Journal of Machine Learning Research 23 (2022) 1-55 Submitted 11/21; Revised 4/22; Published 4/22

Mean-field Analysis of Piecewise Linear Solutions for Wide ReLU Networks

Alexander Shevchenko ALEX.SHEVCHENKO@IST.AC.AT
Institute of Science and Technology Austria
3400 Klosterneuburg, Austria

Vyacheslav Kungurtsev VYACHESLAV.KUNGURTSEV@FEL.CVUT.CZ
Department of Computer Science
Czech Technical University in Prague
166 36 Prague, Czechia

Marco Mondelli MARCO.MONDELLI@IST.AC.AT
Institute of Science and Technology Austria
3400 Klosterneuburg, Austria

Editor: Joan Bruna

Abstract

Understanding the properties of neural networks trained via stochastic gradient descent (SGD) is at the heart of the theory of deep learning. In this work, we take a mean-field view, and consider a two-layer ReLU network trained via noisy-SGD for a univariate regularized regression problem. Our main result is that SGD with vanishingly small noise injected in the gradients is biased towards a simple solution: at convergence, the ReLU network implements a piecewise linear map of the inputs, and the number of "knot" points – i.e., points where the tangent of the ReLU network estimator changes – between two consecutive training inputs is at most three. In particular, as the number of neurons of the network grows, the SGD dynamics is captured by the solution of a gradient flow and, at convergence, the distribution of the weights approaches the unique minimizer of a related free energy, which has a Gibbs form. Our key technical contribution consists in the analysis of the estimator resulting from this minimizer: we show that its second derivative vanishes everywhere, except at some specific locations which represent the "knot" points. We also provide empirical evidence that knots at locations distinct from the data points might occur, as predicted by our theory.

Keywords: Stochastic Gradient Descent, Implicit Bias, ReLU Activation, Overparameterized Models, Mean-Field

1. Introduction

Neural networks are the key ingredient behind many recent advances in machine learning. They achieve state-of-the-art performance on various practical tasks, such as image classification (He et al., 2016) and synthesis (Brock et al., 2019), natural language processing (Vaswani et al., 2017) and reinforcement learning (Silver et al., 2016). However, these results would not be possible without computational advances which enabled the training of highly overparameterized models with billions of weights. Such complex networks are capable of extracting more sophisticated patterns from the data than their less parameter-

heavy counterparts. Nonetheless, in the view of classical learning theory, models with a large number of parameters are prone to over-fitting (Von Luxburg and Schölkopf, 2011). Contrary to the conventional statistical wisdom, overparameterization turns out to be a rather desirable property for neural networks. This was even observed in a classical paper by Bartlett (1998), which demonstrated that in the overparameterized setting, the size of the network is less important than the magnitude of the weights. More recently, phenomena such as double descent (Belkin et al., 2019; Spigler et al., 2019; Nakkiran et al., 2020) and benign overfitting (Bartlett et al., 2020; Li et al., 2021; Bartlett et al., 2021) suggest that understanding the generalization properties of overparameterized models lies beyond the scope of the usual control of capacity via the size of the parameter set (Neyshabur et al., 2015).

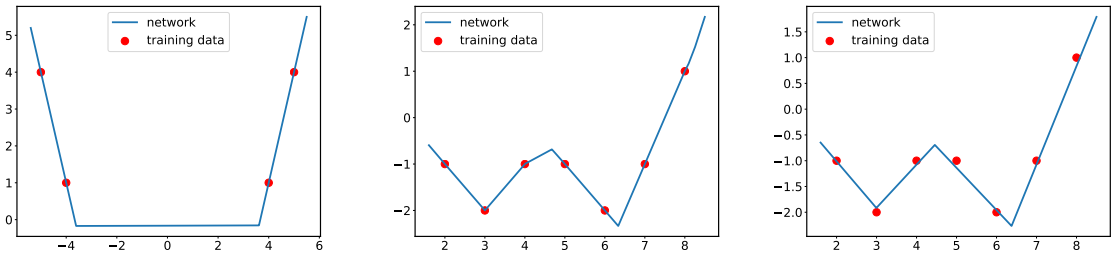
One way to explain the generalization capability of large neural networks lies in characterizing the properties of solutions found by stochastic gradient descent (SGD). In other words, the question is whether the optimization procedure is implicitly selective, i.e., it finds the functionally simple solutions that exhibit superior generalization ability in comparison to other candidates with roughly the same value of the empirical risk. For instance, Chizat and Bach (2020) consider shallow networks minimizing the logistic loss, and show that SGD converges to a max-margin classifier on a certain functional space endowed with the variation norm. In the machine learning literature, it has been suggested that large margin classifiers inherently exhibit better performance on unseen data (Bartlett et al., 2021; Cortes and Vapnik, 1995).

Constraints on the functional class of network solutions can also be imposed explicitly, e.g., via ℓ_2 regularization or by adding label noise. In some cases, it has been shown that the presence of parameter penalties or noise results in surprising implications. Depending on the regime, it biases optimization to find smooth solutions (Sahs et al., 2020; Jin and Montúfar, 2020; Savarese et al., 2019) or piecewise linear functions (Blanc et al., 2020; Ergen and Pilanci, 2021). The study by Balestriero and Baraniuk (2018) proposes an alternative to conventional ℓ_p regularization inspired by max-affine spline operators. It enforces a neural network to learn orthogonal representations, which significantly improves the performance and does not require any modifications of the network architecture.

In this work, we develop a novel approach towards understanding the implicit bias of gradient descent methods applied to overparameterized neural networks. In particular, we focus on the following key questions:

> *Once stochastic gradient descent has converged, how does the distribution of the weights of the neural network look like? What functional properties of the resulting solution are induced by this stationary distribution? Can we quantitatively characterize the trade-off between the complexity of the solution and the size of the training data in the overparameterized regime?*

To answer these questions, we consider training a wide two-layer ReLU (rectified linear unit) network for univariate regression, and we focus on the mean-field regime (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020). In this regime, the idea is that, as the number of neurons of the network grows, the weights obtained via SGD are close to i.i.d. samples coming from the solution of a



(a) **(b)** **(c)**

Figure 1: Example of functions learnt by a two-layer ReLU network with $N = 1000$ neurons on different training data. Solutions (a)-(b) are obtained with *no* regularization and label noise, i.e., $\lambda = 0$ and $\beta = +\infty$, while in (c) we have a sufficiently large regularization coefficient, which does not allow the network to fit the training data perfectly. Note that the piecewise linear solution exhibits tangent changes also at points different from the training data. Furthermore, the number of "knot" points may differ from the minimum required to fit the data: for instance, in (a) the minimum amount of tangent changes is 1, but the solution has two of them.

certain Wasserstein gradient flow. As a consequence, the output of the neural network approaches the following quantity:

$$y_\rho^{\sigma^*}(x) = \int \sigma^*(x, \boldsymbol{\theta})\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

Here, x is the input, σ^* denotes the activation function, and ρ is the solution of the Wasserstein gradient flow minimizing the free energy

$$\mathcal{F}(\rho) = \frac{1}{2}\,\mathbb{E}_{(x,y)\sim\mathbb{P}}\big\{(y - y_\rho^{\sigma^*}(x))^2\big\} + \frac{\lambda}{2}\int \|\boldsymbol{\theta}\|_2^2\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \beta^{-1}\int \rho(\boldsymbol{\theta})\log\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{1.1}$$

The first term corresponds to the expected squared loss (under the data distribution $\mathbb{P}$); the second term comes from the ℓ_2 regularization; the differential entropy term is linked to the noise introduced into the SGD update, and it penalizes non-uniform solutions. The coefficient β is often referred to as *inverse temperature*. In Mei et al. (2018), it is also shown that the minimizer of the free energy, call it ρ^*, has a Gibbs form for a sufficiently regular activation function σ^*. We review the connection between the dynamics of gradient descent and the solution ρ of the Wasserstein gradient flow in Section 3.1.

A number of works has exploited this connection to provide a rigorous justification to various phenomena attributed to neural networks. Mei et al. (2018, 2019) give global convergence guarantees for two-layer networks by studying the energy dissipation along the trajectory of the flow. The paper by Chizat and Bach (2018) takes a different route and exploits a lifting property enabled by a certain type of initialization and regularization, and Javanmard et al. (2020) put forward an argument based on displacement convexity. Nguyen and Pham (2020) and Araújo et al. (2019) tackle the multi-layer case, and, in particular, Nguyen and Pham (2020) establish convergence guarantees for a three-layer network. Fang et al. (2021) introduce a mean-field dynamics capturing the evolution of the features (instead of the network parameters) and show global convergence of ResNet

type of architectures. Shevchenko and Mondelli (2020) prove two properties commonly observed in practice (see e.g. Garipov et al. (2019); Draxler et al. (2018); Kuditipudi et al. (2019)), namely dropout stability and mode connectivity, for multi-layer networks trained under the mean-field regime. De Bortoli et al. (2020) consider different scalings of the step size of SGD, and identify two regimes under which different mean-field limits are obtained. Williams et al. (2019) show that the gradient flow for unregularized objectives forces the neurons of a two-layer ReLU network to concentrate around a subset of the training data points.

In this paper, we take a mean-field view to show that SGD is biased towards functionally simple solutions, namely, piecewise linear functions. Our idea is to analyze the stationary distribution ρ^* minimizing the free energy (1.1). We show that, in the low temperature regime ($\beta \to \infty$), the estimator's curvature vanishes everywhere except for a certain *cluster set*. More precisely, for each interval between two consecutive training inputs, aside for a set of small measure, the second derivative vanishes, i.e.,

$$\frac{\partial^2}{\partial x^2}\, y^{\sigma^*}_{\rho^*}(x) \to 0, \qquad \text{as } \beta \to \infty.$$

Furthermore, we provide a characterization of the cluster set and show that its measure vanishes while it concentrates around at most 3 points per interval. Ultimately, this analysis guarantees that, in the regime of decreasing temperature (corresponding to a small noise injected in the gradient updates), the solution found by SGD is piecewise linear. Our main contribution can be summarized in the following informal statement:

Theorem (Informal)**.** *Under the low temperature regime, i.e., $\beta \to \infty$, the estimator obtained by training a two-layer ReLU network via noisy-SGD converges to a piecewise linear solution. Furthermore, the number of "knot" points – i.e., points at which distinct linear pieces connect – between two consecutive training inputs is at most 3.*

Let us remark on a few important points. In the overparameterized regime, the number of neurons N is significantly larger than the number of training samples M, i.e., $N \gg M$. The output of the two-layer ReLU network is a linear combination of N ReLU units, hence the function implemented by the network is clearly piecewise linear with $\mathcal{O}(N)$ knot points. Here, we show that the number of knot points is actually $\mathcal{O}(M) \ll \mathcal{O}(N)$. Our analysis applies for both constant ($\lambda \to \bar{\lambda} > 0$) and vanishing ($\lambda \to 0$) regularization, and it does not require a specific form for the initialization of the parameters of the networks (as long as some mild technical conditions are satisfied).

In a nutshell, we establish a novel technique that accurately characterizes the solution to which gradient descent methods converge, when training overparameterized two-layer ReLU networks. Our analysis unveils a behaviour which is qualitatively different from that described in recent works (Williams et al., 2019; Blanc et al., 2020; Ergen and Pilanci, 2021) (see also a detailed comparison in Section 8): knot points are not necessarily allocated at the training points, or in a way that results in a function with the minimum number of tangent changes required to fit the data. We provide also numerical simulations to validate our findings (see Section 7 and Figure 1 above). We suggest that this novel behaviour is likely due to the difference in settings and the additional ℓ_2 regularization (including of the bias parameters).

Organization of the paper. The rest of the paper is organized as follows. In Section 2, we review the related work and a more detailed comparison is deferred to Section 8. In Section 3, we provide some preliminaries, including a background on the mean-field analysis in Section 3.1. Our main results are stated in Section 4 and proved in Section 5. In Section 6, we provide an example of a dataset for which the estimator found by SGD has a knot at a location different from the training inputs. We validate our findings with numerical simulations for different regression tasks in Section 7. We conclude and discuss some future directions in Section 9. Some of the technical lemmas and the corresponding proofs are deferred to Appendix A.

Notation. We use bold symbols for vectors $\boldsymbol{a}, \boldsymbol{b}$, and plain symbols for real numbers a, b. We use capitalized bold symbols to denote matrices, e.g., $\boldsymbol{\Theta}$. We denote the ℓ_2 norm of vectors $\boldsymbol{a}, \boldsymbol{b}$ by $\|\boldsymbol{a}\|_2, \|\boldsymbol{b}\|_2$. Given an integer N, we denote $[N] = \{1, \dots, N\}$. Given a discrete set $\mathcal{A}$, $|\mathcal{A}|$ is its cardinality. Similarly, given a Lebesgue measurable set $\mathcal{B} \subset \mathbb{R}^d$ its Lebesgue measure is given by $|\mathcal{B}|$. Given a sequence of distributions $\{\rho_n\}_{n\geq 0}$, we write $\rho_n \rightharpoonup \rho$ to denote the weak L_1 convergence of the corresponding measures. For a sequence of functions $\{f_n\}_{n\geq 0}$ we denote by $f_n \to f$ the *pointwise* convergence to a function f. Given a real number $x \in \mathbb{R}$, the closest integer that is not greater than x is defined by $\lfloor x \rfloor$.

2. Related Work

The line of works (Williams et al., 2019; Jin and Montúfar, 2020) shows that, in the lazy training regime (Chizat et al., 2019; Jacot et al., 2018) and for a uniform initialization, SGD converges to a cubic spline interpolating the data. Furthermore, for multivariate regression in the lazy training regime, Jin and Montúfar (2020) proved that the optimization procedure is biased towards solutions minimizing the 2-norm of the Radon transform of the fractional Laplacian. Similar results (although without the connection to the training dynamics) are obtained in (Savarese et al., 2019; Ongie et al., 2020), which analyze the solutions with zero loss and minimum norm of the parameters. Ergen and Pilanci (2021) develop a convex analytic framework to explain the bias towards simple solutions. In particular, an explicit characterization of the minimizer is provided, which implies that an optimal set of parameters yields linear spline interpolation for regression problems involving one dimensional or rank-one data. Cao et al. (2021) show that, for overparameterized models, the lower degree spherical harmonics are easier to learn. This observation comes from the fact that, in the lazy training regime, the convergence occurs faster along the directions given by the top eigenfunctions of the neural tangent kernel. Classification with linear networks on separable data is considered in (Soudry et al., 2018), where it is shown that gradient descent converges to the max-margin solution. This max-margin behavior is demonstrated in (Chizat and Bach, 2020) for non-linear wide two-layer networks using a mean-field analysis. In particular, in the mean-field regime, optimizing the logistic loss is equivalent to finding the max-margin classifier in a certain functional space. The paper by Zhang et al. (2020) focuses on the lazy training regime, and it shows that the optimization procedure finds a solution that fits the data perfectly and is closest to the starting point of the dynamics in terms of Euclidean distance in the parameter space. Wu et al. (2021) characterize the directional bias of GD and SGD in the case of moderate (but annealing) learning rate.

The behavior of SGD with label noise near the zero-loss manifold is studied in (Blanc et al., 2020). Here, it is shown that the training algorithm implicitly optimizes an auxiliary objective, namely, the sum of squared norms of the gradients evaluated at each training sample. This allows the authors of (Blanc et al., 2020) to show that SGD with label noise for a two-layer ReLU network with skip-connections is biased towards a piecewise linear solution. In particular, this piecewise linear solution has the minimum number of tangent changes required to fit the data. Williams et al. (2019) consider the Wasserstein gradient flow on a certain space of reduced parameters (in polar coordinates), and show that the points where the solution changes tangent are concentrated around a subset of training examples. A trade-off between the scale of the initialization and the training regime is also provided in (Williams et al., 2019; Sahs et al., 2020). Maennel et al. (2018) prove that the gradient flow enforces the weight vectors to concentrate at a small number of directions determined by the input data. Through the lens of spline theory, Parhi and Nowak (2020b) explain that a number of best practices used in deep learning, such as weight decay and path-norm, are connected to the ReLU activation and its smooth counterparts. Neyshabur et al. (2019) suggest a novel complexity measure for neural networks that provides a tighter generalization for the case of ReLU activation.

3. Preliminaries

3.1 Mean-field Background

We consider a two-layer neural network with N neurons and one-dimensional input $x \in \mathbb{R}$:

$$\hat{y}_N(x, \boldsymbol{\Theta}) = \frac{1}{N} \sum_{i=1}^{N} \sigma^*(x, \boldsymbol{\theta}_i), \tag{3.1}$$

where $\hat{y}_N(x, \boldsymbol{\Theta}) \in \mathbb{R}$ is the output of the network, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N) \in \mathbb{R}^{D \times N}$, with $\boldsymbol{\theta}_i \in \mathbb{R}^D$, are the parameters of the network, D is the dimension of parameters of each neuron, and $\sigma^* : \mathbb{R} \times \mathbb{R}^D \to \mathbb{R}$ represents the activation function. A typical example is $\sigma^*(x, \boldsymbol{\theta}) = a(wx + b)_+$, where $\boldsymbol{\theta} = (a, w, b) \in \mathbb{R}^3$ and $(\cdot)_+ : \mathbb{R} \to \mathbb{R}$ is a rectified linear unit activation.

We consider a regression problem for a dataset $\{(x_j, y_j)\}_{j=1}^M$ containing M points, and we aim to minimize the following expected squared loss with ℓ_2 regularization:

$$\mathbb{E}\Big\{(\hat{y}_N(x, \boldsymbol{\Theta}) - y)^2\Big\} + \frac{\lambda}{N} \sum_{i=1}^{N} \|\boldsymbol{\theta}_i\|_2^2 = \frac{1}{M} \sum_{j=1}^{M} (\hat{y}_N(x_j, \boldsymbol{\Theta}) - y_j)^2 + \frac{\lambda}{N} \sum_{i=1}^{N} \|\boldsymbol{\theta}_i\|_2^2. \tag{3.2}$$

On the LHS of (3.2), the expectation is taken over $(x, y) \sim \mathbb{P}$, with $\mathbb{P} = M^{-1} \sum_{j=1}^{M} \delta_{(x_j, y_j)}$ and $x_j < x_{j+1}\ \forall j \in [M-1]$. Here, $\delta_{(a,b)}$ stands for a delta distribution centered at $(a, b) \in \mathbb{R}^2$.

We are given samples $(\tilde{x}_k, \tilde{y}_k)_{k \geq 0} \sim_{\text{i.i.d.}} \mathbb{P}$, and we learn the network parameters $\boldsymbol{\Theta}$ via stochastic gradient descent (SGD) with step size s_k and additive Gaussian noise scaled by a factor $\beta^{-1} > 0$ (often referred to as a temperature):

$$\boldsymbol{\theta}_i^{k+1} = (1 - 2\lambda s_k)\boldsymbol{\theta}_i^k + 2s_k\big(\tilde{y}_k - \hat{y}_N(\tilde{x}_k, \boldsymbol{\Theta}^k)\big)\nabla_{\boldsymbol{\theta}_i}\big(\sigma^*(\tilde{x}_k, \boldsymbol{\theta}_i^k)\big) + \sqrt{2s_k/\beta}\boldsymbol{g}_i^k, \tag{3.3}$$

where $\mathbf{\Theta}^k$ stands for the network parameters after k steps of the optimization procedure, $\{\boldsymbol{g}_i^k\}_{i\in[N],k\geq 0} \sim_{\text{i.i.d.}} \mathcal{N}(0, \boldsymbol{I}_D)$, and the term $-2\lambda s_k \boldsymbol{\theta}_i^k$ corresponds to ℓ_2 regularization. The parameters are initialized independently according to a given distribution ρ_0, i.e., $\{\boldsymbol{\theta}_i^0\}_{i\in[N]} \sim_{\text{i.i.d.}} \rho_0$.

For some $\varepsilon > 0$, we assume that the step size of the noisy SGD update (3.3) is given by $s_k = \varepsilon\xi(\varepsilon k)$, where $\xi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a sufficiently regular function. Let $\hat{\rho}_k^N := \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{\theta}_i^k}$ denote the empirical distribution of weights after k steps of noisy SGD. Then, in Mei et al. (2018), it is proved that the evolution of $\hat{\rho}_k^N$ is well approximated by a certain distributional dynamics. In formulas,

$$\hat{\rho}_{\lfloor t/\varepsilon \rfloor}^{(N)} \rightharpoonup \rho_t,$$

almost surely along any sequence $(N \to \infty, \varepsilon_N \to 0)$ such that $N/\log(N/\varepsilon_N) \to \infty$ and $\varepsilon_N \log(N/\varepsilon_N) \to 0$. (Here, we have put the subscript N in ε_N to emphasize that the choice of the learning rate depends on N.) The distribution ρ_t is the solution of the following partial differential equation (PDE):

$$\begin{aligned} \partial_t \rho_t &= 2\xi(t)\nabla_{\boldsymbol{\theta}} \cdot (\rho_t \nabla_{\boldsymbol{\theta}} \Psi_\lambda(\boldsymbol{\theta}, \rho_t)) + 2\xi(t)\beta^{-1}\Delta_{\boldsymbol{\theta}}\rho_t, \\ \Psi_\lambda(\boldsymbol{\theta}, \rho) :&= \frac{1}{M}\sum_{i=1}^M \big(y_\rho^{\sigma^*}(x_i) - y_i\big)\cdot \sigma^*(x_i, \boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2. \end{aligned} \tag{3.4}$$

Here, $\nabla_\theta \cdot \boldsymbol{v}(\boldsymbol{\theta})$ stands for the divergence of the vector field $\boldsymbol{v}(\boldsymbol{\theta})$, and $\Delta_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \sum_{j=1}^D \partial_{\theta_j}^2 f(\boldsymbol{\theta})$ is the Laplacian of the function $f : \mathbb{R}^D \to \mathbb{R}^D$. To describe the next result, we first introduce a few related quantities. Define the infinite-width network with activation $\sigma^* : \mathbb{R}\times\mathbb{R}^D \to \mathbb{R}$ and weight distribution $\rho : \mathbb{R}^D \to [0, +\infty)$ as follows:

$$y_\rho^{\sigma^*}(x) = \int \sigma^*(x, \boldsymbol{\theta})\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

where the integral is taken over the support of ρ. For the forthcoming analysis, a certain regularity is required for the weight distribution ρ. In particular, the weight distribution is restricted to a set of admissible densities

$$\mathcal{K} := \left\{\rho : \mathbb{R}^D \to [0, +\infty) \text{ measurable: } \int \rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1,\ M(\rho) < \infty\right\},$$

where $M(\rho) = \int \|\boldsymbol{\theta}\|_2^2 \rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$. The expected risk attained on the distribution ρ by the infinite-width network with activation σ^* is defined by

$$R^{\sigma^*}(\rho) := \frac{1}{M}\sum_{i=1}^M \big(y_\rho^{\sigma^*}(x_i) - y_i\big)^2.$$

The quantity

$$H(\rho) := -\int \rho(\boldsymbol{\theta})\log\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

stands for the differential entropy of ρ, which is equal to $-\infty$ if the distribution ρ is singular. In this view, the distributional dynamics (3.4) is the Wasserstein gradient flow that

minimizes the free energy

$$\mathcal{F}^{\sigma^*}(\rho) = \frac{1}{2}R^{\sigma^*}(\rho) + \frac{\lambda}{2}M(\rho) - \beta^{-1}H(\rho), \tag{3.5}$$

over the set of admissible densities $\mathcal{K}$. Furthermore, this free energy has a unique minimizer and the solution of (3.4) converges to it as $t \to \infty$:

$$\rho_t \rightharpoonup \rho^*_{\sigma^*} \in \arg\min_{\rho' \in \mathcal{K}} \mathcal{F}^{\sigma^*}(\rho'), \qquad \text{as } t \to \infty.$$

The unique minimizer $\rho^*_{\sigma^*}$ is absolutely continuous, and it has the Gibbs form

$$\rho^*_{\sigma^*}(\boldsymbol{\theta}) = \frac{\exp\big\{-\beta\Psi_\lambda(\boldsymbol{\theta}, \rho^*_{\sigma^*})\big\}}{Z_{\sigma^*}(\beta, \lambda)}, \tag{3.6}$$

where $Z_{\sigma^*}(\beta, \lambda)$ is the normalization constant, also referred to as partition function.

3.2 Approximation of the ReLU Activation

Let us elaborate on the properties which σ^* should satisfy so that the results of Section 3.1 hold. First, the distributional dynamic (3.4) is known to be well-defined for a smooth and bounded potential Ψ_λ. In particular, it suffices to choose a bounded, Lipschitz σ^* with Lipschitz gradient, see assumptions A2-A3 in Mei et al. (2018). Furthermore, the minimizer of the free energy (3.5) exists and has a Gibbs form even for non-smooth potentials and, in particular, it suffices that σ^* is bounded and Lipschitz (this allows the first derivative to be discontinuous), see Lemmas 10.2-10.4 in Mei et al. (2018).

In the case of a ReLU activation, the corresponding σ^* has the following form

$$\sigma^*(x, \boldsymbol{\theta}) = a(wx + b)_+ = a\max\{0, wx + b\}, \quad \boldsymbol{\theta} = (a, w, b) \in \mathbb{R}^3,$$

which does not satisfy some of the aforementioned conditions. The first salient problem is the lack of continuity of the derivative at zero. This issue can be dealt with by considering a soft-plus activation with scale τ:

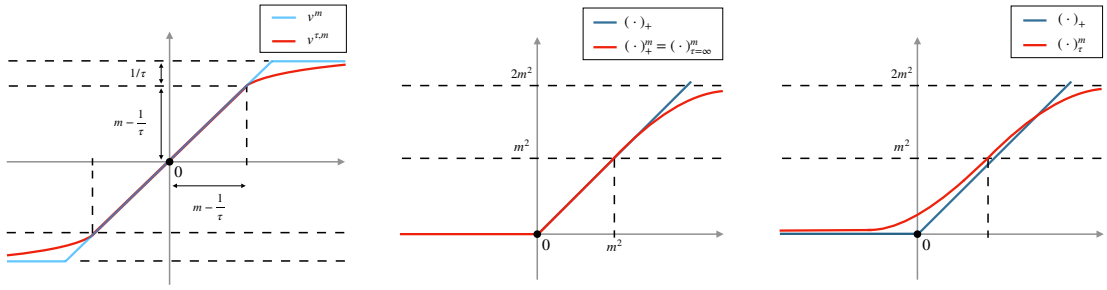
$$(x)_\tau := \frac{\log(1 + e^{\tau x})}{\tau}.$$

Notice that, as τ grows large, we have that $(\cdot)_\tau \to (\cdot)_+$. Another issue is that the function $\sigma^*(x, \boldsymbol{\theta})$ is not Lipschitz in the parameters $\boldsymbol{\theta}$, and it is unbounded. This problem can be solved by an appropriate truncation applied to the parameter a of the activation. The truncation should be Lipschitz and smooth for the dynamics to be well-defined.

In this view, we now provide the details of the approximation of the ReLU activation. For a parameter $v \in \mathbb{R}$, we denote by v^m its m-truncation defined as

$$v^m := \mathbb{1}_{\{|v| > m\}} \cdot m \cdot \text{sign}(v) + \mathbb{1}_{\{|v| \leq m\}} \cdot v.$$

Notice that the function $f(v) = v^m$ is Lipschitz continuous and bounded. For a parameter $v \in \mathbb{R}$, we denote by $v^{\tau,m}$ its τ-smooth m-truncation defined as follows: $v^{\tau,m}$ converges



(a) v^m and $v^{\tau,m}$ **(b)** $(\cdot)_+$ and $(\cdot)_+^m$ **(c)** $(\cdot)_+$ and $(\cdot)_\tau^m$

Figure 2: Visualization of the functions involved in the approximation of the ReLU activation.

pointwise to v^m as $\tau \to \infty$, $v^{\tau,m} = v$ inside the ball $\{v \in \mathbb{R} : |v| < m - \frac{1}{\tau}\}$, and the map $v \mapsto v^{\tau,m}$ is odd and belongs to $C^4(\mathbb{R})$. For a visualization of v^m and $v^{\tau,m}$, see Figure 2a.

We define the smooth m-truncation $(\cdot)_+^m$ of the ReLU activation as

$$(x)_+^m := \mathbb{1}_{\{x \le m^2\}}(x)_+ + \mathbb{1}_{\{x > m^2\}}\phi_m(x),$$

where ϕ_m is chosen so that the following holds: $(x)_+^m \in C^4(\mathbb{R})$, $(x)_+^m \le (x)_+$ for all $x \in \mathbb{R}$, and $|\phi_m''(x)| \le \frac{1}{m^2}$ for $x > m$. Note that these conditions imply that $\phi_m(m^2) = m^2$ and $\phi_m'(m^2) = 1$. Furthermore, in order to enforce the bound on ϕ_m'', we pick ϕ_m so that $\lim_{x\to+\infty} \phi_m(x) = 2m^2$, and $\lim_{x\to+\infty} \phi_m'(x) = \lim_{x\to+\infty} \phi_m''(x) = 0$. For a visualization of $(\cdot)_+^m$, see Figure 2b.

Finally, we define the smooth m-truncation $(\cdot)_\tau^m$ of the softplus activation as

$$(x)_\tau^m := \mathbb{1}_{\{x \le x_m\}}(x)_\tau + \mathbb{1}_{\{x > x_m\}}\phi_{\tau,m}(x), \tag{3.7}$$

where $x_m \in \mathbb{R}$ is such that $(x_m)_\tau = m^2$. As in the truncation of ReLU, we choose $\phi_{\tau,m}$ so that $(x)_\tau^m \in C^4(\mathbb{R})$ and $|\phi_{\tau,m}''(x)| \le \frac{1}{m^2}$ for $x > x_m$. Furthermore, we require that $(\cdot)_\tau^m$ converges pointwise to $(\cdot)_+^m$ as $\tau \to \infty$ (which we can guarantee since $(\cdot)_\tau \to (\cdot)_+$, as $\tau \to \infty$). To enforce these conditions, we pick $\phi_{\tau,m}$ so that $\phi_{\tau,m}(x_m) = m^2$, $\phi_{\tau,m}'(x_m) = (x)_\tau'\big|_{x=x_m}$, $\lim_{x\to+\infty} \phi_{\tau,m}(x) = 2m^2$, and $\lim_{x\to+\infty} \phi_{\tau,m}'(x) = \lim_{x\to+\infty} \phi_{\tau,m}''(x) = 0$. For a visualization of $(\cdot)_\tau^m$, see Figure 2c.

Notice that, for $\tau \ge 1$, the soft-plus activation can be sandwiched as follows:

$$(x)_+ - \frac{1}{\tau} \le (x)_\tau \le (x)_+ + \frac{1}{\tau}.$$

In order to establish the continuity of a certain limit and smoothness properties, we also pick $\phi_{\tau,m}$ such that the smooth m-truncation of soft-plus activation satisfies a similar bound:

$$(x)_+ - \frac{1}{\tau} \le (x)_\tau^m \le (x)_+ + \frac{1}{\tau}. \tag{3.8}$$

At this point, we remark that the activation $(\boldsymbol{\theta}, x) \mapsto a^{\tau,m}(w^{\tau,m}x + b)_\tau^m$ satisfies all the conditions necessary for the results of Section 3.1 to hold. In what follows, we will also use the activation $(\boldsymbol{\theta}, x) \mapsto a^m(w^m x + b)_\tau^m$ as an auxiliary object. This map is not smooth,

but it satisfies all the assumptions required for the existence of a free energy minimizer $\rho^*_{\sigma^*}$. We also note that the truncation of the parameter w might seem unnatural (we are truncating the ReLU activation anyway), but it simplifies our analysis. In particular, it allows us to establish a connection between the derivatives (w.r.t. the input x) of the predictor implemented by the solution of the flow (3.4) and the same quantity evaluated on the minimizer, as t grows large.

We will use the following notation for the values of the risks corresponding to different activations

$$R_i^{\tau,m}(\rho) := -\frac{1}{M}\left(y_i - \int a^{\tau,m}(w^{\tau,m}x_i + b)_\tau^m \rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right), \quad R^{\tau,m}(\rho) := M\sum_{i=1}^M \left(R_i^{\tau,m}(\rho)\right)^2,$$
$$R_i^{m}(\rho) := -\frac{1}{M}\left(y_i - \int a^{m}(w^{m}x_i + b)_+^m \rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right), \qquad R^{m}(\rho) := M\sum_{i=1}^M \left(R_i^{m}(\rho)\right)^2,$$

and for the related free-energies

$$\mathcal{F}^{\tau,m}(\rho) := \frac{1}{2}R^{\tau,m}(\rho) + \frac{\lambda}{2}M(\rho) - \beta^{-1}H(\rho),$$
$$\mathcal{F}^{m}(\rho) := \frac{1}{2}R^{m}(\rho) + \frac{\lambda}{2}M(\rho) - \beta^{-1}H(\rho).$$

Here, $R_i^{\tau,m}$ and R_i^m represent the rescaled error on the i-th training sample, and $R^{\tau,m}$ and R^m are the standard expected square losses. In this way, we can write the Gibbs minimizers in a compact form, namely,

$$\rho^*_{\tau,m}(\boldsymbol{\theta}) = Z^{-1}_{\tau,m}(\beta,\lambda)\exp\left\{-\beta\left[\sum_{i=1}^M R_i^{\tau,m}(\rho^*_{\tau,m})\cdot a^{\tau,m}(w^{\tau,m}x_i+b)_\tau^m + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right]\right\}, \tag{3.9}$$

$$\rho^*_{m}(\boldsymbol{\theta}) = Z^{-1}_{m}(\beta,\lambda)\exp\left\{-\beta\left[\sum_{i=1}^M R_i^{m}(\rho^*_{m})\cdot a^{m}(w^{m}x_i+b)_+^m + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right]\right\}, \tag{3.10}$$

where $Z_{\tau,m}(\beta,\lambda)$ and $Z_m(\beta,\lambda)$ denote the partition functions.

4. Main Results

Before presenting the main results, let us introduce the notion of a *cluster set*. This set allows us to identify the locations of the knot points of an estimator function that is implemented by the neural network. In particular, we consider the second derivative of the predictor evaluated at the Gibbs distribution with activation $(\boldsymbol{\theta}, x) \mapsto a^{\tau,m}(w^{\tau,m}x + b)_\tau^m$, for large τ, i.e.,

$$\lim_{\tau\to\infty}\frac{\partial^2}{\partial x^2}\int a^{\tau,m}(w^{\tau,m}x+b)_\tau^m \rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{4.1}$$

Then, the cluster set is associated to the inputs on which the quantity (4.1) might grow unbounded in absolute value, in the low temperature regime ($\beta^{-1}\to 0$). Intuitively, this indicates that on some points of the cluster set, the tangent of the predictor changes abruptly, resulting in "knots". We denote the cluster set by $\Omega(m,\beta,\lambda)$, and we define it below.

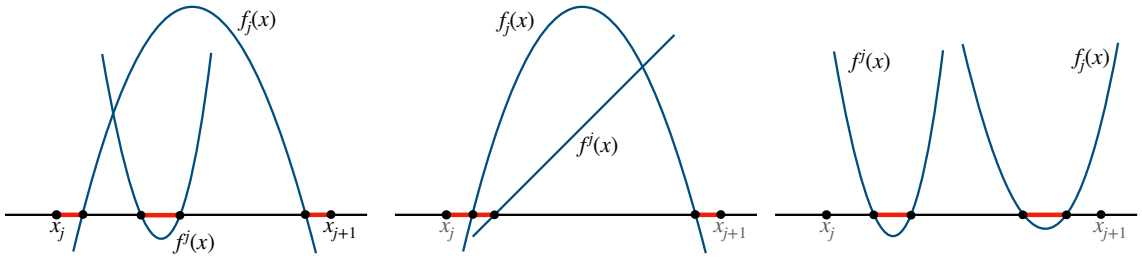


Figure 3: Three different configurations of the polynomials $f^j(x)$ and $f_j(x)$, together with the corresponding cluster set. The dark blue curves show the shape of polynomials, and the red bold intervals indicate the set on which polynomials attain non-positive value.

Let $\mathcal{I}$ be the set of prediction intervals, i.e.,

$$\mathcal{I} = \Big\{[x_0 := -L, x_1], [x_1, x_2], \cdots, [x_{M-1}, x_M], [x_M, x_{M+1} := L]\Big\},$$

where $L > \max\{|x_1|, \cdots, |x_M|\}$ is any fixed positive constant independent of $(\tau, m, \beta, \lambda)$. For each $I_j := [x_j, x_{j+1}] \in \mathcal{I}$, the intersection of the cluster set with the prediction interval I_j is denoted by $\overline{\Omega}_j(m, \beta, \lambda)$, i.e.,

$$\overline{\Omega}_j(m, \beta, \lambda) = \Omega(m, \beta, \lambda) \cap I_j. \tag{4.2}$$

Thus, in order to define the cluster set $\Omega(m, \beta, \lambda)$, it suffices to give the definition of $\overline{\Omega}_j(m, \beta, \lambda)$. To do so, consider the second-degree polynomials $f^j(x)$ and $f_j(x)$ given by

$$\begin{aligned} f^j(x) &:= 1 + x^2 - (A^j x - B^j)^2, \\ f_j(x) &:= 1 + x^2 - (A_j x - B_j)^2, \end{aligned} \tag{4.3}$$

with coefficients

$$\begin{aligned} A^j &:= \frac{1}{\lambda} \sum_{i=j+1}^{M} R_i^m(\rho_m^*), \;\; A_j := \frac{1}{\lambda} \sum_{i=1}^{j} R_i^m(\rho_m^*), \\ B^j &:= \frac{1}{\lambda} \sum_{i=j+1}^{M} R_i^m(\rho_m^*) x_i, \;\; B_j := \frac{1}{\lambda} \sum_{i=1}^{j} R_i^m(\rho_m^*) x_i. \end{aligned} \tag{4.4}$$

Here, if the summation set is empty (e.g., for A_0), the corresponding coefficient is equal to zero. Then, the set $\overline{\Omega}_j(m, \beta, \lambda)$ is defined as the union of the non-positive sets of the second-degree polynomials $f^j(x)$ and $f_j(x)$:

$$\overline{\Omega}_j(m, \beta, \lambda) = \Omega^j(m, \beta, \lambda) \cup \Omega_j(m, \beta, \lambda), \tag{4.5}$$

where

$$\begin{aligned} \Omega^j(m, \beta, \lambda) &:= \{x \in I_j : f^j(x) \le 0\}, \\ \Omega_j(m, \beta, \lambda) &:= \{x \in I_j : f_j(x) \le 0\}. \end{aligned} \tag{4.6}$$

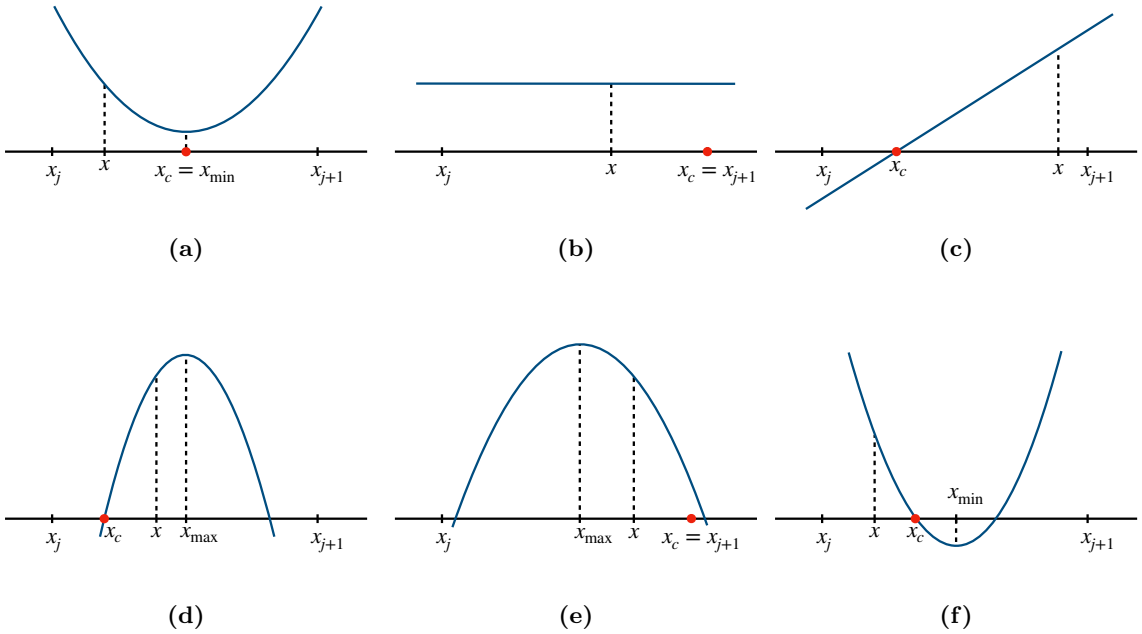


Figure 4: Representation of the critical point x_c for different configurations of the polynomial f^j and evaluation point x. The red dot indicates the location of the critical point. The dashed line indicates the value of f^j attained at the corresponding point. The dark blue curve shows the shape of the polynomial f^j.

We now provide an informal explanation on how the non-positive sets of the second-degree polynomials $f^j(x)$ and $f_j(x)$ come into play. A central object of interest in our analysis is the second derivative of the estimator implemented by the neural network, and our strategy is to bound its magnitude by a particular Gaussian-like integral. This integral does not diverge as long as the corresponding covariance matrix is non-degenerate, i.e., it has strictly positive eigenvalues. In this view, the non-positive sets of the polynomials $f^j(x)$ and $f_j(x)$ precisely characterize the inputs x for which this covariance matrix is degenerate. Hence, for such inputs x, this upper bound on the second derivative of the estimator diverges, which implies that the predictor may have a "knot".

Since $f^j(x)$ and $f_j(x)$ are second-degree polynomials, the set $\overline{\Omega}_j(m, \beta, \lambda)$ can be always written as the union of at most 3 intervals. Moreover, $\overline{\Omega}_j(m, \beta, \lambda)$ depends only on the errors of the estimator at the training points and on the penalty parameter λ. Thus, if one has access to the value of the errors at each training point for the optimal estimator, i.e., $R_i^m(\rho_m^*)$, an explicit expression for the cluster set can be readily obtained. Figure 3 shows three different configurations of the polynomials $f^j(x)$ and $f_j(x)$, together with the corresponding cluster set.

The size of the set $\overline{\Omega}_j(m, \beta, \lambda)$ can be controlled explicitly as a function of the parameters (m, β, λ). More formally, in Lemma 5.3, we show that the Lebesgue measure of $\overline{\Omega}_j(m, \beta, \lambda)$ can be upper bounded as

$$|\overline{\Omega}_j(m, \beta, \lambda)| \leq \frac{e^{C\beta}}{m^2}, \tag{4.7}$$

where $C > 0$ denotes a numerical constant independent of $(\tau, m, \beta, \lambda)$ and we have made the following assumption:

A1. $\tau \geq 1$, $\beta \geq \max\left\{C_1, \frac{1}{\lambda}, \frac{1}{\lambda}\log\frac{1}{\lambda}\right\}$, $m > C_2$ and $\lambda < C_3$ for some numerical constants $C_1, C_2, C_3 > 0$.

In particular, (4.7) implies that the cluster set vanishes as $\beta \to \infty$ and $m = e^{\Theta(\beta)}$. Therefore, as $\overline{\Omega}_j(m, \beta, \lambda)$ is the union of at most 3 intervals, the cluster set concentrates on at most 3 points per prediction interval.

We note that our use of **A1** throughout the sequel is with the flexibility of C_1, C_2, and C_3 in mind; we are interested in the behavior as m and β grow large, so we permit liberty in the determination of the constants implying the formal statements we intend to show.

A key step of our analysis (cf. Theorem 1) consists in showing that, outside the cluster set, the absolute value of the second derivative vanishes. Our bound on this absolute value is connected to the speed of decay to zero of the polynomials $f^j(x)$ and $f_j(x)$, as the input x approaches the cluster set. In order to establish a quantitative bound for such a decay, we introduce an auxiliary quantity, namely, a *critical point*, that is associated to each input point outside of the cluster set. Given the polynomial $f^j(\cdot)$ and the input $x \in I_j \backslash \Omega^j(m, \beta, \lambda)$, the critical point x_c associated to x is defined below.

Definition 4.1 (Critical point)**.** *If $f_j(\tilde{x}) = 0$ has no solutions for $\tilde{x} \in \mathbb{R}$, then the critical point x_c associated to x and $I_j \setminus \Omega^j(m, \beta, \lambda)$ is defined to be the minimizer of $f_j(\cdot)$ on I_j, i.e., $x_c = \arg\min_{\tilde{x} \in I_j} f_j(\tilde{x})$. In case of multiple minimizers, e.g., $(a, b) = (1, 0)$, we set $x_c = x_{j+1}$. If $f_j(\tilde{x}) = 0$ has at least one solution for $\tilde{x} \in \mathbb{R}$, then we let x_r be the root of f_j (in $\mathbb{R}$ and not necessarily in the segment I_j) which is the closest in Euclidean distance to x, and we define the* critical point *x_c to be the closest point to x_r in I_j, i.e., $x_c = x_r$ if $x_r \in I_j$ and x_c is one of the two extremes of the interval otherwise.*

Figure 4 provides a visualization of the critical point associated to x for several configurations of f^j. For the polynomial $f_j(\cdot)$ and an input $x \in I_j \setminus \Omega_j(m, \beta, \lambda)$, the critical point $\bar{x}_c$ is defined in a similar fashion. In this view, we show in Lemma 5.5 that the following lower bounds on f^j, f_j hold for $x \in I_j \setminus \overline{\Omega}(m, \beta, \lambda)$,

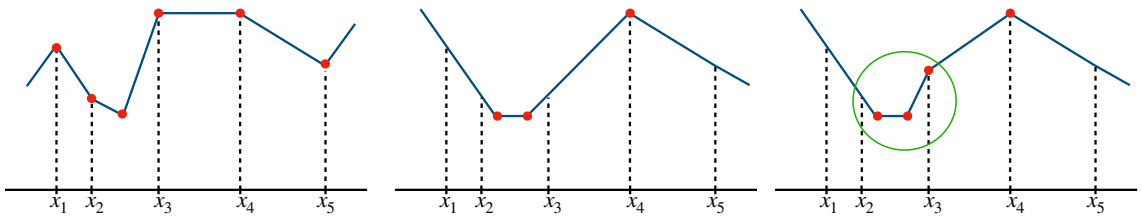
$$C^j(x) := \gamma_1(x - x_c)^2 + \gamma_2 \leq f^j(x), \quad C_j(x) := \gamma_3(x - \bar{x}_c)^2 + \gamma_4 \leq f_j(x). \tag{4.8}$$

The coefficients $\gamma_1, \gamma_2, \gamma_3, \gamma_4 > 0$ satisfy the following condition: either $\gamma_1 > \varepsilon$ or $\gamma_2 > \varepsilon$, and either $\gamma_3 > \varepsilon$ or $\gamma_4 > \varepsilon$, where $\varepsilon > 0$ is a numerical constant independent of the choice of (m, β, λ).

At this point, we are ready to state our upper bound on the second derivative outside the cluster set.

Theorem 1 (Vanishing curvature)**.** *Assume that condition **A1** is satisfied and that $m > e^{K_1\beta}$ for some numerical constant $K_1 > 0$ independent of $(\tau, m, \beta, \lambda)$. Then, for each $x \in I_j \setminus \overline{\Omega}_j(m, \beta, \lambda)$, the following upper bound on the second derivative holds*

$$\lim_{\tau \to +\infty} \left| \frac{\partial^2}{\partial x^2} \int a^{\tau,m}(w^{\tau,m}x + b)_\tau^m \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \right| \leq \mathcal{O}\left(\frac{1}{m\lambda} + \frac{1}{\beta\lambda^{7/4}(\bar{C}^j(x))^2} \right), \tag{4.9}$$



(a) **(b)** **(c)**

Figure 5: Three examples of piecewise linear functions that fit the data with zero squared error. Dashed black line indicates the y value for each training input. Red dots are located at points where the function changes its tangent. (a) and (b) illustrates two admissible piecewise linear solutions, while (c) is not admissible due to the location of break points on interval $[x_2, x_3]$.

where the coefficient $\bar{C}^j(x)$ *is defined as*

$$\bar{C}^j(x) = \min\left\{C^j(x), C_j(x), 1\right\}, \tag{4.10}$$

with $C^j(x)$ *and* $C_j(x)$ *given by (4.8). Furthermore, the following upper-bound on the size of the cluster set holds*

$$|\Omega(m, \beta, \lambda)| \leq \frac{K_2}{m}, \tag{4.11}$$

for some numerical constant $K_2 > 0$ *independent of* $(\tau, m, \beta, \lambda)$.

Some remarks are in order. First, the inequality (4.9) shows that, in the low temperature regime, the curvature vanishes outside the cluster set, and it also provides a decay rate. Second, we will upper bound the measure of the cluster set as in (4.7), thus the condition $m > e^{K_1\beta}$ ensures that the upper bound (4.11) holds. Finally, the presence of the coefficient $\bar{C}^j(x)$ is due to the fact that the second derivative can grow unbounded for points approaching the cluster set. Let us highlight that this growth is solely dictated by the distance to the cluster set, and it does not depend on (m, β, λ). In fact, (4.8) holds, where one of the coefficients in $\{\gamma_1, \gamma_2\}$ and in $\{\gamma_3, \gamma_4\}$ is lower bounded by a strictly positive constant independent of (m, β, λ).

From Theorem 1, we conclude that, as $m\lambda \to \infty$ and $\beta\lambda^{7/4} \to \infty$, the second derivative vanishes for all $x \in I_j \setminus \overline{\Omega}_j(m, \beta, \lambda)$. Furthermore, for $m > e^{C\beta}$ and $\beta \to \infty$, the cluster set concentrates on at most 3 points per interval. Therefore, the estimator $\int a^{\tau,m}(w^{\tau,m}x + b)_\tau^m \rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is piecewise linear with "knot" points given by the cluster set (cf. Theorem 2). To formalize this result, we define the notion of an *admissible piecewise linear solution.*

Definition 4.2 (Admissible piecewise linear solution)**.** *Given a set of prediction intervals* $\mathcal{I}$*, a function* $f : \mathbb{R} \to \mathbb{R}$ *is an admissible piecewise linear solution if* f *is continuous, piecewise linear and has at most 3 knot points (i.e., the points where a change of tangent occurs) per prediction interval* $I_j \in \mathcal{I}$*. Moreover, the only configuration possible for 3 knots to occur is the following: two knots are located strictly at the end points of the interval, and the remaining point lies strictly in the interior of the interval.*

Figure 5 provides some examples of piecewise linear solutions: (a) and (b) are admissible (in the sense of Definition 4.2), while (c) is not admissible, since it has two knots in the interior of the prediction interval and one located at the right endpoint. As mentioned before, the location of the knot points is associated with the limiting behaviour of the corresponding polynomials $f^j(x)$ and $f_j(x)$. For instance, consider the prediction interval $[x_2, x_3] \in \mathcal{I}$. Then, the configuration of Figure 5a corresponds to the case described in Figure 3a. In fact, f^j has a negative leading coefficient, and its roots are converging to the end points of the interval. Moreover, f_j has positive curvature and the minimizer is located inside the interval. The same parallel can be drawn between Figure 5b and Figure 3c. Furthermore, one can verify that the situation described in Figure 5c cannot be achieved for any configuration of $f^j(x)$ and $f_j(x)$.

We are now ready to state our result concerning the structure of the function obtained from the Gibbs distribution $\rho^*_{\tau,m}$.

Theorem 2 (Free energy minimizer solution is increasingly more piecewise linear)**.** *Assume that condition **A1** is satisfied and that $m > e^{K_1\beta}$, where $K_1 > 0$ is a constant independent of $(\tau, m, \beta, \lambda)$. Then, given a set of prediction intervals $\mathcal{I}$, there exists a family of admissible piecewise linear solutions $\{f_{m,\beta,\lambda}\}$ as per Definition 4.2, such that, for any $I \in \mathcal{I}$ and $x \in I$, the following convergence result holds*

$$\lim_{\beta\lambda^{7/4}\to+\infty} \lim_{\tau\to+\infty} \left| f_{m,\beta,\lambda}(x) - \int a^{\tau,m}(w^{\tau,m}x + b)^m_\tau \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \right| = 0.$$

In words, Theorem 2 means that the solution resulting from the minimization of the free energy (3.5) approaches a piecewise linear function, as the noise vanishes. Let us highlight that our result tackles both the regularized case in which λ approaches a fixed positive constant and the un-regularized one in which λ vanishes (as long as its vanishing rate is sufficiently slow to ensure that $\beta\lambda^{7/4} \to \infty$). We also note that that the family $\{f_{m,\beta,\lambda}\}$ is well-behaved, i.e., on each linear region the function $f_{m,\beta,\lambda}$ has the following representation: $f_{m,\beta,\lambda} = ux + v$ for some $u, v \in \mathbb{R}$, and the coefficients $|u|, |v|$ are uniformly bounded in (m, β, λ).

The proof of Theorem 2 crucially relies on the fact that the second moment of $\rho^*_{\tau,m}$ is uniformly bounded along the sequence $\beta\lambda^{7/4} \to \infty$. In fact, the uniform bound on the second moment implies that the *first* derivatives of the predictors w.r.t. the input are uniformly bounded (even for points *inside* the cluster set), and therefore the sequence of predictors is equi-Lipschitz. This, in particular, allows us to show that the limit is well-behaved, as function changes can be controlled via Lipschitz bounds.

Let us clarify that Theorem 2 does not establish the uniqueness of the limit in (m, β, λ), i.e., that the limiting piecewise linear function is the same regardless of the subsequence. Our numerical results reported in Figures 1, 6b, 7 and 8 suggest that the limit is unique. However, a typical line of argument (see e.g. Jordan et al. (1998)) would require the lower-semicontinuity of the free energy (which does not hold for $m = \infty$). Furthermore, even the uniqueness of the minimizer for $\beta = \infty$ remains unclear in our setup. Nevertheless, let us point out that the sequence $\{\rho^*_{\tau,m}\}$ is tight, since the second moments are uniformly bounded by Lemma A.6, and Proposition 2.3 in Hu et al. (2021) suggests that at least the cluster points of the sequence $\{\rho^*_{\tau,m}\}$ as $\beta \to \infty$ coincide with the set of minimizers of

the limiting objective ($\beta = \infty$). Another piece of evidence comes from the fact that the annealed dynamics converges to the minimizers of the noiseless objective (Chizat, 2022). We leave for future work the resolution of these issues.

We remark that providing a quantitative bound on the parameter τ appears to be challenging. The current analysis relies on a dominated convergence argument which does not lead to an explicit convergence rate. Obtaining such a rate requires understanding the trade-off between the terms in the free energy (3.5) for varying τ, and it is also left for future work.

Finally, by combining Theorem 2 with the mean-field analysis in Mei et al. (2018), we obtain the desired result on finite-width networks trained via noisy SGD in the low temperature regime.

Corollary 4.3 (Noisy SGD solution is increasingly more piecewise linear)**.** *Assume that condition **A1** holds and that $m > e^{K_1\beta}$, where $K_1 > 0$ is a constant independent of $(\tau, m, \beta, \lambda)$. Let ρ_0 be absolutely continuous and K_0 sub-Gaussian, where $K_0 > 0$ is some numerical constant. Assume also that $M(\rho_0) < \infty$ and $H(\rho_0) > -\infty$. Let $\sigma^*(x, \boldsymbol{\theta}) = a^{\tau,m}(w^{\tau,m}x + b)_\tau^m$ be the activation function, and let $\boldsymbol{\theta}^k$ be obtained by running $k = \lfloor t/\varepsilon \rfloor$ steps of the noisy SGD algorithm (3.3) with data $(\tilde{x}_k, \tilde{y}_k)_{k\geq 0} \sim_{\text{i.i.d.}} \mathbb{P}$ and initialization ρ_0. Then, given a set of prediction intervals $\mathcal{I}$, there exists a family of admissible piecewise linear solutions $\{f_{m,\beta,\lambda}\}$ as per Definition 4.2, such that, for any $I \in \mathcal{I}$ and $x \in I$, the following convergence result holds almost surely:*

$$\lim_{\beta\lambda^{7/4}\to+\infty} \lim_{\tau\to+\infty} \lim_{t\to+\infty} \lim_{\substack{\varepsilon\to 0 \\ N\to\infty}} \left| f_{m,\beta,\lambda}(x) - \frac{1}{N}\sum_{i=1}^{N} \sigma^*(x, \boldsymbol{\theta}_i^k) \right| = 0,$$

where the limit in N, ε is taken along any subsequence $\{(N, \varepsilon = \varepsilon_N)\}$ with $N/\log(N/\varepsilon_N) \to \infty$ and $\varepsilon_N \log(N/\varepsilon_N) \to 0$.

In words, Corollary 4.3 means that, at convergence, the estimator implemented by a wide two-layer ReLU network approaches a piecewise linear function, in the regime of vanishingly small noise. In fact, as $\tau, m \to \infty$, the activation function $\sigma^*(x, \boldsymbol{\theta}) = a^{\tau,m}(w^{\tau,m}x + b)_\tau^m$ converges pointwise to the ReLU activation $a(wx + b)_+$. We also remark that our result holds for any initialization of the weights of the network, as long as some mild technical conditions are fulfilled (absolute continuity, sub-Gaussian tails, finite second moment and entropy).

Let us clarify some technical aspects of the statement of Corollary 4.3. The result holds for a particular sequence of minimizers, since some of the limits ($t \to \infty$, $(N, \varepsilon^{-1}) \to \infty$, and $\beta \to \infty$) are not interchangeable. Furthermore, it appears to be difficult to prove the same statement directly for the noiseless case ($\beta = \infty$). We also point out that the stochasticity of the gradient descent algorithm does not play a role in our analysis, since its impact is seen to be inconsequential by the usual concentration argument (Mei et al., 2018) when passing to its non-stochastic counterpart.

As concerns the limit in t, describing the dependence of the mixing time of the diffusion dynamics (3.4) on the temperature parameter β is a cumbersome task. In particular, Gayrard et al. (2004) show that an exponentially bad dependence could occur if the target function has multiple small risk regions. However, some recent studies show an exponentially

fast convergence of the noisy dynamics under some reasonable but particular conditions on the objective landscape (Chizat, 2022; Nitanda et al., 2022).

As concerns the limit in (N, ε), the analyses in Mei et al. (2018, 2019) lead to an upper bound on the error term that, with probability at least $1 - e^{-z^2}$, is given by

$$Ce^{Ct}\sqrt{1/N \vee \varepsilon} \cdot \left[\sqrt{1 + \log(N(t/\varepsilon \vee 1))} + z\right], \tag{4.12}$$

where $a \vee b$ denotes the maximum between a and b. The exponential dependence of (4.12) in the time t of the dynamics is a common drawback of existing mean-field analyses, and improving it is an open problem which lies beyond the scope of this work. Let us conclude by mentioning that the numerical results presented in Section 7 suggest that, in practical settings, the convergence to the limit occurs rather quickly in the various parameters.

5. Proof of the Main Results

5.1 Roadmap of the Argument

We start by providing an informal outline of the proof for the main statements. In Section 5.2, we show that, in the low temperature regime, the curvature of the predictor evaluated at the Gibbs distribution $\rho^*_{\tau,m}$ vanishes everywhere except at a small neighbourhood of at most three points per prediction interval $I_j \in \mathcal{I}$ (Theorem 1). This is done in a few steps. First, in Lemma 5.1, we show that, as $\tau \to \infty$, the density $\rho^*_{\tau,m}$ acts similarly to a delta distribution supported on the lower-dimensional linear subspace $\{b \in \mathbb{R} : b = -w^m x\}$, namely,

$$\lim_{\tau\to\infty} \frac{\partial^2}{\partial x^2} \int a^{\tau,m}(w^{\tau,m}x + b)^m_\tau \rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \approx \int a^m (w^m)^2 \rho^*_m(a, w, -w^m x)\mathrm{d}a\mathrm{d}w. \tag{5.1}$$

To do so, in Lemma A.4 we prove that, as $\tau \to \infty$, the sequence $\rho^*_{\tau,m}(\boldsymbol{\theta})$ of minimizers of the free energy $\mathcal{F}^{\tau,m}$ converges pointwise for all $\boldsymbol{\theta}$ to a minimizer $\rho^*_m(\boldsymbol{\theta})$ of the free energy $\mathcal{F}^m$ with truncated ReLU activation. Then, a dominated convergence argument allows us to obtain (5.1). Next, in Lemma 5.7 we show that, as $\beta \to \infty$, the absolute value of the integral

$$\int a^m (w^m)^2 \rho^*_m(a, w, -w^m x)\mathrm{d}a\mathrm{d}w \tag{5.2}$$

can be made arbitrary small for all x except those in the cluster set. The idea is that the absolute value of (5.2) can be bounded by a certain Gaussian integral, and the corresponding covariance matrix is well-defined everywhere except in the cluster set (see Lemmas 5.4 and 5.5). The definition of the cluster set (see (4.2)-(4.6)) together with the fact that the partition function of ρ^*_m is uniformly bounded in m (see Lemma 5.2) allows us to show that the cluster set concentrates on at most three points per interval as $\beta \to \infty$.

In Section 5.3, we show that the predictor evaluated at the Gibbs distribution $\rho^*_{\tau,m}$ can be approximated arbitrarily well by an admissible piecewise linear solution (Theorem 2). First, via a Taylor argument, since the curvature vanishes, the estimator can be approximated by a linear function on each interval of $\mathcal{I}\backslash\Omega(m, \beta, \lambda)$. Since the cluster set vanishes concentrating on at most three points per prediction interval, the predictor converges to an admissible piecewise linear solution. However, there is one technical subtlety to consider before reaching

this conclusion. Namely, we must consider the possibility that the sequence of predictors experiences unbounded oscillations inside the cluster set, which might ultimately result in a discontinuous limit. Fortunately, this scenario is ruled out because the sequence $\rho^*_{\tau,m}$ has uniformly bounded second moments. This fact in conjunction with the structure of the *first* derivative of the predictor yields the conclusion that the sequence of predictors is equi-Lipschitz, and therefore the limit is well-behaved.

Finally, the proof of Corollary 4.3 follows from similar arguments together with the application of the result established in Mei et al. (2018). More specifically, first, the truncation of the parameter w ensures that, as $t \to \infty$, the curvature of the predictor evaluated on the solution ρ_t of the flow (3.4) converges pointwise in x to the corresponding evaluation on the Gibbs distribution $\rho^*_{\tau,m}$. Next, following Mei et al. (2018), we couple the weights obtained after $\lfloor t/\varepsilon \rfloor$ steps of the SGD iteration (3.3) with N i.i.d. particles with distribution ρ_t, thus obtaining that the curvature of the SGD predictor converges to the curvature of the flow predictor. By using this coupling again, together with the fact that along the trajectory of the flow $M(\rho_t) < C$ (see Mei et al. (2018) or Jordan et al. (1998)), we obtain a uniform bound on the second moment of the empirical distribution $\hat{\rho}^N_{\lfloor t/\varepsilon \rfloor}$ of the SGD weights. The final result then follows from the same Lipschitz argument described above.

5.2 Proof of Theorem 1

Let us start with the proof of the vanishing curvature phenomenon. The quantity

$$\frac{\partial^2}{\partial x^2} \int a^{\tau,m} (w^{\tau,m} x + b)^m_\tau \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{5.3}$$

is hard to analyze directly due to the presence of the τ-smoothing in the soft-plus activation. However, the structure of the activation $(\cdot)^m_\tau$ alongside with the pointwise convergence of the minimizers $\rho^*_{\tau,m}$ to ρ^*_m (cf. Lemma A.4) allows us to infer the properties of (5.3) through the analysis of the auxiliary object:

$$\int a^m (w^m)^2 \rho^*_m(a, w, -w^m x) \mathrm{d}a\mathrm{d}w. \tag{5.4}$$

Formally, we show that the approximation result below holds.

Lemma 5.1 (Convergence to delta)**.** *Assume that condition **A1** holds. Let $\rho^*_{\tau,m}$ and ρ^*_m be the minimizers of the free energy for truncated softplus and ReLU activations, respectively, as defined in (3.9)-(3.10). Then,*

$$\lim_{\tau \to \infty} \left| \frac{\partial^2}{(\partial x)^2} \int a^{\tau,m} \left[(w^{\tau,m} x + b)^m_\tau \right] \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - \int a^m (w^m)^2 \rho^*_m(a, w, -w^m x) \mathrm{d}a\mathrm{d}w \right| \leq \frac{C}{m\lambda},$$

where C is a constant independent of $(m, \tau, \beta, \lambda)$.

Proof of Lemma 5.1. First, we show that

$$\begin{aligned} \lim_{\tau \to \infty} \Bigg| \int a^{\tau,m} \left[\frac{\partial^2}{(\partial x)^2} (w^{\tau,m} x + b)^m_\tau \right] \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\ - \int a^m (w^m)^2 \rho^*_m(a, w, -w^m x) \mathrm{d}a\mathrm{d}w \Bigg| \leq \frac{C}{m\lambda}. \end{aligned} \tag{5.5}$$

Recall the definition of the activation $(\cdot)_m^\tau$ provided in (3.7). We can decompose the integral into two pieces with respect to the domain of truncation and obtain

$$\begin{aligned}
&\int a^{\tau,m}\left[\frac{\partial^2}{(\partial x)^2}(w^{\tau,m}x+b)_\tau^m\right]\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\\
=&\int_{w^{\tau,m}x+b\le x_m} a^{\tau,m}\left[\frac{\partial^2}{(\partial x)^2}(w^{\tau,m}x+b)_\tau\right]\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\\
+&\int_{w^{\tau,m}x+b>x_m} a^{\tau,m}(w^{\tau,m})^2\left[\frac{\partial^2}{(\partial u)^2}\phi_{\tau,m}(u)\bigg|_{u=w^{\tau,m}x+b}\right]\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \qquad (5.6)
\end{aligned}$$

Let us focus on the first term in the RHS of (5.6). The second derivative has the following form

$$\frac{\partial^2}{(\partial x)^2}(w^{\tau,m}x+b)_\tau=(w^{\tau,m})^2\cdot\frac{\tau e^{\tau(w^{\tau,m}x+b)}}{\left(e^{\tau(w^{\tau,m}x+b)}+1\right)^2}>0.$$

Thus, the following chain of equalities holds

$$\begin{aligned}
&\int_{w^{\tau,m}x+b\le x_m} a^{\tau,m}\left[\frac{\partial^2}{(\partial x)^2}(w^{\tau,m}x+b)_\tau\right]\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\\
=&\int_{w^{\tau,m}x+b\le x_m} a^{\tau,m}(w^{\tau,m})^2\cdot\frac{\tau e^{\tau(w^{\tau,m}x+b)}}{\left(e^{\tau(w^{\tau,m}x+b)}+1\right)^2}\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\\
=&\int \mathbb{1}_{\{y\le\tau x_m\}}\cdot a^{\tau,m}(w^{\tau,m})^2\cdot\frac{e^y}{(e^y+1)^2}\rho^*_{\tau,m}\left(a,w,\frac{y}{\tau}-w^{\tau,m}x\right)\mathrm{d}(a,w,y),
\end{aligned}$$

where in the last step we have performed the change of variables $y=\tau(w^{\tau,m}x+b)$. By Lemma A.4, we have that, as $\tau\to\infty$, $\rho^*_{\tau,m}(\boldsymbol{\theta})$ converges to $\rho^*_m(\boldsymbol{\theta})$ pointwise in $\boldsymbol{\theta}$. Furthermore, as $\tau\to\infty$, $a^{\tau,m}$ converges to a^m for any a, and $w^{\tau,m}$ converges to w^m for any w. Thus, as the Gibbs distributions $\rho^*_{\tau,m}(\boldsymbol{\theta})$ and $\rho^*_m(\boldsymbol{\theta})$ are continuous with respect to $\boldsymbol{\theta}$, we have that

$$\begin{aligned}
&\lim_{\tau\to\infty}\left[\mathbb{1}_{\{y\le\tau x_m\}}\cdot a^{\tau,m}(w^{\tau,m})^2\cdot\frac{e^y}{(e^y+1)^2}\rho^*_{\tau,m}\left(a,w,\frac{y}{\tau}-w^{\tau,m}x\right)\right]\\
=&\,a^m(w^m)^2\cdot\frac{e^y}{(e^y+1)^2}\rho^*_m\left(a,w,-w^mx\right).
\end{aligned}$$

Furthermore, combining (A.4) and (A.5) from Lemma A.2, we get the following bound

$$\rho^*_{\tau,m}(\boldsymbol{\theta})\le C'\exp\left(-\frac{\beta\lambda\|\boldsymbol{\theta}\|_2^2}{2}\right), \qquad (5.7)$$

for some constant $C'>0$ independent of $\boldsymbol{\theta}$ and τ. Thus, we have

$$\begin{aligned}
&|a^{\tau,m}|(w^{\tau,m})^2\cdot\frac{e^y}{(e^y+1)^2}\rho^*_{\tau,m}\left(a,w,\frac{y}{\tau}-w^{\tau,m}x\right)\\
\le&C'm^3\cdot\frac{e^y}{(e^y+1)^2}\cdot\exp\left(-\frac{\beta\lambda(a^2+w^2)}{2}\right),
\end{aligned}$$

which is integrable in (y, a, w). Hence, by using the Dominated Convergence theorem and integrating out y using Tonelli's theorem, we have

$$\lim_{\tau\to\infty}\left|\int_{w^{\tau,m}x+b\le x_m} a^{\tau,m}\left[\frac{\partial^2}{(\partial x)^2}(w^{\tau,m}x+b)_\tau\right]\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - \int a^m(w^m)^2\rho^*_m(a,w,-w^m x)\mathrm{d}a\mathrm{d}w\right| = 0. \tag{5.8}$$

Now, by triangle inequality, it remains to show that the absolute value of the second term in the RHS of (5.6) can be upper bounded by $\mathcal{O}\left(\frac{1}{m\lambda}\right)$ as $\tau\to\infty$. Recall that, by construction,

$$|\phi''_{\tau,m}(x)| \le \frac{1}{m^2}, \quad |a^{\tau,m}| \le m, \quad |w^{\tau,m}| \le |w|,$$

for any $x > x_m$ and any $(a, w) \in \mathbb{R}^2$. Thus, the following upper bound holds

$$\begin{aligned}&\lim_{\tau\to\infty}\int_{w^{\tau,m}x+b>x_m}|a^{\tau,m}|(w^{\tau,m})^2\left|\frac{\partial^2}{(\partial u)^2}\phi_{\tau,m}(u)\right|_{u=w^{\tau,m}x+b}\right|\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\\ &\le \frac{1}{m}\lim_{\tau\to\infty}\int w^2\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.\end{aligned} \tag{5.9}$$

In addition, we have the following pointwise convergence of the integrand

$$\lim_{\tau\to\infty} w^2\rho^*_{\tau,m}(\boldsymbol{\theta}) = w^2\rho^*_m(\boldsymbol{\theta}).$$

Furthermore, by using (5.7), we conclude that the integrand can be dominated by an integrable function. Hence, an application of the Dominated Convergence theorem gives that

$$\frac{1}{m}\lim_{\tau\to\infty}\int w^2\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \frac{1}{m}\int w^2\rho^*_m(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \le \frac{C''}{m\lambda}, \tag{5.10}$$

where the last inequality follows from Lemma A.2, which gives that $M(\rho^*_m) < C''/\lambda$ for some $C'' > 0$ that is independent of (m, λ). By combining (5.6), (5.8), (5.9) and (5.10), we conclude that (5.5) holds. Finally, by using a standard line of arguments, i.e., Mean Value theorem and Dominated Convergence, the derivative can be pushed inside the integral sign, which finishes the proof. ■

Next, we study the set on which (5.4) might grow unbounded. In particular, in Lemma 5.3, we provide an upper bound on the measure of the set $\overline{\Omega}_j(m, \beta, \lambda)$ defined in (4.5)-(4.6). To do so, we will first show that the partition function of ρ^*_m is uniformly bounded in m, as stated and proved below.

Lemma 5.2 (Uniform bound on partition function)**.** *Consider $\sigma^*(\boldsymbol{\theta}, x) = a^{\tau,m}(w^{\tau,m}x+b)^m_\tau$ or $\sigma^*(\boldsymbol{\theta}, x) = a^m(w^m x+b)^m_+$, and let $\rho^*_{\sigma^*}$ be the Gibbs distribution with activation σ^*. Then, the following upper bound holds for its partition function $Z_{\sigma^*}(\beta, \lambda)$:*

$$\ln Z_{\sigma^*}(\beta, \lambda) \le \beta C + 1 + 3\log\frac{8\pi}{\beta\lambda},$$

where $C > 0$ is a constant independent of $(m, \tau, \beta, \lambda)$.

Proof of Lemma 5.2. Let $R_i^{\sigma^*}(\rho^*_{\sigma^*})$ be defined as follows

$$R_i^{\sigma^*}(\rho^*_{\sigma^*}) := -\frac{1}{M}\left(y_i - y^{\sigma^*}_{\rho^*_{\sigma^*}}(x_i)\right).$$

By substituting the form (3.6) of the Gibbs distribution into the free energy functional (3.5), we have that

$$\begin{aligned}\mathcal{F}^{\sigma^*}(\rho) = \frac{1}{2M}\sum_{i=1}^{M}\left(y_i - y^{\sigma^*}_{\rho^*_{\sigma^*}}(x_i)\right)^2 + \frac{\lambda}{2}M(\rho^*_{\sigma^*}) \\ - \int \sum_{i=1}^{M}\left[R_i^{\sigma^*}(\rho^*_{\sigma^*})\cdot\sigma^*(x_i,\boldsymbol{\theta})\right]\rho^*_{\sigma^*}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - \frac{\lambda}{2}\int\|\boldsymbol{\theta}\|_2^2\rho^*_{\sigma^*}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - \frac{1}{\beta}\ln Z_{\sigma^*}(\beta,\lambda).\end{aligned}$$

Note that, by Fubini's theorem, we can interchange summation and integration in the first integral, since the activation and the labels are bounded. By using also the definition of $R_i^{\sigma^*}(\rho^*_{\sigma^*})$, we have that

$$\begin{aligned}\mathcal{F}^{\sigma^*}(\rho) &= \frac{1}{2M}\sum_{i=1}^{M}y_i^2 + \frac{1}{2M}\sum_{i=1}^{M}\left(y^{\sigma^*}_{\rho^*_{\sigma^*}}(x_i)\right)^2 - \frac{1}{M}\sum_{i=1}^{M}y_i\cdot y^{\sigma^*}_{\rho^*_{\sigma^*}}(x_i) + \frac{\lambda}{2}M(\rho^*_{\sigma^*}) \\ &\quad - \frac{1}{M}\sum_{i=1}^{M}\left(y^{\sigma^*}_{\rho^*_{\sigma^*}}(x_i)\right)^2 + \frac{1}{M}\sum_{i=1}^{M}y_i\cdot y^{\sigma^*}_{\rho^*_{\sigma^*}}(x_i) - \frac{\lambda}{2}M(\rho^*_{\sigma^*}) - \frac{1}{\beta}\ln Z_{\sigma^*}(\beta,\lambda) \\ &= -\frac{1}{\beta}\ln Z_{\sigma^*}(\beta,\lambda) - \frac{1}{2M}\sum_{i=1}^{M}\left(y^{\sigma^*}_{\rho^*_{\sigma^*}}(x_i)\right)^2 + \frac{1}{2M}\sum_{i=1}^{M}y_i^2 \\ &\le -\frac{1}{\beta}\ln Z_{\sigma^*}(\beta,\lambda) + \frac{1}{2M}\sum_{i=1}^{M}y_i^2 \\ &\le -\frac{1}{\beta}\ln Z_{\sigma^*}(\beta,\lambda) + C,\end{aligned}$$

where $C > 0$ is independent of (m,τ,β,λ). From Lemma 10.2 in Mei et al. (2018), we obtain that, for any $\rho\in\mathcal{K}$,

$$\mathcal{F}(\rho) \ge \frac{1}{2}R(\rho) + \frac{\lambda}{4}M(\rho) - \frac{1}{\beta}\left[1 + 3\log\frac{8\pi}{\beta\lambda}\right] \ge -\frac{1}{\beta}\left[1 + 3\log\frac{8\pi}{\beta\lambda}\right],$$

where the last inequality follows from non-negativity of $R(\rho)$ and $M(\rho)$. Combining the upper and lower bounds gives

$$-\frac{1}{\beta}\ln Z_{\sigma^*}(\beta,\lambda) + C \ge -\frac{1}{\beta}\left[1 + 3\log\frac{8\pi}{\beta\lambda}\right].$$

After a rearrangement, we have

$$\ln Z_{\sigma^*}(\beta,\lambda) \le \beta C + 1 + 3\log\frac{8\pi}{\beta\lambda},$$

which concludes the proof. ■

In order to bound the measure of $\overline{\Omega}_j(m,\beta,\lambda)$, the idea is to combine the upper bound on the partition function of Lemma 5.2 with a lower bound that diverges in m unless $|\overline{\Omega}_j(m,\beta,\lambda)|$ vanishes. In particular, we derive a lower bound with the structure of a Gaussian integral which grows unbounded for a certain set of inputs. This set of inputs corresponds to the scenario when the Gaussian covariance has non-positive eigenvalues, and it can be expressed as the set in which the polynomials f_j and f^j defined in (4.3) are non-negative. For brevity, we suppress the dependence of Ω_j and Ω^j on (m,β,λ) in the proofs below.

Lemma 5.3 (Bound on measure of cluster set)**.** *Assume that condition **A1** holds. For $j \in \{0,\dots,M\}$, let Ω^j and Ω_j be defined as in (4.6). Then,*

$$|\Omega^j|,\ |\Omega_j| \le K_1 \frac{e^{\beta K_2}}{m^2}, \tag{5.11}$$

where $K_1, K_2 > 0$ is independent of (m,β,λ).

Proof of Lemma 5.3. We start with the proof for Ω^j. For $j = M$, the corresponding polynomial $f^M(x)$ is equal to $1+x^2$ and therefore $|\Omega^M| = 0$. Let us now consider the case $j \neq M$, and assume that $\mu(\Omega^j) > 0$. (If that's not the case, the claim trivially holds.)

Note that, as $f^j(x)$ is a polynomial of degree at most two in x, Ω^j is the union of at most two intervals. Then, the following set has a non-zero Lebesgue measure in $\mathbb{R}^2$:

$$\Omega := \{(w,b) \in \mathbb{R}_+ \times \mathbb{R} : b = -w^m x,\ 0 < w < m,\ x \in \Omega^j\}.$$

Now, we can lower bound the partition function as

$$\begin{aligned} Z_m(\beta,\lambda) &\ge \int_{\{|a|<m\}\times\Omega} \exp\left\{-\frac{\beta\lambda}{2}\left[\frac{2}{\lambda}\sum_{i=1}^{M} R_i^m(\rho_m^*)\cdot a^m (w^m x_i + b)_+^m + \|\boldsymbol{\theta}\|_2^2\right]\right\} \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\{|a|<m\}\times\Omega} \exp\left\{-\frac{\beta\lambda}{2}\left[\frac{2}{\lambda}\sum_{i=j+1}^{M} R_i^m(\rho_m^*)\cdot a^m (w^m x_i + b) + \|\boldsymbol{\theta}\|_2^2\right]\right\} \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{5.12}$$

Here, the equality in the second line follows from the following observation: if $i \in [j]$ and $(w,b) \in \Omega$, then $w^m x_i + b \le 0$ and therefore $(w^m x_i + b)_+^m = 0$; if $i > j$ and $(w,b) \in \Omega$, then $0 < w^m x_i + b < m^2$ ($|x|, |x_i| \le L$, hence $|x_i - x| \le m$, as L is a numerical constant independent of m and m is sufficiently large by assumption **A1**) and therefore $(w^m x_i + b)_+^m = w^m x_i + b$ for all $(w,b) \in \Omega$. Thus, after the change of variables $(a,w,b) \mapsto (a,w,-w^m x)$ and an application of Tonelli's theorem, the RHS in (5.12) reduces to

$$\int_{x\in\Omega^j}\int_{\{|a|<m\}\times\{0<w<m\}} w \cdot \exp\left\{-\frac{\beta\lambda}{2}\left[2aw(B^j - A^j x) + a^2 + w^2(1+x^2)\right]\right\} \mathrm{d}(a,w)\mathrm{d}x. \tag{5.13}$$

Here the coefficients A^j and B^j are defined as per (4.4). The term under the exponent can be rewritten as

$$2aw(B^j - A^j x) + a^2 + w^2(1+x^2) = \begin{bmatrix} a & w \end{bmatrix} \Sigma^{-1} \begin{bmatrix} a \\ w \end{bmatrix},$$

with

$$\Sigma^{-1} = \begin{bmatrix} 1 & (B^j - A^j x) \\ (B^j - A^j x) & 1 + x^2 \end{bmatrix}.$$

By definition of Ω^j in conjunction with Sylvester's criterion, we have that Σ^{-1} has a non-positive eigenvalue with corresponding eigenvector

$$\lambda_- = \frac{1}{2}\left(-\sqrt{4(B^j - A^j x)^2 + x^4} + x^2 + 2\right) \leq 0, \quad v_- = \left(-\frac{x^2 + \sqrt{4(B^j - A^j x)^2 + x^4}}{2(B^j - A^j x)}, 1\right).$$

Furthermore, the other eigenvalue with corresponding eigenvector is given by

$$\lambda_+ = \frac{1}{2}\left(\sqrt{4(B^j - A^j x)^2 + x^4} + x^2 + 2\right) > 0, \quad v_+ = \left(-\frac{x^2 - \sqrt{4(B^j - A^j x)^2 + x^4}}{2(B^j - A^j x)}, 1\right).$$

Note that v_- and v_+ are orthogonal, and consider the following change of variables for the integral

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} v_-/\|v_-\|_2 \\ v_+/\|v_+\|_2 \end{bmatrix} \begin{bmatrix} a \\ w \end{bmatrix} = Q^T \begin{bmatrix} a \\ w \end{bmatrix} \Leftrightarrow Q\mathbf{z} = \begin{bmatrix} a \\ w \end{bmatrix} \Leftrightarrow \begin{bmatrix} a(\mathbf{z}) \\ w(\mathbf{z}) \end{bmatrix} := Q\mathbf{z}.$$

As the matrix Q is unitary, the quantity in (5.13) can be rewritten as

$$\int_{x \in \Omega^j} \int_{\{|a(\mathbf{z})| < m\} \times \{0 < w(\mathbf{z}) < m\}} w(\mathbf{z}) \cdot \exp\left\{-\frac{\beta\lambda}{2}\left[\lambda_- z_1^2 + \lambda_+ z_2^2\right]\right\} \mathrm{d}\mathbf{z}\mathrm{d}x,$$

as the determinant of the Jacobian is 1 for any unitary linear transformation. As $\lambda_- \leq 0$, this quantity is lower bounded by

$$\int_{x \in \Omega^j} \int_{\{|a(\mathbf{z})| < m\} \times \{0 < w(\mathbf{z}) < m\}} w(\mathbf{z}) \cdot \exp\left\{-\frac{\beta\lambda}{2}\left[\lambda_+ z_2^2\right]\right\} \mathrm{d}\mathbf{z}\mathrm{d}x. \tag{5.14}$$

Notice that $\|v_-\| \geq 1$, $\|v_+\| \geq 1$ and $w(\mathbf{z}) = z_1/\|v_-\|_2 + z_2/\|v_+\|_2$. Thus, picking $z_1 \in (0, m/2]$ and $z_2 \in (0, m/2]$ ensures that $0 < w(\mathbf{z}) < m$. Furthermore, these conditions on $\mathbf{z}$ do not violate the requirement on $a(\mathbf{z})$, since $|a(\mathbf{z})| \leq |z_1| + |z_2| \leq m$. Consequently, as the integrand is non-negative, the integral in (5.14) is lower bounded by

$$\int_{x \in \Omega^j} \int_{\{0 < z_1 < m/2\} \times \{0 < z_2 < m/2\}} w(\mathbf{z}) \cdot \exp\left\{-\frac{\beta\lambda}{2}\left[\lambda_+ z_2^2\right]\right\} \mathrm{d}\mathbf{z}\mathrm{d}x. \tag{5.15}$$

By Lemma A.5, $|R_i^m(\rho_m^*)|$ is bounded by a constant independent of (m, β, λ), since $\lambda < C_3$ from condition **A1**. Hence, $\lambda|A^j x - B^j|$ is also uniformly bounded in (m, β, λ). This, in particular, implies that

$$\lambda \cdot \lambda_+ \leq K_1,$$

where $K_1 > 0$ is independent of (m, β, λ). Furthermore, by definition of Ω^j, $|B^j - A^j x| > 1$, which implies that $\|v_+\|_2$ and $\|v_-\|_2$ are also upper bounded by a constant $K_2 > 0$ independent of (m, β, λ), and therefore

$$w(z) \leq \frac{z_1 + z_2}{K_2}.$$

With this in mind, we can then further lower bound the integral in (5.15) by

$$\begin{aligned}
&\int_{x\in\Omega^j}\int_{\{0<z_1<m/2\}\times\{0<z_2<m/2\}} \frac{1}{K_2}(z_1+z_2)\cdot\exp\left\{-\frac{K_1\beta}{2}\cdot z_2^2\right\}\mathrm{d}\mathbf{z}\mathrm{d}x \\
&= |\Omega^j|\int_{\{0<z_1<m/2\}\times\{0<z_2<m/2\}} \frac{1}{K_2}(z_1+z_2)\cdot\exp\left\{-\frac{K_1\beta}{2}\cdot z_2^2\right\}\mathrm{d}\mathbf{z} \\
&\geq |\Omega^j|\int_{\{0<z_1<m/2\}\times\{0<z_2<m/2\}} \frac{1}{K_2}z_1\cdot\exp\left\{-\frac{K_1\beta}{2}\cdot z_2^2\right\}\mathrm{d}\mathbf{z} \qquad (5.16)\\
&= \frac{|\Omega^j|}{K_2}\left[\frac{m^2}{8}\sqrt{\frac{\pi}{2K_1\beta}}\mathrm{erf}\left(\frac{m\sqrt{K_1\beta}}{2\sqrt{2}}\right)\right] \\
&\geq |\Omega^j|\frac{K_3\, m^2}{\sqrt{\beta}},
\end{aligned}$$

where $K_3 > 0$ is independent of (m,β,λ) and in the last passage we have used that $\mathrm{erf}\left(\frac{m\sqrt{K_1\beta}}{2\sqrt{2}}\right) \geq 1/10$ for sufficiently large m and β. By combining (5.16) with the upper bound on the partition function given by Lemma 5.2, the desired result immediately follows and the proof for Ω^j is complete.

In regards to the argument for Ω_j, for $j=0$ the result trivially holds, since $f_0(x) = 1+x^2$ and, thus, $|\Omega_0| = 0$. For $j > 0$, the partition function can be lower bounded by

$$\int_{\{|a|<m\}\times\Omega}\exp\left\{-\frac{\beta\lambda}{2}\left[\frac{2}{\lambda}\sum_{i=1}^{j}R_i^m(\rho_m^*)\cdot a^m(w^m x_i+b)+\|\boldsymbol{\theta}\|_2^2\right]\right\}\mathrm{d}\boldsymbol{\theta}, \qquad (5.17)$$

where the set Ω is defined on non-positive w and $x\in\Omega_j$, i.e.,

$$\Omega := \{(w,b)\in\mathbb{R}_+\times\mathbb{R} : b = -w^m x,\ -m < w < 0,\ x\in\Omega_j\}.$$

The rest of the argument remains the same by noting that with the change of variable

$$(a,w,b)\mapsto(-a,-w,w^m x)$$

the quantity in (5.17) is equal to

$$\int_{x\in\Omega_j}\int_{\{|a|<m\}\times\{0<w<m\}} w\cdot\exp\left\{-\frac{\beta\lambda}{2}\left[2aw(B_j-A_j x)+a^2+w^2(1+x^2)\right]\right\}\mathrm{d}(a,w)\mathrm{d}x,$$

which is exactly as in (5.13), but with $x\in\Omega_j$ and the polynomial $(B_j - A_j x)$ in place of $x\in\Omega^j$ and the polynomial $(B^j - A^j x)$. ■

In order to control the magnitude of (5.4), it is also necessary to understand the behavior of the polynomials defined in (4.3). The worst case scenario, in terms of presenting a challenge to bounding the curvature, corresponds to f^j or f_j being arbitrarily close to zero on the whole area outside of cluster set. In fact, this would imply that the Gaussian-like integral arising in the computation of (5.4) has arbitrary small eigenvalues. More specifically,

our plan is to exploit the following bound for $x \in I_j \setminus \overline{\Omega}_j(m, \beta, \lambda)$:

$$|(5.4)| \leq C \int |a| w^2 \Bigg[\exp\left\{ -\frac{\beta\lambda}{2} \cdot f^j(x) \cdot (a^2 + w^2) \right\} + \exp\left\{ -\frac{\beta\lambda}{2} \cdot f_j(x) \cdot (a^2 + w^2) \right\} \Bigg] \mathrm{d}\theta. \tag{5.18}$$

Now, the RHS of (5.18) diverges (and, therefore, the bound is useless), if either of the polynomials is arbitrarily close to zero outside of the cluster set. Fortunately, we are able to prove that this cannot happen: in Lemma 5.5 we show that $f^j(x)$ and $f_j(x)$ can be small only when x approaches the cluster set, i.e.,

$$f^j(x), f_j(x) \geq \min\{C^j(x), C_j(x), 1\},$$

where $C^j(x), C_j(x)$ are defined in (4.8) and, because of the condition on their coefficients $\{K_i\}_{i=1}^4$, they cannot be arbitrarily close to 0 in any interval I_j.

As a preliminary step towards the proof of Lemma 5.5, we show an auxiliary result for polynomials of a certain form. Fix some interval $I = [I_l, I_r] \subset \mathbb{R}$. Given two quantities $a, b \in \mathbb{R}$, consider the following polynomial of degree at most two

$$P_2(x) := (1 - a^2) \cdot x^2 + 2ab \cdot x + (1 - b^2), \quad x \in I, \tag{5.19}$$

where we suppress the dependence on (a, b), i.e., $P_2(x; a, b) = P_2(x)$, for more compact notation. In addition, let Ω_+ be the subset of I on which P_2 is strictly positive, i.e.,

$$\Omega_+ := \{x \in I : P_2(x) > 0\}.$$

For a fixed small constant $C_\Omega > 0$, define the set of admissible coefficients as follows

$$\mathcal{U} := \{(a, b) \in \mathbb{R}^2 : |\Omega_+| \geq C_\Omega\}. \tag{5.20}$$

Given $(a, b) \in \mathcal{U}$ and $x \in \Omega_+$, we define the critical point x_c of the polynomial P_2 associated with x and Ω_+ in the same fashion as in Definition 4.1, after replacing $f^j(\cdot)$ with $P_2(\cdot)$ and $I_j \setminus \Omega^j(m, \beta, \lambda)$ with Ω_+. Notice that, since Ω_+ has strictly positive Lebesgue measure for $(a, b) \in \mathcal{U}$, the critical point is well-defined and, in particular, $x_c \in I$ always holds.

Lemma 5.4 (Lower bound on polynomial)**.** *Fix some C_Ω such that $\mathcal{U}$, as defined in (5.20), is of positive measure. Pick some interval $(a, b) \in \mathcal{U}$. Let $x \in \Omega_+$ and x_c be the critical point associated to x. Then, the following holds*

$$P_2(x) \geq \alpha_2 (x - x_c)^2 + \alpha_1 |x - x_c| + \alpha_0, \tag{5.21}$$

where $\alpha_0, \alpha_1, \alpha_2 \geq 0$ and at least one of them is lower bounded by a strictly positive constant depending on C_Ω but independent of the choice of $(a, b) \in \mathcal{U}$.

We defer the proof of Lemma 5.4 to Appendix A.3. Recall the definition of the polynomial $f^j(x)$ given in (4.3), and notice that expression can be rearranged such that $f^j(x)$ is in the form of (5.19), namely

$$f^j(x) = 1 + x^2 - (A^j x - B^j)^2 = (1 - (A^j)^2) x^2 + 2A^j B^j x + (1 - (B^j)^2).$$

In this view, the following result follows from Lemma 5.4.

Lemma 5.5 (Well-defined quadratic form)**.** *Assume that $(A^j, B^j) \in \mathcal{U}$, i.e., $|I_j \setminus \Omega^j|$ is lower bounded by a positive constant. Given $x \in I_j \setminus \Omega^j$, let x_c be the critical point associated to x. Then, we have that*

$$f^j(x) \geq C^j(x) := \gamma_1(x - x_c)^2 + \gamma_2, \tag{5.22}$$

where $\gamma_1, \gamma_2 > 0$ and either $\gamma_1 > \varepsilon$ or $\gamma_2 > \varepsilon$ for some $\varepsilon > 0$ that is independent of (A^j, B^j) but depending on C_Ω as appearing in the definition of $\mathcal{U}$.

Proof of Lemma 5.5. Note that $I_j \setminus \Omega^j$ is the set in which f^j is strictly positive. Hence, since $|I_j \setminus \Omega^j|$ is lower bounded by a positive constant independent of A^j, B^j, we can apply Lemma 5.4 to get

$$f^j(x) \geq \alpha_2(x - x_c)^2 + \alpha_1|x - x_c| + \alpha_0,$$

where $\alpha_0, \alpha_1, \alpha_2 \geq 0$ and at least one of them is lower bounded by a strictly positive constant independent of (A^j, B^j). Thus, since each term of the RHS above is non-negative, we get

$$f^j(x) \geq \alpha_i|x - x_c|^i + \alpha_0,$$

where $i = \arg\max_{j \in \{1,2\}} \alpha_j$. Furthermore, as $|x - x_c| \leq |I_j|$, we have

$$f^j(x) \geq \frac{\alpha_i}{|I_j|^{2-i}}|x - x_c|^2 + \alpha_0.$$

Now, either α_i or α_0 as well as $1/|I_j|$ are lower bounded by strictly positive constants independent of (A^j, B^j). Thus, taking $\gamma_1 = \alpha_i/|I_j|^{2-i}$ and $\gamma_2 = \alpha_0$ concludes the proof. ■

Let us point out that, although ε does not depend on the values of $(A^j, B^j) \in \mathcal{U}$, the position of a critical point x_c *depends* on (A^j, B^j).

In a similar fashion, we define $\bar{\mathcal{U}}$ to be the set of admissible (A_j, B_j) as in (5.20), and given $x \in I_j \setminus \Omega_j$, we let $\bar{x}_c$ be the critical point associated to x and Ω_j. Then, a result analogous to Lemma 5.5 holds for $f_j(x)$:

$$f_j(x) \geq C_j(x) := \gamma_3(x - \bar{x}_c)^2 + \gamma_4, \tag{5.23}$$

where $\gamma_3, \gamma_4 > 0$ and either $\gamma_3 > \varepsilon$ or $\gamma_4 > \varepsilon$ for some $\varepsilon > 0$ that is independent of the choice of $(A_j, B_j) \in \bar{\mathcal{U}}$.

The last ingredient for the proof of the vanishing curvature phenomenon is the control of the decay of the partition function $Z_m(\beta, \lambda)$ as $\beta \to 0$.

Lemma 5.6 (Lower bound on partition function independent of m)**.** *Assume that condition **A1** holds. Then,*

$$Z_m(\beta, \lambda) \geq \frac{C}{\sqrt{\beta^3 \lambda^{3/2}}},$$

for some $C > 0$ that is independent of (m, β, λ).

The proof of Lemma 5.6 is deferred to Appendix A.2. At this point, we are ready to provide an upper bound on the magnitude of (5.4).

Lemma 5.7 (Integral upper bound)**.** *Assume that condition **A1** holds. Furthermore, assume that $m > e^{\beta K_2}$, where K_2 is given in (5.11). Fix $j \in \{0, \dots, M\}$. Then, for any $x \in I_j \setminus \big(\Omega^j \cup \Omega_j\big)$,*

$$\left| \int a^m (w^m)^2 \rho_m^*(a, w, -w^m x) \mathrm{d}a\mathrm{d}w \right| \leq \frac{K}{\beta \lambda^{7/4} (\bar{C}^j(x))^2},$$

where $K > 0$ is independent of (m, β, λ), $\bar{C}^j(x) := \min\big\{C_j(x), C^j(x), 1\big\}$, and $C^j(x), C_j(x)$ are given by (5.22) and (5.23), respectively.

Proof of Lemma 5.7. Note that the following upper bound holds

$$\left| \int a^m (w^m)^2 \rho_m^*(a, w, -w^m x) \mathrm{d}a\mathrm{d}w \right| \leq I(x) := \int |a^m| (w^m)^2 \rho_m^*(a, w, -w^m x) \mathrm{d}a\mathrm{d}w.$$

Let us now decompose the integral $I(x)$ depending on the sign of w, i.e.,

$$Z_m(\beta, \lambda) \cdot I(x) = I^j(x) + I_j(x),$$

where

$$I^j(x) := \int_{\{a \in \mathbb{R}\} \times \{w \geq 0\}} |a^m| (w^m)^2 \exp\left\{-\beta \Psi^j(a, w, \rho_m^*)\right\} \mathrm{d}a\mathrm{d}w,$$
$$I_j(x) := \int_{\{a \in \mathbb{R}\} \times \{w < 0\}} |a^m| (w^m)^2 \exp\left\{-\beta \Psi_j(a, w, \rho_m^*)\right\} \mathrm{d}a\mathrm{d}w,$$

and, recalling the form of $\rho_m^*(a, w, -w^m x)$ from (3.10), the corresponding potentials are given by

$$\Psi^j(a, w, \rho) = \sum_{i=j+1}^{M} R_i^m(\rho) \cdot a^m w^m (x_i - x) + \frac{\lambda}{2}\left\{a^2 + w^2 + (w^m)^2 x^2\right\},$$
$$\Psi_j(a, w, \rho) = \sum_{i=1}^{j} R_i^m(\rho) \cdot a^m w^m (x_i - x) + \frac{\lambda}{2}\left\{a^2 + w^2 + (w^m)^2 x^2\right\}.$$

By recalling from (4.4) the definitions of A^j, A_j, B^j and B_j, we obtain the following upper bounds.

$$I^j(x) \leq 2 \int_{\{a \geq 0\} \times \{w \geq 0\}} a w^2 \exp\left\{-\frac{\beta\lambda}{2}\left[-2aw^m |B^j - A^j x| + a^2 + w^2 + (w^m)^2 x^2\right]\right\} \mathrm{d}a\mathrm{d}w, \tag{5.24}$$

$$I_j(x) \leq 2 \int_{\{a \geq 0\} \times \{w < 0\}} a w^2 \exp\left\{-\frac{\beta\lambda}{2}\left[2aw^m |B_j - A_j x| + a^2 + w^2 + (w^m)^2 x^2\right]\right\} \mathrm{d}a\mathrm{d}w. \tag{5.25}$$

Let us analyze the RHS of (5.24). This term can be rewritten as

$$\begin{aligned} 2 \int_{\{a \geq 0\} \times \{w \geq 0\}} a w^2 \exp\left\{-\frac{\beta\lambda}{2}\left[-2aw^m |A^j x - B^j| + a^2 + (w^m)^2 (A^j x - B^j)^2\right]\right\} \\ \cdot \exp\left\{-\frac{\beta\lambda}{2}\left[w^2 + (w^m)^2 x^2 - (w^m)^2 (A^j x - B^j)^2\right]\right\} \mathrm{d}a\mathrm{d}w. \end{aligned} \tag{5.26}$$

Note that

$$|\Omega^j| \leq \frac{K_1\, e^{\beta K_2}}{m^2} \leq \frac{K_1}{e^{\beta K_2}},$$

where the first inequality follows from Lemma 5.3, and the second inequality uses that $m > e^{\beta K_2}$. Therefore, for sufficiently large β, $|\Omega^j|$ is smaller than $|I_j|/2$, and therefore $|I_j \setminus \Omega^j|$ is lower bounded by $|I_j|/2$. At this point, we can apply Lemma 5.5 which gives that $1 + x^2 - (A^j x - B^j)^2 \geq C^j(x) \geq \bar{C}^j(x) := \min\left\{C^j(x), C_j(x), 1\right\}$. Thus, (5.26) is upper bounded by

$$\begin{aligned} &2\int_{\{a\geq 0\}\times\{w\geq 0\}} a w^2 \exp\left\{-\frac{\beta\lambda}{2}\left(a - |B^j - A^j x| w^m\right)^2\right\} \\ &\qquad\qquad \cdot \exp\left\{-\frac{\beta\lambda}{2}\left[w^2 - (w^m)^2(1-\bar{C}^j(x))\right]\right\} \mathrm{d}a\mathrm{d}w \qquad (5.27) \\ &= 2\int_{\{w\geq 0\}} w^2 \exp\left\{-\frac{\beta\lambda}{2}\left[w^2 - (w^m)^2(1-\bar{C}^j(x))\right]\right\} \sqrt{\frac{2\pi}{\beta\lambda}}\mathbb{E}\left[(A)_+\right] \mathrm{d}w, \end{aligned}$$

where $A \sim \mathcal{N}(|B^j - A^j x| w^m, (\beta\lambda)^{-1})$. Furthermore, the following chain of inequalities hold:

$$\mathbb{E}\left[(A)_+\right] \leq \mathbb{E}\left[|A|\right] \leq \sqrt{\mathbb{E}\left[A^2\right]} = \sqrt{|B^j - A^j x|^2 (w^m)^2 + \frac{1}{\beta\lambda}}, \qquad (5.28)$$

where the second passage follows from Jensen's inequality. By using (5.28), the RHS of (5.27) is upper bounded by

$$\begin{aligned} &\frac{2\sqrt{2\pi}}{\sqrt{\beta\lambda}} \int_{\{w\geq 0\}} \sqrt{(B^j - A^j x)^2 (w^m)^2 + \frac{1}{\beta\lambda}} \\ &\qquad\qquad \cdot\, w^2 \exp\left\{-\frac{\beta\lambda}{2}\left[w^2 - (w^m)^2(1-\bar{C}^j(x))\right]\right\} \mathrm{d}w. \end{aligned}$$

Applying Lemma 5.5 again to obtain $(A^j x - B^j)^2 \leq 1 + x^2 - \bar{C}^j(x) \leq 1 + x^2$ and noting by definition that $(w^m)^2 \leq w^2$, we now upper bound this last term by

$$\begin{aligned} &2\sqrt{2\pi}\int_{\{w\geq 0\}} \sqrt{\frac{w^2(1+x^2)}{\beta\lambda} + \frac{1}{\beta^2\lambda^2}} \;\cdot\; w^2 \exp\left\{-\frac{\beta\lambda}{2}\left[w^2 - (w^m)^2(1-\bar{C}^j(x))\right]\right\} \mathrm{d}w \\ &\leq 2\sqrt{2\pi}\int_{\{w\in\mathbb{R}\}} \sqrt{\frac{w^2(1+x^2)}{\beta\lambda} + \frac{1}{\beta^2\lambda^2}} \;\cdot w^2 \exp\left\{-\frac{\beta\lambda}{2}\left[\bar{C}^j(x)\cdot w^2\right]\right\} \mathrm{d}w \qquad (5.29) \\ &\leq 2\sqrt{2\pi}\int_{\{w\in\mathbb{R}\}} \left(\sqrt{\frac{w^2(1+x^2)}{\beta\lambda}} + \sqrt{\frac{1}{\beta^2\lambda^2}}\right) \cdot w^2 \exp\left\{-\frac{\beta\lambda}{2}\left[\bar{C}^j(x)\cdot w^2\right]\right\} \mathrm{d}w, \end{aligned}$$

where in the second line we use that $1 - \bar{C}^j(x) \geq 0$ and again that $(w^m)^2 \leq w^2$, and in the third line we use that $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$.

Finally, computing explicitly the last integral gives the following upper bound on the RHS of (5.24) and consequently on $I^j(x)$:

$$I^j(x) \leq 4\pi\sqrt{\frac{1}{\beta^2\lambda^2}} \;\cdot\; \sqrt{\frac{1}{(\bar{C}^j(x))^3\beta^3\lambda^3}} + 2\sqrt{2\pi}\sqrt{\frac{1+x^2}{\beta\lambda}}\sqrt{\frac{1}{(\bar{C}^j(x))^4\beta^4\lambda^4}}.$$

By following the similar passages, we obtain the same upper bound for $I_j(x)$. By using the lower bound on the partition function shown in Lemma 5.6, we conclude that

$$I(x) = \frac{I^j(x) + I_j(x)}{Z_m(\beta, \lambda)} \leq \frac{K}{\beta \lambda^{7/4} (\bar{C}^j(x))^2},$$

where $K > 0$ is independent of (m, β, λ), and the proof is complete. ■

The proof of Theorem 1 is an immediate consequence of the results presented so far.

Proof of Theorem 1. The proof of (4.9) follows from Lemmas 5.1 and 5.7, and the proof of (4.11) follows from Lemma 5.3. ■

5.3 Proof of Theorem 2

To summarize, at this point we have shown that as $\beta \to \infty$ the second derivative of the predictor vanishes outside the cluster set, and that the size of the cluster set shrinks to concentrate on at most 3 points per prediction interval. With these results in mind, we are ready to provide the proof for Theorem 2.

Proof of Theorem 2. The predictor evaluated at the Gibbs distribution is given by

$$y_n(x) = \int a^{\tau,m}(w^{\tau,m}x + b)_\tau^m \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta},$$

where $n = (\tau, m, \beta, \lambda)$ denotes the aggregated index and we suppress the dependence on (β, λ) in $\rho^*_{\tau,m}$ for convenience. By Lemma A.6, there exists $\tau(m, \beta, \lambda)$ such that, for any $\tau > \tau(m, \beta, \lambda)$,

$$M(\rho^*_{\tau,m}) \leq C, \tag{5.30}$$

for some $C > 0$ independent of $(\tau, m, \beta, \lambda)$. We start by showing that the family of predictors $\{y_n\}$ is equi-Lipschitz for $\infty > \tau > \tau(m, \beta, \lambda)$. First, note that

$$\frac{\partial}{\partial x} y_n(x) = \int \frac{\partial}{\partial x} \Big[a^{\tau,m}(w^{\tau,m}x + b)_\tau^m \Big] \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \tag{5.31}$$

since the derivative can be pushed inside by the same line of arguments as given in the proof of Lemma 5.1. Next, we have that, by construction of the activation, the following holds

$$\int \frac{\partial}{\partial x} \Big[a^{\tau,m}(w^{\tau,m}x + b)_\tau^m \Big] \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq C_1 \int |a^{\tau,m} w^{\tau,m}| \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta},$$

where, from here on, $C_1 > 0$ denotes a generic constant which might change from line to line, but is independent of $(\tau, m, \beta, \lambda)$. By construction, for any $u \in \mathbb{R}$, it holds that $|u^{\tau,m}| \leq |u|$. Thus, we have that

$$\int \frac{\partial}{\partial x} \Big[a^{\tau,m}(w^{\tau,m}x + b)_\tau^m \Big] \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq C_1 \int |aw| \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

Using the Cauchy-Schwartz inequality and (5.30), we obtain that

$$\int \frac{\partial}{\partial x} \Big[a^{\tau,m}(w^{\tau,m}x + b)_\tau^m \Big] \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq C_1 M(\rho^*_{\tau,m}) \leq C_1. \tag{5.32}$$

By combining (5.31) and (5.32), we have shown that the family $\{y_n\}$ for $\tau > \tau(m, \beta, \lambda)$ is equi-Lipschitz, as the derivatives are uniformly bounded. By using a similar argument, we can show that the same result holds for the predictor itself, i.e., for all $x \in \bigcup_{j=0}^{M} I_j$, $y_n(x)$ is uniformly bounded.

Note that Theorem 1 considers the curvature of points outside the cluster set, and it gives an upper bound which diverges when $\bar{C}^j(x)$ approaches 0 for some $j \in [M]$. Thus, our next step is to develop the analytical machinery to make this scenario impossible. Let us recall Definitions (4.8) and (4.10). Then, by Lemma 5.5, we have that

$$\bar{C}^j(x) \geq \min\{\gamma_1(x - x_c)^2 + \gamma_2,\ \gamma_3(x - \bar{x}_c)^2 + \gamma_4\},$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4 > 0$ and $\min\{\max\{\gamma_1, \gamma_2\}, \max\{\gamma_3, \gamma_4\}\} > \varepsilon$, for some $\varepsilon > 0$ that is independent of (m, β, λ). Let us focus on the term $\gamma_1(x - x_c)^2 + \gamma_2$. If $\gamma_2 = 0$ or it approaches 0 (as $m, \beta \to \infty$), then we extend $\Omega^j(m, \beta, \lambda)$ as

$$\mathrm{ext}_\delta(\Omega^j(m, \beta, \lambda)) := \left\{ x \in I_j : \min_{x' \in \Omega^j(m,\beta,\lambda) \cup \{x_c\}} |x - x'| < \delta \right\}.$$

Note that adding the singleton $\{x_c\}$ to the argument of the min allows us to also cover the case in which $\Omega^j(m, \beta, \lambda)$ is empty. Otherwise, i.e., if $\gamma_2 > \varepsilon$ for some $\varepsilon > 0$ that is independent of (m, β, λ), the upper bound on the curvature does not diverge and we set $\mathrm{ext}_\delta(\Omega^j(m, \beta, \lambda)) := \Omega^j(m, \beta, \lambda)$. In a similar fashion, we define the extension of $\Omega_j(m, \beta, \lambda)$ by $\mathrm{ext}_\delta(\Omega_j(m, \beta, \lambda))$.

Let $\bar{\Omega}^j_{\mathrm{ext}}$ be the union of $\mathrm{ext}_\delta(\Omega^j(m, \beta, \lambda))$ and $\mathrm{ext}_\delta(\Omega_j(m, \beta, \lambda))$, where we drop the explicit dependence of $\bar{\Omega}^j_{\mathrm{ext}}$ on $(\delta, m, \beta, \lambda)$ for convenience. Then, since f^j and f_j are polynomials of degree two, the extended set $\bar{\Omega}^j_{\mathrm{ext}}$ (just like $\bar{\Omega}^j$) is the union of at most three disjoint open intervals, i.e.,

$$\bar{\Omega}^j_{\mathrm{ext}} = A^j_1 \cup A^j_2 \cup A^j_3,$$

where $\{A^j_i\}_{i=1}^3$ denote such (possibly empty) open intervals. Furthermore, $I_j \setminus \bar{\Omega}^j_{\mathrm{ext}}$ is the union of at most three disjoint closed intervals, i.e.,

$$I_j \setminus \bar{\Omega}^j_{\mathrm{ext}} = B^j_1 \cup B^j_2 \cup B^j_3,$$

where $\{B^j_i\}_{i=1}^3$ denote such (possibly empty) closed intervals.

At this point, we are ready to show that, for all closed intervals $\{B^j_i\}_{i=1}^3$, the predictor y_n can be approximated arbitrarily well by a linear function (which may be different in different closed intervals). Note that y_n is twice continuously differentiable for $\tau < \infty$, and fix $\tilde{x} \in B^j_i$. Then, by combining Taylor's theorem with the result of Theorem 1, we obtain that, for any $x \in B^j_i$,

$$\lim_{\tau \to \infty} |y_n(x) - y_n(\tilde{x}) - y_n'(\tilde{x})(x - \tilde{x})| \leq \mathcal{O}\left(\frac{1}{m\lambda} + \frac{1}{\delta^4 \cdot \beta\lambda^{7/4}}\right), \tag{5.33}$$

where we use that $|x - x_c| \geq \delta$ by construction of the extended set $\bar{\Omega}^j_{\mathrm{ext}}$. Let us define

$$f^i_n(x) = y_n(\tilde{x}) - y_n'(\tilde{x})(x - \tilde{x}).$$

Then, by picking a sufficiently small δ, (5.33) implies that, as $m\lambda \to \infty$ and $\beta\lambda^{7/4} \to \infty$, for all $x \in B_i^j$,

$$|y_n(x) - f_n^i(x)| \to 0. \tag{5.34}$$

We remark that, as shown previously, the coefficients $y_n(\tilde{x})$ and $y_n'(\tilde{x})$ are uniformly bounded in absolute value.

Let us now consider the open intervals $\{A_i^j\}_{i=1}^3$. For any $x \in A_i^j$, let

$$x' = \arg\min_{y \notin A_i^j} |x - y|,$$

and note that, by definition, $x' \in B_{\tilde{i}}^j$ for some $\tilde{i} \in \{1, 2, 3\}$. By picking the linear approximation f_n^i that corresponds to B_i^j and by using the triangle inequality, we obtain that

$$\begin{aligned} |y_n(x) - f_n^i(x)| &\leq |y_n(x) - y_n(x')| + |y_n(x') - f_n^i(x')| + |f_n^i(x') - f_n^i(x)| \\ &\leq \mathcal{O}\left(|x - x'| + |y_n(x') - f_n^i(x')|\right), \end{aligned} \tag{5.35}$$

where the second inequality is due to the fact that the families $\{y_n\}$ and $\{f_n^i\}$ are equi-Lipschitz. From (5.34) the second term in the RHS in (5.35) vanishes. As for the first term, by construction of the extension, together with the result of Lemma 5.3, we have that

$$|x - x'| \leq \mathcal{O}\left(\frac{e^{\beta K_2}}{m^2} + \delta\right),$$

for some $K_2 > 0$ independent of (m, β, λ). Thus, by picking a sufficiently small δ and $m > e^{\beta K_2}$, we conclude that the first term in the RHS in (5.35) also vanishes.

So far, we have showed that, both inside and outside of the extension of the cluster set, the predictor y_n is well approximated by linear functions. It remains to prove that the linear pieces connect, i.e., there exists $\hat{x} \in \bar{\Omega}_{\text{ext}}^j$ such that, for two neighboring linearities f_n^i and f_n^{i+1} (possibly belonging to different intervals), the following holds

$$f_n^i(\hat{x}) - f_n^{i+1}(\hat{x}) = 0.$$

This claim follows from Lipschitz arguments similar to those presented above, and the proof is complete. ■

5.4 Proof of Corollary 4.3

At this point, we have proved a result about the structure of the predictor coming from the minimizer of the free energy (3.5). By using the mean-field analysis in Mei et al. (2018), we finally show that this structural result holds for the predictor obtained from a wide two-layer ReLU network.

Proof of Corollary 4.3. First, we show that, as $t \to \infty$, the second derivative of the predictor evaluated on the solution ρ_t of the flow (3.4) converges to the same quantity evaluated on the Gibbs minimizer $\rho_{\tau,m}^*$. To do so, we decompose the integral involving ρ_t as in Lemma

5.1 (cf. (5.6)):

$$\int a^{\tau,m}\left[\frac{\partial^2}{(\partial x)^2}(w^{\tau,m}x+b)_\tau^m\right]\rho_t(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$
$$=\int_{w^{\tau,m}x+b\le x_m} a^{\tau,m}\left[\frac{\partial^2}{(\partial x)^2}(w^{\tau,m}x+b)_\tau\right]\rho_t(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$
$$+\int_{w^{\tau,m}x+b>x_m} a^{\tau,m}(w^{\tau,m})^2\left[\frac{\partial^2}{(\partial u)^2}\phi_{\tau,m}(u)\bigg|_{u=w^{\tau,m}x+b}\right]\rho_t(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{5.36}$$

Next, we show that a technical condition bounding the free energy at initialization appearing in the statement of Theorem 4 in Mei et al. (2018) is satisfied under the assumption $M(\rho_0)<\infty$ and $H(\rho_0)>-\infty$. Recalling the sandwich bound for the truncated soft-plus activation (3.8) and the fact that that $\tau\ge 1$ by condition **A1**, an application of Cauchy-Schwarz inequality gives

$$R^{\tau,m}(\rho_0)<CM(\rho_0)+C'<\infty,$$

where $C,C'>0$ are some numerical constants independent of (τ,m). This readily implies that

$$\mathcal{F}^{\tau,m}(\rho_0)<\infty,$$

since λ and β^{-1} are upper-bounded by assumption **A1**.

Now we can apply Theorem 4 in Mei et al. (2018) to conclude that, as $t\to\infty$,

$$\rho_t \rightharpoonup \rho^*_{\tau,m}.$$

Thus, as the terms inside the integrals in (5.36) are all bounded for fixed (τ,m,β,λ), by definition of weak convergence, we get that, as $t\to\infty$,

$$\int a^{\tau,m}\left[\frac{\partial^2}{(\partial x)^2}(w^{\tau,m}x+b)_\tau^m\right]\rho_t(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\to\int a^{\tau,m}\left[\frac{\partial^2}{(\partial x)^2}(w^{\tau,m}x+b)_\tau^m\right]\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

Consequently, since the derivative operator can be pushed inside by the same arguments as in Lemma 5.1, we have that, as $t\to\infty$, the following pointwise convergence holds

$$\frac{\partial^2}{(\partial x)^2}\int a^{\tau,m}(w^{\tau,m}x+b)_\tau^m\rho_t(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\to\frac{\partial^2}{(\partial x)^2}\int a^{\tau,m}(w^{\tau,m}x+b)_\tau^m\rho^*_{\tau,m}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{5.37}$$

Next, we show that the second derivative of the predictor obtained from the two-layer ReLU network also converges to the same limit. Recall that $\sigma^*(x,\boldsymbol{\theta})=a^{\tau,m}(w^{\tau,m}x+b)_\tau^m$. Then, by Theorem 3 in Mei et al. (2018), we have that, almost surely, as $N\to\infty,\ \varepsilon_N\to 0$

$$\frac{\partial^2}{(\partial x)^2}\left[\frac{1}{N}\sum_{i=1}^N\sigma^*\left(x,\boldsymbol{\theta}_i^{\lfloor t/\varepsilon\rfloor}\right)\right]\to\frac{\partial^2}{(\partial x)^2}\int a^{\tau,m}(w^{\tau,m}x+b)_\tau^m\rho_t(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{5.38}$$

along any sequence $\{\varepsilon_N\}$ such that $\varepsilon_N\log(N/\varepsilon_N)\to 0$ and $N/\log(N/\varepsilon_N)\to\infty$. By combining (5.37) and (5.38), we obtain that the desired convergence result holds for the LHS of (5.37).

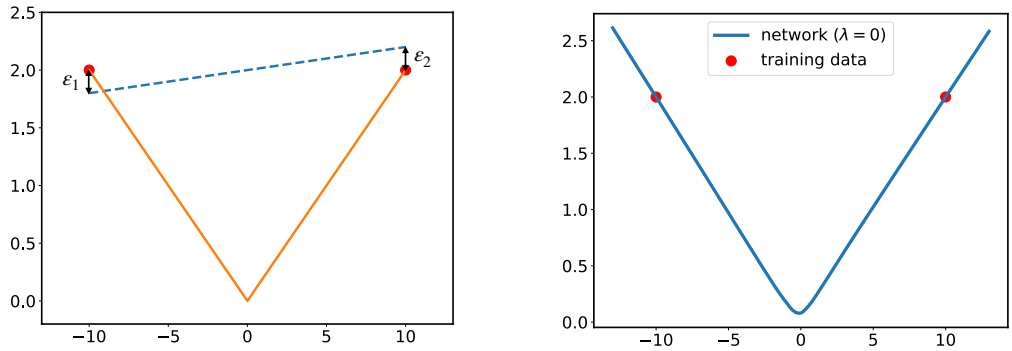


Figure 6: (a) The orange curve represents the function $f^*(x)$ which interpolates the training data (red dots) and exhibits a knot at the point $(0, 0)$; the blue dashed curve is linear in the interval between the training points ($\varepsilon_1 = -0.2$, $\varepsilon_2 = 0.2$). (b) We run noiseless SGD ($\beta = \infty$) with no regularization ($\lambda = 0$) for a two-layer ReLU network with $N = 500$ neurons, trained on the dataset (6.2). The resulting estimator (in blue) approaches the piecewise linear function $f^*(x)$ with a knot between the two training data points.

Another application of Theorem 3 of Mei et al. (2018), together with the fact that the second moment of the flow solution ρ_t is uniformly bounded along the sequence $t \to \infty$ (cf. Lemma 10.2 in Mei et al. (2018), following Proposition 4.1 in Jordan et al. (1998)), gives that the gradients

$$\frac{\partial}{\partial x}\left[\frac{1}{N}\sum_{i=1}^{N}\sigma^*\left(x, \boldsymbol{\theta}_i^{\lfloor t/\varepsilon \rfloor}\right)\right]$$

are almost surely uniformly bounded. This fact, in turn, implies that the corresponding predictor is almost surely equi-Lipschitz. In a similar fashion, we also have that the predictor itself is almost surely uniformly bounded in absolute value.

At this point, the desired result follows from the same line of arguments as in the proof of Theorem 2. ■

6. Knots Inside the Interval

In this section, we provide an explicit example of a 2-point dataset such that the SGD solution exhibits a change of tangent (or "knot") *inside* the training interval. To do so, we will show that neural networks implementing a linear function without knots on the prediction interval *cannot* minimize the free energy (3.5). To simplify the analysis, throughout the section we omit the limits in (τ, m), i.e., we consider directly ReLU activations (this corresponds to taking $\tau = m = \infty$). Similar arguments apply to the case of sufficiently large parameters τ and m.

6.1 Noiseless Regime

We start with the case of noiseless SGD training, i.e., $\beta = +\infty$. Here, the free energy has no entropy penalty and it can be expressed as

$$\mathcal{F}_\infty(\rho) = \frac{1}{2}R(\rho) + \frac{\lambda}{2}M(\rho). \tag{6.1}$$

We consider the following dataset which consists of two points:

$$\mathcal{D} = \{(-\bar{x}, \bar{y}), (\bar{x}, \bar{y})\} = \{(-10, 2), (10, 2)\}. \tag{6.2}$$

Let $f^*(x)$ be the piecewise linear function that interpolates the training data $\{(-\bar{x}, \bar{y}), (\bar{x}, \bar{y})\}$ and passes through the point $(0, 0)$, where it exhibits a knot (see the orange curve in Figure 6a). Note that

$$f^*(x) = \int a(wx + b)_+ \rho^*(a, w, b)\mathrm{d}a\mathrm{d}w\mathrm{d}b,$$

where

$$\rho^*(a, b, w) = \frac{1}{2}\left[\delta_{\left(\sqrt{2\frac{\bar{y}}{\bar{x}}}, -\sqrt{2\frac{\bar{y}}{\bar{x}}}, 0\right)}(a, w, b) + \delta_{\left(\sqrt{2\frac{\bar{y}}{\bar{x}}}, \sqrt{2\frac{\bar{y}}{\bar{x}}}, 0\right)}(a, w, b)\right], \tag{6.3}$$

and $\delta_{(a_0, w_0, b_0)}$ denotes the Dirac delta function centered at (a_0, w_0, b_0). Note that $R(\rho^*) = 0$ and $M(\rho^*) = \frac{2}{5}$. Thus, the free energy is given by

$$\mathcal{F}_\infty(\rho^*) = \frac{1}{2}R(\rho^*) + \frac{\lambda}{2}M(\rho^*) = \frac{\lambda}{5}. \tag{6.4}$$

Let $f(x)$ be a linear function on the interval $[-\bar{x}, \bar{x}]$ such that $f(-\bar{x}) = \bar{y} + \varepsilon_1$ and $f(\bar{x}) = \bar{y} + \varepsilon_2$ (see the blue dashed line in Figure 6a), and let ρ be the corresponding distribution of the parameters, i.e.,

$$f(x) = \int a(wx + b)_+ \rho(a, w, b)\mathrm{d}a\mathrm{d}w\mathrm{d}b. \tag{6.5}$$

In the rest of this section, we will show that, for all $\lambda \leq 1$,

$$\min_{\varepsilon_1, \varepsilon_2} \mathcal{F}_\infty(\rho) > \mathcal{F}_\infty(\rho^*). \tag{6.6}$$

In words, the minimizer of the free energy cannot be a linear function on the interval $[-\bar{x}, \bar{x}]$. As f is linear, we have that

$$f(x) = \frac{\varepsilon_2 - \varepsilon_1}{2\bar{x}}(x - \bar{x}) + \bar{y} + \varepsilon_2,$$

which implies that

$$f(0) = \bar{y} + \frac{\varepsilon_1 + \varepsilon_2}{2} = \int a(b)_+ \rho(a, b)\mathrm{d}a\mathrm{d}b. \tag{6.7}$$

First, we consider the case $f(0) = 0$. From (6.7), we have that $\varepsilon_1 + \varepsilon_2 = -2\bar{y}$. Hence,

$$\mathcal{F}_\infty(\rho) \geq \frac{1}{2}R(\rho) = \frac{1}{4}(\varepsilon_1^2 + \varepsilon_2^2) \geq \frac{1}{8}(\varepsilon_1 + \varepsilon_2)^2 = \frac{\bar{y}^2}{2} = 2. \tag{6.8}$$

By combining (6.8) and (6.4), we conclude that (6.6) holds for all $\lambda \leq 1$ (under the additional restriction $f(0) = 0$).

Next, we consider the case $f(0) \neq 0$. By using (6.7) and applying Cauchy-Schwarz inequality, we have that

$$|f(0)| = \left| \int a(b)_+ \rho(a,b) \mathrm{d}a\mathrm{d}b \right| = |\mathbb{E}[a(b)_+]| \leq \sqrt{\mathbb{E}[a^2]\,\mathbb{E}[(b)_+^2]} \quad \Longrightarrow \quad \mathbb{E}[a^2] \geq \frac{(f(0))^2}{\mathbb{E}[(b)_+^2]}.$$

With this in mind, we can lower bound the regularization term as

$$M(\rho) \geq \mathbb{E}[a^2] + \mathbb{E}[b^2] \geq \frac{(f(0))^2}{\mathbb{E}[(b)_+^2]} + \mathbb{E}[b^2] \geq \frac{(f(0))^2}{\mathbb{E}[(b)_+^2]} + \mathbb{E}[(b)_+^2] \geq 2|f(0)| = 2\left|\bar{y} + \frac{\varepsilon_1 + \varepsilon_2}{2}\right|,$$

where the last inequality follows from the fact that $g(t) = (f(0))^2/t + t$ is minimized over $t \geq 0$ by taking $t = |f(0)|$. Therefore, we have that

$$\mathcal{F}_\infty(\rho) \geq \frac{1}{4}(\varepsilon_1^2 + \varepsilon_2^2) + \lambda\left|\bar{y} + \frac{\varepsilon_1 + \varepsilon_2}{2}\right|.$$

Note that, for a fixed value of the sum $\varepsilon_1 + \varepsilon_2$, the quantity $\varepsilon_1^2 + \varepsilon_2^2$ is minimized when $\varepsilon_1 = \varepsilon_2$. Thus, by recalling that $\bar{y} = 2$, we have

$$\mathcal{F}_\infty(\rho) \geq \min_\varepsilon \left\{ \frac{1}{2}\varepsilon^2 + \lambda|2 + \varepsilon| \right\}. \tag{6.9}$$

One can readily verify that, for any $\lambda \leq 2$, the minimizer is given by $\varepsilon^* = -\lambda$. Thus,

$$\mathcal{F}_\infty(\rho) \geq 2\lambda - \frac{\lambda^2}{2} \geq \frac{3\lambda}{2} > \frac{\lambda}{5} = \mathcal{F}_\infty(\rho^*), \tag{6.10}$$

where the first inequality uses (6.9) and that the minimizer is $\varepsilon^* = -\lambda$, and the next two inequalities use that $\lambda \geq 1$. Merging two cases regarding $f(0)$, we conclude that (6.6) holds, as desired.

6.2 Low Temperature Regime

We now focus on the case of noisy SGD with temperature β^{-1}. Here, the free energy can be expressed as

$$\mathcal{F}_\beta(\rho) = \frac{1}{2}R(\rho) + \frac{\lambda}{2}M(\rho) - \beta^{-1}H(\rho). \tag{6.11}$$

We consider the two-point dataset (6.2) and we recall that $f^*(x)$ has a knot inside the training interval. In this section we will show that the following two results hold for all $\lambda \leq 1$:

(i) There exists a sequence of distributions $\{\rho_\beta^*\}_\beta$ such that, for any $x \in [-\bar{x}, \bar{x}]$,

$$\lim_{\beta\to\infty} \int a(wx+b)_+ \rho_\beta^*(a,w,b)\mathrm{d}a\mathrm{d}w\mathrm{d}b = f^*(x), \tag{6.12}$$

and

$$\limsup_{\beta\to\infty} \mathcal{F}_\beta(\rho_\beta^*) \leq \frac{\lambda}{5}. \tag{6.13}$$

(ii) Let ρ be a distribution such that the function $f(x)$ given by (6.5) is linear in the interval $[-\bar{x}, \bar{x}]$. Pick a sequence of distributions $\{\rho_\beta\}_\beta$ such that $\rho_\beta \rightharpoonup \rho$ and for any $x \in [-\bar{x}, \bar{x}]$,

$$\lim_{\beta\to\infty} \int a(wx+b)_+ \rho_\beta(a,w,b)\mathrm{d}a\mathrm{d}w\mathrm{d}b = f(x). \tag{6.14}$$

Then, we have that

$$\liminf_{\beta\to\infty} \mathcal{F}_\beta(\rho_\beta) > \frac{\lambda}{5}. \tag{6.15}$$

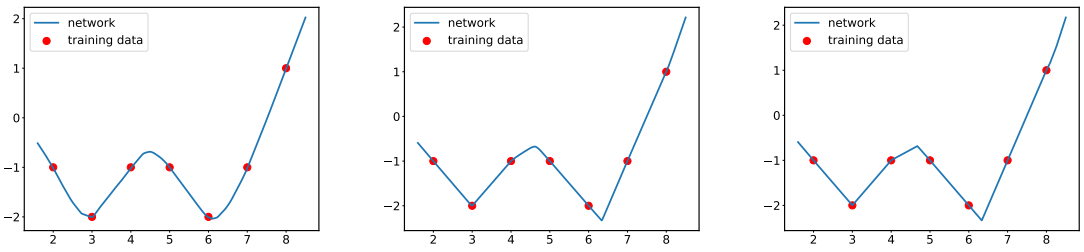
Combining these two results gives that, for sufficiently large β, the minimizer of the free energy (6.11) cannot yield a linear estimator on the interval between the two data points. In Figure 6b, we represent the function obtained by training via SGD a two-layer ReLU network with 500 neurons on the dataset (6.2). Clearly, the blue curve approaches the piecewise linear function $f^*(x)$, which contains a knot inside the interval $[-10, 10]$. The plot represented in the Figure corresponds to the case with no regularization ($\lambda = 0$), but similar results are obtained for small (but non-zero) regularization.

Proof of *(i)*. Let ρ_β^* be defined as

$$\rho_\beta^* = \frac{1}{2}\left[\mathcal{N}\left(\left[\sqrt{2\frac{\bar{y}}{\bar{x}}}, -\sqrt{2\frac{\bar{y}}{\bar{x}}}, 0\right], \beta^{-1} I_{3\times 3}\right) + \mathcal{N}\left(\left[\sqrt{2\frac{\bar{y}}{\bar{x}}}, \sqrt{2\frac{\bar{y}}{\bar{x}}}, 0\right], \beta^{-1} I_{3\times 3}\right)\right],$$

where $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate Gaussian distribution with mean μ and covariance Σ. As $\beta \to \infty$, we have that $\rho_\beta^* \rightharpoonup \rho^*$, where ρ^* is given by (6.3). However, weak convergence does not suffice for pointwise convergence of the corresponding estimators, since the function $\sigma^*(x) = a(wx+b)_+$ is unbounded (in x). To solve this issue, we observe that the fourth moment of ρ_β^* is uniformly bounded as $\beta \to \infty$. Thus, by the de la Vallée Poussin criterion (see e.g. Hu and Rosalsky (2011)), we have that the sequence of random variables $\{\|X_\beta\|_2^2\}_\beta$ is uniformly integrable, with $X_\beta \sim \rho_\beta^*$. Consider a ball $B_r = \{\mathbf{v} \in \mathbb{R}^3 : \|\mathbf{v}\|_2 \leq r\}$, for $r > \sqrt{4\bar{y}/\bar{x}}$. Then, we have

$$\begin{aligned}
&\left|\int_{\mathbb{R}^3} a(wx+b)_+(\rho_\beta^*(a,w,b) - \rho^*(a,w,b))\mathrm{d}a\mathrm{d}w\mathrm{d}b\right| \\
&\qquad \leq \left|\int_{B_r} a(wx+b)_+(\rho_\beta^*(a,w,b) - \rho^*(a,w,b))\mathrm{d}a\mathrm{d}w\mathrm{d}b\right| \\
&\qquad\qquad + \left|\int_{\mathbb{R}^3\setminus B_r} a(wx+b)_+ \rho_\beta^*(a,w,b)\mathrm{d}a\mathrm{d}w\mathrm{d}b\right|,
\end{aligned} \tag{6.16}$$



(a) $\beta^{-1} = 0.005$ **(b)** $\beta^{-1} = 0.0001$ **(c)** $\beta^{-1} = 0$

Figure 7: Functions learnt by a two-layer ReLU network with $N = 500$ neurons, for different values of the temperature parameter β^{-1}. The regularization coefficient λ is set to zero.

where we have used that the support of ρ^* lies inside the ball B_r. The first term in the RHS of (6.16) vanishes as $\beta \to \infty$ by weak convergence, since the function $a(wx+b)_+$ is bounded inside B_r. For the second term, we have that, for any $x \in [-\bar{x}, \bar{x}]$,

$$\left| \int_{\mathbb{R}^3 \setminus B_r} a(wx+b)_+ \rho^*_\beta(a,w,b) \mathrm{d}a\mathrm{d}w\mathrm{d}b \right| \leq \int_{\mathbb{R}^3 \setminus B_r} (|aw| \cdot |x| + |ab|) \rho^*_\beta(a,w,b) \mathrm{d}a\mathrm{d}w\mathrm{d}b$$
$$\leq C \int_{\mathbb{R}^3 \setminus B_r} (a^2 + b^2 + w^2) \rho^*_\beta(a,w,b) \mathrm{d}a\mathrm{d}w\mathrm{d}b,$$

where $C > 0$ is a constant independent of (β, r). Since the sequence $\{\|X_\beta\|_2^2\}_\beta$ is uniformly integrable, we can make the RHS arbitrary small by picking a sufficiently large r (uniformly for all β). As a result, (6.12) readily follows. Note that (6.12) immediately implies that, as $\beta \to \infty$, $R(\rho^*_\beta) \to R(\rho^*) = 0$. Furthermore, with similar arguments we obtain that, as $\beta \to \infty$, $M(\rho^*_\beta) \to M(\rho^*)$. By convexity of the differential entropy, we have that $H(\frac{1}{2}\rho_1 + \frac{1}{2}\rho_2) \geq \frac{1}{2}H(\rho_1) + \frac{1}{2}H(\rho_2)$. Hence, $H(\rho^*_\beta) \geq C \log(2\pi e/\beta)$, where $C > 0$ is independent of β. By combining these bounds on $R(\rho^*_\beta)$, $M(\rho^*_\beta)$ and $H(\rho^*_\beta)$, we conclude that

$$\limsup_{\beta \to \infty} \mathcal{F}_\beta(\rho^*_\beta) \leq \mathcal{F}_\infty(\rho^*),$$

which, combined with (6.4), completes the proof of (6.13).

Proof of *(ii)*. From (6.14), we obtain that $\lim_{\beta\to\infty} R(\rho_\beta) = R(\rho)$. As the second moment is lower-semicontinuous and bounded from below, we have that $\liminf_{\beta\to\infty} M(\rho_\beta) \geq M(\rho)$. Furthermore, Lemma 10.2 in Mei et al. (2018) implies that

$$\mathcal{F}_\beta(\rho_\beta) \geq \frac{1}{2}R(\rho_\beta) + \frac{\lambda}{4}M(\rho_\beta) - \beta^{-1}(1 + 3\log 8\pi) + \beta^{-1}\log(\beta\lambda).$$

By combining these bounds, we have that

$$\liminf_{\beta \to \infty} \mathcal{F}_\beta(\rho_\beta) \geq \frac{1}{2}R(\rho) + \frac{\lambda}{4}M(\rho). \tag{6.17}$$

By replicating the argument leading to (6.10) (but now with regularization coefficient $\lambda/2$ instead of λ), we obtain that the RHS of (6.17) can be lower bounded as

$$\frac{1}{2}R(\rho) + \frac{\lambda}{4}M(\rho) \geq \lambda - \frac{\lambda^2}{8} \geq \frac{7\lambda}{8} > \frac{\lambda}{5}, \tag{6.18}$$

for all $\lambda \leq 1$. Then, the desired result follows from (6.17) and (6.18).

7. Numerical Simulations

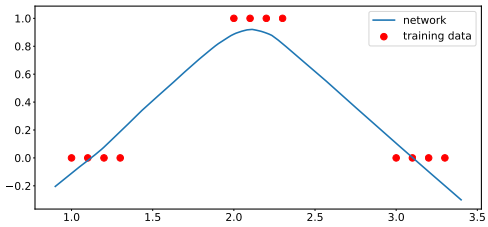
We consider training the two-layer neural network (3.1) with N neurons and ReLU activation functions, i.e., $\sigma^*(x, \boldsymbol{\theta}) = a(wx+b)_+$, with $\boldsymbol{\theta} = (a, w, b)$. We run the SGD iteration (3.3) (no momentum or weight decay, batch size equal to 1), and we plot the resulting predictor once the algorithm has converged. The results for two different unidimensional datasets are reported in Figures 7 and 8. In these experiments, we set $N = 500$ and we remark that the plots for wider networks ($N \in \{1000, 2000, 5000\}$) look identical. We also point out that the shape of the predictor does not change for different runs of the SGD algorithm (with different initializations, and order of the training samples). This is in agreement with the mean-field predictions when $\beta < \infty$, $\lambda > 0$ and the variance of the initialization does not depend on N. The same setup is employed to obtain the numerical results of Figure 1 and 6b, discussed in Section 1 and 6, respectively.

In Figure 7, we plot the shape of the function learnt by the network for different values of the temperature parameter β^{-1}. The learning rate is $s_k = 1$, the total number of training epochs required for SGD to converge is roughly 5×10^4, and no ℓ_2 regularization is enforced ($\lambda = 0$). As predicted by our theoretical findings, the predictor approaches a piecewise linear function whose number of tangent changes (or knots) is proportional to the number of training samples (and not to the width of the network): if $\beta^{-1} = 0.005$, the predictor is still rather smooth; if $\beta^{-1} = 10^{-4}$, the predictor sharpens, except for a smoother tangent change in the interval $[4, 5]$; and finally if $\beta = 0$, the predictor is piecewise linear. Let us highlight that the knots sometimes do not coincide with the training data points, as suggested by the results of Section 4 and demonstrated in the example of Section 6.

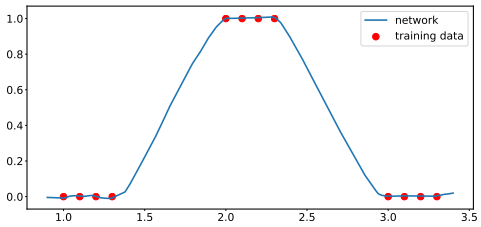
In Figure 8, we consider another dataset and plot the neural network predictor for four different pairs of (β^{-1}, λ). By comparing (a) with (b) and with the bottom plots (c)-(d), it is clear that the solution becomes increasingly piecewise linear as the noise decreases. Furthermore, the effect of regularization can be noticed by comparing plots (a)-(c) on the left with plots (b)-(d) on the right: adding an ℓ_2 penalty implies that the network does not fit the data and therefore the location of the knots changes.

8. Comparison with Related Work

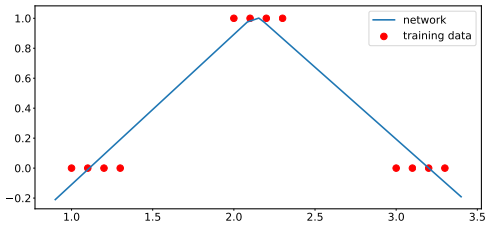
The line of works (Savarese et al., 2019; Ergen and Pilanci, 2021; Ongie et al., 2020; Parhi and Nowak, 2020a) studies the properties of the minimizers of certain optimization objectives, and therefore these results are not directly connected to the dynamics of gradient descent algorithms. On the contrary, the goal of this paper is to understand the implicit bias due to gradient descent, namely, to characterize the structure of the neural network predictor once the algorithm has converged. Another important difference lies in the fact



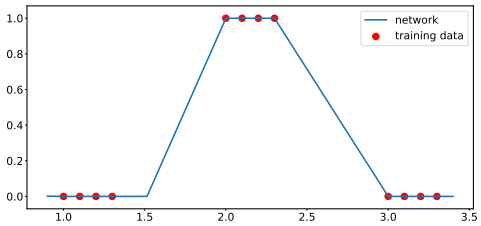
(a) $\beta^{-1} = 0.01,\ \lambda = 0.003$



(b) $\beta^{-1} = 0.001,\ \lambda = 0$



(c) $\beta^{-1} = 0,\ \lambda = 0.003$



(d) $\beta^{-1} = 0,\ \lambda = 0$

Figure 8: Functions learnt by a two-layer ReLU network with $N = 500$ neurons, for different values of the temperature parameter β^{-1} and the regularization coefficient λ.

that our ℓ_2 regularization involves all the parameters, including the bias b, while existing work does not regularize the biases of the network. This fact may lead to the qualitatively different behavior unveiled by our study. Going into detail, Ergen and Pilanci (2021) show that the network that minimizes a regularized objective implements a linear spline. In contrast, our analysis suggests that the knots (i.e., abrupt changes in the tangent of the predictor) can occur at points different from the training samples. Let us also mention that Savarese et al. (2019) and Ongie et al. (2020) give an explicit form of the functional regularizer of the neural network solution, but it is not clear how to characterize the function class to which the solution belongs, e.g., whether the function implemented by the neural network is a cubic or linear spline. Furthermore, the upper bound on the number of knot points appearing in (Parhi and Nowak, 2020a) depends on the null space of a certain operator, and computing the dimension of this null space explicitly appears to be difficult.

The work by Williams et al. (2019) considers a noiseless setting with no regularization, and it studies the properties of gradient flow on the space of reduced parameters. In particular, the initial ReLU neurons depending on three parameters (a, b and w, in our notation) are mapped to a two-dimensional space, where each neuron is defined by its magnitude and angle. Then, it is proven that the Wasserstein gradient flow on this reduced space drives the activation points of the ReLU neurons to the training data. As a consequence, the solution found by SGD is piecewise linear and the knot points are located at a subset of the training samples. Blanc et al. (2020) consider SGD with label noise and no regularization, and show that, once the squared loss is close to zero, the algorithm minimizes an auxiliary quantity, i.e., the sum of the squared norms of the gradients evaluated at each training point. By

instantiating this result in the case of a two-layer ReLU network with a skip connection, the authors show that the solution found by SGD is piecewise linear with the minimum amount of knots required to fit the data.

While our result shares some similarities with (Williams et al., 2019) and (Blanc et al., 2020), let us highlight some crucial differences. First, we note that Blanc et al. (2020) consider a two-layer network with a skip connection which fits the training data perfectly. In contrast, our two-layer model is standard (no skip connections) and the analysis does not require a perfect fit of the data, as we allow for non-vanishing ℓ_2 regularization. Furthermore, even when the regularization term is vanishing, our characterization does not lead to the minimum number of knots required to fit the data (as in Blanc et al., 2020), and the knots are not necessarily located at the training points (as in Williams et al., 2019). In fact, our theoretical results suggest the presence of additional knot points, a feature that is confirmed in numerical simulations. The novel behavior that we unveil appears to be due to the differences in the setting and to the addition of (a possibly vanishing) ℓ_2 regularization term in the optimization. Concerning the proof techniques, the work by Blanc et al. (2020) exploits an Ornstein-Uhlenbeck like analysis, while this work tackles the increasingly popular mean-field regime. Our key technical contribution is to analyze the Gibbs minimizer of a certain free energy, while Williams et al. (2019) consider the gradient flow on reduced parameters and connect it to the flow on the full parameters via a specific type of initialization. Our analysis directly establishes a result for the full parameters, and it requires mild technical assumptions on the initialization. Finally, let us point out that it is an open problem to extend the approach of Williams et al. (2019) to a regularized objective, because of the non-injectivity of the mapping to the canonical parameters.

9. Concluding Remarks

We develop a new technique to characterize the implicit bias of gradient descent methods used to train overparameterized neural networks. In particular, we consider training a wide two-layer ReLU network via SGD for a univariate regression task and, by taking a mean field view, we show that the predictor obtained at convergence has a simple piecewise linear form. Our results hold in the regime of vanishingly small noise added to the SGD gradients, and handle both constant and vanishing ℓ_2 regularization. The analysis leads to an exact characterization of the number and location of the tangent changes (or knots) in the predictor: on each interval between consecutive training inputs, the number of knots is at most three. To obtain the desired result, we relate the distribution of the weights of the network once SGD has converged to the minimizer of a certain free energy. Then, we prove that the curvature of the predictor resulting from this minimizer vanishes everywhere except in a *cluster set*, which concentrates on at most three points per prediction interval. This novel strategy opens the way to several interesting directions. We discuss them below.

We focus on ReLU networks. However, only the following two properties of the activation appear to be crucial for the analysis: *(i)* its second derivative behaves like a Dirac delta, and *(ii)* its growth is at most linear. In fact, the first property reduces the computation of the curvature to an integral over a lower-dimensional subspace; and the second property leads to a uniform bound on the second moment of the network parameters. Hence, our approach may be extendable to a more general class of piecewise linear activations, although

this would come at the cost of a more intricate structure for the cluster set containing the location of the tangent changes.

We focus on univariate regression. The natural ordering on one-dimensional features allows for a convenient characterization of the activation regions that correspond to each input conditioned on the sign of w. For larger input dimension, such a characterization appears to be cumbersome, as the structure of these regions is induced by the intersection of hyperplanes. Furthermore, in the setting considered in this work, the cluster set is the union of intervals where certain second-degree polynomials are non-positive. For multivariate regression, we expect the cluster set to be connected to the non-positive set of quadratic forms. Hence, the structure of the cluster set may be highly non-linear, and its concentration can occur on subspaces which are hard to define explicitly.

We provide an upper bound on the number of tangent changes of the predictor. The numerical simulations of Section 6 suggest that one and two knots between consecutive training inputs can occur. Showing whether our theoretical bound of three knots is tight by providing an explicit example, or by proving a tighter bound of two, is an open question for possible future work. We also remark that, given the errors R_i of the neural network estimator at the data points, one can deduce the location of the knot points. Such implicit characterization is similar in spirit to the attractive/repulsive condition on the training points of Williams et al. (2019).

In conclusion, in this work we demonstrate how to exploit the Gibbs form of the minimizer in order to accurately characterize a functional property of the predictor learnt by the neural network using limiting arguments of the training process. The general spirit of this technique could potentially be informative in additional ways. For instance, utilizing the properties of the Gibbs distribution reached at convergence may be of additional interest for future study. We conjecture that this could yield insight into the stability of the predictor with respect to perturbations in the training data at *finite* temperature β.

Acknowledgements

We would like to thank Mert Pilanci for several exploratory discussions in the early stage of the project, Jan Maas for clarifications about Jordan et al. (1998), and Max Zimmer for suggestive numerical experiments. A. Shevchenko and M. Mondelli are partially supported by the 2019 Lopez-Loreta Prize. V. Kungurtsev acknowledges support to the OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765 Research Center for Informatics.

References

Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*, 2019.

Randall Balestriero and Richard G. Baraniuk. A spline theory of deep learning. In *International Conference on Machine Learning*, pages 374–383. PMLR, 2018.

Peter L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998. doi: 10.1109/18.661502.

Peter L. Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: A statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL `https://www.pnas.org/content/116/32/15849`.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Conference on Learning Theory*, pages 483–513. PMLR, 2020.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In *International Joint Conference on Artificial Intelligence*, 2021.

Lénaïc Chizat. Mean-field langevin dynamics: Exponential convergence and annealing. *arXiv preprint arXiv:2202.01009*, 2022.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 32, 2018.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 33, 2019.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Valentin De Bortoli, Alain Durmus, Xavier Fontaine, and Umut Simsekli. Quantitative propagation of chaos for SGD in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2020.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, pages 1309–1318. PMLR, 2018.

Tolga Ergen and Mert Pilanci. Convex geometry and duality of over-parameterized neural networks. *The Journal of Machine Learning Research*, 22, 2021.

Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on Learning Theory*, pages 1887–1936. PMLR, 2021.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Véronique Gayrard, Anton Bovier, Michael Eckhoff, and Markus Klein. Metastability in reversible diffusion processes I: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 57, pages 2043–2065. Institut Henri Poincaré, 2021.

Tien-Chung Hu and Andrew Rosalsky. A note on the de La Vallée Poussin criterion for uniform integrability. *Statistics & Probability Letters*, 81(1):169–174, 2011.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *The Annals of Statistics*, 48(6):3619–3642, 2020.

Hui Jin and Guido Montúfar. Implicit bias of gradient descent for mean squared error regression with wide neural networks. *arXiv preprint arXiv:2006.07356*, 2020.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Sanjeev Arora, and Rong Ge. Explaining landscape connectivity of low-cost solutions for multilayer nets. In *Advances in Neural Information Processing Systems*, volume 33, 2019.

Philippe Laurençot. Weak compactness techniques and coagulation equations. In *Evolutionary Equations with Applications in Natural Sciences*, pages 199–253. Springer, 2015.

Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes ReLU network features. *arXiv preprint arXiv:1803.08367*, 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *Workshop Contribution at International Conference on Learning Representations*, 2015.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.

Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv preprint arXiv:2001.11443*, 2020.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. *arXiv preprint arXiv:2201.10469*, 2022.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *International Conference on Learning Representations*, 2020.

Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *The Journal of Machine Learning Research*, 22, 2020a.

Rahul Parhi and Robert D Nowak. The role of neural network activation functions. *IEEE Signal Processing Letters*, 27:1779–1783, 2020b.

Grant M. Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. In *Advances in Neural Information Processing Systems*, volume 32, 2018.

Justin Sahs, Ryan Pyle, Aneel Damaraju, Josue Ortega Caro, Onur Tavaslioglu, Andy Lu, and Ankit Patel. Shallow univariate ReLU networks as splines: initialization, loss surface, hessian, & gradient flow dynamics. *arXiv preprint arXiv:2008.01772*, 2020.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

Alexander Shevchenko and Marco Mondelli. Landscape connectivity and dropout stability of sgd solutions for over-parameterized neural networks. In *International Conference on Machine Learning*, pages 8773–8784. PMLR, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under- to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001, Oct. 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Ulrike Von Luxburg and Bernhard Schölkopf. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, pages 651–706. Elsevier, 2011.

Francis Williams, Matthew Trager, Claudio Silva, Daniele Panozzo, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate ReLU networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. In *International Conference on Learning Representations*, 2021.

Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. A type of generalization error induced by initialization in deep neural networks. In *Mathematical and Scientific Machine Learning*, pages 144–164. PMLR, 2020.

Appendix A. Technical Results

In this appendix, we prove a few technical results which are used in the arguments of Section 5.2. More specifically, in Section A.1 we show that, as $\tau \to \infty$, the minimizer $\rho^*_{\tau,m}(\boldsymbol{\theta})$ of the free energy $\mathcal{F}^{\tau,m}$ converges pointwise in $\boldsymbol{\theta}$ to the minimizer $\rho^*_m(\boldsymbol{\theta})$ of the free energy $\mathcal{F}^m$. This pointwise convergence is needed to establish the result of Lemma 5.1. In Section A.2, we derive upper bounds on the risk of the minimizer (used in Lemma 5.3) and on its second moment (which implies that the sequence of predictors is equi-Lipschitz), and we also prove the lower bound on the partition function in Lemma 5.6. Finally, in Section A.3

we give the proof of Lemma 5.4, which lower bounds the growth of the polynomials f^j and f_j.

A.1 Convergence of Minimizers

Lemma A.1 (Convergence of densities)**.** *Let $\{\rho_n\}_n$ be a sequence of densities in $\mathcal{K}$ with uniformly bounded truncated entropy, that is*

$$\int \max\{\rho_n(\boldsymbol{\theta})\log\rho_n(\boldsymbol{\theta}), 0\}\,\mathrm{d}\boldsymbol{\theta} \leq C, \quad \forall n,$$

for some $C > 0$ that is independent of n, and uniformly bounded second moment, i.e., $M(\rho_n) \leq C$ for all n. Then, there exists a subsequence $\{\rho_{n'}\}_{n'}$ of $\{\rho_n\}_n$ and $\rho \in \mathcal{K}$ such that $\rho_{n'} \rightharpoonup \rho$ and

$$C \geq \liminf_{n'\to\infty} M(\rho_{n'}) \geq M(\rho) \geq 0.$$

Proof of Lemma A.1. Since $z \mapsto \max\{z\log z, 0\}$, $z \in [0, +\infty)$, has super-linear growth, this result in conjunction with the de la Vallée Poussin criterion (see for instance Hu and Rosalsky (2011)) guarantees that the sequence of densities $\{\rho_n\}_n$ is uniformly integrable. By Dunford-Pettis Theorem (for σ-finite measure spaces, see for instance Laurençot (2015)), relative weak compactness in L_1 is equivalent to uniform integrability. Hence, there exists a density ρ and a subsequence $\{\rho_{n'}\}_{n'}$ of $\{\rho_n\}_n$ such that $\rho_{n'} \rightharpoonup \rho$.

As $M(\cdot)$ is lower-semicontinuous with respect to the topology of weak convergence in L_1 and bounded from below, we have that $\liminf_{n'\to\infty} M(\rho_{n'}) \geq M(\rho)$. Furthermore, as $M(\rho_n) \leq C$, we get that $M(\rho) \leq C$ and, thus, $\rho \in \mathcal{K}$. ■

Lemma A.2 (Uniformly bounded $M(\rho^*_{\tau,m})$ and limit of $\rho^*_{\tau,m}$)**.** *Assume that condition **A1** holds. Consider the sequence of minimizing Gibbs distributions $\{\rho^*_{\tau,m}\}_\tau$. The following results hold:*

1. *$M(\rho^*_{\tau,m})$ is uniformly bounded in (τ, m). Moreover, if $\beta\lambda > 1$,*

$$M(\rho^*_m),\ M(\rho^*_{\tau,m}) \leq \frac{C_3}{\lambda}, \quad \forall\tau \in (0, +\infty),$$

   where $C_3 > 0$ is independent of $(\tau, m, \beta, \lambda)$.

2. *Given any m consistent with **A1**, there exists $\rho_m \in \mathcal{K}$ and a subsequence $\{\rho^*_{\tau',m}\}_{\tau'}$ (which with an abuse of notation we identify with $\{\rho^*_{\tau,m}\}_\tau$) such that $\rho^*_{\tau,m} \rightharpoonup \rho_m$ as $\tau \to \infty$.*

3. *Given any m consistent with **A1**, $\lim_{\tau\to\infty} R^{\tau,m}_i(\rho^*_{\tau,m}) = R^m_i(\rho_m)$ for all $i \in [M]$, and $\liminf_{\tau\to\infty} \mathcal{F}^{\tau,m}(\rho^*_{\tau,m}) \geq \mathcal{F}^m(\rho_m)$.*

Proof of Lemma A.2. We provide the proof of the first result for $\rho^*_{\tau,m}$. The arguments for ρ^*_m are the same after changing the notation from $\rho^*_{\tau,m}$ to ρ^*_m. Let $\rho = \mathcal{N}(0, \mathbb{I}_{3\times3})$. Then, we have that

$$R^{\tau,m}(\rho) = \frac{1}{M}\sum_{i=1}^{M} y_i^2,\ M(\rho) = 3,\ H(\rho) = \frac{3}{2}\ln(2\pi e). \tag{A.1}$$

Note that for this ρ, $R^{\tau,m}(\rho)$, in fact, does not depend on $(\tau, m, \beta, \lambda)$.

From Lemma 10.2 in Mei et al. (2018), since $\rho^*_{\tau,m}$ is the unique minimizer of the free energy $\mathcal{F}^{\tau,m}$, we have that the following inequalities hold

$$\mathcal{F}^{\tau,m}(\rho) \geq \mathcal{F}^{\tau,m}(\rho^*_{\tau,m}) \geq R^{\tau,m}(\rho^*_{\tau,m}) + \lambda/4 \cdot M(\rho^*_{\tau,m}) - 1/\beta \cdot [1 + 3 \cdot \log(8\pi/(\beta\lambda))]. \quad \text{(A.2)}$$

Furthermore, by using (A.1) and the fact that $\beta > C_1$ and $\lambda < C_2$, we obtain

$$\mathcal{F}^{\tau,m}(\rho) \leq K_1 + K_1\lambda - \beta^{-1}K_1 \leq K_2, \quad \text{(A.3)}$$

for some $K_1, K_2 > 0$ that are independent of $(\tau, m, \beta, \lambda)$. By combining (A.3) and (A.2) and using that $R^{\tau,m}(\rho^*_m) \geq 0$, we conclude that

$$\lambda \cdot M(\rho^*_{\tau,m}) \leq K_3 + 1/\beta \cdot [1 + 3 \cdot \log(8\pi/(\beta\lambda))],$$

where $K_3 > 0$ is independent of $(\tau, m, \beta, \lambda)$. As $\beta\lambda > 1$, the first claim immediately follows.

Since the activation and the labels are uniformly bounded in τ and $\{i\}_{i\in[M]}$ is finite, $|R^{\tau,m}_i(\rho^*_{\tau,m})|$ is uniformly bounded in (τ, i). Hence, the following lower bound on the partition function $Z_{\tau,m}(\beta, \lambda)$ holds

$$\begin{aligned} Z_{\tau,m}(\beta,\lambda) &= \int \exp\left\{-\beta\left[\sum_{i=1}^{M} R^{\tau,m}_i(\rho^*_{\tau,m}) \cdot a^{\tau,m}(w^{\tau,m}x_i + b)^m_\tau + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right]\right\} \mathrm{d}\boldsymbol{\theta} \\ &\geq \int \exp\left\{-\beta\left[\sum_{i=1}^{M} |R^{\tau,m}_i(\rho^*_{\tau,m})| \cdot 2m^3 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right]\right\} \mathrm{d}\boldsymbol{\theta} \\ &\geq K_4 \int \exp\left\{-\frac{\beta\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right\} \mathrm{d}\boldsymbol{\theta} = \frac{K_5}{\sqrt{\beta^3\lambda^3}} \geq K_6, \end{aligned} \quad \text{(A.4)}$$

for some $K_4, K_5, K_6 > 0$ independent of τ (but dependent on (m, β, λ)). In the same way, one can upper bound $\rho^*_{\tau,m} \cdot Z_{\tau,m}(\beta, \lambda)$ as

$$\exp\left\{-\beta\left[\sum_{i=1}^{M} R^{\tau,m}_i(\rho^*_{\tau,m}) \cdot a^{\tau,m}(w^{\tau,m}x_i + b)^m_\tau + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right]\right\} \leq K_7 \exp\left\{-\frac{\beta\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right\}, \quad \text{(A.5)}$$

where $K_7 > 0$ is independent of τ (but dependent on (m, β, λ)). Notice that we can increase K_7 to be arbitrarily large and still satisfy (A.5), and in particular, increase it to satisfy $K_7/K_6 > 1$. Thus, by combining (A.4) and (A.5), we get

$$\begin{aligned} &\int \max\{\rho^*_{\tau,m}(\boldsymbol{\theta}) \ln \rho^*_{\tau,m}(\boldsymbol{\theta}), 0\} \mathrm{d}\boldsymbol{\theta} \\ &\leq \int \max\left\{\frac{K_7}{K_6} \exp\left\{-\frac{\beta\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right\} \cdot \left(\ln\frac{K_7}{K_6} - \frac{\beta\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right), 0\right\} \mathrm{d}\boldsymbol{\theta} \\ &= \int_\Omega \frac{K_7}{K_6} \exp\left\{-\frac{\beta\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right\} \cdot \left(\ln\frac{K_7}{K_6} - \frac{\beta\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right) \mathrm{d}\boldsymbol{\theta} \leq \int_\Omega \frac{K_7}{K_6} \ln\frac{K_7}{K_6} \mathrm{d}\boldsymbol{\theta}, \end{aligned}$$

where

$$\Omega = \left\{\boldsymbol{\theta} \in \mathbb{R}^3 : \|\boldsymbol{\theta}\|_2^2 \leq \ln\left(\frac{K_7}{K_6}\right)\frac{2}{\beta\lambda}\right\}.$$

Since $\mathrm{vol}(\Omega) < K_8$ for some $K_8 \geq 0$ independent of τ, we get that

$$\int \max\{\rho^*_{\tau,m}(\boldsymbol{\theta}) \ln \rho^*_{\tau,m}(\boldsymbol{\theta}), 0\} \mathrm{d}\boldsymbol{\theta} \leq K_8 \cdot \frac{K_7}{K_6} \cdot \ln \frac{K_7}{K_6},$$

where the RHS is independent of τ. As $M(\rho^*_{\tau,m})$ is uniformly bounded in τ, we can invoke Lemma A.1 to finish the proof of the second statement.

We now prove the third statement. By the triangle inequality, we have that, for all $i \in [M]$,

$$\begin{aligned}
&\lim_{\tau\to\infty} \left| \int a^{\tau,m}(w^{\tau,m}x_i + b)^m_\tau \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - \int a^m (w^m x_i + b)^m_+ \rho_m(\mathrm{d}\boldsymbol{\theta}) \right| \\
\leq &\lim_{\tau\to\infty} \left| \int a^{\tau,m}(w^{\tau,m}x_i + b)^m_\tau \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - \int a^m (w^m x_i + b)^m_+ \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \right| \\
&+ \lim_{\tau\to\infty} \left| \int a^m (w^m x_i + b)^m_+ \rho^*_{\tau,m}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - \int a^m (w^m x_i + b)^m_+ \rho_m(\mathrm{d}\boldsymbol{\theta}) \right| := A_1 + A_2.
\end{aligned}$$

By upper bounding $\rho^*_{\tau,m}$ as in (A.4)-(A.5), we have

$$A_1 \leq K_9 \lim_{\tau\to\infty} \int |a^{\tau,m}(w^{\tau,m}x_i + b)^m_\tau - a^m (w^m x_i + b)^m_+| \exp\left\{ -\frac{\beta\lambda}{2} \|\boldsymbol{\theta}\|_2^2 \right\} \mathrm{d}\boldsymbol{\theta},$$

where $K_9 > 0$ is independent of τ. Thus, an application of the Dominated Convergence theorem gives that the term A_1 vanishes. Furthermore, the term A_2 vanishes by weak convergence of $\rho^*_{\tau,m}$ to ρ_m. This proves that, as $\tau \to \infty$, $y^{\sigma^*}_{\rho^*_{\tau,m}}(x_i) \to y^{\sigma^*}_{\rho_m}(x_i)$ and so $R^{\tau,m}_i(\rho^*_{\tau,m}) \to R^m_i(\rho_m)$.

Note that $-H(\cdot)$ and $M(\cdot)$ are lower-semicontinuous in $\mathcal{K}$. Furthermore, $M(\cdot)$ is lower bounded and $-H(\cdot)$ is lower bounded by Lemma 10.1 in Mei et al. (2018) on the subsequence $\{\rho^*_{\tau,m}\}_\tau$, as $M(\rho^*_{\tau,m})$ is uniformly bounded in τ. Hence, as $\rho^*_{\tau,m}$ converges weakly to $\rho_m \in \mathcal{K}$, we conclude that

$$\liminf_{\tau\to\infty} -H(\rho^*_{\tau,m}) \geq -H(\rho_m), \quad \liminf_{\tau\to\infty} M(\rho^*_{\tau,m}) \geq M(\rho_m),$$

which, combined with $R^{\tau,m}_i(\rho^*_{\tau,m}) \to R^m_i(\rho_m)$, implies the desired result. ■

Lemma A.3 (Pointwise convergence of free-energies)**.** *Fix some distribution $\rho \in \mathcal{K}$, then we have the following pointwise convergence:*

$$\lim_{\tau\to\infty} \mathcal{F}^{\tau,m}(\rho) = \mathcal{F}^m(\rho).$$

Proof of Lemma A.3. By construction, we have that $(x)^m_\tau$ converges to $(x)^m_+$, for all $x \in \mathbb{R}$. It is clear that

$$|a^{\tau,m}|(w^{\tau,m}x + b)^m_\tau \rho(\boldsymbol{\theta}) \leq 2m^3 \rho(\boldsymbol{\theta}),$$

and the RHS is integrable. Thus, an application of the Dominated Convergence theorem gives that

$$\lim_{\tau\to\infty} R^{\tau,m}(\rho) = R^m(\rho).$$

This concludes the proof since $M(\rho)$ and $H(\rho)$ are independent of τ. ■

Lemma A.4 (Pointwise convergence of minimizers)**.** *Assume that condition **A1** holds and consider any satisfactory m. Then, as $\tau \to \infty$, the minimizer $\rho^*_{\tau,m}$ of the free energy $\mathcal{F}^{\tau,m}$ converges pointwise in $\boldsymbol{\theta}$ to the minimizer ρ^*_m of the free energy $\mathcal{F}^m$, i.e.,*

$$\lim_{\tau\to\infty} \rho^*_{\tau,m}(\boldsymbol{\theta}) = \rho^*_m(\boldsymbol{\theta}), \qquad \forall\, \boldsymbol{\theta} \in \mathbb{R}^3.$$

Proof of Lemma A.4. From Lemma A.2, we have that there exists a subsequence $\{\rho^*_{\tau,m} \in \mathcal{K}\}$ and $\rho_m \in \mathcal{K}$ such that the following holds

$$\liminf_{\tau\to\infty} \mathcal{F}^{\tau,m}(\rho^*_{\tau,m}) \geq \mathcal{F}^m(\rho_m). \tag{A.6}$$

Since $\rho^*_{\tau,m} \in \mathcal{K}$ minimizes $\mathcal{F}^{\tau,m}$, we have

$$\mathcal{F}^{\tau,m}(\rho^*_{\tau,m}) \leq \mathcal{F}^{\tau,m}(\rho^*_m).$$

By taking the liminf on both sides, using Lemma A.3 and (A.6), we have

$$\mathcal{F}^m(\rho_m) \leq \liminf_{\tau\to\infty} \mathcal{F}^{\tau,m}(\rho^*_{\tau,m}) \leq \liminf_{\tau\to\infty} \mathcal{F}^{\tau,m}(\rho^*_m) = \mathcal{F}^m(\rho^*_m).$$

Since ρ^*_m is the unique minimizer of $\mathcal{F}^m$ (see Lemma 10.2 of Mei et al. (2018)), ρ^*_m and ρ_m coincide almost everywhere, which implies that

$$R^m_i(\rho_m) = R^m_i(\rho^*_m).$$

Hence, by Lemma A.2, we have that

$$\lim_{\tau\to\infty} R^{\tau,m}_i(\rho^*_{\tau,m}) = R^m_i(\rho^*_m).$$

Recall that, by construction, for any parameter $v \in \mathbb{R}$, the τ-smooth m-truncation $v^{\tau,m}$ converges to v^m as $\tau \to \infty$. Furthermore, as $\tau \to \infty$, the smooth m-truncation $(\cdot)^m_\tau$ of the softplus activation converges pointwise to the smooth m-truncation $(\cdot)^m_+$ of the ReLU activation. Thus,

$$\lim_{\tau\to\infty} \Psi_\tau(\boldsymbol{\theta}) = \lim_{\tau\to\infty} \sum_{i=1}^M R^{\tau,m}_i(\rho^*_{\tau,m}) \cdot a^{\tau,m}(w^{\tau,m}x_i + b)^m_\tau = \sum_{i=1}^M R^m_i(\rho^*_m) \cdot a^m(w^m x_i + b)^m_+ = \Psi(\boldsymbol{\theta}),$$

where the convergence is intended to be pointwise in $\boldsymbol{\theta}$. Note that $\Psi_\tau(\boldsymbol{\theta})$ is uniformly bounded in τ, hence

$$\lim_{\tau\to\infty} \exp\left\{-\beta\Psi_\tau(\boldsymbol{\theta}) - \frac{\beta\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right\} = \exp\left\{-\beta\Psi(\boldsymbol{\theta}) - \frac{\beta\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right\},$$

which implies that $Z_{\tau,m}\rho^*_{\tau,m}(\boldsymbol{\theta})$ converges pointwise to $Z_m\rho^*_m(\boldsymbol{\theta})$. Furthermore, as $\tau \to \infty$, $Z_{\tau,m}$ converges to Z_m by Dominated Convergence, which concludes the proof. ■

A.2 Bounds on Risk of Minimizer, Second Moment and Partition Function

Lemma A.5 (Bound on risk of the minimizer)**.** *Assume that condition **A1** holds. Then,*

$$R^m(\rho_m^*) \leq C\lambda,$$

where $C > 0$ is a constant independent of (m, β, λ). In addition, for any $\varepsilon > 0$, there exists $\bar{\tau}(\varepsilon, m, \beta, \lambda)$ such that for any $\tau > \bar{\tau}(\varepsilon, m, \beta, \lambda)$ we have

$$R^{\tau,m}(\rho_{\tau,m}^*) \leq C\lambda + \varepsilon.$$

Proof of Lemma A.5. Consider a "saw-tooth" function centered at x_i with height y_i and width $\varepsilon > 0$, namely,

$$\mathrm{ST}_{x_i,y_i}(x) := \begin{cases} 0, & x < x_i - \varepsilon \text{ or } x > x_i + \varepsilon, \\ \frac{y_i}{\varepsilon}(x - x_i + \varepsilon), & x_i - \varepsilon \leq x \leq x_i, \\ \frac{y_i}{\varepsilon}(x_i - x + \varepsilon), & x_i < x \leq x_i + \varepsilon, \end{cases}$$

Notice that this function can be implemented by the following $\hat{\rho}_i$:

$$\hat{\rho}_i = \frac{1}{3}\left(\delta_{\left(\frac{3y_i}{\varepsilon}, 1, \varepsilon - x_i\right)} + \delta_{\left(-\frac{6y_i}{\varepsilon}, 1, -x_i\right)} + \delta_{\left(\frac{3y_i}{\varepsilon}, 1, -\varepsilon - x_i\right)}\right),$$

in the sense that

$$\mathrm{ST}_{x_i,y_i}(x) = \int a(wx + b)_+ \hat{\rho}_i(\mathrm{d}\boldsymbol{\theta}),$$

where $\delta_{\boldsymbol{\theta}}$ stands for a delta distribution centered at the point $\boldsymbol{\theta} = (a, w, b) \in \mathbb{R}^3$. Let us pick ε such that $\varepsilon < \min_{i \in [M-1]} \{|x_i - x_{i+1}|/2\}$. This condition on ε guarantees that

$$\left\{x \in \mathbb{R} : \int a(wx + b)_+ \hat{\rho}_i(\mathrm{d}\boldsymbol{\theta}) \neq 0\right\} \cap \left\{x \in \mathbb{R} : \int a(wx + b)_+ \hat{\rho}_j(\mathrm{d}\boldsymbol{\theta}) \neq 0\right\} = \emptyset, \ \forall i \neq j,$$

which ensures that the "saw-tooth" functions are not intersecting. Define

$$\hat{\rho} = \frac{1}{3M} \sum_{i=1}^{M} \left[\delta_{\left(\frac{3My_i}{\varepsilon}, 1, \varepsilon - x_i\right)} + \delta_{\left(-\frac{6My_i}{\varepsilon}, 1, -x_i\right)} + \delta_{\left(\frac{3My_i}{\varepsilon}, 1, -\varepsilon - x_i\right)}\right].$$

Then, one immediately has that, for all $i \in [M]$,

$$\int a(wx_i + b)_+ \hat{\rho}(\mathrm{d}\boldsymbol{\theta}) = y_i.$$

Furthermore, by taking a sufficiently large m, in particular, taking $m > \max_i\{6M|y_i|/\varepsilon\} + 3|x_M| + 3|x_1| + 2$ suffices, we get that, for all $x \in [x_1, x_M]$,

$$\int a^m (w^m x + b)_+^m \hat{\rho}(\mathrm{d}\boldsymbol{\theta}) = \int a(wx + b)_+ \hat{\rho}(\mathrm{d}\boldsymbol{\theta}),$$

which implies that $R^m(\hat{\rho}) = 0$.

Let $\mathcal{N}(\mu, \sigma^2)$ denote a Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}$, and let $U(\mu, \sigma^2)$ denote the uniform distribution with mean μ and variance $\sigma^2/12$. Given $(\mu_1, \mu_2, \mu_3) \in \mathbb{R}^3$ and $\sigma^2 \in \mathbb{R}$, let $\rho_{((\mu_1,\mu_2,\mu_3),\sigma^2)}$ denote the following product distribution

$$\rho_{((\mu_1,\mu_2,\mu_3),\sigma^2)} := U(\mu_1, \sigma^2) \times \mathcal{N}(\mu_2, \sigma^2) \times \mathcal{N}(\mu_3, \sigma^2),$$

and define

$$\tilde{\rho} = \frac{1}{3M} \sum_{i=1}^{M} \left[\rho_{\left(\left(\frac{3My_i}{\varepsilon}, 1, \varepsilon - x_i\right), \sigma^2\right)} + \rho_{\left(\left(-\frac{6My_i}{\varepsilon}, 1, -x_i\right), \sigma^2\right)} + \rho_{\left(\left(\frac{3My_i}{\varepsilon}, 1, -\varepsilon - x_i\right), \sigma^2\right)} \right]. \tag{A.7}$$

Note that, for $\sigma^2 < 1$ and m chosen sufficiently large as mentioned previously,

$$\int a^m (w^m x + b)_+^m \tilde{\rho}(\mathrm{d}\boldsymbol{\theta}) = \int a (w^m x + b)_+^m \tilde{\rho}(\mathrm{d}\boldsymbol{\theta}).$$

Thus, by computing the integral w.r.t. a, we have that

$$\begin{aligned}
&\int a^m (w^m x + b)_+^m \hat{\rho}(\mathrm{d}\boldsymbol{\theta}) - \int a^m (w^m x + b)_+^m \tilde{\rho}(\mathrm{d}\boldsymbol{\theta}) \\
&= \sum_{i=1}^{M} \left[\frac{y_i}{\varepsilon} \left(\int (w^m x + b)_+^m \delta_{(1,\varepsilon - x_i)}(\mathrm{d}w\, \mathrm{d}b) - \int (w^m x + b)_+^m \rho_{((1,\varepsilon - x_i),\sigma^2)}(\mathrm{d}w\, \mathrm{d}b) \right) \right] \\
&- \sum_{i=1}^{M} \left[\frac{2y_i}{\varepsilon} \left(\int (w^m x + b)_+^m \delta_{(1,-x_i)}(\mathrm{d}w\, \mathrm{d}b) - \int (w^m x + b)_+^m \rho_{((1,-x_i),\sigma^2)}(\mathrm{d}w\, \mathrm{d}b) \right) \right] \\
&+ \sum_{i=1}^{M} \left[\frac{y_i}{\varepsilon} \left(\int (w^m x + b)_+^m \delta_{(1,-\varepsilon - x_i)}(\mathrm{d}w\, \mathrm{d}b) - \int (w^m x + b)_+^m \rho_{((1,-\varepsilon - x_i),\sigma^2)}(\mathrm{d}w\, \mathrm{d}b) \right) \right],
\end{aligned} \tag{A.8}$$

where, with an abuse of notation, we denote by $\rho_{((\mu_2,\mu_3),\sigma^2)}$ the marginal of $\rho_{((\mu_1,\mu_2,\mu_3),\sigma^2)}$ with respect to the last two components. By applying to Kantorovich-Rubinstein theorem (see, for instance, Villani (2009)), we have that

$$K \cdot W_1(p, q) = \sup_{\|f\|_{\mathrm{Lip}} \le K} |\mathbb{E}_{x \sim p} f(x) - \mathbb{E}_{y \sim q} f(y)|, \tag{A.9}$$

for two densities p and q, where W_1 is the 1-Wasserstein distance and $\|f\|_{\mathrm{Lip}}$ denotes the Lipschitz constant of f. Notice that $(w^m x + b)_+^m$ is Lipschitz in (w, b) with Lipschitz constant upper bounded by $\max(|x|, 1)$. Hence, combining (A.8) and (A.9), we have that

$$\begin{aligned}
&\left(\int a^m (w^m x + b)_+^m \hat{\rho}(\mathrm{d}\boldsymbol{\theta}) - \int a^m (w^m x + b)_+^m \tilde{\rho}(\mathrm{d}\boldsymbol{\theta}) \right)^2 \\
&\qquad \le K_1 \Bigg(\sum_{i=1}^{M} W_1(\delta_{(1,\varepsilon - x_i)}, \rho_{((1,\varepsilon - x_i),\sigma^2)}) + W_1(\delta_{(1,-x_i)}, \rho_{((1,-x_i),\sigma^2)}) \\
&\qquad\qquad + W_1(\delta_{(1,-\varepsilon - x_i)}, \rho_{((1,-\varepsilon - x_i),\sigma^2)}) \Bigg)^2,
\end{aligned} \tag{A.10}$$

where $K_1 > 0$ is a constant independent of m. Recalling the form of the 2-Wasserstein distance between a delta and a Gaussian distribution, we have that

$$W_2^2(\delta_{(w,b)}, \rho_{((w,b),\sigma^2)}) \leq K_2\sigma^2, \tag{A.11}$$

for some constant $K_2 > 0$. As the W_1 distance is upper bounded by the W_2 distance (via Hölder's inequality), by combining (A.10) and (A.11), we conclude that

$$\left(\int a^m(w^m x + b)_+^m \hat{\rho}(\mathrm{d}\boldsymbol{\theta}) - \int a^m(w^m x + b)_+^m \tilde{\rho}(\mathrm{d}\boldsymbol{\theta})\right)^2 \leq K_3\sigma^2,$$

where $K_3 > 0$ is a constant independent of m. Hence, by taking $\sigma^2 = \min(\lambda, 1/2)$, we have

$$R^m(\tilde{\rho}) \leq K_4\lambda,$$

where $K_4 > 0$ is a constant independent of m.

Now recall that the differential entropy is a concave function of the distribution. Hence, by using the fact that $\rho_{((\mu_1,\mu_2,\mu_3),\sigma^2)}$ is a product distribution and by explicitly computing the entropy of a Gaussian and a uniform random variable, we conclude that

$$H(\tilde{\rho}) \geq K_5(-1 + \log\lambda),$$

where $K_5 > 0$ is a constant independent of m. As $M(\tilde{\rho})$ is upper bounded by a constant independent of m, we conclude that

$$\mathcal{F}^m(\tilde{\rho}) \leq K_6\lambda + \frac{K_5}{\beta}(1 - \log\lambda), \tag{A.12}$$

with $K_6 > 0$ independent of m. Hence, since ρ_m^* is the minimizer of the free energy, by using the bound from Lemma 10.2 in Mei et al. (2018), we get that

$$\frac{1}{2}R^m(\rho_m^*) \leq K_6\lambda + \frac{K_5}{\beta}(1 - \log\lambda) + \frac{1}{\beta}\left[1 + 3\log\frac{8\pi}{\beta\lambda}\right]. \tag{A.13}$$

Since $\beta > -\frac{1}{\lambda}\log\lambda$ and $\beta\lambda > 1$, (A.13) implies that

$$R^m(\rho_m^*) \leq K_7\lambda, \tag{A.14}$$

for $K_7 > 0$ independent of (m, β, λ). This finishes the proof of the first part of the statement. The second part of the statement follows by combining (A.14) with Lemma A.4. ■

Lemma A.6 (Second moment is uniformly bounded)**.** *Assume that condition **A1** holds. It holds that there exists $\tau(m, \beta, \lambda)$ such that for any $\tau > \tau(m, \beta, \lambda)$ the following upper bound holds:*

$$M(\rho_{\tau,m}^*) \leq C,$$

for some $C > 0$ that is independent of $(\tau, m, \beta, \lambda)$.

Proof of Lemma A.6. Let $\tilde{\rho}$ be defined as in (A.7). Then, by combining (A.12) with Lemma A.3, we have that, for $\tau > \tau(m, \beta, \lambda)$,

$$\mathcal{F}^{\tau,m}(\tilde{\rho}) \leq K_1\lambda + \frac{K_2}{\beta}(1 - \log\lambda),$$

where $K_1, K_2 > 0$ are independent of m. Hence, by using (A.2) with $\tilde{\rho}$ in place of ρ and by recalling that $R^{\tau,m}(\rho^*_{\tau,m}) \geq 0$ and the existence of constants C_1 and C_2 such that $\beta > C_1$ and $\lambda < C_2$, the result readily follows. ∎

We conclude this part of the appendix by providing the proof of Lemma 5.6.

Proof of Lemma 5.6. Consider the following lower bound

$$\begin{aligned} Z_m(\lambda,\beta) &\geq \int \exp\left\{-\beta\left[\sum_{i=1}^{M}|R_i^m(\rho_m^*)|\cdot|a^m|(w^m x_i + b)_+^m + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right]\right\}\mathrm{d}\boldsymbol{\theta} \\ &\geq \int \exp\left\{-\beta\left[\sum_{i=1}^{M}|R_i^m(\rho_m^*)|\cdot|a|(w^m x_i + b)_+ + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right]\right\}\mathrm{d}\boldsymbol{\theta} \\ &\geq \int \exp\left\{-\beta\left[\sum_{i=1}^{M}|x_i R_i^m(\rho_m^*)|\cdot|aw| + \sum_{i=1}^{M}|R_i^m(\rho_m^*)|\cdot|ab| + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\right]\right\}\mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{A.15}$$

Define $A = \sum_{i=1}^{M}|x_i R_i^m(\rho_m^*)|$ and $B = \sum_{i=1}^{M}|R_i^m(\rho_m^*)|$. By Lemma A.5, $|R_i^m(\rho_m^*)| \leq K_1\sqrt{\lambda}$, where $K_1 > 0$ is independent of (m, β, λ, i). Therefore, $|A|, |B| \leq K_2\sqrt{\lambda}$ for some $K_2 > 0$ independent of (m, β, λ). Using the inequalities $2|aw| \leq a^2 + w^2$ and $2|ab| \leq a^2 + b^2$, the RHS of (A.15) can be lower bounded by

$$\int_{\mathbb{R}^3} \exp\left\{-\frac{\beta}{2}\left[(2K_2\sqrt{\lambda} + \lambda)\cdot a^2 + (K_2\sqrt{\lambda} + \lambda)\cdot w^2 + (K_2\sqrt{\lambda} + \lambda)\cdot b^2\right]\right\} d\boldsymbol{\theta}.$$

By explicitly computing the integral above, the desired result immediately follows. ∎

A.3 Lower Bound on Polynomials

Proof of Lemma 5.4. We start by rewriting P_2 as

$$\begin{aligned} P_2(x) &= (1-a^2)\cdot(x - x_c)^2 + \left[2(1-a^2)x_c + 2ab\right]\cdot(x - x_c) \\ &\qquad + (1-a^2)x_c^2 + 2abx_c + 1 - b^2 \\ &= \frac{1}{2}P_2''(x_c)\cdot(x - x_c)^2 + P_2'(x_c)\cdot(x - x_c) + P_2(x_c). \end{aligned} \tag{A.16}$$

By definition of x_c, one can immediately verify that $P_2(x_c) \geq 0$. Notice that, if $P_2''(x_c)$ is close to 0, then a^2 is close to 1, which implies that (since $x_c \in I$ and, thus, bounded in absolute value) $|P_2'(x_c)|$ is close to $2|b|$ and $P_2(x_c)$ is close to $\mathrm{sign}(a)\cdot 2bx_c + 1 - b^2$. Therefore, at least one of the coefficients $P_2(x_c), |P_2'(x_c)|, |P_2''(x_c)|$ is lower bounded by a constant that is independent of (a, b).

Next, we distinguish two cases depending on the sign of $P_2''(x_c)$. First, assume that $P_2''(x_c) \geq 0$. We now show that $P_2'(x_c) \cdot (x - x_c) \geq 0$.

In case of a degenerate polynomial, i.e., $P_2''(x_c) = 0$, we distinguish two sub-cases: either $P_2'(x_c) > 0$ or $P_2'(x_c) < 0$ holds. (The case corresponding to $P_2'(x_c) = 0$ is trivial.) If $P_2'(x_c) > 0$, then by definition of x_c and the fact that $x \in \Omega_+$ by assumption, we have that $(x - x_c) > 0$. In fact, recalling the definition of x_r in Definition 4.1, as Ω_+ has non-zero Lebesgue measure, x_c is either the left extreme of I (i.e., $x_c = \inf_{\tilde{x} \in I} \tilde{x}$) or $x_c = x_r \in I$ (i.e., in the interior) and, hence, $\Omega_+ = (x_r, I_r]$ with $x_r < I_r$. This gives that $P_2'(x_c)(x - x_c) \geq 0$. The case $P_2'(x_c) < 0$ follows from similar arguments.

Now assume that $P_2''(x_c) > 0$ and let $x_{\min}$ be the minimizer of P_2 on the interval I. If $x \geq x_{\min}$ then, by definition of a critical point, $x_c \geq x_{\min}$ which means that x_c is located on the *right* branch of the parabola and, hence, $P_2'(x_c) \geq 0$. Furthermore, x belongs to the interval $[x_c, I_r]$ by definition of x_c. These facts imply that $P_2'(x_c) \cdot (x - x_c) \geq 0$. The case $x < x_{\min}$ is treated in a similar fashion.

As it was shown, at least one of the coefficients $P_2(x_c), |P_2'(x_c)|, P_2''(x_c)$ is lower bounded by a constant that is independent of (a, b), and $P_2'(x_c)(x - x_c) \geq 0$, hence, choosing

$$\alpha_2 = P_2''(x_c), \quad \alpha_1 = |P_2'(x_c)|, \quad \alpha_0 = P_2(x_c),$$

concludes the proof for the case of non-negative curvature.

Assume now that $P_2''(x_c) < 0$. As $|\Omega_+|$ is lower bounded by a strictly positive constant, we can pick $\tilde{x} \in \Omega_+$ such that $|\tilde{x} - x_c| = C$, for some $C > 0$ which is independent of (a, b). As $\tilde{x} \in \Omega_+$, we have that $P_2(\tilde{x}) \geq 0$. Furthermore, by rewriting $P_2(\tilde{x})$ as in (A.16), we obtain that

$$\frac{1}{2} P_2''(x_c) \cdot (\tilde{x} - x_c)^2 + P_2'(x_c) \cdot (\tilde{x} - x_c) + P_2(x_c) \geq 0,$$

which implies that

$$|P_2'(x_c)||\tilde{x} - x_c| + P_2(x_c) \geq -\frac{1}{2} P_2''(x_c)(\tilde{x} - x_c)^2. \tag{A.17}$$

As $|\tilde{x} - x_c| = C$, (A.17) is equivalent to

$$|P_2'(x_c)| \cdot C + P_2(x_c) \geq -\frac{1}{2} P_2''(x_c) \cdot C^2. \tag{A.18}$$

Now, if both $|P_2'(x_c)|$ and $P_2(x_c)$ are close to 0, then (A.18) immediately implies that $P_2''(x_c)$ is also close to 0. However, following the argument above, it is not possible that $-P_2''(x_c)$, $|P_2'(x_c)|$ and $P_2(x_c)$ are simultaneously close to 0. This proves that $\max(|P_2'(x_c)|, P_2(x_c))$ is lower bounded by a constant that is independent of (a, b).

Let $x_{\max}$ be the maximizer of P_2 and, without loss of generality, assume that $x_c < x_{\max}$ (the case $x_c \geq x_{\max}$ is handled in a similar way). Note that, by definition of x_c, the point x lies in the interval $[x_c, x_{\max}]$. To show this, let us assume the contrary, i.e., $x > x_{\max}$ (the case $x < x_c < x_{\max}$ is ruled out by the assumption that $x \in \Omega_+$). Then, the root of P_2 which is the closest in Euclidean distance to x is located to the right of $x_{\max}$, hence $x_c < x_{\max}$ cannot be a critical point for x, which leads to a contradiction. This proves that x lies in the interval $[x_c, x_{\max}]$ and in particular, $x \leq x_{\max}$. Furthermore, by concavity, the

parabola $P_2(\tilde{x})$ is lower bounded by the line that connects $(x_c, P_2(x_c))$ and $(x_{\max}, P_2(x_{\max}))$ for $\tilde{x} \in [x_c, x_{\max}]$. By the focal property of the parabola, this line has angular coefficient $|P_2'(x_c)|/2$. Therefore,

$$P_2(\tilde{x}) \geq (\tilde{x} - x_c) \cdot |P_2'(x_c)|/2 + P_2(x_c), \quad \tilde{x} \in [x_c, x_{\max}].$$

Picking $\tilde{x} = x$ and

$$\alpha_2 = 0, \quad \alpha_1 = |P_2'(x_c)|/2, \quad \alpha_0 = P_2(x_c),$$

gives the desired result in the case $P_2''(x_c) < 0$ and concludes the proof. ■