^2}^2 + \tau_1\|K(0)\|^2 \\ R_\infty^{(2)}(b) &= \|\dot{b}\|_{L^2}^2 + \tau_2\|b(0)\|^2 \end{aligned} \tag{11}$$

and $R^{(3)}$, $R^{(4)}$ are unchanged. We note that $R_\infty^{(i)}$, $i = 1, 2$ are well defined on H^1, since by regularity properties of Sobolev spaces any $u \in H^1$ is continuous and therefore pointwise evaluation is well defined; in particular we may define $\|K(0)\|$, $\|b(0)\|$ for H^1 functions. In fact, see the discussion in Sect. 3.4, $R_\infty^{(i)}$, $i = 1, 2$, are equivalent to the H^1 norm whenever $\tau_i > 0$.

Once we append the classification layer to the neural network, we arrive at the limiting objective functional

$$\begin{aligned} \mathcal{E}_\infty(K, b, W, c) = \sum_{s=1}^{S} E_\infty(K, b, W, c; x_s, y_s) + \alpha_1 R_\infty^{(1)}(K) \\ + \alpha_2 R_\infty^{(2)}(b) + \alpha_3 R^{(3)}(W) + \alpha_4 R^{(4)}(c) \end{aligned} \tag{12}$$

where

$$E_\infty(K, b, W, c; x, y) = \mathcal{L}\left(h(WX(1; x, K, b) + c), y\right).$$

The main result of the paper is to show that minimisers of $\mathcal{E}_n$ converge to minimisers of $\mathcal{E}_\infty$.

2 Main results

Our main results concern the convergence of the variational problem $\min \mathcal{E}_n$ to $\min \mathcal{E}_\infty$. In particular we show

$$\begin{aligned} \min_{\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c} \mathcal{E}_n(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c) &\to \min_{K, b, W, c} \mathcal{E}_\infty(K, b, W, c), \\ \operatorname*{argmin}_{\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c} \mathcal{E}_n(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c) &\to \operatorname*{argmin}_{K, b, W, c} \mathcal{E}_\infty(K, b, W, c), \end{aligned}$$

as $n \to \infty$. At this point we have not specified the topology on which we define the discrete-to-continuum convergence. For now it is enough to say that the distance is given by a function $d : \Theta^{(n)} \times \Theta \to [0, +\infty)$ where $\Theta^{(n)}$ is the parameter space of $\mathcal{E}_n$ and Θ is the parameter space of $\mathcal{E}_\infty$. The topology is described in detail in Sect. 3.2. We do want to emphasise at this point that although d appears to depend on n —and in fact appears not to be a distance at all, due to a lack of symmetry between its two arguments— it is in fact a restriction of an n-independent metric on a higher-dimensional space to an n-dependent subset.

We will use the following assumptions for our results.

Assumptions 1 The following assumptions will be used in our main convergence result:

1. $\alpha_i > 0$ for $i = 1, 2, 3, 4$ and $\tau_j > 0$ for $j = 1, 2$;
2. $h \in C^0(\mathbb{R}^m; \mathbb{R}^m)$;
3. σ is Lipschitz continuous and acts componentwise;
4. $\sigma(0) = 0$;
5. $\mathcal{L} \geq 0$ and is continuous in its first argument;

We note that the condition on $\mathcal{L}$ does not restrict it to be a typical loss function. There is no requirement for $\mathcal{L}$ to be a norm on the difference between its two arguments.

Our main result is the convergence of optimal parameters.

Theorem 2.1 *Let $\Theta^{(n)}$ and Θ be given by (14) and (15) respectively. Define $\mathcal{E}_n$, $\mathcal{E}_\infty$, E_n, E_∞, $R_n^{(i)}$, $R_\infty^{(i)}$, $R^{(j)}$ for $i = 1, 2$, $j = 3, 4$ as in Sect. 1.1-1.3. Let Assumptions 1 hold. Let $\{(x_s, y_s)\}_{s=1}^S$ be any given set of training data ($S \geq 1$). Then minimisers of $\mathcal{E}_n$ and $\mathcal{E}_\infty$ exist in $\Theta^{(n)}$ and Θ respectively. Furthermore let $\theta^{(n)} \subset \Theta^{(n)}$ be any sequence of minimisers of $\mathcal{E}_n$, then*

$$\min_{\Theta^{(n)}} \mathcal{E}_n = \mathcal{E}_n(\theta^{(n)}) \to \min_{\Theta} \mathcal{E}_\infty, \qquad \text{as } n \to \infty,$$

$\{\theta^{(n)}\}_{n\in\mathbb{N}}$ is relatively compact, and any limit point of $\{\theta^{(n)}\}_{n\in\mathbb{N}}$ is a minimiser of $\mathcal{E}_\infty$.

Through the following proposition, and under the additional and stronger assumptions in Assumptions 2, we remark that one obtains extra regularity on minimisers to the deep limit variational problem (as is to be expected based on elliptic regularity).

Assumptions 2 The following additional assumptions will be used in our regularity result:

1. $\sigma \in C^2(\mathbb{R}^d; \mathbb{R}^d)$;
2. $h \in C^2(\mathbb{R}^m; \mathbb{R}^m)$;
3. $\mathcal{L}(\cdot, y) \in C^2(\mathbb{R}^m; \mathbb{R})$ for all $y \in \mathbb{R}^m$;
4. all norms on $\mathbb{R}^d$ and $\mathbb{R}^{d\times d}$ are induced by inner products.

Proposition 2.2 *Let Assumptions 1 and 2 hold. Then any minimiser $\theta = (K, b, W, c) \in \Theta$ of $\mathcal{E}_\infty$ satisfies $K \in H^2_{\text{loc}}([0, 1]; \mathbb{R}^{d\times d})$ and $b \in H^2_{\text{loc}}([0, 1]; \mathbb{R}^d)$.*

The proof of the proposition is given in Sect. 4.4.

Theorem 2.1 states that, up to subsequences, minimisers of $\mathcal{E}_n$ converge to minimisers of $\mathcal{E}_\infty$. If the minimiser of $\mathcal{E}_\infty$ is unique then we have that the sequence of minimisers converges (without recourse to a subsequence) to the minimiser of $\mathcal{E}_\infty$. The proof of the

theorem relies on variational methods and is given in Sects. 4.1-4.3. We do not prove a convergence rate for the minimisers, but we conjecture a convergence rate of $\frac{1}{n}$. The conjecture is motivated by considering Taylor expansions for a fixed $\theta = (K, b, W, c) \in \Theta$; indeed one can show that for $K, b \in H^2$ the recovery sequence $\theta^{(n)}$ given by (26-29) satisfies

$$|\mathcal{E}_n(\theta^{(n)}) - \mathcal{E}_\infty(\theta)| \sim \frac{C(\theta)}{n},$$

where $C(\theta)$ is a constant that depends on $\|\ddot{K}\|_{L^2}$ and $\|\ddot{b}\|_{L^2}$. Assuming that this can be extended to minimising sequences (i.e. the above holds for any sequence of minimisers $\theta^{(n)} \to \theta$) one can conclude that the rate of convergence of the minima is $O(n^{-1})$. Making another conjecture that one can show a local bound of the form $d(\theta^{(n)}, \theta) \leq C \left|\mathcal{E}_n(\theta^{(n)}) - \mathcal{E}_\infty(\theta)\right|$ whenever $d(\theta^{(n)}, \theta)$ is small implies

$$d(\theta^{(n)}, \theta) = O\left(\frac{1}{n}\right).$$

We do not prove a rate of convergence for either the minimisers or the minimum here. However, we are able to show a rate of convergence for the forward pass through the Neural Network; more precisely, the output of the ResNet model is converging to the output of a dynamical system with the rate given by the following corollary.

Corollary 2.3 *Let Assumptions 1 hold. We use the matrix operator norm on* $\mathbf{K}^{(n)}$ *in* $R_n^{(1)}$ *and* $R_\infty^{(1)}$*, and let* $L_\sigma > 0$ *be a Lipschitz constant for* σ*. Let* $\{(x_s, y_s)\}_{s=1}^S \subset \mathbb{R}^d \times \mathbb{R}^m$ *be a set of training data and, for all* $n \in \mathbb{N}$*, let* $(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W^{(n)}, c^{(n)}) \in \operatorname{argmin}_{(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c)} \mathcal{E}_n(\mathbf{K}^{(n)}, \mathbf{b}^{(n)}, W, c)$*. Let* $x \in \mathbb{R}^d$*. Let* (K, b, W, c) *be the minimiser of* $\mathcal{E}_\infty$ *which we assume is unique. For all* $n \in \mathbb{N}$ *and for all* $i \in \{1, \dots, n\}$*, let* $X_i^{(n)}$ *be the solution to* (6) *with* $X_0^{(n)} = x$*, and* $X : [0, 1] \to \mathbb{R}^d$ *be the solution to the ODE in* (10) *(with coefficients* K *and* b*) with initial condition* $X(0) = x$*. Then, for all* $\delta > 0$*, there exists an* $N \in \mathbb{N}$ *such that, for all* $n \geq N$*, there exists an* $\varepsilon_n \in \mathbb{R}$ *such that, for all* $i \in \{0, 1, \dots, n\}$*,*

$$\|X(i/n) - X_i^{(n)}\| \leq \frac{n}{L_\sigma(\|K\|_{L^\infty} + \delta)} A_n \left[\exp\left(\frac{i}{n} L_\sigma(\|K\|_{L^\infty} + \delta)\right) - 1\right], \tag{13}$$

where

$$A_n = \frac{1}{n}(1 + \|X\|_{L^\infty}) L_\sigma \delta + \varepsilon_n.$$

Moreover, $\varepsilon_n = o\left(\frac{1}{n}\right)$ *as* $n \to \infty$*.*

We provide the proof of Corollary 2.3 in Sect. 4.5.

Remark 2.4 In Corollary 2.3 we made the assumption that the minimiser (K, b, W, c) of $\mathcal{E}_\infty$ is unique, mainly to keep the notation as simple as possible. If minimisers of $\mathcal{E}_\infty$ are not unique, we need to be more careful in our statement of the corollary. In that case, by Theorem 2.1 there exists a minimiser (K, b, W, c) of $\mathcal{E}_\infty$ such that, *up to subsequences*, $K^{(n)} \to K$ and $b^{(n)} \to b$ as $n \to \infty$ in the topology of Sect. 3.2 (where the same indices can be chosen for both subsequences). Taking $\mathcal{N}$ to be the (infinite) set containing the (common) indices n of these subsequences, the statement of the corollary still holds if we restrict the indices n to the set $\mathcal{N}$ instead of allowing them to vary over all of $\mathbb{N}$.

3 Background material

In this section we give background material necessary to present the proofs of the main results. In particular, we start by clarifying our notation. We then give a description on the discrete-to-continuum topology. Finally, for the convenience of the reader, we give a brief overview on Γ-convergence.

3.1 Notation

Let Ω be an open subset of a Euclidean space. Given a probability measure $\mu \in \mathcal{P}(\Omega)$ on Ω we write $L^p(\mu; \Xi)$ for the set of functions from Ω to Ξ that are L^p integrable with respect to μ, when appropriate we will shorten notation to $L^p(\mu)$. The $L^p(\mu)$ norm for a function $f : \Omega \to \Xi$ is denoted by $\|f\|_{L^p(\mu)}$. When μ is the Lebesgue measure on Ω we will also write L^p or $L^p(\Omega)$ for $L^p(\mu; \Xi)$, and $\|f\|_{L^p}$ or $\|f\|_{L^p(\Omega)}$ for $\|f\|_{L^p(\mu)}$. The L^2 inner product with respect to the Lebesgue measure is denoted by $\langle \cdot, \cdot, \rangle_{L^2}$. The Sobolev space of functions that are k-times weakly differentiable and with each weak derivative in L^2 is denoted by H^k. In order to make clear the domain Ω and range Ξ of H^k, we will also write $H^k(\Omega; \Xi)$ (in order to avoid complications defining derivatives, the underlying measure in Sobolev spaces when $k > 0$ is always the Lebesgue measure). For functions $f : \Omega \to \mathbb{R}$ that are k times continuously differentiable, we write $f \in C^k(\Omega)$, and if all its kth partial derivatives are Hölder continuous with exponent γ —i.e. if $\max_{\alpha:|\alpha|=k} \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{\|x-y\|^\gamma} < \infty$, where the α are multi-indices with sum $|\alpha|$ equal to k and $D^\alpha f$ denotes the corresponding partial derivative of f of order k— we write $f \in C^{k,\gamma}(\Omega)$.

We often do not specify a matrix or vector norm, clearly these are finite-dimensional spaces and therefore all norms are topologically equivalent. If $b \in \mathbb{R}^d$ is a vector and $K \in \mathbb{R}^{d\times d}$ is a matrix then we will write $\|b\|$ and $\|K\|$ for both the vector norm and the matrix norm. In particular we point out that we only use subscripts for L^p norms. Sometimes we will need that the norms are induced by inner products, we will write when we need this additional structure.

We use superscripts on the parameters $\mathbf{K}^{(n)}$ and $\mathbf{b}^{(n)}$ (later denoted $K^{(n)}$ and $b^{(n)}$) in order to clearly denote the dependence of the number of layers on the parameters themselves (this is particularly important as we take the limit $n \to \infty$). The parameters W, c are respectively a $m \times d$ matrix and a m-dimensional vector and therefore we do not include any reference to n unless we are considering sequences.

Vectors are always column vectors. For two vectors $A, B \in \mathbb{R}^\kappa$ we use $\odot$ to denote componentwise multiplication, i.e. $A \odot B = [A_1B_1, A_2B_2, \ldots, A_\kappa B_\kappa]^\top$. When $A \in \mathbb{R}^\kappa$ and $C \in \mathbb{R}^{\kappa\times d}$ then $\odot$ represents row-wise multiplication, i.e.

$$A \odot C = C \odot A = \begin{bmatrix} A_1C_{11} & A_1C_{12} & \cdots & A_1C_{1d} \\ A_2C_{21} & A_2C_{22} & \cdots & A_2C_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ A_\kappa C_{\kappa 1} & A_\kappa C_{\kappa 2} & \cdots & A_\kappa C_{\kappa d} \end{bmatrix}.$$

We can also interpret this product as $A \odot C = \mathrm{diag}(A)C$, where diag is the diagonal $\kappa \times \kappa$ matrix with the vector A on its diagonal.

We use the convention that $0 \notin \mathbb{N}$.

3.2 Discrete-to-continuum topology

We introduced the parameters for the ResNet model with n layers $\mathbf{K}^{(n)}$ and $\mathbf{b}^{(n)}$ as sets of matrices/vectors, i.e. $\mathbf{K}^{(n)} = \{K_i^{(n)}\}_{i=0}^{n-1} \subset \mathbb{R}^{d\times d}$ and $\mathbf{b}^{(n)} = \{b_i^{(n)}\}_{i=0}^{n-1} \subset \mathbb{R}^d$. In fact it is more convenient to think of them as functions with respect to the discrete measure $\mu_n = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\frac{i}{n}}$ on $[0,1]$. More precisely, for $\mathbf{K}^{(n)}$ we make the identification with $K^{(n)} \in L^0(\mu_n;\mathbb{R}^{d\times d})$ by $K^{(n)}(i/n) = K_i^{(n)}$. In the sequel we will, with a small abuse of notation, write both $K^{(n)}$ and $K_i^{(n)}$, where the former is understood as a function in $L^0(\mu_n;\mathbb{R}^{d\times d})$ and the latter as the matrix $K_i^{(n)} = K^{(n)}(i/n) \in \mathbb{R}^{d\times d}$. Similarly for $b^{(n)}$ and $b_i^{(n)}$.

With this notation we can define the finite layer parameter space by

$$\Theta^{(n)} = L^2(\mu_n;\mathbb{R}^{d\times d}) \times L^2(\mu_n;\mathbb{R}^d) \times \mathbb{R}^{m\times d} \times \mathbb{R}^m. \tag{14}$$

For any $p, q > 0$ the discrete spaces $L^p(\mu_n), L^q(\mu_n)$ are topologically equivalent. Since we apply a discrete analogue of an H^1 regularisation penalty to parameters of the neural network (see also Sect. 3.4), it will transpire that the natural limiting space to work in is given by

$$\Theta = H^1([0,1];\mathbb{R}^{d\times d}) \times H^1([0,1];\mathbb{R}^d) \times \mathbb{R}^{m\times d} \times \mathbb{R}^m. \tag{15}$$

Given $K \in L^2([0,1];\mathbb{R}^{d\times d})$ and $K^{(n)} \in L^2(\mu_n;\mathbb{R}^{d\times d})$ we define a distance by extending $K^{(n)}$ to a function on $[0,1]$ by $\tilde{K}^{(n)}(t) = K^{(n)}(t_i)$ (for $t \in (t_{i-1}, t_i)$, $t_i = i/n$, $i = 1,\dots,n$) and comparing in L^2; that is

$$d_1(K^{(n)}, K) = \|\tilde{K}^{(n)} - K\|_{L^2}.$$

We note the distance d_1 is closely related to the TL^2 distance, see [30], when the discrete measure is of the form $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{t_i}$ and the domain is $[0,1]$. The TL^2 distance (see (16) below), or more generally the TL^p distance, is a topology useful for metrising discrete-to-continuum convergence on a general domain $\Omega \subseteq \mathbb{R}^d$. The idea is to think of functions on different domains as the coupling of a function with a measure; that is the TL^p space is the space of pairs (μ, K) where $K \in L^p(\mu)$ and μ is a probability measure on Ω with finite pth moment. The TL^p space is metrised by a Wasserstein distance on the graphs of functions. To compare a discrete function $K^{(n)} : \{x_i\}_{i=1}^n \to \mathbb{R}$ with associated discrete measure $\mu_n = \sum_{i=1}^n \delta_{x_i}$ to a continuum function $K : \Omega \to \mathbb{R}$ with the continuum measure $\mu \in \mathcal{P}(\Omega)$ associated with Ω, we perform the following steps. Firstly, we find an "optimal" partitioning of equal mass of the underlying state space, $T^{(n)} : \Omega \to \{x_i\}_{i=1}^n$ (where optimal is in the sense of solving an optimal transport problem between the discrete measure μ_n and the continuum measure μ.). Secondly, we extend $K^{(n)}$ to Ω to be piecewise constant, i.e. $\tilde{K}^{(n)} = K^{(n)} \circ T^{(n)} : \Omega \to \mathbb{R}$. Lastly we compare $\tilde{K}^{(n)}$ to K in an L^p norm. The notation TL^p stands for "transport" and "L^p". We leave further details of the topology that is constructed in this way to [30].

To make the connection between d_1 and TL^2 precise consider the following. For pairs $(\mu, K), (\nu, L)$ where $\mu, \nu \in \mathcal{P}(\Omega)$ and $K \in L^2(\mu)$, $L \in L^2(\nu)$ the TL^2 distance is defined (in the Kantorovich formulation) by

$$d_{TL^2}^2((\mu, K), (\nu, L)) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} \|x - y\|^2 + \|K(x) - L(y)\|^2 \, \mathrm{d}\pi(x, y). \tag{16}$$

Here $\Pi(\mu, \nu)$ denotes the set of all Borel probability measures on $\Omega \times \Omega$ whose marginals on the first and second variable are μ and ν, respectively (so-called couplings). We say that $(\mu_n, K^{(n)}) \to (\mu, K)$ in TL^2 if $d_{TL^2}^2((\mu_n, K^{(n)}), (\mu, K)) \to 0$. In the Monge formulation

we write

$$d^2_{TL^2}((\mu, K), (\nu, L)) = \inf_{T : T_\# \mu = \nu} \int_\Omega \|x - T(x)\|^2 + \|K(x) - L(T(x))\|^2 \, \mathrm{d}\mu(x).$$

We note that the Monge formulation is not always defined as there may not exist transport maps T between μ and ν. Comparing the metric in TL^2 to the Wasserstein distance

$$d^2_W(\mu, \nu) = \begin{cases} \inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} \|x - y\|^2 \, \mathrm{d}\pi(x, y) & \text{in the Kantorovich formulation} \\ \inf_{T : T_\# \mu = \nu} \int_\Omega \|x - T(x)\|^2 \, \mathrm{d}\mu(x) & \text{in the Monge formulation,} \end{cases}$$

we can see that $d_{TL^2}((\mu, K), (\nu, L)) = d_W((\mathrm{Id} \times K)_\# \mu, (\mathrm{Id} \times L)_\# \nu)$.

In our case we choose μ to be the Lebesgue measure on $[0, 1]$, $\nu = \mu_n$ the discrete measure defined above, and $L = K^{(n)}$. It is a consequence of results in [30] (since μ_n converges weakly* to μ—we say weak* to be consistent with notation in functional analysis rather than weak which is often the notation in statistics) that

$$d_{TL^2}((\mu, K), (\mu_n, K^{(n)})) \to 0 \quad \Leftrightarrow \quad d_1(K^{(n)}, K) \to 0.$$

More precisely, in our setting $d_{TL^2}((\mu, K), (\mu_n, K^{(n)})) \to 0$ is equivalent to $\mu_n \overset{*}{\rightharpoonup} \mu$ and the existence of a sequence of transport maps $T^{(n)}$ (between μ_n and μ) such that $T^{(n)} \to \mathrm{Id}$ in L^2 and $K^{(n)} \circ T^{(n)} \to K$ in L^2. The existence of such a sequence $T^{(n)}$ is guaranteed in our case, since we can choose $T^{(n)}(t) = \bar{t}_i$ for $t \in (\bar{t}_{i-1}, \bar{t}_i)$ where $\bar{t}_i = \frac{i}{n-1}$, which leads to $K^{(n)} \circ T^{(n)}$ being a piecewise constant interpolation of $K^{(n)}$. Hence, we can use the simpler function d_1. We note that d_1 is not a metric (for example $d_1(K, K^{(n)})$ does not make sense; hence, d_1 is not symmetric), however due to the relationship of d_1 with d_{TL^2} we can still take advantage of metric properties.

Similarly, we define $d_2(b^{(n)}, b) = \|\tilde{b}^{(n)} - b\|_{L^2}$ and the distance between $\theta = (K, b, W, c)$ and $\theta^{(n)} = (K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)})$ is given by

$$\begin{aligned} &d : \Theta^{(n)} \times \Theta \mapsto [0, \infty) \\ &\quad d(\theta^{(n)}, \theta) = d_1(K^{(n)}, K) + d_2(b^{(n)}, b) + \|W^{(n)} - W\| + \|c^{(n)} - c\|. \end{aligned} \tag{17}$$

We could also have used a piecewise linear interpolation rather than the piecewise constant interpolation we use, i.e. we could have defined

$$\begin{aligned} \bar{K}^{(n)}(t) &= \frac{\bar{t}_i - t}{\bar{t}_i - \bar{t}_{i-1}} K^{(n)}_{i-1} + \frac{t - \bar{t}_{i-1}}{\bar{t}_i - \bar{t}_{i-1}} K^{(n)}_i \quad \text{for } t \in (\bar{t}_{i-1}, \bar{t}_i) \\ &\quad i = 1, \ldots, n-1 \text{ where } \bar{t}_i = \frac{i}{n-1}, \end{aligned}$$

and compared $\bar{d}_1(K^{(n)}, K) := \|\bar{K}^{(n)} - K\|_{L^2}$. However, under appropriate conditions (which are satisfied in this paper)

$$\bar{d}_1(K^{(n)}, K) \to 0 \quad \Leftrightarrow \quad d_1(K^{(n)}, K) \to 0.$$

Since the piecewise constant and piecewise linear constructions both generate the TL^2 topology, we choose the simpler former one. This also gives us a metric space structure that we can use to establish Γ-limits.

3.3 Γ-Convergence

Recall that we wish to show minimisers of $\mathcal{E}_n$ converge to minimisers of $\mathcal{E}_\infty$. In particular, we want to show that $\mathcal{E}_\infty$ is the variational limit of $\mathcal{E}_n$. To characterise variational convergence we first define the Γ-limit in a general metric space setting.

Definition 3.1 Let $\mathcal{E}_n : \Omega \to \mathbb{R} \cup \{\pm\infty\}$, $\mathcal{E}_\infty : \Omega \to \mathbb{R} \cup \{+\infty\}$ where (Ω, d) is a metric space. Then $\mathcal{E}_n$ Γ-converges to $\mathcal{E}_\infty$, and we write $\mathcal{E}_\infty = 0\text{-}\lim_{n\to\infty} \mathcal{E}_n$, if for all $u \in \Omega$ the following holds:

1. (the liminf inequality) for any $u_n \to u$

$$\liminf_{n\to\infty} \mathcal{E}_n(u_n) \geq \mathcal{E}_\infty(u);$$

2. (the recovery sequence) there exists $u_n \to u$ such that

$$\limsup_{n\to\infty} \mathcal{E}_n(u_n) \leq \mathcal{E}_\infty(u).$$

For brevity we focus only on the key property of Γ-convergence, and the property that justifies the term variational convergence. For a more substantial introduction to Γ-convergence, we refer to [5,18].

Theorem 3.2 *Let (Ω, d) be a metric space and $\mathcal{E}_n$ a proper sequence of functionals on Ω. Let u_n be a sequence of almost minimisers for $\mathcal{E}_n$, i.e. $\mathcal{E}_n(u_n) \leq \max\{\inf_{u\in\Omega} \mathcal{E}_n(u_n) + \varepsilon_n, -\frac{1}{\varepsilon_n}\}$ for some $\varepsilon_n \to 0^+$. Assume that $\mathcal{E}_\infty = 0\text{-}\lim_{n\to\infty} \mathcal{E}_n$ and $\{u_n\}_{n=1}^\infty$ are relatively compact. Then,*

$$\inf_{u\in\Omega} \mathcal{E}_n(u) \to \min_{u\in\Omega} \mathcal{E}_\infty(u)$$

where the minimum of $\mathcal{E}_\infty$ exists. Moreover if $u_{n_m} \to u_\infty$ is a convergent subsequence then u_∞ minimises $\mathcal{E}_\infty$.

Clearly if one assumes that the minimum of $\mathcal{E}_\infty$ is unique then, by the above theorem, $u_n \to u_\infty$ (without recourse to subsequences) where u_∞ is the unique minimiser of $\mathcal{E}_\infty$.

Theorem 3.2 forms the basis for our proof of Theorem 2.1. In order to apply Theorem 3.2, we must show that minimisers are relatively compact and $\mathcal{E}_\infty = 0\text{-}\lim_{n\to\infty} \mathcal{E}_n$.

We note that Definition 3.1 and Theorem 3.2 are in the context of metric spaces. As we described in Sect. 3.2 we can describe the convergence of $K^{(n)}$ in terms of the TL^2 distance d_{TL^2} which is a metric on the space $\Omega = \{(\mu, f) : f \in L^2(\mu; \mathbb{R}^{d\times d}), \mu \in \mathcal{P}([0,1])\}$ (and similarly for $b^{(n)}$). Hence, we can use the distance

$$\tilde{d}\left((\mu, K, b, W, c), (\nu, L, a, V, d)\right) = d_{TL^2}((\mu, K), (\nu, L)) + d_{TL^2}((\mu, b), (\nu, a)) + \|W - V\| + \|c - d\|$$

which is a metric on the space

$$\left\{(\mu, K, b, W, c) : K \in L^2(\mu; \mathbb{R}^{d\times d}), b \in L^2(\mu; \mathbb{R}^d), W \in \mathbb{R}^{d\times d}, c \in \mathbb{R}^d, \mu \in \mathcal{P}([0,1])\right\}.$$

Since convergence in $\tilde{d}$ is equivalent to convergence in d, we can simplify our notation by considering sequences that converge in d whilst still being able to apply Theorem 3.2.

3.4 Sobolev spaces

For readers unfamiliar with Sobolev spaces, in this section we provide some definitions and results that are needed to read the remainder of the current paper. For a more detailed introduction and further in-depth study of these concepts we refer the reader to [1,62].

We define the Sobolev space $H^k([0,1])$ recursively: for $k \geq 2$, $f \in H^k([0,1])$ if $\dot{f} \in H^{k-1}([0,1])$ where $\dot{f}$ is the weak derivative of f, and for $k = 1$, $f \in H^1([0,1])$ if $f \in L^2([0,1])$ and $\dot{f} \in L^2([0,1])$. We can replace H^k with H^k_{loc} by replacing L^2 with L^2_{loc} in the

previous definition (where $L^2_{\text{loc}}([0,1])$ is the set of functions that are in $L^2([a,b])$ for every $0 < a < b < 1$). The Sobolev norm in $H^1([0,1])$ is defined by

$$\|f\|_{H^1([0,1])} = \|f\|_{L^2([0,1])} + \|\dot{f}\|_{L^2([0,1])}.$$

Of course these definitions extend to p-norms and functions of several variables [62].

Morrey's inequality in one dimension [62, Theorem 11.34] implies that there exists a constant C such that $\|f\|_{C^{0,\frac{1}{2}}} \leq C\|f\|_{H^1}$ for all $f \in H^1$. We note in particular that such an f has a continuous representative, so that the pointwise evaluation $f(0)$ is well-defined. Therefore,

$$|f(0)| + \|\dot{f}\|_{L^2} \leq \|f\|_{C^0} + \|\dot{f}\|_{L^2} \leq (C+1)\|f\|_{H^1}.$$

Moreover, for any $f \in H^1$ we have $|f(x) - f(y)| \leq C\sqrt{|x-y|}\|\dot{f}\|_{L^2}$ [62, Remark 11.35] so that $|f(x)| \leq |f(0)| + C\sqrt{|x|}\|\dot{f}\|_{L^2} \leq |f(0)| + C\|\dot{f}\|_{L^2}$ implying

$$\|f\|_{H^1} \leq |f(0)| + (C+1)\|\dot{f}\|_{L^2}.$$

It follows that $|f(0)| + \|\dot{f}\|_{L^2}$ and $\|f\|_{H^1}$ are equivalent norms. In particular, our regularisation terms $R^{(1)}_\infty$ and $R^{(2)}_\infty$ in the deep layer limit (see (11)) are equivalent to H^1 norms. We furthermore have the Rellich–Kondrachov type embedding result that $H^1([0,1])$ is compactly embedded in both $C^{0,\frac{1}{2}}([0,1])$ and $L^2([0,1])$ [62, Section 11.3].

Although the finite layer regularisation uses a finite difference approximation of the derivative (as one cannot use usual derivatives in discrete spaces), one can expect minimisers of $\mathcal{E}_n$ to enjoy similar regularity properties in the deep layer limit when the scale in the discretisation goes to zero, as minimisers of $\mathcal{E}_\infty$ have.

Non-local characterisations of Sobolev spaces are possible, see for example [62, Theorem 10.55], and we utilise such ideas to prove Γ-convergence.

4 Proofs

The proof of Theorem 2.1 is a straightforward application of the following theorem, Theorem 4.1, with Theorem 3.2. This section is devoted to the proofs of Theorem 4.1, Proposition 2.2 and Corollary 2.3.

Theorem 4.1 *Under the assumptions of Theorem 2.1, the following holds:*

1. *for every $n \in \mathbb{N}$ there exists a minimiser of $\mathcal{E}_n$ in $\Theta^{(n)}$,*
2. *any sequence $\{(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)})\}_{n\in\mathbb{N}}$ which is bounded in $\mathcal{E}_n$, i.e.*

$$\sup_{n\in\mathbb{N}} \mathcal{E}_n(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)}) < \infty,$$

   is relatively compact, and
3. $0\text{-}\lim_{n\to\infty} \mathcal{E}_n = \mathcal{E}_\infty$.

The first three subsections are each dedicated to the proof of one part of the above theorem. In Sect. 4.1 we show that sequences bounded in $\mathcal{E}_n$ are relatively compact. The argument relies on approximating discrete sequences $\theta^{(n)} = (K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)}) \in \Theta^{(n)}$ with a continuum sequence $\tilde{\theta}^{(n)} = (\tilde{K}^{(n)}, \tilde{b}^{(n)}, W^{(n)}, c^{(n)}) \in \Theta$ and using standard Sobolev embedding arguments to deduce the compactness of $\tilde{\theta}^{(n)}$, and therefore, $\theta^{(n)}$.

In Sect. 4.2 we prove the existence of minimisers. The strategy is to apply the direct method from the calculus of variations. That is, we show that $\mathcal{E}_n$ is lower semi-continuous (in fact continuous). For compactness of minimising sequences it is enough to show

bounded in norm (since for finite n parameters are finite-dimensional). Compactness plus lower semi-continuity is enough to imply the existence of minimisers.

In the third subsection we prove the Γ-convergence of $\mathcal{E}_n$ to $\mathcal{E}_\infty$. This relies on a variational convergence of finite differences.

In Sect. 4.4 we analyse the regularity of minimisers of $\mathcal{E}_\infty$ and prove Proposition 2.2. To show this we compute the Gâteaux derivative then apply methods from elliptic regularity theory to infer additional smoothness. In this section we assume that the norms $\|\cdot\|$ on $\mathbb{R}^d$ and $\mathbb{R}^{d\times d}$ are induced by an inner product $\langle\cdot,\cdot\rangle$.

Finally, in Sect. 4.5 we prove the uniform convergence of the parameters of the neural network to parameters of the continuum model (Corollary 2.3).

4.1 Proof of compactness

We start with a preliminary result which implies that $\|K^{(n)}\|_{L^\infty(\mu_n)} \leq CR_n^{(1)}(K^{(n)})$, this is a discrete analogue of the well-known Morrey's inequality. We include the proof as it is important that the constant C can be chosen independently of μ_n.

In the following, where we write $\mathbb{R}^\kappa$, κ can be any integer. Specific choices for κ will be made when the result is applied.

Proposition 4.2 *Fix $n \in \mathbb{N}$ and let $t_i = \frac{i}{n}$, $\mu_n = \frac{1}{n}\sum_{i=0}^{n-1}\delta_{t_i}$, and $f_n : \{t_i\}_{i=0}^{n-1} \to \mathbb{R}^\kappa$. Then*

$$\|f_n\|^2_{L^\infty(\mu_n)} \leq 2\left(\|f_n(t_0)\|^2 + n\sum_{j=1}^{n-1}\|f_n(t_j) - f_n(t_{j-1})\|^2\right).$$

Proof We note that

$$\|f_n(t_i) - f_n(t_0)\|^2 \leq \left(\sum_{j=1}^{i}\|f_n(t_j) - f_n(t_{j-1})\|\right)^2 \leq n\sum_{j=1}^{n-1}\|f_n(t_j) - f_n(t_{j-1})\|^2$$

by Jensen's inequality. Hence,

$$\begin{aligned}\|f_n(t_i)\|^2 &\leq 2\left(\|f_n(t_i) - f_n(t_0)\|^2 + \|f_n(t_0)\|^2\right)\\ &\leq 2\left(\|f_n(t_0)\|^2 + n\sum_{j=1}^{n-1}\|f_n(t_j) - f_n(t_{j-1})\|^2\right).\end{aligned}$$

Taking the supremum over $i \in \{0, 1, \ldots, n-1\}$ proves the proposition. □

Let $(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)}) \in \Theta^{(n)}$ be a sequence such that $\sup_{n\in\mathbb{N}} \mathcal{E}_n(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)}) < +\infty$. Then compactness of $\{W^{(n)}\}_{n\in\mathbb{N}}$ and $\{c^{(n)}\}_{n\in\mathbb{N}}$ is immediate from the regularisation functionals $R^{(3)}$ and $R^{(4)}$. For $K^{(n)}$ and $b^{(n)}$ we deduce compactness by using a smooth continuum approximation. In particular, let $f_n : \{t_i\}_{i=0}^{n-1} \to \mathbb{R}^\kappa$, where $t_i = \frac{i}{n}$, be a sequence of discrete functions that are bounded in the discrete H^1 norm $\mathcal{R}_n$ given by

$$\mathcal{R}_n(f_n) = \sqrt{\|f_n(t_0)\|^2 + n\sum_{j=1}^{n-1}\|f_n(t_j) - f_n(t_{j-1})\|^2}.$$

We compare f_n to a smooth continuum function $g_n : [0,1] \to \mathbb{R}^\kappa$ with the property $\|g_n\|_{H^1} \lesssim \mathcal{R}_n(f_n)$. By Sobolev embedding arguments we have that $\{g_n\}_{n\in\mathbb{N}}$ is relatively

compact in $L^2([0.1];\mathbb{R}^\kappa)$. Compactness of $\{f_n\}_{n\in\mathbb{N}}$ follows from $\|f_n \circ T_n - g_n\|_{L^2} \to 0$ where T_n is the map $T_n(t) = t_i$ if $t \in [t_i, t_{i+1})$.

In the following proposition $\mathrm{Leb}\lfloor_{[0,1]}$ is the Lebesgue measure on $\mathbb{R}$ restricted to the interval $[0, 1]$.

Proposition 4.3 *For each $n \in \mathbb{N}$ let $t_i^{(n)} = \frac{i}{n}$, $\mu_n = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{t_i^{(n)}}$, and $f_n : \{t_i^{(n)}\}_{i=0}^{n-1} \to \mathbb{R}^\kappa$. If*

$$\sup_{n\in\mathbb{N}} \left(\|f_n(0)\|^2 + n\sum_{j=1}^{n-1} \|f_n(t_j^{(n)}) - f_n(t_{j-1}^{(n)})\|^2 \right) < +\infty \tag{18}$$

then $\{(\mu_n, f_n)\}_{n\in\mathbb{N}}$ is relatively compact in TL^2 and any cluster point (μ, f) satisfies $\mu = \mathrm{Leb}\lfloor_{[0,1]}$ and $f \in C^{0,\gamma}([0, 1];\mathbb{R}^\kappa)$ for any $\gamma < \frac{1}{2}$. Furthermore, for any converging subsequence there exists a further subsequence (which we relabel), and a $f \in C^{0,\gamma}([0, 1];\mathbb{R}^\kappa)$, such that

$$\max_{i\in\{0,1,\ldots,n-1\}} \left\| f_n(t_i^{(n)}) - f(t_i^{(n)}) \right\| \to 0. \tag{19}$$

Proof First note that

$$\begin{aligned}
\|f_n(t_i^{(n)})\| &\le \|f_n(0)\| + \sum_{j=1}^{i-1} \|f_n(t_j^{(n)}) - f_n(t_{j-1}^{(n)})\| \\
&\le \|f_n(0)\| + \sum_{j=1}^{n-1} \|f_n(t_j^{(n)}) - f_n(t_{j-1}^{(n)})\| \\
&\le 1 + \frac{1}{2}\left(\|f_n(0)\|^2 + n\sum_{j=1}^{n-1} \|f_n(t_j^{(n)}) - f_n(t_{j-1}^{(n)})\|^2 \right),
\end{aligned}$$

with the last line following from Young's inequality, so by (18) $\|f_n\|_{L^\infty(\mu_n)}$ is bounded. In particular, there exists $M < +\infty$ such that

$$\|f_n(0)\|^2 + n\sum_{j=1}^{n-1} \|f_n(t_j^{(n)}) - f_n(t_{j-1}^{(n)})\|^2 \le M, \qquad \|f_n\|_{L^\infty(\mu_n)} \le M.$$

Let $\tilde{f}_n$ be the continuum extension of f_n defined by

$$\tilde{f}_n(t) = \begin{cases} f_n(0) & \text{if } t < 0 \\ f_n(t_i^{(n)}) & \text{if } t \in [t_i^{(n)}, t_{i+1}^{(n)}) \text{ for some } i = 0, \ldots, n-1 \\ f_n(t_{n-1}^{(n)}) & \text{if } t \ge 1. \end{cases}$$

Define $g_n = J_{\varepsilon_n} * \tilde{f}_n$ where $J \in C^\infty(\mathbb{R})$ is a standard mollifier [62, Remark C.18(ii)] with $\|J\|_{L^1} = 1$, $\|J\|_{L^\infty} \le \beta$, for some $\beta > 0$, $J_\varepsilon = \frac{1}{\varepsilon}J(t/\varepsilon)$, for all $\varepsilon > 0$, and $\varepsilon_n = \frac{1}{2n}$. We recall the following facts about mollifiers (which are stated in the domain $\mathbb{R}$, but hold for higher-dimensional Euclidean spaces as well):

(M1) $\|J_\varepsilon * f\|_{L^\infty} \le \|f\|_{L^\infty}$, for any $f \in L^\infty$ and any $\varepsilon > 0$ (by Young's inequality [62, Theorem C.15]);

(M2) $\frac{\mathrm{d}}{\mathrm{d}t}(J_\varepsilon * f) = \frac{1}{\varepsilon}(\dot{J})_\varepsilon * f$, for any $f \in L^1$ and any $\varepsilon > 0$, where $(\dot{J})_\varepsilon(t) = \frac{1}{\varepsilon}\dot{J}(t/\varepsilon)$ [62, Theorem C.20];

(M3) $\int_{\mathbb{R}} (\dot{J})_\varepsilon(s)\,\mathrm{d}s = 0$ (since the order of integration and differentiation can be reversed, as $(\dot{J})_\varepsilon$ is continuous and supported on a compact subset of $\mathbb{R}$);

(M4) $\|(\dot{J})_\varepsilon\|_{L^\infty} \leq \frac{\|J\|_{L^\infty}}{\varepsilon}$, for any $\varepsilon > 0$.

We first show that g_n is bounded in $H^1([0,1];\mathbb{R}^\kappa)$. As $\|f_n\|_{L^\infty(\mu_n)}$ is bounded then $\|\tilde{f}_n\|_{L^\infty([0,1])}$ is bounded, so by (M1) g_n is bounded in $L^\infty([0,1];\mathbb{R}^\kappa)$. It is therefore sufficient to show that $\sup_{n\in\mathbb{N}} \|\dot{g}_n\|_{L^2} < +\infty$. For $t \in [t_i^{(n)}, t_i^{(n)} + \varepsilon_n]$ and $i \geq 1$ we have,

$$\begin{aligned}
\|\dot{g}_n(t)\| &= \frac{1}{\varepsilon_n} \left\| (\dot{J})_{\varepsilon_n} * \tilde{f}_n(t) \right\| \qquad \text{by (M2)} \\
&= \frac{1}{\varepsilon_n} \left\| \int_{\mathbb{R}} (\dot{J})_{\varepsilon_n}(t-s) \tilde{f}_n(s)\,\mathrm{d}s \right\| \\
&= \frac{1}{\varepsilon_n} \left\| \int_{\mathbb{R}} (\dot{J})_{\varepsilon_n}(t-s) \left(\tilde{f}_n(s) - \tilde{f}_n(t) \right) \mathrm{d}s \right\| \qquad \text{by (M3)} \\
&\leq \frac{\beta}{\varepsilon_n^2} \int_{\mathbb{R}} \left\| \tilde{f}_n(s) - \tilde{f}_n(t) \right\| \mathrm{d}s \qquad \text{by (M4)} \\
&= 4\beta n \left\| f_n(t_i^{(n)}) - f_n(t_{i-1}^{(n)}) \right\|.
\end{aligned}$$

Similarly, for $t \in [t_i^{(n)} + \varepsilon_n, t_{i+1}^{(n)}]$ and $i \leq n-2$ we have,

$$\|\dot{g}_n(t)\| \leq 4n\beta \|f_n(t_{i+1}^{(n)}) - f_n(t_i^{(n)})\|.$$

From the definition of $\tilde{f}_n$ we have that $\dot{g}_n(t) = 0$ for all $t \leq \varepsilon_n$ or $t \geq 1 - \varepsilon_n$. Squaring and integrating the above inequality over $t \in [0,1]$ implies

$$\|\dot{g}_n\|_{L^2}^2 \leq 16\beta^2 n \sum_{i=1}^{n-1} \|f_n(t_i^{(n)}) - f_n(t_{i-1}^{(n)})\|^2.$$

Hence, g_n is bounded in $H^1([0,1];\mathbb{R}^\kappa)$.

By Morrey's inequality [62, Theorem 11.34], g_n is relatively compact in $C^{0,\gamma}([0,1];\mathbb{R}^\kappa)$ for any $\gamma \in (0, \frac{1}{2})$. In particular g_n is relatively compact in $L^\infty([0,1];\mathbb{R}^\kappa)$. Hence, we may assume that there exists a subsequence (which we relabel) and $g \in C^{0,\gamma}$ such that $g_n \to g$ in $L^\infty([0,1];\mathbb{R}^\kappa)$. The proposition is proved once we show $\|\tilde{f}_n - g_n\|_{L^\infty} \to 0$. For $t \in [t_i^{(n)}, t_i^{(n)} + \varepsilon_n]$ we have

$$\begin{aligned}
\|\tilde{f}_n(t) - g_n(t)\| &= \left\| \int_{\mathbb{R}} J_{\varepsilon_n}(s-t) \left(\tilde{f}_n(s) - \tilde{f}_n(t) \right) \mathrm{d}s \right\| \\
&\leq \frac{\beta}{\varepsilon_n} \int_{t_{i-1}^{(n)}}^{t_{i+1}^{(n)}} \|\tilde{f}_n(s) - \tilde{f}_n(t)\|\,\mathrm{d}s \\
&= \begin{cases} 2\beta \|f_n(t_i^{(n)}) - f_n(t_{i-1}^{(n)})\| & \text{if } i \geq 1 \\ 0 & \text{if } i = 0. \end{cases}
\end{aligned}$$

Similarly, for $t \in [t_i^{(n)} + \varepsilon_n, t_{i+1}^{(n)}]$ we have

$$\|\tilde{f}_n(t) - g_n(t)\| \leq \begin{cases} 2\beta \|f_n(t_{i+1}^{(n)}) - f_n(t_i^{(n)})\| & \text{if } i \leq n-2 \\ 0 & \text{if } i = n-1. \end{cases}$$

Hence

$$\|\tilde{f}_n - g_n\|_{L^\infty}^2 \leq 4\beta^2 \sup_{i\in\{1,\ldots,n-1\}} \|f_n(t_i^{(n)}) - f_n(t_{i-1}^{(n)})\|^2$$

$$\leq 4\beta^2 \sum_{i=1}^{n-1} \|f_n(t_i^{(n)}) - f_n(t_{i-1}^{(n)})\|^2$$
$$= O\left(\frac{1}{n}\right).$$

It follows that $\tilde{f}_n \to g$ in $L^\infty([0,1];\mathbb{R}^\kappa)$ (and therefore in $L^2([0,1];\mathbb{R}^\kappa)$) which proves (19). Clearly $\mu_n \overset{*}{\rightharpoonup} \text{Leb}\lfloor_{[0,1]}$; hence, $(\mu_n, f_n) \to (\text{Leb}\lfloor_{[0,1]}, g)$ in the TL^2 topology, from which it follows that $\{(\mu_n, f_n)\}_{n\in\mathbb{N}}$ is relatively compact in TL^2. □

Compactness of sequences bounded in $\mathcal{E}_n$ is now a simple corollary of the above proposition.

Corollary 4.4 *Let $\Theta^{(n)}$ and Θ be given by (14) and (15) respectively. Define $\mathcal{E}_n$, $\mathcal{E}_\infty$, E_n, E_∞, $R_n^{(i)}$, $R_\infty^{(i)}$, $R^{(j)}$ for $i = 1, 2$, $j = 3, 4$ as in Sects. 1.1-1.3. Assume that $\alpha_i > 0$ for $i = 1, 2, 3, 4$, $\tau_j > 0$ for $j = 1, 2$, $h(x) \in (-\infty, +\infty)$ for all $x \in \mathbb{R}^d$, $\mathcal{L}(z, y) \in [0, +\infty)$ for all $x, z \in \mathbb{R}^d$, and σ is Lipschitz continuous with $\sigma(0) = 0$. If*

$$\sup_{n\in\mathbb{N}} \mathcal{E}_n(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)}) < +\infty$$

then there exists a subsequence n_m and $(K, b, W, c) \in \Theta$ such that

$$d\left((K^{(n_m)}, b^{(n_m)}, W^{(n_m)}, c^{(n_m)}), (K, b, W, c)\right) \to 0.$$

Furthermore, $\mathcal{E}_\infty(K, b, W, c) < +\infty$.

Proof Relative compactness in TL^2 of $\{(K^{(n)}\}_{n=1}^\infty$ and $\{b^{(n)}\}_{n=1}^\infty$ follows from Proposition 4.3 and compactness of $\{W^{(n)}\}_{n=1}^\infty$ and $\{c^{(n)}\}_{n=1}^\infty$ is immediate from the bounds on $R^{(3)}(W^{(n)})$ and $R^{(4)}(c^{(n)})$. To see that $\mathcal{E}_\infty(K, b, W, c) < +\infty$, we note that, by the bound on σ we must have that $X(1; x, K, b)$ is finite for any x; hence, $E_\infty(K, b, W, c; x, y) < +\infty$ for any (x, y). □

In fact one can obtain compactness in a stronger sense; in particular one can show that, if

$$\sup_{n\in\mathbb{N}} \mathcal{E}_n(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)}) < +\infty$$

then there exists a subsequence such that

$$\max_{i\in\{0,\ldots,n_m-1\}} \left\| K\left(\frac{i}{n_m}\right) - K_i^{(n_m)} \right\| \to 0 \quad \text{and} \quad \max_{i\in\{0,\ldots,n_m-1\}} \left\| b\left(\frac{i}{n_m}\right) - b_i^{(n_m)} \right\| \to 0.$$

See Lemma 4.17.

4.2 Proof of existence of minimisers

The existence of minimisers is a straightforward application of the direct method from the calculus of variations. In particular, for $n \in \mathbb{N}$ all parameters are finite-dimensional; hence, it is enough to show that minimising sequences are bounded. For W, c this is clear from the regularisation, for $K^{(n)}, b^{(n)}$ this follows from Proposition 4.2. Lower semi-continuity then implies that converging minimising sequences converges to minimisers.

Proposition 4.5 *Let $n \in \mathbb{N}$ and $\Theta^{(n)}$ be given by (14). Define $\mathcal{E}_n$, E_n, $R_n^{(i)}$, $R^{(j)}$ for $i = 1, 2$, $j = 3, 4$ as in Sect. 1.1 and 1.2. Assume that $\alpha_i > 0$ for $i = 1, 2, 3, 4$ and $\tau_j > 0$ for $j = 1, 2$.*

Further assume that σ and h are continuous, that $\sigma(0) = 0$, and that $\mathcal{L}$ is non-negative and continuous in its first argument. Then, there exists a minimiser of $\mathcal{E}_n$ in $\Theta^{(n)}$.

Proof Let $\theta_m^{(n)} = (K_m^{(n)}, b_m^{(n)}, W_m, c_m) \in \Theta^{(n)}$ be a minimising sequence, i.e.

$$\mathcal{E}_n(\theta_m^{(n)}) \to \inf_{\Theta^{(n)}} \mathcal{E}_n \quad \text{as } m \to \infty.$$

Since $\mathcal{E}_n(\underline{0}) = \sum_{s=1}^{S} \mathcal{L}(h(\underline{0}), y_i) =: C < \infty$, we can assume that $\mathcal{E}_n(\theta_m^{(n)}) \leq C$ for all m. Hence, $\sup_{m\in\mathbb{N}} \max\{R_n^{(1)}(K_m^{(n)}), R_n^{(2)}(b_m^{(n)}), R^{(3)}(W_m), R^{(4)}(c_m)\} \leq C$. We emphasise that all the parameters are finite-dimensional. Since $\|W_m\| \leq \sqrt{C}$ and $\|c_m\| \leq \sqrt{C}$, we immediately have that $\{W_m\}_{m\in\mathbb{N}}$ and $\{c_m\}_{m\in\mathbb{N}}$ are bounded, hence relatively compact. By Proposition 4.2 $\{K_m^{(n)}\}_{m\in\mathbb{N}}$ and $\{b_m^{(n)}\}_{m\in\mathbb{N}}$ are also bounded in the supremum norm, hence relatively compact.

With recourse to a subsequence, we assume that $(K_m^{(n)}, b_m^{(n)}, W_m, c_m) \to (K^{(n)}, b^{(n)}, W, c) = \theta^{(n)} \in \Theta^{(n)}$. By induction on i it is easy to see that $X_i[x, K_m^{(n)}, b_m^{(n)}] \to X_i[x, K^{(n)}, b^{(n)}]$ as $m \to \infty$ (by continuity of σ). Hence, by continuity of h and $\mathcal{L}(\cdot, y_i)$, it follows that $\mathcal{E}_n(\theta_m^{(n)}) \to \mathcal{E}_n(\theta^{(n)})$. Now since,

$$\mathcal{E}_n(\theta^{(n)}) = \lim_{m\to\infty} \mathcal{E}_n(\theta_m^{(n)}) = \inf_{\Theta^{(n)}} \mathcal{E}_n$$

it follows that $\mathcal{E}_n(\theta^{(n)}) = \inf_{\Theta^{(n)}} \mathcal{E}_n$. □

4.3 Γ-Convergence of $\mathcal{E}_n$

In this section we prove the Γ-convergence of $\mathcal{E}_n$ to $\mathcal{E}_\infty$. We divide the result into two parts: the liminf inequality is in Lemma 4.9, and the existence of a recovery sequence is given in Lemma 4.11. Before getting to these results we start with some preliminary results, the first is that, for any $K^{(n)} \to K$ and $b^{(n)} \to b$, the discrete model (6) converges uniformly to the continuum model (10). The next preliminary result uses this to infer the convergence of $E_n(\theta^{(n)}; x, y) \to E(\theta; x, y)$.

Lemma 4.6 *Consider sequences $K^{(n)} \in L^2(\mu_n; \mathbb{R}^{d\times d})$, $b^{(n)} \in L^2(\mu_n; \mathbb{R}^d)$ where $\mu_n = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{t_i}$ and $t_i = \frac{i}{n}$. Let $d_1(K^{(n)}, K) \to 0$ and $d_1(b^{(n)}, b) \to 0$ where $K \in H^1([0,1]; \mathbb{R}^{d\times d})$ and $b \in H^1([0,1]; \mathbb{R}^d)$. Define $R_n^{(i)}$, $i = 1, 2$, as in Sect. 1.2 with $\tau_i > 0$. Assume that σ is Lipschitz continuous with constant L_σ, $\sigma(0) = 0$, $\max\{\sup_{n\in\mathbb{N}} R_n^{(1)}(K^{(n)}), \sup_{n\in\mathbb{N}} R_n^{(2)}(b^{(n)})\} < +\infty$ and $x \in \mathbb{R}^d$. Then $\|X(\cdot; x, K, b)\|_{L^\infty} \leq C$ where C depends only on L_σ, $\|x\|$, $\|K\|_{L^\infty}$, and $\|b\|_{L^\infty}$ and furthermore*

$$\sup_{i\in\{0,1,\ldots,n-1\}} \sup_{t\in[t_i, t_{i+1}]} \left\| X(t; x, K, b) - X_i^{(n)}[x; K^{(n)}, b^{(n)}] \right\| \to 0$$

where $X(t; x, K, b)$ and $X_i^{(n)}[x, K^{(n)}, b^{(n)}]$ are determined by (10) and (6) respectively.

Proof Let $X_i^{(n)} = X_i^{(n)}[x; K^{(n)}, b^{(n)}]$ and $X(t) = X(t; x, K, b)$. We have

$$\begin{aligned}\left\| X\left(\frac{i}{n}\right) - X_i^{(n)} \right\| &= \left\| X\left(\frac{i-1}{n}\right) + \int_{\frac{i-1}{n}}^{\frac{i}{n}} \dot{X}(t)\,\mathrm{d}t - X_{i-1}^{(n)} - \left(X_i^{(n)} - X_{i-1}^{(n)}\right) \right\| \\ &\leq \left\| X\left(\frac{i-1}{n}\right) - X_{i-1}^{(n)} \right\| + \left\| \int_{\frac{i-1}{n}}^{\frac{i}{n}} \dot{X}(t)\,\mathrm{d}t - \left(X_i^{(n)} - X_{i-1}^{(n)}\right) \right\|.\end{aligned}$$

Using the iterative update for $X_i^{(n)}$, i.e. (6), the continuum differential equation governing the dynamics of $X(t)$, i.e. (10), and the Lipschitz bound on σ, we may bound the second term above by the following:

$$\begin{aligned}
&\left\| \int_{\frac{i-1}{n}}^{\frac{i}{n}} \dot{X}(t)\,\mathrm{d}t - \left(X_i^{(n)} - X_{i-1}^{(n)}\right) \right\| \\
&\quad = \left\| \int_{\frac{i-1}{n}}^{\frac{i}{n}} \sigma(K(t)X(t) + b(t)) - \sigma\left(K_{i-1}^{(n)} X_{i-1}^{(n)} + b_{i-1}^{(n)}\right) \mathrm{d}t \right\| \\
&\quad \leq L_\sigma \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| K(t)X(t) - b(t) - K_{i-1}^{(n)} X_{i-1}^{(n)} - b_{i-1}^{(n)} \right\| \mathrm{d}t \\
&\quad \leq L_\sigma \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| b(t) - b_{i-1}^{(n)} \right\| + \left\| K(t)X(t) - K_{i-1}^{(n)} X_{i-1}^{(n)} \right\| \mathrm{d}t. \qquad (20)
\end{aligned}$$

By Proposition 4.2 we can show that $\|K\|_{L^\infty}, \|b\|_{L^\infty}$ is finite (since $K^{(n)} \to K$ and $K^{(n)}$ is uniformly bounded in $L^\infty(\mu_n; \mathbb{R}^{d\times d})$, analogously for b). Now we show that $\sup_{n\in\mathbb{N}} \|X^{(n)}\|_{L^\infty(\mu_n)} < +\infty$. We have, by (6) and the Lipschitz assumption on σ,

$$\|X_{i+1}^{(n)} - X_i^{(n)}\| \leq \frac{L_\sigma}{n} \|K_i^{(n)} X_i^{(n)} + b_i^{(n)}\| \leq \frac{L_\sigma M_1}{n} \left(\|X_i^{(n)}\| + 1\right)$$

where $M_1 = \sup_{n\in\mathbb{N}} \max\{\|K^{(n)}\|_{L^\infty(\mu_n)}, \|b^{(n)}\|_{L^\infty(\mu_n)}\} < +\infty$ by Proposition 4.2. Hence,

$$\|X_{i+1}^{(n)}\| \leq \left(1 + \frac{L_\sigma M_1}{n}\right) \|X_i^{(n)}\| + \frac{L_\sigma M_1}{n}.$$

Since $X_j^{(n)} = x$, by induction it follows that, for $j \in \{1, \ldots, n\}$,

$$\begin{aligned}
\|X_j^{(n)}\| &\leq \|x\| \left(1 + \frac{L_\sigma M_1}{n}\right)^j + \frac{L_\sigma M_1}{n} \sum_{i=1}^{j} \left(1 + \frac{L_\sigma M_1}{n}\right)^{i-1} \\
&\leq (\|x\| + L_\sigma M_1) \left(1 + \frac{L_\sigma M_1}{n}\right)^n \\
&\to (\|x\| + L_\sigma M_1)\, e^{L_\sigma M_1}, \quad \text{as } n \to \infty.
\end{aligned}$$

Hence, $\sup_{n\in\mathbb{N}} \|X^{(n)}\|_{L^\infty(\mu_n)} < +\infty$.

Now consider, for $0 \leq s_1 < s_2 \leq 1$,

$$\begin{aligned}
\|X\|_{L^\infty([s_1,s_2])} &= \sup_{s\in[s_1,s_2]} \|X(s)\| \\
&= \sup_{s\in[s_1,s_2]} \left\| \int_{s_1}^{s} \dot{X}(r)\,\mathrm{d}r + X(s_1) \right\| \\
&= \sup_{s\in[s_1,s_2]} \left\| \int_{s_1}^{s} \sigma\left(K(r)X(r) + b(r)\right)\,\mathrm{d}r + X(s_1) \right\| \quad \text{by (10)} \\
&\leq \sup_{s\in[s_1,s_2]} \int_{s_1}^{s} L_\sigma \left\| K(r)X(r) + b(r) \right\| \mathrm{d}r \\
&\quad + \|X(s_1)\| \text{ as } \sigma \text{ is Lipschitz and } \sigma(0) = 0 \\
&\leq \sup_{s\in[s_1,s_2]} L_\sigma (s - s_1) \|X\|_{L^\infty([s_1,s])} \|K\|_{L^\infty} + L_\sigma (s - s_1) \|b\|_{L^\infty} + \|X(s_1)\| \\
&= L_\sigma (s_2 - s_1) \|X\|_{L^\infty([s_1,s_2])} \|K\|_{L^\infty} + L_\sigma (s_2 - s_1) \|b\|_{L^\infty} + \|X(s_1)\|.
\end{aligned}$$

Therefore, if we choose $s_2 = \min\{1, s_1 + \frac{1}{2L_\sigma \|K\|_{L^\infty}}\}$ we have $\|X\|_{L^\infty([s_1,s_2])} \leq 2L_\sigma \|b\|_{L^\infty} + 2\|X(s_1)\|$. Let $s_i = \min\{1, \frac{i}{2L_\sigma \|K\|_{L^\infty}}\}$ and $N = \lceil 2L_\sigma \|K\|_{L^\infty} \rceil$ (we note that $s_{N-1} < 1 = s_N$). For $i \in \{2, \dots, N\}$ we have

$$\begin{aligned}
\|X\|_{L^\infty([0,s_i])} &\leq \max\left\{\|X\|_{L^\infty([0,s_{i-1}])}, \|X\|_{L^\infty([s_{i-1},s_i])}\right\} \\
&\leq \max\left\{\|X\|_{L^\infty([0,s_{i-1}])}, 2L_\sigma \|b\|_{L^\infty} + 2\|X(s_{i-1})\|\right\} \\
&\leq 2L_\sigma \|b\|_{L^\infty} + 2\|X\|_{L^\infty([0,s_{i-1}])}.
\end{aligned}$$

For $i = 1$ we have $\|X\|_{L^\infty([0,s_1])} \leq 2L_\sigma \|b\|_{L^\infty} + \|x\|$, and by induction for $i \in \{2, \dots, N\}$ we have

$$\begin{aligned}
\|X\|_{L^\infty([0,s_i])} &\leq 2^i \|X\|_{L^\infty([0,s_1])} + 2L_\sigma \|b\|_{L^\infty} \sum_{k=0}^{i-1} 2^k \\
&\leq 2^{i+1} L_\sigma \|b\|_{L^\infty} + 2^i \|x\| + 2(2^i - 1) L_\sigma \|b\|_{L^\infty} \\
&= 2(2^i - 1) L_\sigma \|b\|_{L^\infty} + 2^i \|x\|.
\end{aligned}$$

In particular, for $i = N$ we have

$$\|X\|_{L^\infty([0,1])} \leq 2(2^N - 1) L_\sigma \|b\|_{L^\infty} + 2^N \|x\|.$$

Now, using $\sigma(0) = 0$ and the Lipschitz assumption on σ, we have

$$\|\dot{X}\|_{L^\infty} = \|\sigma(KX + b)\|_{L^\infty} \leq L_\sigma \left(\|K\|_{L^\infty} \|X\|_{L^\infty} + \|b\|_{L^\infty}\right),$$

hence, X is Lipschitz. Let L_X be the Lipschitz constant for X.

Returning to (20) we concentrate on the second term, we bound

$$\begin{aligned}
&\int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| K(t)X(t) - K_{i-1}^{(n)} X_{i-1}^{(n)} \right\| \mathrm{d}t \\
&\quad \leq \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| X(t) - X_{i-1}^{(n)} \right\| \left\| K(t) \right\| \mathrm{d}t + \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| K(t) - K_{i-1}^{(n)} \right\| \left\| X_{i-1}^{(n)} \right\| \mathrm{d}t \\
&\quad \leq M_1 \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| X(t) - X_{i-1}^{(n)} \right\| \mathrm{d}t + \|X^{(n)}\|_{L^\infty(\mu_n)} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| K(t) - K_{i-1}^{(n)} \right\| \mathrm{d}t \\
&\quad \leq M_2 \left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| X(t) - X_{i-1}^{(n)} \right\| \mathrm{d}t + \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| K(t) - K_{i-1}^{(n)} \right\| \mathrm{d}t \right)
\end{aligned} \tag{21}$$

where $M_2 = \max\{M_1, \|X^{(n)}\|_{L^\infty(\mu_n)}\}$. Continuing to manipulate the first term on the right-hand side of the above expression, we have,

$$\begin{aligned}
\int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| X(t) - X_{i-1}^{(n)} \right\| \mathrm{d}t &\leq \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| X_{i-1}^{(n)} - X\left(\frac{i-1}{n}\right) \right\| + \left\| X\left(\frac{i-1}{n}\right) - X(t) \right\| \mathrm{d}t \\
&= \frac{1}{n} \left\| X_{i-1}^{(n)} - X\left(\frac{i-1}{n}\right) \right\| + \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\| X\left(\frac{i-1}{n}\right) - X(t) \right\| \mathrm{d}t \\
&\leq \frac{1}{n} \left\| X_{i-1}^{(n)} - X\left(\frac{i-1}{n}\right) \right\| + L_X \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left(t - \frac{i-1}{n} \right) \mathrm{d}t \\
&= \frac{1}{n} \left\| X_{i-1}^{(n)} - X\left(\frac{i-1}{n}\right) \right\| + \frac{L_X}{2n^2}.
\end{aligned} \tag{22}$$

Combining the bounds (20), (21) and (22), we have

$$\left\|X\left(\frac{i}{n}\right)-X_i^{(n)}\right\| \le \left(1+\frac{L_\sigma M_2}{n}\right)\left\|X\left(\frac{i-1}{n}\right)-X_{i-1}^{(n)}\right\| + L_\sigma \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\|b(t)-b_{i-1}^{(n)}\right\| \,\mathrm{d}t + L_\sigma M_2 \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left\|K(t)-K_{i-1}^{(n)}\right\| \,\mathrm{d}t + \frac{L_X L_\sigma M_2}{2n^2}.$$

By induction, for any $k \in \{0, 1, \ldots, n\}$, we have

$$\begin{aligned}\left\|X\left(\frac{k}{n}\right)-X_k^{(n)}\right\| &\le L_\sigma \sum_{i=1}^{n}\left(1+\frac{L_\sigma M_2}{n}\right)^{n-i}\int_{\frac{i-1}{n}}^{\frac{i}{n}} \|b(t)-b_{i-1}^{(n)}\| \,\mathrm{d}t \\ &\quad + L_\sigma M_2 \sum_{i=1}^{n}\left(1+\frac{L_\sigma M_2}{n}\right)^{n-i}\int_{\frac{i-1}{n}}^{\frac{i}{n}} \|K(t)-K_{i-1}^{(n)}\| \,\mathrm{d}t \\ &\quad + \frac{L_X L_\sigma M_2}{2n^2}\sum_{i=1}^{n}\left(1+\frac{L_\sigma M_2}{n}\right)^{n-i} \\ &\le \frac{\varepsilon L_\sigma}{2n}\sum_{i=1}^{n}\left(1+\frac{L_\sigma M_2}{n}\right)^{2(n-i)} + \frac{L_\sigma n}{2\varepsilon}\sum_{i=1}^{n}\left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} \|b(t)-b_{i-1}^{(n)}\| \,\mathrm{d}t\right)^2 \\ &\quad + \frac{\varepsilon L_\sigma M_2}{2n}\sum_{i=1}^{n}\left(1+\frac{L_\sigma M_2}{n}\right)^{2(n-i)} + \frac{L_\sigma M_2 n}{2\varepsilon}\sum_{i=1}^{n}\left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} \|K(t)-K_{i-1}^{(n)}\| \,\mathrm{d}t\right)^2 \\ &\quad + \frac{L_X L_\sigma M_2}{2n^2}\sum_{i=0}^{n-1}\left(1+\frac{L_\sigma M_2}{n}\right)^{i}\end{aligned}$$

for any $\varepsilon > 0$ by employing Young's inequality, $\alpha\beta \le \frac{\varepsilon\alpha^2}{2} + \frac{\beta^2}{2\varepsilon}$, for appropriately chosen α and β, on the first two terms for the final inequality (notice that the right hand side is independent of k). By Hölder's inequality and the assumption that $d_1(K^{(n)}, K) \to 0$ we have

$$n\sum_{i=1}^{n}\left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} \|K(t)-K_{i-1}^{(n)}\| \,\mathrm{d}t\right)^2 \le \sum_{i=1}^{n}\int_{\frac{i-1}{n}}^{\frac{i}{n}} \|K(t)-K_{i-1}^{(n)}\|^2 \,\mathrm{d}t \to 0$$

(and similarly for the sequence $b^{(n)}$). Hence, to show

$$\sup_{k\in\{0,1,\ldots,n\}}\left\|X\left(\frac{k}{n}\right)-X_k^{(n)}\right\| \to 0, \tag{23}$$

it is enough to show (i) $\frac{1}{n^2}\sum_{i=0}^{n-1}\left(1+\frac{L_\sigma M_2}{n}\right)^i \to 0$ and (ii) $\sup_n \frac{1}{n}\sum_{i=1}^{n}\left(1+\frac{L_\sigma M_2}{n}\right)^{2(n-i)} < \infty$.

For (i) we have that

$$0 \le \frac{1}{n^2}\sum_{i=0}^{n-1}\left(1+\frac{L_\sigma M_2}{n}\right)^i \le \frac{1}{n}\left(1+\frac{L_\sigma M_2}{n}\right)^n \to 0.$$

And for (ii) we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(1+\frac{L_\sigma M_2}{n}\right)^{2(n-i)} \le \left(1+\frac{L_\sigma M_2}{n}\right)^{2n} \to e^{2L_\sigma M_2}.$$

Hence, if we replace ε by a sequence ε_n that converges to zero sufficiently slowly and that satisfies

$$\frac{1}{\varepsilon_n}\sum_{i=1}^{n}\int_{\frac{i-1}{n}}^{\frac{i}{n}} \|K(t)-K_{i-1}^{(n)}\|^2\,\mathrm{d}t \to 0 \qquad \text{and} \qquad \frac{1}{\varepsilon_n}\sum_{i=1}^{n}\int_{\frac{i-1}{n}}^{\frac{i}{n}} \|b(t)-b_{i-1}^{(n)}\|^2\,\mathrm{d}t \to 0,$$

then we have that (23) holds.

Finally,

$$\begin{aligned}\sup_{t\in[t_k,t_{k+1}]}\|X(t)-X_k^{(n)}\| &\leq \sup_{t\in[t_k,t_{k+1}]}\left(\|X(t)-X(t_k)\|+\|X(t_k)-X_k^{(n)}\|\right)\\ &\leq \frac{L_X}{n}+\|X(t_k)-X_k^{(n)}\|\to 0\end{aligned}$$

where the convergence is uniform over $k\in\{0,1,\ldots,n\}$ as required. □

We say that $\Theta\ni\theta_n=(K^{(n)},b^{(n)},W^{(n)},c^{(n)})\to\theta=(K,b,W,c)\in\Theta$ if $d_1(K^{(n)},K)\to 0$, $d_2(b^{(n)},b)\to 0$ and $W^{(n)}\to W$, $c^{(n)}\to c$ (where, since $W^{(n)}$ and $c^{(n)}$ are sequences in $\mathbb{R}^{d\times d}$ and $\mathbb{R}^d$, we choose any norm induced-topology for the latter). The above result implies the following lemma.

Lemma 4.7 *In addition to the assumptions of Theorem 2.1 let $\Theta^{(n)}\ni\theta^{(n)}\to\theta\in\Theta$, with $\max\{\sup_{n\in\mathbb{N}}R_n^{(1)}(K^{(n)}),\sup_{n\in\mathbb{N}}R_n^{(2)}(b^{(n)})\}<+\infty$ and $x\in\mathbb{R}^d$, $y\in\mathbb{R}^m$, then*

$$\lim_{n\to\infty}E_n(\theta^{(n)};x,y)=E_\infty(\theta;x,y).$$

Proof By continuity of h and $\mathcal{L}$ (in its first argument), convergence of $W^{(n)}\to W$, $c^{(n)}\to c$ and $X_n^{(n)}[x,K^{(n)},b^{(n)}]\to X(1;x,K,b)$ (with the latter following from Lemma 4.6) we can easily conclude the result. □

The following is a small generalisation of Theorem 10.55 in [62]. The difference between the results stated here and the result in [62] is that here we treat sequences of functions f_n, whilst in [62] $f_n=f$. We also only state the result on the domain $[0,1]$ and for L^2 convergence (the result generalises to bounded sets in higher dimensions and L^p convergence where $p>1$).

Proposition 4.8 *Let $f_n\in L^2([0,1];\mathbb{R}^\kappa)$, $f\in L^2([0,1];\mathbb{R}^\kappa)$ and $\varepsilon_n\to 0^+$. Assume that $f_n\to f$ in $L^2([0,1];\mathbb{R}^\kappa)$. If*

$$\liminf_{n\to\infty}\frac{1}{\varepsilon_n^2}\int_{\varepsilon_n}^{1}\left\|f_n(t)-f_n(t-\varepsilon_n)\right\|^2\,\mathrm{d}t<+\infty,$$

then $f\in H^1([0,1];\mathbb{R}^\kappa)$ and

$$\liminf_{n\to\infty}\frac{1}{\varepsilon_n^2}\int_{\varepsilon_n}^{1}\|f_n(t)-f_n(t-\varepsilon_n)\|^2\,\mathrm{d}t\geq\int_0^1\|\dot f(t)\|^2\,\mathrm{d}t.$$

Proof The strategy is to show the following two inequalities:

$$\int_{\delta'}^{1-\delta'}\left\|J_\delta*\tilde g(t)-J_\delta*\tilde g(t-\varepsilon_n)\right\|^2\,\mathrm{d}t\leq\int_{\varepsilon_n}^{1}\left\|\tilde g(t)-\tilde g(t-\varepsilon_n)\right\|^2\,\mathrm{d}t, \tag{24}$$

for any $\tilde g\in L^2([0,1];\mathbb{R}^\kappa)$ and any $\delta,\delta'>0$ that satisfy $\varepsilon_n+\delta<\delta'$, and where J_δ is a standard mollifier; and

$$\int_{2\delta'}^{1-2\delta'}\left\|\dot g(t)\right\|^2\,\mathrm{d}t\leq\liminf_{n\to\infty}\frac{1}{\varepsilon_n^2}\int_{2\delta'}^{1-2\delta'}\left\|g_n(t)-g_n(t-\varepsilon_n)\right\|^2\,\mathrm{d}t \tag{25}$$

for any $g, g_n \in C^\infty([\delta', 1-\delta']; \mathbb{R}^\kappa)$ with $\dot{g}_n \to \dot{g}$ in $L^\infty([\delta', 1-\delta']; \mathbb{R}^\kappa)$ and $\sup_n \|\ddot{g}_n\|_{L^\infty([\delta', 1-\delta'])} < \infty$.

Before we prove these two inequalities, we use them to prove the result of the lemma. We note that $\|\frac{\mathrm{d}}{\mathrm{d}t} J_\delta * f - \frac{\mathrm{d}}{\mathrm{d}t} J_\delta * f_n\|_{L^\infty([\delta', 1-\delta'])} \leq \|\frac{\mathrm{d}}{\mathrm{d}t} J_\delta\|_{L^2(\mathbb{R})} \|f_n - f\|_{L^2([0,1])}$ and $\|\frac{\mathrm{d}^2}{\mathrm{d}t^2} J_\delta * f_n\|_{L^\infty([\delta', 1-\delta'])} \leq \|\frac{\mathrm{d}^2}{\mathrm{d}t^2} J_\delta\|_{L^2(\mathbb{R})} \|f_n\|_{L^2([0,1])}$. Therefore, we may apply (25) to $g = J_\delta * f$ and $g_n = J_\delta * f_n$.

To show existence of $\dot{f} \in L^2([2\delta', 1-2\delta']; \mathbb{R}^\kappa)$ we assume the above inequalities hold, then by (24) and the assumptions on f_n there exists M such that

$$\liminf_{n\to\infty} \frac{1}{\varepsilon_n^2} \int_{\delta'}^{1-\delta'} \left\| J_\delta * f_n(t) - J_\delta * f_n(t-\varepsilon_n) \right\|^2 \mathrm{d}t \leq M.$$

Furthermore, by (25) $\int_{2\delta'}^{1-2\delta'} \|\frac{\mathrm{d}}{\mathrm{d}t} J_\delta * f(t)\|^2 \mathrm{d}t \leq M$. In addition to the four standard properties of mollifiers which we recalled in the proof of Proposition 4.3, we list a fifth one here:

(M5) $J_\delta * \tilde{g} \to \tilde{g}$ in $L^2([0,1]; \mathbb{R}^\kappa)$, as $\delta \to 0^+$, for any $\tilde{g} \in L^2([0,1]; \mathbb{R}^\kappa)$ [62, Theorem C.19].

By (M5) and since $\frac{\mathrm{d}}{\mathrm{d}t} J_\delta * f$ is bounded in $L^2([2\delta', 1-2\delta']; \mathbb{R}^\kappa)$, uniformly in δ, there exists an $h \in L^2([2\delta', 1-2\delta']; \mathbb{R}^\kappa)$ such that, after potentially passing to a subsequence, $J_\delta * f \to f$ and $\frac{\mathrm{d}}{\mathrm{d}t}(J_\delta * f) \rightharpoonup h$ in $L^2([0,1]; \mathbb{R}^\kappa)$, as $\delta \to 0^+$. Therefore, for any differentiable φ with compact support in $[2\delta', 1-2\delta']$,

$$\int_{2\delta'}^{1-2\delta'} \varphi h \leftarrow \int_{2\delta'}^{1-2\delta'} \varphi \frac{\mathrm{d}}{\mathrm{d}t}(J_\delta * f) = -\int_{2\delta'}^{1-2\delta'} \frac{\mathrm{d}}{\mathrm{d}t} \varphi J_\delta * f \to -\int_{2\delta'}^{1-2\delta'} \dot{\varphi} f = \int_{2\delta'}^{1-2\delta'} \varphi \dot{f}.$$

Hence, $h = \dot{f}$ and in particular $\dot{f} \in L^2([2\delta', 1-2\delta']; \mathbb{R}^\kappa)$. Since $\frac{\mathrm{d}}{\mathrm{d}t}(J_\delta * f) = J_\delta * \dot{f}$, we can use again the convergence of mollifiers to infer $\frac{\mathrm{d}}{\mathrm{d}t}(J_\delta * f) \to \dot{f}$ (*strongly*) in $L^2([2\delta', 1-2\delta']; \mathbb{R}^\kappa)$.

Applying (25) followed by (24) we have

$$\begin{aligned}\int_{2\delta'}^{1-2\delta'} \left\| \frac{\mathrm{d}}{\mathrm{d}t} J_\delta * f(t) \right\|^2 \mathrm{d}t &\leq \liminf_{n\to\infty} \int_{2\delta'}^{1-2\delta'} \left\| \frac{J_\delta * f_n(t) - J_\delta * f_n(t-\varepsilon_n)}{\varepsilon_n} \right\|^2 \mathrm{d}t \\ &\leq \liminf_{n\to\infty} \int_{\varepsilon_n}^{1} \left\| \frac{f_n(t) - f_n(t-\varepsilon_n)}{\varepsilon_n} \right\|^2 \mathrm{d}t.\end{aligned}$$

By $L^2([2\delta', 1-2\delta']; \mathbb{R}^\kappa)$ convergence of $\frac{\mathrm{d}}{\mathrm{d}t} J_\delta * f$ as $\delta \to 0$ we have

$$\int_{2\delta'}^{1-2\delta'} \left\| \dot{f}(t) \right\|^2 \mathrm{d}t \leq \liminf_{n\to\infty} \int_{\varepsilon_n}^{1} \left\| \frac{f_n(t) - f_n(t-\varepsilon_n)}{\varepsilon_n} \right\|^2 \mathrm{d}t,$$

with the inequality above valid for any $\delta' > 0$ (since the additional constraint imposed by $\delta' > \delta + \varepsilon_n$ vanishes when taking $\delta \to 0$ and $\varepsilon_n \to 0$). Taking $\delta' \to 0$ proves the lemma under the assumption of (24) and (25).

To show (24) we have, assuming $\delta + \varepsilon_n < \delta'$,

$$\begin{aligned}&\left(\int_{\delta'}^{1-\delta'} \left\| J_\delta * \tilde{g}(t) - J_\delta * \tilde{g}(t-\varepsilon_n) \right\|^2 \mathrm{d}t \right)^{\frac{1}{2}} \\ &\qquad = \left(\int_{\delta'}^{1-\delta'} \left\| \int_{-\delta}^{\delta} J_\delta(s) \left[\tilde{g}(t-s) - \tilde{g}(t-\varepsilon_n - s) \right] \mathrm{d}s \right\|^2 \mathrm{d}t \right)^{\frac{1}{2}} \\ &\qquad \leq \int_{-\delta}^{\delta} J_\delta(s) \left(\int_{\delta'}^{1-\delta'} \left\| \tilde{g}(t-s) - \tilde{g}(t-\varepsilon_n - s) \right\|^2 \mathrm{d}t \right)^{\frac{1}{2}} \mathrm{d}s\end{aligned}$$

$$\leq \int_{-\delta}^{\delta} J_\delta(s) \left(\int_{\varepsilon_n}^{1} \left\| \tilde{g}(t) - \tilde{g}(t-\varepsilon_n) \right\|^2 \mathrm{d}t \right)^{\frac{1}{2}} \mathrm{d}s$$
$$= \left(\int_{\varepsilon_n}^{1} \left\| \tilde{g}(t) - \tilde{g}(t-\varepsilon_n) \right\|^2 \mathrm{d}t \right)^{\frac{1}{2}},$$

where the antepenultimate line follows from Minkowski's inequality for integrals.

For inequality (25), by Taylor's theorem we have

$$g_n(t) - g_n(t-\varepsilon_n) = \varepsilon_n \dot{g}_n(t) + \varepsilon_n^2 \ddot{g}_n(z) \qquad \text{for some } z \in [t-\varepsilon_n, t].$$

Therefore, for $t \in [2\delta', 1-2\delta']$, when $\varepsilon_n < \delta'$,

$$\frac{\|g_n(t) - g_n(t-\varepsilon_n)\|}{\varepsilon_n} \geq \|\dot{g}_n(t)\| - \varepsilon_n \|\ddot{g}_n\|_{L^\infty([\delta', 1-\delta'])}.$$

For any $\eta > 0$ there exists $C_\eta > 0$ such that $|a+b|^2 \leq (1+\eta)|a|^2 + C_\eta |b|^2$ for any $a, b \in \mathbb{R}$ (a consequence of Young's inequality, and one can show that $C_\eta = 1 + \frac{1}{\eta}$), hence

$$\|\dot{g}_n(t)\|^2 \leq (1+\eta) \left\| \frac{g_n(t) - g_n(t-\varepsilon_n)}{\varepsilon_n} \right\|^2 + C_\eta \varepsilon_n^2 \|\ddot{g}_n\|^2_{L^\infty([\delta', 1-\delta'])}.$$

In particular, by Lebesgue's dominated convergence theorem and $\sup_{n\in\mathbb{N}} \|\ddot{g}_n\|_{L^\infty([\delta', 1-\delta'])} < \infty$,

$$\int_{2\delta'}^{1-2\delta'} \|\dot{g}(t)\|^2 \mathrm{d}t = \lim_{n\to\infty} \int_{2\delta'}^{1-2\delta'} \|\dot{g}_n(t)\|^2 \mathrm{d}t$$
$$\leq (1+\eta) \liminf_{n\to\infty} \int_{2\delta'}^{1-2\delta'} \left\| \frac{g_n(t) - g_n(t-\varepsilon_n)}{\varepsilon_n} \right\|^2 \mathrm{d}t.$$

Taking $\eta \to 0$ proves (25). □

By application of the preceeding lemma we can now prove the liminf inequality for the Γ-convergence of $\mathcal{E}_n$.

Lemma 4.9 *Under the assumptions of Theorem 2.1 let $\Theta^{(n)} \ni \theta^{(n)} \to \theta \in \Theta$, then,*

$$\liminf_{n\to\infty} \mathcal{E}_n(\theta^{(n)}) \geq \mathcal{E}_\infty(\theta).$$

Proof Let $\theta^{(n)} = (K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)})$ and $\theta = (K, b, W, c)$. We only need to consider the case when $\liminf_{n\to\infty} \mathcal{E}_n(\theta^{(n)}) < +\infty$. Hence, we assume that $\mathcal{E}_n(\theta^{(n)})$ is bounded and therefore by the compactness property (Corollary 4.4) $\mathcal{E}_\infty(\theta) < +\infty$. We will show the following

(A) $\lim_{n\to\infty} E_n(\theta^{(n)}; x, y) = E_\infty(\theta; x, y)$,

(B) $\liminf_{n\to\infty} R_n^{(1)}(K^{(n)}) \geq R_\infty^{(1)}(K)$, and

(C) $\liminf_{n\to\infty} R_n^{(2)}(b^{(n)}) \geq R_\infty^{(2)}(b)$.

Indeed (A) holds by Lemma 4.7 and since $R_n^{(1)}(K^{(n)})$, $R_n^{(2)}(b^{(n)})$ are uniformly (in n) bounded.

Parts (B) and (C) are analogous, so we only show (B). Let $\tilde{K}^{(n)}(t) = K_i^{(n)}$ for $t \in \left(\frac{i}{n}, \frac{i+1}{n}\right]$, for $i = 0, \dots, n-1$, then

$$\begin{aligned}
\liminf_{n\to\infty} R_n^{(1)}(K^{(n)}) &= \liminf_{n\to\infty} \left(n \sum_{i=1}^{n} \|K_i^{(n)} - K_{i-1}^{(n)}\|^2 + \tau_1 \|K_0^{(n)}\|^2 \right) \\
&\geq \liminf_{n\to\infty} n^2 \int_{\frac{1}{n}}^{1} \left\| \tilde{K}^{(n)}(t) - \tilde{K}^{(n)}\left(t - \frac{1}{n}\right) \right\|^2 \mathrm{d}t + \tau_1 \liminf_{n\to\infty} \|K_0^{(n)}\|^2 \\
&\geq \int_0^1 \|\dot{K}(t)\|^2 \,\mathrm{d}t + \tau_1 \|K(0)\|^2,
\end{aligned}$$

with the last inequality holding as, by Proposition 4.3, $\tilde{K}_n \to K$ in $L^2([0,1])$. Hence, we may apply Proposition 4.8. □

We now turn our attention to the recovery sequence. For any $\theta \in \Theta$ we define a sequence $\theta^{(n)} \in \Theta^{(n)}$ by

$$K_i^{(n)} = n \int_{\frac{i}{n}}^{\frac{i+1}{n}} K(t) \,\mathrm{d}t, \qquad \text{for } i = 0, \dots, n-1, \tag{26}$$

$$b_i^{(n)} = n \int_{\frac{i}{n}}^{\frac{i+1}{n}} b(t) \,\mathrm{d}t, \qquad \text{for } i = 0, \dots, n-1, \tag{27}$$

$$W^{(n)} = W, \tag{28}$$

$$c^{(n)} = c. \tag{29}$$

The above sequence is our candidate recovery sequence. We first show that $\theta^{(n)} \to \theta$ in the TL^2 topology.

Lemma 4.10 *Under the assumptions of Theorem 2.1 let $\theta = (K, b, W, c) \in \Theta$ and define $\theta^{(n)} = (K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)}) \in \Theta^{(n)}$ by (26-29). Then $\theta^{(n)} \to \theta$ in the TL^2 topology.*

Proof We show that $K^{(n)} \to K$; the argument for $b^{(n)} \to b$ is analogous and $W^{(n)} = W$, $c^{(n)} = c$ so there is nothing to show for these parts. Let $\tilde{K}^{(n)}(t) = K_i^{(n)}$ for $t \in \left[\frac{i}{n}, \frac{i+1}{n}\right)$ for $i = 0, \dots, n-1$ and $\tilde{K}^{(n)}(1) = K_{n-1}^{(n)}$. Since $K \in H^1([0,1]; \mathbb{R}^{d\times d})$, by Morrey's inequality we have that $K \in C^{0,\frac{1}{2}}([0,1]; \mathbb{R}^{d\times d})$. In particular, $\|K(s) - K(t)\| \leq L_K \sqrt{|t-s|}$ for some L_K. So,

$$\begin{aligned}
\|\tilde{K}^{(n)} - K\|_{L^2}^2 &= \sum_{i=0}^{n-1} \int_{\frac{i}{n}}^{\frac{i+1}{n}} \left\| K_i^{(n)} - K(t) \right\|^2 \mathrm{d}t \\
&= \sum_{i=0}^{n-1} \int_{\frac{i}{n}}^{\frac{i+1}{n}} \left\| n \int_{\frac{i}{n}}^{\frac{i+1}{n}} K(s) - K(t) \,\mathrm{d}s \right\|^2 \mathrm{d}t \\
&\leq n \sum_{i=0}^{n-1} \int_{\frac{i}{n}}^{\frac{i+1}{n}} \int_{\frac{i}{n}}^{\frac{i+1}{n}} \left\| K(s) - K(t) \right\|^2 \mathrm{d}s \,\mathrm{d}t \\
&\leq L_K^2 n \sum_{i=0}^{n-1} \int_{\frac{i}{n}}^{\frac{i+1}{n}} \int_{\frac{i}{n}}^{\frac{i+1}{n}} |s-t| \,\mathrm{d}s \,\mathrm{d}t \\
&= \frac{L_K^2}{3n} \to 0.
\end{aligned}$$

Therefore, $K^{(n)} \to K$. □

We now prove that the sequence from Lemma 4.10 is a recovery sequence.

Lemma 4.11 *Under the assumptions of Theorem 2.1 for any $\theta \in \Theta$ we define $\theta^{(n)} \in \Theta^{(n)}$ as in Lemma 4.10. Then $\theta^{(n)} \to \theta$ and*

$$\limsup_{n\to\infty} \mathcal{E}_n(\theta^{(n)}) \leq \mathcal{E}_\infty(\theta).$$

Proof Let $\theta = (K, b, W, c) \in \Theta$ and assume $\mathcal{E}_\infty(\theta) < \infty$ (else the result is trivial). By Lemma 4.10 we already have that $\theta^{(n)} \to \theta$.

We show that $\theta^{(n)}$ is a recovery sequence. It is enough to show the following.

(A) $\lim_{n\to\infty} E_n(\theta^{(n)}; x, y) = E_\infty(\theta; x, y)$,

(B) $\limsup_{n\to\infty} R_n^{(1)}(K^{(n)}) \leq R_\infty^{(1)}(K)$, and

(C) $\limsup_{n\to\infty} R_n^{(2)}(b^{(n)}) \leq R_\infty^{(2)}(b)$.

Part (A) follows from Lemma 4.7 once we show parts (B) and (C). Since (B) and (C) are analogous, we only show (B).

Let $\varepsilon > 0$ and $C_\varepsilon = 1 + \frac{1}{\varepsilon}$, then $\|a+b\|^2 \leq (1+\varepsilon)\|a\|^2 + C_\varepsilon \|b\|^2$ (as a consequence of Young's inequality). So,

$$\begin{aligned}
R_n^{(1)}(K^{(n)}) &= n \sum_{i=1}^{n-1} \left\| n \int_{\frac{i}{n}}^{\frac{i+1}{n}} K(t) - K\left(t - \frac{1}{n}\right) \mathrm{d}t \right\|^2 + \tau_1 \left\| n \int_0^{\frac{1}{n}} K(t)\, \mathrm{d}t \right\|^2 \\
&\leq n^2 \sum_{i=1}^{n-1} \int_{\frac{i}{n}}^{\frac{i+1}{n}} \left\| K(t) - K\left(t - \frac{1}{n}\right) \right\|^2 \mathrm{d}t + (1+\varepsilon)\tau_1 \|K(0)\|^2 \\
&\qquad + C_\varepsilon \tau_1 n \int_0^{\frac{1}{n}} \|K(t) - K(0)\|^2 \, \mathrm{d}t \\
&\leq n^2 \int_{\frac{1}{n}}^1 \left\| K(t) - K\left(t - \frac{1}{n}\right) \right\|^2 \mathrm{d}t + (1+\varepsilon)\tau_1 \|K(0)\|^2 + C_\varepsilon L_K^2 \tau_1 n \int_0^{\frac{1}{n}} t \, \mathrm{d}t \\
&\leq \int_0^1 \left\| \dot{K}(t) \right\|^2 \mathrm{d}t + (1+\varepsilon)\tau_1 \|K(0)\|^2 + \frac{C_\varepsilon L_K^2 \tau_1}{2n},
\end{aligned}$$

where the last line follows from [62, Theorem 10.55] (we note that we cannot use Proposition 4.8 here, since it gives the lower bound $\liminf_{n\to\infty} n^2 \int_{\frac{1}{n}}^1 \left\| K(t) - K\left(t - \frac{1}{n}\right) \right\|^2 \mathrm{d}t \geq \int_0^1 \left\| \dot{K}(t) \right\|^2 \mathrm{d}t$, rather than an upper bound). Taking $n \to \infty$ we have

$$\limsup_{n\to\infty} R_n^{(1)}(K^{(n)}) \leq \int_0^1 \left\| \dot{K}(t) \right\|^2 \mathrm{d}t + (1+\varepsilon)\tau_1 \|K(0)\|^2 \leq (1+\varepsilon) R_\infty^{(1)}(K).$$

Taking $\varepsilon \to 0^+$ proves (B). □

4.4 Regularity of minimisers

The aim of this section is to show the higher regularity (i.e. H^2_{loc} rather than H^1) of minimisers. The strategy is to apply elliptic regularity techniques. For this we need to compute the Euler–Lagrange equation for $\mathcal{E}_\infty$. We start by showing how the finite layer model (6) behaves when the parameters $K^{(n)}$ and $b^{(n)}$ are perturbed. By taking the limit $n \to \infty$ we can then infer the corresponding result for the ODE limit (10).

Lemma 4.12 *Let $n \in \mathbb{N}$, $t_i = \frac{i}{n}$, $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{t_i}$ and $K^{(n)}, L^{(n)} \in L^2(\mu_n; \mathbb{R}^{d\times d})$ and $b^{(n)}, \beta^{(n)} \in L^2(\mu_n; \mathbb{R}^d)$. Assume*

$$\max\left\{R_n^{(1)}(K^{(n)}), R_n^{(1)}(L^{(n)}), R_n^{(2)}(b^{(n)}), R_n^{(2)}(\beta^{(n)})\right\} \leq C$$

where $R_n^{(j)}$, $j = 1, 2$ are defined in Sect. 1.2 with $\tau_i > 0$. Furthermore, we assume that $\sigma \in C^2$, $\sigma(0) = 0$, and σ acts componentwise. Let $\theta^{(n)} = (K^{(n)}, b^{(n)})$ and $\xi^{(n)} = (L^{(n)}, \beta^{(n)})$ and define $X_i^{(n)}[x; \theta^{(n)}]$, $i \in \{0, \dots, n-1\}$, as a solution to (6) with initial condition $X_0^{(n)} = x$. We define, for $r > 0$ and $i \in \{0, \dots, n-1\}$,

$$D_{r,i}^{(n)}(x, \theta^{(n)}, \xi^{(n)}) = \frac{1}{r}\left(X_i^{(n)}[x; \theta^{(n)} + r\xi^{(n)}] - X_i^{(n)}[x; \theta^{(n)}]\right). \tag{30}$$

Then,

$$\begin{aligned} D_{r,n}^{(n)}(x, \theta^{(n)}, \xi^{(n)}) = \frac{1}{n}\sum_{i=0}^{n-1}\Bigg\{ &\left[\prod_{j=i+1}^{n-1}\left(\mathrm{Id} + \frac{1}{n}\dot\sigma\left(K_j^{(n)}X_j^{(n)}[x;\theta^{(n)}] + b_j^{(n)}\right)\odot K_j^{(n)}\right)\right] \\ &\times\left(\left[L_i^{(n)}X_i^{(n)}[x;\theta^{(n)}] + \beta_i^{(n)}\right]\odot\dot\sigma\left(K_i^{(n)}X_i^{(n)}[x;\theta^{(n)}] + b_i^{(n)}\right)\right)\Bigg\} \\ &+ O(r). \end{aligned} \tag{31}$$

where the $O(r)$ term depends on $K^{(n)}, L^{(n)}, b^{(n)}, \beta^{(n)}$ only through the parameter C and does not depend on n in any other way.

Remark 4.13 Not that for vectors $A, C \in \mathbb{R}^d$ and a matrix $B \in \mathbb{R}^{d\times d}$ we have $[BC]\odot A = A \odot [BC] = [A \odot B]C = [B \odot A]C$ where $A \odot B$ is understood to be taken componentwise in each row, i.e. $(A \odot B)_{ij} = A_i B_{ij}$. Hence, the usual matrix multiplication $\times$ and componentwise multiplication $\odot$ commute.

Proof of Lemma 4.12 Since $\theta^{(n)}$, $\xi^{(n)}$ and x are fixed, we may shorten our notation by writing

$$D_{r,i}^{(n)} = D_{r,i}^{(n)}(x_s, \theta^{(n)}, \xi^{(n)}), \tag{32}$$

$$X_i^{(n)}(r) = X_i^{(n)}[x; \theta^{(n)} + r\xi^{(n)}], \quad \text{and} \tag{33}$$

$$X_i^{(n)} = X_i^{(n)}(0) \tag{34}$$

throughout the proof.

Fix $i \in \{0, 1, \dots, n\}$. Then, where we understand the square of the brackets below to be taken componentwise,

$$\begin{aligned} D_{r,i}^{(n)} &= \frac{1}{rn}\left(\sigma\left(\left(K_{i-1}^{(n)} + rL_{i-1}^{(n)}\right)X_{i-1}^{(n)}(r) + b_{i-1}^{(n)} + r\beta_{i-1}^{(n)}\right) - \sigma\left(K_{i-1}^{(n)}X_{i-1}^{(n)} + b_{i-1}^{(n)}\right)\right) \\ &\quad + D_{r,i-1}^{(n)} \\ &= \frac{1}{rn}\left[\left(K_{i-1}^{(n)} + rL_{i-1}^{(n)}\right)X_{i-1}^{(n)}(r) + r\beta_{i-1}^{(n)} - K_{i-1}^{(n)}X_{i-1}^{(n)}\right]\odot\dot\sigma\left(K_{i-1}^{(n)}X_{i-1}^{(n)} + b_{i-1}^{(n)}\right) \\ &\quad + \frac{1}{2rn}\left[\left(K_{i-1}^{(n)} + rL_{i-1}^{(n)}\right)X_{i-1}^{(n)}(r) + r\beta_{i-1}^{(n)} - K_{i-1}^{(n)}X_{i-1}^{(n)}\right]^2\odot\ddot\sigma(\xi_i) + D_{r,i-1}^{(n)} \\ &= \frac{1}{n}\left[K_{i-1}^{(n)}D_{r,i-1}^{(n)} + L_{i-1}^{(n)}X_{i-1}^{(n)} + \beta_{i-1}^{(n)}\right]\odot\dot\sigma\left(K_{i-1}^{(n)}X_{i-1}^{(n)} + b_{i-1}^{(n)}\right) \end{aligned}$$

$$+\frac{r}{2n}\left[K_{i-1}^{(n)}D_{r,i-1}^{(n)}+L_{i-1}^{(n)}X_{i-1}^{(n)}(r)+\beta_{i-1}^{(n)}\right]^2\odot\ddot{\sigma}(\xi_i)+D_{r,i-1}^{(n)} \tag{35}$$

where the first equality follows from the definitions of $D_{r,i}^{(n)}$, $X_{i-1}^{(n)}(r)$ and $X_{i-1}^{(n)}$, the second equality follows from Taylor's theorem for some $\xi_i\in\mathbb{R}^d$, and the third equality follows from the definition of $D_{r,i-1}^{(n)}$. We can bound ξ_i by

$$\xi_i\geq\min\left\{K_{i-1}^{(n)}X_{i-1}^{(n)}+b_{i-1}^{(n)},\left(K_{i-1}^{(n)}+rL_{i-1}^{(n)}\right)X_{i-1}^{(n)}(r)+b_{i-1}^{(n)}+r\beta_{i-1}^{(n)}\right\}$$
$$\xi_i\leq\max\left\{K_{i-1}^{(n)}X_{i-1}^{(n)}+b_{i-1}^{(n)},\left(K_{i-1}^{(n)}+rL_{i-1}^{(n)}\right)X_{i-1}^{(n)}(r)+b_{i-1}^{(n)}+r\beta_{i-1}^{(n)}\right\}$$

where we understand the inequalities, minimum and maximum to hold componentwise.

By Lemma 4.6 $X^{(n)}$, $X^{(n)}(r)$ are uniformly bounded by a constant depending only on C (for $r\leq 1$ say), and so we can assume ξ_i is uniformly bounded independent of i and n. Hence, if we can show that $\sup_{r\in\{0,1\}}\sup_{i\in\{0,1,\ldots,n\}}\|D_{r,i}^{(n)}\|\leq C'$ where C' depends only on C, in particular is independent of n, then

$$D_{r,i}^{(n)}=D_{r,i-1}^{(n)}+\frac{1}{n}\left[K_{i-1}^{(n)}D_{r,i-1}^{(n)}+L_{i-1}^{(n)}X_{i-1}^{(n)}+\beta_{i-1}^{(n)}\right]\odot\dot{\sigma}\left(K_{i-1}^{(n)}X_{i-1}^{(n)}+b_{i-1}^{(n)}\right)+O\left(\frac{r}{n}\right).$$

By induction the above implies (31).

We are left to show that $D_{r,i}^{(n)}$ is uniformly bounded in i and r. From (35) we may infer the existence of constants c_1 and c_2, that are independent of r and n and, given C can also be made independent of $K^{(n)}, L^{(n)}, b^{(n)}, \beta^{(n)}$, such that

$$\|D_{r,i}^{(n)}\|\leq\left(\frac{c_1(1+r)}{n}+1\right)\|D_{r,i-1}^{(n)}\|+\frac{c_2}{n}.$$

Hence, by induction,

$$\|D_{r,i}^{(n)}\|\leq\sum_{k=0}^{i-1}\left(1+\frac{c_1(1+r)}{n}\right)^k\frac{c_2}{n}\leq c_2\left(1+\frac{c_1(1+r)}{n}\right)^n\to c_2e^{c_1(1+r)}\quad\text{as } n\to\infty.$$

It follows that $\sup_{r\in\{0,1\}}\sup_{i\in\{0,1,\ldots,n\}}\|D_{r,i}^{(n)}\|$ can be bounded as claimed. □

We now use the above result to deduce the behaviour of the output of the ODE model (10) when the parameters K and b are perturbed.

Lemma 4.14 *Assume $\sigma\in C^2$, $\sigma(0)=0$ and σ acts componentwise. Let $\theta=(K,b)$ and $\xi=(L,\beta)$ where $K,L\in H^1([0,1];\mathbb{R}^{d\times d})$ and $b,\beta\in H^1([0,1];\mathbb{R}^d)$. Furthermore, let $X(t;x,\theta)$ be defined as a solution to (10) for the input θ and initial condition $X(0)=x$. Define, for $r>0$,*

$$D_r(t;x,\theta,\xi)=\frac{1}{r}\left(X(t;x,\theta+r\xi)-X(t;x,\theta)\right). \tag{36}$$

Then,

$$\lim_{r\to0^+}D_r(1;x,\theta,\xi)=\int_0^1\left[\exp\left(\int_t^1\dot{\sigma}\left(K(s)X(s;x,\theta)+b(s)\right)\odot K(s)\,\mathrm{d}s\right)\right.$$
$$\left.\times\left(L(t)X(t;x,\theta)+\beta(t)\right)\odot\dot{\sigma}\left(K(t)X(t;x,\theta)+b(t)\right)\right]\mathrm{d}t.$$

Proof Let $K^{(n)}, L^{(n)}, b^{(n)}, \beta^{(n)}$ be any discrete sequences converging to K, L, b, β respectively with

$$\sup_{n\in\mathbb{N}} \max\left\{R_n^{(1)}(K^{(n)}), R_n^{(1)}(L^{(n)}), R_n^{(2)}(b^{(n)}), R_n^{(2)}(\beta^{(n)})\right\} < +\infty \tag{37}$$

and where the convergence is uniform:

$$\max_{i\in\{0,1,\dots,n-1\}} \sup_{t\in[t_i,t_{i+1}]} \max\left\{\|K_i^{(n)} - K(t)\|, \|L_i^{(n)} - L(t)\|, \|b_i^{(n)} - b(t)\|, \|\beta_i^{(n)} - \beta(t)\|\right\} \to 0.$$

For example, the recovery sequences, as defined by (26) and (27), are sufficient. To shorten notation we write

$$\begin{aligned} D_r &= D_r(1; x, \theta^{(n)}, \xi^{(n)}) \\ X_r(t) &= X(t; x, \theta + r\xi) \\ X(t) &= X_0(t) \end{aligned}$$

and again use the abbreviations in (32)-(34) where $D_{r,i}^{(n)}(x, \theta, \xi)$ is defined by (30).

By Lemma 4.6 we have $X_n^{(n)}(r) \to X_r(1)$ as $n \to \infty$ for all $r \geq 0$. Hence, $\lim_{n\to\infty} D_{r,n}^{(n)} = D_r$. By Lemma 4.12 (and in particular using that the $O(r)$ term in (31) is independent of n given the bound (37) on $K^{(n)}, L^{(n)}, b^{(n)}, \beta^{(n)}$) we have that

$$\lim_{r\to 0^+} D_r = \lim_{r\to 0^+} \lim_{n\to\infty} D_{r,n}^{(n)} = \lim_{r\to 0^+} \lim_{n\to\infty} \left(\frac{1}{n}\sum_{i=0}^{n-1} A_i^{(n)} B_i^{(n)} + O(r)\right) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} A_i^{(n)} B_i^{(n)}$$

where

$$\begin{aligned} A_i^{(n)} &= \prod_{j=i+1}^{n-1} \left(\mathrm{Id} + \frac{1}{n}\dot\sigma\left(K_j^{(n)} X_j^{(n)} + b_j^{(n)}\right) \odot K_j^{(n)}\right) \text{ and} \\ B_i^{(n)} &= \left[L_i^{(n)} X_i^{(n)} + \beta_i^{(n)}\right] \odot \dot\sigma\left(K_i^{(n)} X_i^{(n)} + b_i^{(n)}\right). \end{aligned}$$

Convergence in TL^∞ implies convergence of the empirical integral, i.e. a relatively standard argument implies that, if $\max_{i\in\{0,1,\dots,n-1\}} \sup_{t\in[t_i,t_{i+1}]} \|F(t) - F_n(t_i)\| \to 0$, then $\frac{1}{n}\sum_{i=0}^{n-1} F_n(t_i) \to \int_0^1 F(t)\,\mathrm{d}t$ (with the result also being true for weaker assumptions, e.g. convergence in TL^1). By assumptions on the sequences $K^{(n)}, L^{(n)}, b^{(n)}, \beta^{(n)}$ we easily have that

$$\max_{i\in\{0,1,\dots,n-1\}} \sup_{t\in[t_i,t_{i+1}]} \left\|B_i^{(n)} - [L(t)X(t) + \beta(t)] \odot \dot\sigma(K(t)X(t) + b(t))\right\| \to 0.$$

We are left to find the uniform limit of $A_i^{(n)}$.

If $\max_{i\in\{0,1,\dots,n-1\}} \sup_{t\in[t_i,t_{i+1}]} \|F(t) - F_n(t_i)\| \leq \varepsilon$ and $\|F\|_{L^\infty} \leq M$, then

$$\begin{aligned} \left\|\frac{1}{n}\sum_{i=\lfloor tn\rfloor+1}^{n-1} F_n(t_i) - \int_t^1 F(s)\,\mathrm{d}s\right\| &\leq \int_t^{t_{\lfloor tn\rfloor+1}} \|F\|\,\mathrm{d}s + \sum_{i=\lfloor tn\rfloor+1}^{n-1} \int_{t_i}^{t_{i+1}} \|F_n(s) - F(s)\|\,\mathrm{d}s \\ &\leq \varepsilon + \frac{M}{n}, \end{aligned}$$

for any $t \in [0, 1]$. Hence,

$$\left\|\prod_{i=\lfloor tn\rfloor+1}^{n-1} \exp\left(\frac{1}{n}F_n(t_i)\right) - \exp\left(\int_t^1 F(s)\,\mathrm{d}s\right)\right\|$$

$$
\begin{aligned}
&= \left\| \exp\left(\frac{1}{n} \sum_{i=\lfloor tn \rfloor + 1}^{n-1} F_n(t_i) \right) - \exp\left(\int_t^1 F(s)\,\mathrm{d}s \right) \right\| \\
&\leq \left\| \frac{1}{n} \sum_{i=\lfloor tn \rfloor + 1}^{n-1} F_n(t_i) - \int_t^1 F(s)\,\mathrm{d}s \right\| e^{\frac{1}{n}\sum_{i=\lfloor tn \rfloor+1}^{n-1} F_n(t_i) - \int_t^1 F(s)\,\mathrm{d}s} e^{\int_t^1 F(s)\,\mathrm{d}s} \\
&\leq \left(\varepsilon + \frac{M}{n} \right) e^{M+\varepsilon+\frac{M}{n}}
\end{aligned}
$$

using the inequality $\|e^{X+Y} - e^X\| \leq \|Y\| e^{\|X\|} e^{\|Y\|}$ (for any square matrices X, Y; see Appendix A) applied to $X = \int_t^1 F(s)\,\mathrm{d}s$ and $Y = \frac{1}{n}\sum_{i=\lfloor tn \rfloor+1}^{n-1} F_n(t_i) - \int_t^1 F(s)\,\mathrm{d}s$. We define $F_n : \{t_j\}_{j=0}^{n-1} \to \mathbb{R}^{d\times d}$ and $F : [0,] \to \mathbb{R}^{d\times d}$ by

$$
\begin{aligned}
F_n(t_j) &= \log\left(\mathrm{Id} + \frac{1}{n} C_j^{(n)} \right)^n, & C_j^{(n)} &= \dot\sigma\left(K_j^{(n)} X_j^{(n)} + b_j^{(n)} \right) \odot K_j^{(n)}, \\
F(s) &= C(s), \text{ and} & C(s) &= \dot\sigma\left(K(s)X(s) + b(s) \right) \odot K(s).
\end{aligned}
$$

By construction $\prod_{j=i+1}^{n-1} \exp\left(\frac{1}{n} F_n(t_j)\right) = A_i^{(n)}$. The L^∞ bound, M, on F is readily verified from the L^∞ bounds on each of K, X and b. We show the uniform convergence of F_n to F shortly. For now we assume this is true, so we can fix an arbitrary $\varepsilon > 0$ and have an N such that

$$
\begin{aligned}
&\max_{i\in\{0,1,\ldots,n-1\}} \sup_{t\in[t_i,t_{i+1}]} \|F(t) - F_n(t_i)\| \\
&\quad = \max_{i\in\{0,1,\ldots,n-1\}} \sup_{t\in[t_i,t_{i+1}]} \left\| C(s) - n\log\left(\mathrm{Id} + \frac{1}{n} C_j^{(n)} \right) \right\| \leq \varepsilon,
\end{aligned} \tag{38}
$$

for all $n \geq N$. For $t \in [t_i, t_{i+1}]$ we have $\lfloor tn \rfloor = i$, and so

$$
\max_{i\in\{0,1,\ldots,n-1\}} \sup_{t\in[t_i,t_{i+1}]} \left\| A_i^{(n)} - \exp\left(\int_t^1 F(s)\,\mathrm{d}s \right) \right\| \leq \left(\varepsilon + \frac{M}{n} \right) e^{M+\varepsilon+\frac{M}{n}}.
$$

Hence, $A_i^{(n)}$ converges uniformly to $\exp\left(\int_t^1 F(s)\,\mathrm{d}s\right)$.

To complete the proof, we show that (38) holds. Analogously to when we considered the sequence B_n, we can infer the existence of N such that, if $n \geq N$, then

$$
\max_{i\in\{0,1,\ldots,n-1\}} \sup_{t\in[t_i,t_{i+1}]} \left\| C(s) - C_j^{(n)} \right\| \leq \varepsilon. \tag{39}
$$

By [40, Proposition 2.9], there exists a constant c (independent of all parameters) such that (assuming $\|C_j^{(n)}\| \leq \frac{n}{2}$)

$$
\begin{aligned}
\left\| C(s) - n\log\left(\mathrm{Id} + \frac{1}{n} C_j^{(n)} \right) \right\| &\leq \left\| C(s) - C_j^{(n)} \right\| + n \left\| \frac{1}{n} C_j^{(n)} - \log\left(\mathrm{Id} + \frac{1}{n} C_j^{(n)} \right) \right\| \\
&\leq \left\| C(s) - C_j^{(n)} \right\| + \frac{c}{n} \|C_j^{(n)}\|^2.
\end{aligned} \tag{40}
$$

Since $\|C_j^{(n)}\|$ is uniformly bounded in j and n, (39) and (40) imply (38). □

The previous result shows the limit $\lim_{r\to 0^+} D_r(t; x, \theta, \xi)$ exists for $t = 1$. Whilst this is all we require in the sequel, we note that a rescaling argument implies that the limit exists for all $t > 0$. In particular, if we fix $t > 0$ and let $\hat X(\cdot; x, \hat\theta)$ satisfy $\frac{\mathrm{d}}{\mathrm{d}s}\hat X(s) =$

$\hat{\sigma}(\hat{K}(s)\hat{X}(s) + \hat{b}(s))$ where $\hat{\sigma}(\cdot) = t\sigma(\cdot)$ and $\hat{\theta} = (\hat{K}(\cdot), \hat{b}(\cdot)) = (K(\cdot t), b(\cdot t))$, then we can apply the above lemma directly to $\hat{X}$ to deduce the existence of $\lim_{r\to 0^+} \hat{D}_r(1; x, \hat{\theta}, \hat{\xi})$, where $\hat{D}_r(s; x, \hat{\theta}, \hat{\xi}) = \frac{1}{r}(\hat{X}(s; x, \hat{\theta} + r\hat{\xi}) - \hat{X}(s; x, \hat{\theta}))$. Because $\hat{X}(s; x, \hat{\theta}) = X(st; x, \theta)$, we have $\hat{D}_r(1; x, \hat{\theta}, \hat{\xi}) = D_r(t; x, \theta, \xi)$.

Using the above result we can compute the Gâteaux derivative of $\mathcal{E}_\infty$, defined as

$$\mathrm{d}\mathcal{E}_\infty(\theta; \xi) = \lim_{r\to 0^+} \frac{\mathcal{E}_\infty(\theta + r\xi) - \mathcal{E}_\infty(\theta)}{r}.$$

Lemma 4.15 *Define $\mathcal{E}_\infty$, E_∞, $R^{(i)}_\infty$, for $i = 1, 2$, and $R^{(j)}$, for $j = 3, 4$, as in Sects. 1.2-1.3. In addition to the assumptions in Lemma 4.14 we assume that $h \in C^2(\mathbb{R}^m; \mathbb{R}^m)$, $\mathcal{L}(\cdot, y) \in C^2(\mathbb{R}^m; \mathbb{R})$ for all $y \in \mathbb{R}^m$, and all norms $\|\cdot\|$ on $\mathbb{R}^d$ and $\mathbb{R}^{d\times d}$ are induced by inner products. Let $\{(x_s, y_s)\}_{s=1}^S \subset \mathbb{R}^d \times \mathbb{R}^m$, $\theta = (K, b, W, c) \in \Theta$ and $\xi = (L, \beta, V, \gamma) \in \Theta$ where Θ is given by (15). We define $D_r(t; x, \theta, \xi)$ by (36) for $r > 0$ and*

$$D_0(t; x, \theta, \xi) = \lim_{r\to 0^+} D_r(t; x, \theta, \xi).$$

Then,

$$\begin{aligned}\mathrm{d}\mathcal{E}_\infty(\theta; \xi) = \sum_{s=1}^S \nabla_z \mathcal{L}(h(WX(1; x_s, \theta) + c), y_s) \cdot \Big[\dot{h}(WX(1; x_s, \theta) + c) \\ \odot (WD_0(1; x_s, \theta, \xi) + VX(1; x_s, \theta) + \gamma) \Big] \\ + \alpha_1 \mathrm{d}R^{(1)}_\infty(K; L) + \alpha_2 \mathrm{d}R^{(2)}_\infty(b; \beta) + \alpha_3 \mathrm{d}R^{(3)}(W; V) + \alpha_4 \mathrm{d}R^{(4)}(c; \gamma),\end{aligned}$$

where with a small abuse of notation we wrote $X(t; x, \theta) = X(t; x, K, b)$, ∇_z is the derivative with respect to the first argument, and

$$\begin{aligned}\mathrm{d}R^{(1)}_\infty(K; L) &= 2\langle \dot{K}, \dot{L}\rangle_{L^2} + 2\tau_1\langle K(0), L(0)\rangle, & \mathrm{d}R^{(3)}(W; V) &= 2\langle W, V\rangle, \\ \mathrm{d}R^{(2)}_\infty(b; \beta) &= 2\langle \dot{b}, \dot{\beta}\rangle_{L^2} + 2\tau_2\langle b(0), \beta(0)\rangle, & \mathrm{d}R^{(4)}(c; \gamma) &= 2\langle c, \gamma\rangle.\end{aligned}$$

Proof We consider the derivative of each term in $\mathcal{E}_\infty$ separately. For ease of notation let

$$\begin{aligned}X_r(t) &= X(t; x_s, \theta + r\xi) \\ X(t) &= X_0(t) \\ \delta h_r &= h((W + rV)X_r(1) + c + r\gamma) - h(WX(1) + c) \\ D_r(t) &= D_r(t, x_s, \theta, \xi).\end{aligned}$$

Applying this new notation to (36), we get $D_r(t) = \frac{1}{r}(X_r(t) - X(t))$.

By Taylor's theorem, for each r there exists $z_r \in \mathbb{R}^m$ such that

$$\begin{aligned}&\mathcal{L}(h(WX_r(1) + c + r\gamma), y_s) - \mathcal{L}(h(WX(1) + c), y_s) \\ &\quad = \nabla_z \mathcal{L}(h(WX(1) + c), y_s) \cdot \delta h_r + \frac{1}{2}(\delta h_r)^\top \nabla^2 \mathcal{L}(z_r, y_s)\delta h_r.\end{aligned}$$

Similarly, by another application of Taylor's theorem, for each r there exists $t_r \in \mathbb{R}^m$ such that

$$\delta h_r = \dot{h}(WX(1) + c) \odot [(W + rV)X_r(1) + r\gamma - WX(1)]$$

$$
\begin{aligned}
&\quad + \frac{1}{2}\left[(W + rV)X_r(1) + r\gamma - WX(1)\right]^2 \odot \ddot{h}(t_r) \\
&= \dot{h}(WX(1) + c) \odot \left[W(X_r(1) - X(1)) + r(VX_r(1) + \gamma)\right] \\
&\quad + \frac{1}{2}\left[W(X_r(1) - X(1)) + r(VX_r(1) + \gamma)\right]^2 \odot \ddot{h}(t_r) \\
&= r\dot{h}(WX(1) + c) \odot \left[WD_r(1) + VX_r(1) + \gamma\right] + O(r^2),
\end{aligned}
$$

where the square is to be evaluated componentwise and we use that $h \in C^2$ and t_r is bounded as function of r. Hence,

$$
\begin{aligned}
&\frac{1}{r}\left[\mathcal{L}(h(WX_r(1) + c + r\gamma), y_s) - \mathcal{L}(h(WX(1) + c), y_s)\right] \\
&\qquad = \nabla_z \mathcal{L}(h(WX(1) + c), y_s) \cdot \left[\dot{h}(WX(1) + c) \odot (WD_r(1) + VX_r(1) + \gamma)\right] + O(r).
\end{aligned}
$$

Taking $r \to 0$ and applying Lemma 4.14 gives

$$
\begin{aligned}
\mathrm{d}E_\infty(\theta; x_s, y_s; \xi) &= \lim_{r \to 0^+} \frac{1}{r}\left[\mathcal{L}(h(WX(1) + c + r\gamma), y_s) - \mathcal{L}(h(WX(1) + c), y_s)\right] \\
&= \nabla_z \mathcal{L}(h(WX(1; x_s, \theta) + c), y_s) \cdot \Bigg[\dot{h}(WX(1; x_s, \theta) + c) \\
&\quad \odot (WD_0(1; x_s, \theta, \xi) + VX(1; x_s, \theta) + \gamma)\Bigg].
\end{aligned}
$$

It is straightforward to show that the Gâteaux derivatives of the regularisation functionals $R_\infty^{(1)}$, $R_\infty^{(2)}$, $R^{(3)}$, and $R^{(4)}$ are as claimed. Summing the individual terms completes the proof. □

Finally we can deduce the regularity of minimisers of $\mathcal{E}_\infty$ by applying techniques from the study of elliptic differential equations (see for example [34, Section 2.2.2] for the same techniques).

Proof of Proposition 2.2 Assume that $\theta = (K, b, W, c) \in \Theta$ be a minimiser of $\mathcal{E}_\infty$. We will show that $K \in H^2_{\mathrm{loc}}([0, 1]; \mathbb{R}^{d \times d})$ (the argument for $b \in H^2_{\mathrm{loc}}([0, 1]; \mathbb{R}^d)$ is analogous).

Since θ is a minimiser of $\mathcal{E}_\infty$, $\mathrm{d}\mathcal{E}_\infty(\theta; \xi) = 0$ for all $\xi \in \Theta$. Let $\Omega_N = [1/N, 1 - 1/N]$ and $\gamma_N \in C^\infty$ be a cut-off function that has support in Ω_{2N} and is identically one on Ω_N. Let $K_N = \gamma_N \odot K$. We extend K_N to the whole of $\mathbb{R}$ by setting $K_N(t) = 0$ for all $t \in \mathbb{R} \setminus [0, 1]$. Clearly $K_N \in H^1(\mathbb{R}; \mathbb{R}^{d \times d})$, $K_N = K$ on Ω_N and K_N has support in Ω_{2N}. Let $\xi = (L, 0, 0, 0)$ where $L \in H^1([0, 1]; \mathbb{R}^{d \times d})$ satisfies $L(0) = 0$, then $\mathrm{d}\mathcal{E}_\infty(\theta; \xi) = 0$ implies

$$
\begin{aligned}
\left\langle \dot{K}, \dot{L} \right\rangle_{L^2} = -\frac{1}{\alpha_1} \sum_{s=1}^{S} &\nabla_z \mathcal{L}(h(WX(1; x_s, \theta) + c), y_s) \\
&\cdot \left[\dot{h}(WX(1; x_s, \theta) + c) \odot (WD_0(1; x_s, \theta, \xi))\right].
\end{aligned}
$$

Using the equality above with $\gamma_N \odot L$ in place of L implies,

$$
\begin{aligned}
\langle \dot{K}_N, \dot{L} \rangle_{L^2} &= \langle \dot{\gamma}_N \odot K + \gamma_N \odot \dot{K}, \dot{L} \rangle_{L^2} \\
&= \langle \dot{\gamma}_N \odot K, \dot{L} \rangle_{L^2} + \langle \dot{K}, \gamma_N \odot \dot{L} \rangle_{L^2}
\end{aligned}
$$

$$
\begin{aligned}
&= -\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{\gamma}_N \odot K\right), L\right\rangle_{L^2} + \left\langle \dot{K}, \frac{\mathrm{d}}{\mathrm{d}t}\left(\gamma_N \odot L\right)\right\rangle_{L^2} - \langle \dot{K}, \dot{\gamma}_N \odot L\rangle_{L^2} \\
&= -\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{\gamma}_N \odot K\right) + \dot{\gamma}_N \odot \dot{K}, L\right\rangle_{L^2} + \left\langle \dot{K}, \frac{\mathrm{d}}{\mathrm{d}t}(\gamma_N \odot L)\right\rangle_{L^2} \\
&= -\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{\gamma}_N \odot K\right) + \dot{\gamma}_N \odot \dot{K}, L\right\rangle_{L^2} - \frac{1}{\alpha_1}\sum_{s=1}^{S} \nabla_z \mathcal{L}(h(WX(1; x_s, \theta) + c), y_s) \\
&\cdot \left[\dot{h}(WX(1; x_s, \theta) + c) \odot (WD_0(1; x_s, \theta, (\gamma_N \odot L, 0, 0, 0)))\right]. \qquad (41)
\end{aligned}
$$

We choose $L = L_{N,r}$ where

$$L_{N,r}(t) = \frac{2K_N(t) - K_N(t+r) - K_N(t-r)}{r^2}.$$

Clearly $L_{N,r} \in H^1(\mathbb{R}; \mathbb{R}^{d\times d})$ for every $r > 0$ and all $N > 2$. Furthermore, $L_{N,r}$ has support in $[\frac{1}{2N} - r, 1 - \frac{1}{2N} + r]$. Since the support of $\dot{K}_N(\cdot - r)$ and $\dot{K}_N$ is contained in $[r, 1]$ for $r \le \frac{1}{N}$,

$$
\begin{aligned}
\langle \dot{K}_N, \dot{L}_{N,r}\rangle_{L^2} &= \frac{1}{r^2}\int_0^1 \dot{K}_N(t)\left(2\dot{K}_N(t) - \dot{K}_N(t+r) - \dot{K}_N(t-r)\right)\mathrm{d}t \\
&= \frac{1}{r^2}\int_0^1 \dot{K}_N(t)\left(\dot{K}_N(t) - \dot{K}_N(t+r)\right)\mathrm{d}t \\
&\quad + \frac{1}{r^2}\int_0^1 \dot{K}_N(t)\left(\dot{K}_N(t) - \dot{K}_N(t-r)\right)\mathrm{d}t \\
&= \frac{1}{r^2}\int_r^{1+r} \dot{K}_N(s-r)\left(\dot{K}_N(s-r) - \dot{K}_N(s)\right)\mathrm{d}s \\
&\quad + \frac{1}{r^2}\int_0^1 \dot{K}_N(t)\left(\dot{K}_N(t) - \dot{K}_N(t-r)\right)\mathrm{d}t \\
&= \frac{1}{r^2}\int_r^1 \dot{K}_N(s-r)\left(\dot{K}_N(s-r) - \dot{K}_N(s)\right)\mathrm{d}s \\
&\quad + \frac{1}{r^2}\int_r^1 \dot{K}_N(t)\left(\dot{K}_N(t) - \dot{K}_N(t-r)\right)\mathrm{d}t \\
&= \frac{1}{r^2}\int_r^1 \left\|\dot{K}_N(t) - \dot{K}_N(t-r)\right\|^2 \mathrm{d}t,
\end{aligned}
$$

where $\hat{\xi}_{N,r} = (\gamma_N \odot L_{N,r}, 0, 0, 0)$. From (41), it follows that

$$\int_r^1 \left\|\frac{\dot{K}_N(t) - \dot{K}_N(t-r)}{r}\right\|^2 \mathrm{d}t \le C_1 \|L_{N,r}\|_{L^2([0,1])} + C_2 \sum_{s=1}^{S} \|D_0(1; x_s, \theta, \hat{\xi}_{N,r})\|, \qquad (42)$$

where

$$
\begin{aligned}
C_1 &= \left\|\frac{\mathrm{d}}{\mathrm{d}t}(\gamma_N \odot K)\right\|_{L^2} + \|\dot{\gamma}_N \odot \dot{K}\|_{L^2} \text{ and} \\
C_2 &= \frac{1}{\alpha_1}\left\|\dot{h}(WX(1; x_s, \theta) + c)\right\| \|W\| \max_{s=1,\ldots,S} \left\|\nabla_z \mathcal{L}(h(WX(1; x_s, \theta) + c), y_s)\right\|
\end{aligned}
$$

(we note that the constants C_1, C_2 may depend on N, but do not depend on r).

We note that we can rewrite $L_{N,r} = \frac{2-\tau_r-\tau_{-r}}{r^2} K_N = \frac{(1-\tau_r)(1-\tau_{-r})}{r^2} K_N$, where τ_r is the shift operator defined by $\tau_r \varphi(x) = \varphi(x+r)$. By [62, Theorem 10.55] for any $\psi \in H^1([0,1];\mathbb{R}^{d\times d})$ we have $\left\| \frac{\tau_r - 1}{r} \psi(t+r) \right\|_{L^2([r,1-r])} \leq \|\dot{\psi}\|_{L^2([r,1])}$. Applying this to $\psi = \frac{(1-\tau_{-r})K_N}{r}$ we have, for r sufficiently small,

$$\|L_{N,r}\|^2_{L^2([0,1])} = \left\| \frac{1-\tau_r}{r} \psi \right\|^2_{L^2([r,1-r])} \leq \|\dot{\psi}\|^2_{L^2([r,1])} = \int_r^1 \left\| \frac{\dot{K}_N(t) - \dot{K}_N(t-r)}{r} \right\|^2 \mathrm{d}t. \tag{43}$$

We can write $D_0(1; x_s, \theta, \hat{\xi}_{N,r}) = \int_0^1 A_{N,s}(t) \odot L_{N,r}(t)\, \mathrm{d}t$ where

$$\begin{aligned} A_{N,s}(t) &= B_s(t)\gamma_N(t), \\ B_s(t) &= \exp\left(\int_t^1 \dot{\sigma}\left(K(u)X_s(u) + b(u)\right) \odot K(u)\, \mathrm{d}u \right) \odot \dot{\sigma}\left(K(t)X_s(t) + b(t)\right) X_s(t), \text{ and} \\ X_s(t) &= X(t; x_s, \theta). \end{aligned}$$

Hence,

$$\|D_0(1; x_s, \theta, \hat{\xi}_{N,r})\| \leq C_3 \|L_{N,r}\|_{L^2([0,1])}. \tag{44}$$

Combining (44) with (42) and (43) and Young's inequality we obtain

$$\int_r^1 \left\| \frac{\dot{K}_N(t) - \dot{K}_N(t-r)}{r} \right\|^2 \mathrm{d}t \leq C_4.$$

Hence, by [62, Theorem 10.55] $\dot{K}_N \in H^1([0,1];\mathbb{R}^{d\times d})$. Since this is true for all N, we have that $\dot{K} \in H^1_{\mathrm{loc}}([0,1];\mathbb{R}^{d\times d})$. Hence, $K \in H^2_{\mathrm{loc}}([0,1];\mathbb{R}^{d\times d})$.

The argument for $b \in H^2_{\mathrm{loc}}([0,1];\mathbb{R}^d)$ is analogous. □

4.5 The forward pass as a discretised ODE

In this section we prove Corollary 2.3.

Lemma 4.16 *Let $K \in H^1([0,1];\mathbb{R}^{d\times d})$, $b \in H^1([0,1];\mathbb{R}^d)$, and let $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ be Lipschitz continuous with Lipschitz constant $L_\sigma > 0$. Let $x \in \mathbb{R}^d$ and suppose that $X : [0,1] \to \mathbb{R}^d$ is the solution to the ODE in (10) with initial condition $X(0) = x$. Let $n \in \mathbb{N}$ and let $K^{(n)} \in L^0(\mu_n;\mathbb{R}^{d\times d})$, $b^{(n)} \in L^0(\mu_n;\mathbb{R}^d)$ be such that there exists a $\delta_n > 0$ such that, for all $i \in \{0,1,\ldots,n-1\}$, $\|K_i^{(n)} - K(i/n)\| < \delta_n$ in matrix operator norm and $\|b_i^{(n)} - b(i/n)\| < \delta_n$. Moreover, let $X_i^{(n)}$ $(i = 0,1,\ldots,n)$ be the solutions to (6) with $X_0^{(n)} = x$. Then there exists an $\varepsilon_n \in \mathbb{R}$ such that, for all $i \in \{0,1,\ldots,n\}$, (13) is satisfied with $\delta = \delta_n$ and $A_n = \frac{1}{n}(1 + \|X\|_{L^\infty}) L_\sigma \delta_n + \varepsilon_n$. Moreover, $\varepsilon_n = o\left(\frac{1}{n}\right)$ as $n \to \infty$.*

Proof We follow closely standard proofs of the convergence of the explicit Euler scheme for well-posed ODEs, [7, Theorem 5.9], [63, Section 6.3.3].

First we note that the case $i = 0$ is trivial, so we will consider $i \geq 1$ from here on.

By the Sobolev embedding theorem [1] K and b are continuous. Since $y \mapsto \sigma(K(t)y + b(t))$ is Lipschitz continuous, by standard ODE theory [39] there is a unique solution X to (10) and this X is continuous. Moreover, $t \mapsto \sigma\left(K(t)X(t) + b(t)\right)$ is continuous and thus $\dot{X}$ is continuous. In particular, $\dot{X}$ is bounded on $[0,1]$. Let $n, k \in \mathbb{N}$ with $k \leq n$. We compute, using Taylor's theorem, for C^1 functions [21, (2.22)],

$$X(k/n) = X((k-1)/n) + \frac{1}{n}\dot{X}((k-1)/n) + r_{k,n}$$

$$= X((k-1)/n) + \frac{1}{n}\sigma(K((k-1)/n)X((k-1)/n) + b((k-1)/n)) + r_{k,n},$$

where $r_{k,n} \in \mathbb{R}^d$ is such that $\|r_{k,n}\| = o\left(\frac{1}{n}\right)$ as $n \to \infty$. Moreover

$$X_k^{(n)} = X_{k-1}^{(n)} + \frac{1}{n}\sigma(K_{k-1}^{(n)}X_{k-1}^{(n)} + b_{k-1}^{(n)})$$

and thus

$$\begin{aligned} X(k/n) - X_k^{(n)} &= X((k-1)/n) - X_{k-1}^{(n)} \\ &\quad + \frac{1}{n}\left(\sigma(K((k-1)/n)X((k-1)/n) + b((k-1)/n)) \right. \\ &\quad \left. -\sigma(K_{k-1}^{(n)}X_{k-1}^{(n)} + b_{k-1}^{(n)})\right) + r_{k,n}. \end{aligned}$$

Using $\|K_i^{(n)} - K(i/n)\| < \delta_n$ and $\|b_i^{(n)} - b(i/n)\| < \delta_n$ and the fact that $L_\sigma > 0$ is a Lipschitz constant for σ, we find

$$\begin{aligned} &\left\|\sigma(K((k-1)/n)X((k-1)/n) + b((k-1)/n)) - \sigma(K_{k-1}^{(n)}X_{k-1}^{(n)} + b_{k-1}^{(n)})\right\| \\ &\leq L_\sigma \left\|K((k-1)/n)X((k-1)/n) + b((k-1)/n) - (K_{k-1}^{(n)}X_{k-1}^{(n)} + b_{k-1}^{(n)})\right\| \\ &\leq L_\sigma \left\|K((k-1)/n)\left(X((k-1)/n) - X_{k-1}^{(n)}\right)\right\| + L_\sigma \|\left(K((k-1)/n) - K_{k-1}^{(n)}\right)X_{k-1}^{(n)}\| \\ &+ L_\sigma \|b((k-1)/n) - b_{k-1}^{(n)}\| \\ &\leq L_\sigma \|K\|_{L^\infty}\left\|X((k-1)/n) - X_{k-1}^{(n)}\right\| + L_\sigma\delta_n\|X_{k-1}^{(n)}\| + L_\sigma\delta_n \\ &\leq L_\sigma(\|K\|_{L^\infty} + \delta_n)\left\|X((k-1)/n) - X_{k-1}^{(n)}\right\| + L_\sigma\delta_n\|X\|_{L^\infty} + L_\sigma\delta_n. \end{aligned}$$

For the final inequality, we used

$$\begin{aligned} \|X_{k-1}^{(n)}\| &\leq \left\|X((k-1)/n) - X_{k-1}^{(n)}\right\| + \|X((k-1)/n)\| \\ &\leq \left\|X((k-1)/n) - X_{k-1}^{(n)}\right\| + \|X\|_{L^\infty}. \end{aligned}$$

Since K is continuous on $[0, 1]$, we have $\|K\|_{L^\infty} < \infty$. Similarly, since X is continuous, we have $\|X\|_{L^\infty} < \infty$. Combining the above we get

$$\begin{aligned} \left\|X(k/n) - X_k^{(n)}\right\| &\leq \left(1 + \frac{1}{n}L_\sigma(\|K\|_{L^\infty} + \delta_n)\right)\left\|X((k-1)/n) - X_{k-1}^{(n)}\right\| \\ &\quad + \frac{1}{n}(1 + \|X\|_{L^\infty})L_\sigma\delta_n + \varepsilon_n, \end{aligned} \tag{45}$$

where we have defined $\varepsilon_n = \max_{1\leq k\leq n}\|r_{k,n}\|$. We note that $\varepsilon_n = o\left(\frac{1}{n}\right)$ as $n \to \infty$ and recall that $A_n = \frac{1}{n}(1 + \|X\|_{L^\infty})L_\sigma\delta_n + \varepsilon_n$. Write $a_k = \left\|X(k/n) - X_k^{(n)}\right\|$ and $C = 1 + \frac{1}{n}L_\sigma(\|K\|_{L^\infty} + \delta_n)$, where we have repressed the dependency on n for notational simplicity. Let $i \in \mathbb{N}$ with $i \leq n$. We claim that

$$a_i \leq A_n \sum_{j=0}^{i-1} C^j. \tag{46}$$

We prove this claim by induction. Since (45) holds for arbitrary k, we have $a_1 \leq Ca_0 + A_n$ directly from (45). Since $a_0 = \|x - x\| = 0$, (46) holds for $i = 1$. Now let $k \in \mathbb{N}$ with $k \leq n$

and assume that (46) holds for $i = k-1$. Then, combining (46) with (45) we deduce that

$$a_k \leq Ca_{k-1} + A_n \leq C\left(A_n \sum_{j=0}^{k-1} C^j\right) + A_n = A_n\left(1 + \sum_{j=0}^{k-1} C^{j+1}\right) = A_n \sum_{j=0}^{k} C^j.$$

Thus claim (46) is proven. Since $C > 1$, we compute

$$\sum_{j=0}^{i-1} C^j = \frac{1-C^i}{1-C} = \frac{n}{L_\sigma(\|K\|_{L^\infty} + \delta_n)}\left[\left(1 + \frac{1}{n}L_\sigma(\|K\|_{L^\infty} + \delta_n)\right)^i - 1\right].$$

Using that $\left(1 + \frac{1}{n}L_\sigma(\|K\|_{L^\infty} + \delta_n)\right)^i \leq \exp\left(\frac{i}{n}L_\sigma(\|K\|_{L^\infty} + \delta_n)\right)$, we find that

$$a_i \leq \frac{n}{L_\sigma(\|K\|_{L^\infty} + \delta_n)} A_n\left[\exp\left(\frac{i}{n}L_\sigma(\|K\|_{L^\infty} + \delta_n)\right) - 1\right],$$

as required. □

We now check that the conditions of Lemma 4.16 hold.

Lemma 4.17 *Let $\Theta^{(n)}$ and Θ be given by (14) and (15) respectively. Define $\mathcal{E}_n$, $\mathcal{E}_\infty$, E_n, E_∞, $R_n^{(i)}$, $R_\infty^{(i)}$, $R^{(j)}$ for $i = 1, 2$, $j = 3, 4$ as in Sects. 1.1-1.3. Assume that the assumptions of Theorem 2.1 hold. If $\{(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)})\} \subset \Theta^{(n)}$ is a sequence of minimisers of $\mathcal{E}_n$ and $(K, b, W, c) \in \Theta$ is the minimiser of $\mathcal{E}_\infty$ which we assume to be unique then we have*

$$\max_{i\in\{0,\ldots,n-1\}} \left\| K\left(\frac{i}{n}\right) - K_i^{(n)} \right\| \to 0 \quad \text{and} \quad \max_{i\in\{0,\ldots,n-1\}} \left\| b\left(\frac{i}{n}\right) - b_i^{(n)} \right\| \to 0,$$

as $n \to \infty$.

Proof Let $(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)})$ minimise $\mathcal{E}_n$. Choose any subsequence $\{n_m\}_{m\in\mathbb{N}}$ of $\mathbb{N}$. By Theorem 2.1 there exists a further subsequence that converges to a minimiser (K, b, W, c) of $\mathcal{E}_\infty$. Since the minimiser is unique, we have $(K^{(n_m)}, b^{(n_m)}, W^{(n_m)}, c^{(n_m)}) \to (K, b, W, c)$. Furthermore, since $\mathcal{E}_{n_m}(K^{(n_m)}, b^{(n_m)}, W^{(n_m)}, c^{(n_m)}) < +\infty$, we have, by Proposition 4.3, that there exists a further subsequence $\{n_{m_k}\}_{k\in\mathbb{N}}$ such that

$$\max_{i\in\{0,\ldots,n_{m_k}-1\}} \left\| K\left(\frac{i}{n_{m_k}}\right) - K_i^{(n_{m_k})} \right\| \to 0, \qquad \max_{i\in\{0,\ldots,n_{m_k}-1\}} \left\| b\left(\frac{i}{n_{m_k}}\right) - b_i^{(n_{m_k})} \right\| \to 0,$$

as $k \to \infty$. We have that any subsequence of $(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)})$ contains a further subsequence that converges uniformly. Now if we suppose that $(K^{(n)}, b^{(n)}, W^{(n)}, c^{(n)})$ does not converge uniformly to (K, b, W, c), then there exists an $\varepsilon > 0$ and a subsequence (which we index by n_m) such that the L^∞ norm of $(K^{(n_m)} - K, b^{(n_m)} - b, W^{(n_m)} - W, c^{(n_m)} - c)$ is bounded from below by ε. But this subsequence cannot now contain a further subsequence that converges uniformly; a contradiction. It follows that uniform convergence holds across the whole sequence as required. □

The proof of Corollary 2.3 follows directly from Lemmas 4.16 and 4.17.

5 Discussion and conclusions

In this paper we proved that the variational limit of the residual neural network is an ODE system, thereby rigorously justifying the observations in [23,37]. These and similar observations have already inspired new architectures for neural networks, e.g. [36,67,80, 87] and the hope is that this work can help in the justification and analysis of these new

architectures. In addition, we proved a regularity result for the coefficients obtained by ResNet training.

We left the question of rates of convergence for the minimisers open (see after Proposition 2.2). We believe this can be approached through a higher order Γ-convergence argument, for example see [6, Theorem 1.5.1], and a coercivity argument, but it falls outside the scope of this current paper.

An interesting open question which the authors intend to field in future work is to recover partial differential equations by simultaneously taking $d \to \infty$ (where d is the number of neurons per layer) and $n \to \infty$. This will mean imposing certain restrictions on the inter-layer connections; in particular, the choice of inter-layer connections is expected to alter the continuum partial differential equation limit.

Another open question concerns our use of explicit regularisation terms in the cost function. In practice often implicit regularisation techniques are used, such as dropout or stochastic gradient descent [38,69,73,85,91]. Incorporating these methods into our setting requires the rigorous mathematical establishment of their regularising effects, which to the best of our knowledge, has not been accomplished as of yet. However, recent work [14] shows that, at least in certain circumstances, the deep layer limit in the absence of explicit regularisation results in a stochastic limit.

In this paper we have established convergence at a variational level. A third open question of interest relates to the convergence of the corresponding gradient flow for the parameters. Except in certain special circumstances, gradient flow convergence does not follow directly from Γ-convergence; in this case an additional difficulty that needs to be taken into account is the ODE constraint. The authors are planning to address this question in future work.

Acknowledgements
This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements 777826 (NoMADS) and 647812). The authors would like to thank Martin Benning, Jeff Calder, Matthias J. Ehrhardt, Nicolás García Trillos and Lukas Lang for enlightening exchanges regarding the work in this paper. We also thank the anonymous referees for their very insightful comments on an earlier version of the manuscript, which have led to improvements in the final paper. MT is grateful for the support of the Cantab Capital Institute for the Mathematics of Information (CCIMI) and the Cambridge Image Analysis group at the University of Cambridge. YvG did a significant part of the work which has contributed to this paper at the University of Nottingham. The authors state that there are no conflicts of interest.

Data Availability
Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Author details
[1]Department of Mathematics, University of Manchester, Manchester M13 9PL, UK, [2] The Alan Turing Institute, London NW1 2DB, UK, [3]Delft Institute of Applied Mathematics, Delft University of Technology, 2628 CD Delft, The Netherlands.

Appendix A: The matrix exponential

For completeness we include a short proof of the inequality

$$\left\| e^{X+Y} - e^X \right\| \leq \|Y\| e^{\|X\|} e^{\|Y\|},$$

for any square matrices X, Y, which is used in the proof of Lemma 4.14. Let $k \in \mathbb{N} \setminus \{0\}$. Then, CA

$$\left\| (X+Y)^k - X^k \right\| \leq \sum_{j=0}^{k-1} \binom{k}{j} \|X\|^j \|Y\|^{k-j} = (\|X\| + \|Y\|)^k - \|X\|^k,$$

where we obtained the inequality by expanding $(X+Y)^k$, applying the triangle inequality and using that $\|XY\| \leq \|X\|\|Y\|$. Then the equality follows from the binomial theorem.

Hence,

$$\begin{aligned}\left\|e^{X+Y}-e^{X}\right\| &= \left\|\sum_{k=0}^{\infty}\frac{1}{k!}\left((X+Y)^k-X^k\right)\right\| \\ &\leq \sum_{k=1}^{\infty}\frac{1}{k!}\left\|(X+Y)^k-X^k\right\| \\ &\leq \sum_{k=1}^{\infty}\frac{1}{k!}\left(\|X\|+\|Y\|\right)^k-\sum_{k=1}^{\infty}\frac{1}{k!}\|X\|^k \\ &= e^{\|X\|+\|Y\|}-e^{\|X\|} \\ &= e^{\|X\|}\left(e^{\|Y\|}-1\right).\end{aligned}$$

Since $e^c - 1 \leq ce^c$ for all $c \geq 0$, we conclude that the inequality holds.

Received: 13 October 2020 Accepted: 18 November 2022
Published online: 16 December 2022

References

1. Adams, R. A., Fournier, J. J. F.: Sobolev spaces, volume 140. Elsevier, (2003)
2. Anthony, M.: Discrete mathematics of neural networks: selected topics, volume 8. SIAM, (2001)
3. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994)
4. Bo, L., Capponi, A., Liao, H.: Deep residual learning via large sample mean-field optimization. preprint arXiv:1906.08894v3, (2020)
5. Braides, A.: Γ-Convergence for Beginners. Oxford University Press, (2002)
6. Braides, A.: Local minimization, variational evolution and Γ-convergence. Springer, (2014)
7. Burden, R. L., Faires, J. D.: Numerical analysis. Cengage Learning, 10th edition, (2010)
8. Celledoni, E., Ehrhardt, M.J., Etmann, C., McLachlan, R.I., Owren, B., Schönlieb, C.-B., Sherry, F.: Structure-preserving deep learning. Eur. J. Appl. Math. **32**(5), 888–936 (2021)
9. Chang, B., Meng, L., Haber, E., Ruthotto, L., Begert, D., Holtham, E.: Reversible architectures for arbitrarily deep residual neural networks. In; Thirty-Second AAAI Conference on Artificial Intelligence, (2018)
10. Chaudhari, P., Oberman, A., Osher, S., Soatto, S., Carlier, G.: Deep relaxation: partial differential equations for optimizing deep neural networks. Res. Math. Sci. **5**(3), 30 (2018)
11. Chen, T. Q., Rubanova, Y., Bettencourt, J., Duvenaud, D. K.: Neural ordinary differential equations. In: Advances in Neural Information Processing Systems, pages 6571–6583, (2018)
12. Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., LeCun, Y.: The loss surfaces of multilayer networks. In: Artificial Intelligence and Statistics, pages 192–204, (2015)
13. Chung, K.L.: On a stochastic approximation method. Ann. Math. Stat. **25**(3), 463–483 (1954)
14. Cohen, A.-S., Cont, R., Rossier, A., Xu, R.: Scaling properties of deep residual networks. In International Conference on Machine Learning, pages 2039–2048, (2021)
15. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Sig. Syst. **2**(4), 303–314 (1989)
16. Cybenko, G.: Neural networks in computational science and engineering. IEEE Comput. Sci. Eng. **3**(1), 36–42 (1996)
17. Dahl, G. E., Sainath, T. N., Hinton, G. E.: Improving deep neural networks for LVCSR using rectified linear units and dropout. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 8609–8613. IEEE, (2013)
18. Dal Maso, G.: An Introduction to Γ-Convergence. Springer, (1993)
19. Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in neural information processing systems, pages 2933–2941. Curran Associates, Inc., (2014)
20. Drucker, H., Le Cun, Y.: Improving generalization performance using double backpropagation. IEEE Trans. Neural Netw. **3**(6), 991–997 (1992)
21. Duistermaat, J. J., Kolk, J. A. C.: Multidimensional real analysis I: differentiation, volume 86. Cambridge University Press, (2004)
22. Dunlop, M. M., Slepčev, D., Stuart, A. M., Thorpe, M.: Large data and zero noise limits of graph-based semi-supervised learning algorithms. Appl. Comput. Harmonic Anal., (2019)
23. E, W.: A proposal on machine learning via dynamical systems. Commun. Math. Statistics **5**(1), 1–11 (2017)
24. E, W., Han, J., Li, Q.: A mean-field optimal control formulation of deep learning. Res. Math. Sci. **6**(1), 10 (2019)
25. E, W., Han, J., Li, Q.: A mean-field optimal control formulation of deep learning. Res. Math. Sci. **6**(1), 10 (2019)
26. Finlay, C., Calder, J., Abbasi, B., Oberman, A.L: Lipschitz regularized deep neural networks generalize and are adversarially robust. preprint arXiv:1808.09540v4 (2019)
27. García Trillos, N., Kaplan, Z., Samakhoana, T., Sanz-Alonso, D.: On the consistency of graph-based Bayesian semi-supervised learning of the scalability of sampling algorithms. J. Mach. Learn. Res. **21**(28), 1–47 (2020)

28. García Trillos, N., Slepčev, D.: A variational approach to the consistency of spectral clustering. Appl. Comput. Harmon. Anal. **45**, 239–281 (2018)
29. García Trillos, N., Slepčev, D.,J. Von Brecht, Laurent, T., Bresson, X.: Consistency of cheeger and ratio graph cuts. J. Mach. Learn. Res., 17(1):6268–6313, (2016)
30. García Trillos, N., Slepčev, D.: Continuum limit of total variation on point clouds. Arch. Ration. Mech. Anal. **220**(1), 193–241 (2016)
31. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 249–256, (2010)
32. Goodfellow, I., Lee, H., Le, Q. V., Saxe, A., Ng, A. Y.: Measuring invariances in deep networks. In: Advances in neural information processing systems, pages 646–654, (2009)
33. Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I., Duvenaud, D.: Ffjord: Free-form continuous dynamics for scalable reversible generative models. preprint arXiv:1810.01367, (2018)
34. Grisvard, P.: Elliptic problems in nonsmooth domains. Pitman Publishing Inc., (1985)
35. Haber, E., Lucka, F., Ruthotto, L.: Never look back - a modified EnKF method and its application to the training of neural networks without back propagation. preprint arXiv:1805.08034, (2018)
36. Haber, E., Ruthotto, L.: Stable architectures for deep neural networks. Inverse Prob. **34**(1), 014004 (2017)
37. Haber, E., Ruthotto, L., Holtham, E., Jun, S.-H.: Learning across scales—multiscale methods for convolution neural networks. In: Thirty-Second AAAI Conference on Artificial Intelligence, (2018)
38. Haeffele, B. D., Vidal, R.: Global optimality in neural network training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7331–7339, (2017)
39. Hale, J.K.: Ordinary Differential Equations. Dover Publications Inc, Mineola, New York, second edition (2009)
40. Hall, B.C.: Lie Groups, Lie Algebras, and Representations: An Elementary Introduction. Springer, New Delhi (2003)
41. Hassoun, M. H.: Fundamentals of artificial neural networks. MIT press, (1995)
42. Haykin, S.: Neural networks: a comprehensive foundation. Prentice Hall PTR, second edition, (1999)
43. He, K., Sun, J.: Convolutional neural networks at constrained time cost. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5353–5360, (2015)
44. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV), pages 1026–1034, (2015)
45. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, (2016)
46. Hertz, J., Krogh, A., Palmer, R. G.: Introduction to the theory of neural computation. Westview Press, (1991)
47. Higham, C.F., Higham, D.J.: Deep learning: an introduction for applied mathematicians. SIAM Rev. **61**(4), 860–891 (2019)
48. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. R.: Improving neural networks by preventing co-adaptation of feature detectors. preprint arXiv:1207.0580, (2012)
49. Hochreiter, S.: Untersuchungen zu dynamischen neuronalen netzen. Diploma, Technische Universität München, 91(1), (1991)
50. Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural Netw. **4**(2), 251–257 (1991)
51. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
52. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In Comput. Vis.- ECCV **2016**, 646–661 (2016)
53. Hunt, K. J., Irwin, G. R., Warwick, K.: Neural network engineering in dynamic control systems. Springer Science & Business Media, (2012)
54. Janocha, K., Czarnecki, W.M.: On loss functions for deep neural networks in classification. Schedae Informaticae **25**, 49–59 (2016)
55. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: Computer Vision, 2009 IEEE 12th International Conference on, pages 2146–2153. IEEE, (2009)
56. Kovachki, N.B., Stuart, A.M.: Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. Inverse Prob. **35**(9), 095005 (2019)
57. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pages 1097–1105, (2012)
58. Kung, S. Y.: Digital Neural Networks. Prentice-Hall, Inc., (1993)
59. Kuo, C., Jay C.: Understanding convolutional neural networks with a mathematical model. J Vis Commun Image Represent **41**, 406–413 (2016)
60. Laurent, T., Brecht, J.: Deep linear networks with arbitrary loss: All local minima are global. In: International Conference on Machine Learning, pages 2908–2913, (2018)
61. Lee, J. D., Simchowitz, M., Jordan, M. I., Recht, B.: Gradient descent only converges to minimizers. In: Conference on Learning Theory, pages 1246–1257, (2016)
62. Leoni, G.: A First Course in Sobolev Spaces, volume 105. American Mathematical Society, (2009)
63. LeVeque, R. J.: Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems, vol. 98. Soc. Ind. Appl. Math, (2007)
64. Li, Q., Chen, L., Tai, C., E, W.: Maximum principle based algorithms for deep learning. J. Mach. Learn. Res. **18**(165), 1–29 (2018)
65. Ljung, L.: Analysis of recursive stochastic algorithms. IEEE Trans. Autom. Control **22**(4), 551–575 (1977)
66. Lu, Y., Ma, C., Lu, Y., Lu, J., Ying, L.: A mean field analysis of deep ResNet and beyond: Towards provably optimization via overparameterization from depth. In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 6426–6436. PMLR, 13–18 Jul (2020)
67. Lu, Y., Zhong, A., Li, Q., Dong, B.: Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In: International Conference on Machine Learning, pages 5181–5190, (2018)

68. Maas, A. L., Hannun, A. Y., Ng, A. Y.: Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning, vol.30, p. 3, (2013)
69. Mianjy, P., Arora, R., Vidal, R.: On the implicit bias of dropout. In: International Conference on Machine Learning, pages 3540–3548, (2018)
70. Mocanu, D.C., Mocanu, E., Stone, P., Nguyen, P.H., Gibescu, M., Liotta, A.: Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. Nat. Commun. **9**(1), 2383 (2018)
71. Nair, V., Hinton, G. E.: Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning, pages 807–814, (2010)
72. Narendra, K.S., Parthasarathy, K.: Identification and control of dynamical systems using neural networks. IEEE Trans. Neural Netw. **1**(1), 4–27 (1990)
73. Neyshabur, B., Tomioka, R., Srebro, N.: In search of the real inductive bias: On the role of implicit regularization in deep learning. preprint arXiv:1412.6614, (2014)
74. Ng, A. Y.: Feature selection, L1 vs. L2 regularization, and rotational invariance. In: International Conference on Machine Learning, (2004)
75. Nielsen, M. A.: Neural Networks and Deep Learning. Determination Press, (2015)
76. Oberman, A.M.: Partial differential equation regularization for supervised machine learning. In: Brenner, S.C., Shparlinski, I., Shu, C.-W., Szyld, D.B. (eds.) 75 Years of Mathematics of Computation. American Mathematical Society (2020)
77. Orponen, P.: Neural networks and complexity theory. In International Symposium on Mathematical Foundations of Computer Science, pages 50–61. Springer, (1992)
78. Ranzato, M. A., Boureau, Y.-L., LeCunn, Y.: Sparse feature learning for deep belief networks. In: Advances in neural information processing systems, (2008)
79. Robinson, A. J., Fallside, F.: The utility driven dynamic error propagation network. Technical Report CUED/F-INFENG/TR.1, University of Cambridge Department of Engineering, (1987)
80. Ruthotto, L., Haber, E.: Deep neural networks motivated by partial differential equations. J. Math. Imag. Vis., pages 1–13, (2018)
81. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
82. Siegelmann, H. T.: Neural networks and analog computation: beyond the Turing limit. Springer Science & Business Media, (2012)
83. Slepčev, D., Thorpe, M.: Analysis of p-Laplacian regularization in semi-supervised learning. SIAM J. Math. Anal. **51**(3), 2085–2120 (2019)
84. Smets, B. M. N., Portegies, J., Bekkers, E. J., Duits, R.: PDE-based group equivariant convolutional neural networks. J. Math. Imaging Vis. (2022). https://doi.org/10.1007/s10851-022-01114-x
85. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–2958 (2014)
86. Srivastava, R. K., Greff, K., Schmidhuber, J.: Highway networks. preprint arXiv:1505.00387, (2015)
87. Treister, E., Ruthotto, L., Sharoni, M., Zafrani, S., Haber, E.: Low-cost parameterizations of deep convolution neural networks. preprint arXiv:1805.07821, (2018)
88. van Gennip, Y., Bertozzi, A.L.: Γ-convergence of graph Ginzburg-Landau functionals. Adv. Differ. Equs. **17**(11–12), 1115–1180 (2012)
89. Vidal, R., Bruna, J., Giryes, R., Soatto, S.: Mathematics of deep learning. preprint arXiv:1712.04741, (2017)
90. Wan, E. A.: Finite impulse response neural networks for autoregressive time series prediction. Time Series Prediction: Forecast. Future and Understanding the Past, 2, (1993)
91. Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., Fergus, R.: Regularization of neural networks using dropconnect. In: International Conference on Machine Learning, pages 1058–1066, (2013)
92. Wiatowski, T., Bölcskei, H.: A mathematical theory of deep convolutional neural networks for feature extraction. IEEE Trans. Inf. Theory **64**(3), 1845–1866 (2018)
93. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural Comput. **1**(2), 270–280 (1989)
94. Yin, P., Zhang, S., Lyu, J., Osher, S., Qi, Y., Xin, J.: Blended coarse gradient descent for full quantization of deep neural networks. Res. Math. Sci. **6**(1), 14 (2019)
95. Zaeemzadeh, A., Rahnavard, N., Shah, M.: Norm-preservation: why residual networks can become extremely deep? IEEE Trans. Pattern Anal. Mach. Intell. **43**(11), 3980–3990 (2020)
96. Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., Hinton, G. E.: On rectified linear units for speech processing. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 3517–3521. IEEE, (2013)
97. Zhang, L., Schaeffer, H.: Forward stability of ResNet and its variants. J. Math. Imag. Vis., pages 1–24, (2019)
98. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Trans. Comput. Imag. **3**(1), 47–57 (2017)
99. Zurada, J.M.: Introduction to Artificial Neural Systems, vol. 8. West publishing company St, Paul (1992)

Publisher's Note